МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
(МОСКОВСКИЙ ПОЛИТЕХ)

Факультет информационных технологий

Кафедра «Прикладная информатика»

Форма обучения: очная

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по направлению 01.03.02 «Прикладная математика и информатика»

на тему: «Анализ показателей и уменьшение рисков хронических заболеваний с помощью технологий больших данных»

Студентка _____ Юлия Григорьевна Бакулина

Руководитель работы
доцент, к.п.н. _____ Наталья Ивановна Царькова

ДОПУСКАЕТСЯ К ЗАЩИТЕ

Заведующий кафедрой
профессор, к.э.н. _____ Станислав Вадимович Суворов

Москва 2020

УТВЕРЖДАЮ

Заведующий кафедрой

«Прикладная информатика»

_____С. В. Суворов

ЗАДАНИЕ

на выпускную квалификационную работу (ВКР)

Студента Бакулиной Юлии Григорьевны **группы** 161-381

1. **Тема**: «Анализ показателей и уменьшение рисков хронических заболеваний с помощью технологий больших данных»
2. **Утверждена** приказом ФГБОУ ВО «Московский политехнический университет» от 22.05.2020 № 301-с
3. **Исходные данные к работе**: Индикаторы хронических заболеваний центра по контролю и профилактике хронических заболеваний США (CDC), научная и методическая литература

4. **Содержание ВКР (**перечень подлежащих разработке вопросов**)**

№№ п/п	Наименование раздела	Содержание раздела
1.	Аналитическая часть	Анализ информации о хронических заболеваний
		Исследования в сфере хронических заболеваний
		Борьба с НИЗ
		Данные для исследований
2.	Теоретическая часть	Технология Big Data
		Python
3.	Проектная часть	Подготовка и обзор данных
		Построение первоначальной модели
		Итоговая модель
		Практическая значимость модели
		Структура модели

5. **Календарный график выполнения ВКР**

№№ п/п	Наименование разделов	Дата проведения консультаций
1.	Аналитическая часть	15.05.2020
2.	Теоретическая часть	30.05.2020
3.	Проектная часть	08.06.2020

6. **Срок сдачи студентом законченной работы 30.06.2020**

Задание выдал 20.04.20
Руководитель
Наталья Ивановна Царькова

Задание получил 20.04.2020
Студентка
Юлия Григорьевна Бакулина

АННОТАЦИЯ

Тема выпускной квалификационной работы: «Анализ показателей и уменьшение рисков хронических заболеваний с помощью технологий больших данных».

Работа содержит 126 страниц, 43 рисунков, 6 таблиц и 21 источник.

Цель ВКР: проанализировать показатели хронических заболеваний США, взятые с сайта центра хронических заболеваний (CDC) уменьшить риск распространенности хронических неинфекционных заболеваний (ХНИЗ), а также сформулировать практическую значимость модели, применив к имеющимся данным технологии больших данных.

Данная работа состоит из трех частей:

В аналитической части был произведен анализ информационных мировых медицинский ресурсов, а также исследования в сфере НИЗ. Был произведен анализ данных и сайта CDC.

В теоретической части была рассмотрены технологии больших данных (Big Data) и их основные методы. Технология Data Mining, ее основные процессы, задачи и методы. Помимо этого, был выбран метод для дальнейшего анализа данных и основной инструмент для анализа (Python 3).

В проектной части исходные данные были очищены и подготовлены для анализа. Далее была применена технология Data Mining и визуального анализа к измененным данным. После подготовки, анализа и проверки моделей были сформированы рекомендации, оценены практические показатели, проанализированы риски.

Ключевые слова: BIG DATA (большие данные), DATA MINING (Интеллектуальный анализ данных), PYTHON 3, НИЗ, МЕТОД ВИЗУАЛИЗАЦИИ ДАННЫХ, ХРОНИЧЕСКИЕ ЗАБОЛЕВАНИЯ.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Здоровье человека – очень важный показатель экологической, социальной и экономической сферы. В зависимости от того, как эти сферы балансируют в государстве меняются и факторы, влияющие на здоровье человека. Результат такого воздействия факторов, как генетические, физиологические и поведенческие можно наблюдать у людей с хроническими неинфекционными заболеваниями (ХНИЗ). Например, астма из-за загрязненного воздуха, рак, развившийся из-за постоянного стресса.

Одну из главных проблем сегодня заняла распространённость неинфекционных заболеваний. Более двух третей всех смертей вызваны одним или несколькими из этих пяти хронических заболеваний: болезни сердца, рак, инсульт, хроническая болезнь легких и диабет. Они уносят каждый год миллионы жизней, а также приводят к тяжелым осложнениям и инвалидности. В свою очередь осложнения связаны с потерей трудоспособности и необходимостью высоко затратного лечения. Сегодня перед системами здравоохранения стоит задача уменьшить показатели хронических заболеваний за счет улучшения качества жизни, увеличения периода активной жизни, повышения и сохранения трудоспособность у пациентов с данными заболеваниями.

Не смотря на, все попытки стран в борьбе с НИЗ, хронические заболевания до сих пор являются главной причиной смерти в мире. Для более эффективной борьбы с НИЗ, а также для оптимизации работы органы здравоохранения стали активно внедрять в свою деятельность различные средства по цифровой обработке баз данных. Применив к имеющимся данным интеллектуальный анализ, можно не только узнать сколько человек болеет определенным заболеванием, но и выявить предрасположенность группы людей к определенной болезни. Полученные в результате анализа выводы помогут медицинским работникам, а также органам здравоохранения в борьбе с хроническими заболеваниями и уменьшить риск их распространенности.

Интеллектуальный анализ данных (Data Mining) – это процесс сортировки больших наборов, для выявления закономерностей и установления взаимосвязей для решения проблем посредством анализа данных. Инструменты интеллектуального анализа данных позволяют предприятиям прогнозировать будущие тенденции.

Обширные объемы данных, хранящихся в медицинских базах данных, структурированных и неструктурированных наборов показателей здоровья требуют разработки и анализа с помощью специализированных инструментов.

Таким образом, интеллектуальный анализ показателей хронических заболеваний США будет иметь практическую значимость и применение.

Цель настоящей работы – анализ показателей и уменьшение рисков хронических заболеваний с помощью технологий больших данных.

Данная цель реализуется с помощью следующих задач:

1. Изучить информацию о хронических неинфекционных заболеваниях.
2. Изучить исследования в сфере НИЗ.
3. Борьба с НИЗ.
4. Подготовка и обзор данных для анализа.
5. Построение первоначальной модели.
6. Построение итоговой модели.
7. Формирование практической значимости и уменьшение рисков.

Объектом работы являются показатели хронических заболеваний (CDI) США, взятые с веб-сайта Центра Хронических Заболеваний (CDC).

Предметом исследования являются показатели хронических заболеваний (CDI) США, взятые с веб-сайта Центра Хронических Заболеваний (CDC). Они представлены с помощью набора из 124 показателей предоставленные CDC's Division of Population Health (отделом здоровья населения CDC), которые были разработаны на основе консенсуса. Он позволяет штатам и территориям единообразно определять, собирать и представлять данные о хронических

заболеваниях, которые важны для практики общественного здравоохранения и доступны для штатов, и крупных столичные районы.

В рамках данной работы использовались такие инструменты анализа, как технологии Big Data, а точнее интеллектуальный анализ больших данных (Big Data Mining), методология визуализации данных и кросс-факторный анализ данных.

Основными источниками данных для анализа, использованными в работе, являются данные показателей хронических заболеваний (CDI), отчет о заболеваемости и смертности (Morbidity and Mortality Weekly Report (MMWR)) и официальный сайт центра хронических заболеваний США (CDC).

1. АНАЛИТИЧЕСКАЯ ЧАСТЬ

1.1 Анализ информации о хронических заболеваний

Нынешний год стал самым настоящим показателем того, как важно относиться серьезно к своему здоровью. Здоровье человека, определяемое как полное состояние физического, социального и психического благополучия, а не просто отсутствие болезни, так же важно, как вода, пища или энергия. Здоровье лежит в основе экономики современных обществ. Для поддержки здорового общества необходима надежная система здравоохранения и увеличение эффективности действий по снижению социального риска. Множество факторов играют важную роль в поддержании здоровья. В свою очередь, хорошее здоровье может снизить риск развития определенных заболеваний. Все болезни можно разделить на две большие группы: инфекционные заболевания (ИЗ) и неинфекционные заболевания (НИЗ), которые так же имеют название, как хронические заболевания.

Хронические заболевания в широком смысле определяются как состояния, которые длятся один год или более и требуют постоянной медицинской помощи или ограничивают повседневную жизнь или и то, и другое. НИЗ не могут быть предотвращены с помощью вакцин или вылечены с помощью лекарств, естественно сами собой они не исчезают. Нездоровые привычки - особенно употребление табака, отсутствие физической активности и плохое питание - являются основными причинами хронических заболеваний. [1]

Основными хроническими заболеваниями являются:

- Сердечно - сосудистые заболевания, главным образом болезни сердца и инсульт;
- рак;
- хронические респираторные заболевания;
- диабет.

Существует много других хронических состояний и болезней, которые вносят значительный вклад в бремя болезней для отдельных людей, семей, обществ и стран. Примерами могут служить такие болезни, как:

- психические расстройства;
- нарушения зрения и слуха;
- заболевания полости рта;
- заболевания костей и суставов;
- генетические нарушения и др.

Каждый год больше двух третей мировых расходов на здравоохранение уходит на лечение и поддержание уровня жизни больных с неинфекционными заболеваниями. Это более 3,5 трлн долларов в год. Для сравнения, суммарные расходы всего мира на лекарственные препараты не дотягивают до 1 трлн. Крайне важно, чтобы возрастающий показатель неинфекционных заболеваний был предвиден, понятен и как можно быстрее решен. Это требует нового подхода со стороны государств, которые в состоянии усилить работу по профилактике хронических заболеваний и борьбе с ними, а также со стороны всемирной организации здравоохранения (ВОЗ). [2]

1.1.1 Анализ увеличения распространенности хронических заболеваний

В настоящее время можно наблюдать существенный рост увеличения распространенности хронических заболеваний. Проблема распространенности НИЗ актуальна не только для России, но и для всего мира без исключения. Хронические заболевания являются главными причинами смертности в мире уже более 20 лет. С 2000 по 2015 год смертность от этих заболеваний возросла более чем в 1,5 раза. [3]

Если рассматривать отчеты и прогнозы Организации Объединенных Наций за 2015 – 2019 гг., то становится ясно, что за последние 4 года уровень смертности снизился (Рисунок 1.1). Снижение уровня смертности было достигнуто благодаря мероприятиям по снижению смертности от ишемической

болезни сердца в 2017 году, а также за счет увеличения периода активной жизни и уменьшения смертности детей в возрасте до 5 лет.

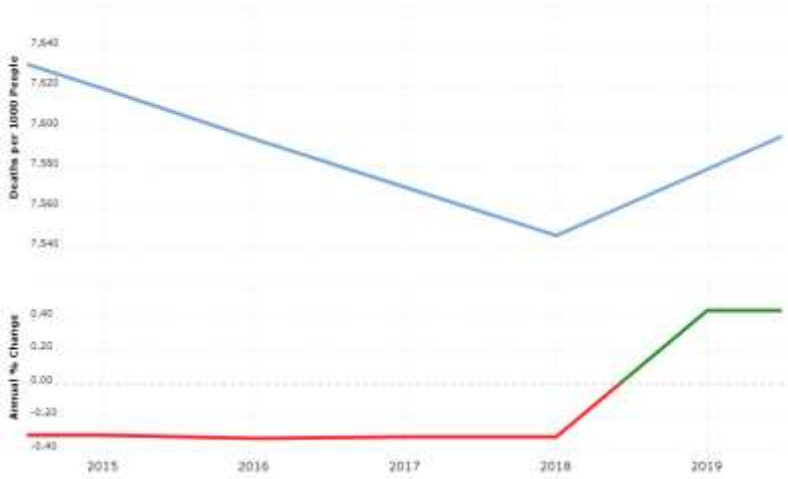

Рисунок 1.1 Отчет уровня смертности за 2015-2019 гг.

Не смотря на то, что мировая смертность уменьшилась, смерть от хронических заболеваний только увеличивается. Если рассматривать показатели смертности от НИЗ в процентах, то можно заметить, что с каждым годом процент увеличивается. За 2015 год от НИЗ умерло 70% от общего числа смертей в мире, в 2016 году 71%, за 2017 более 73%. По оценкам специалистов, последние несколько лет показатели смертности от хронических заболеваний показывают стабильные результаты. В 2018-2019 гг. от хронических заболеваний погибли около 70% людей от общего числа смертей в мире, что говорит об уменьшении динамики смертности от НИЗ, но не о полной победе.

В настоящее время в первую десятку проблем со здоровьем в мире относятся болезни сердца, рак, инсульт, респираторные заболевания, травмы, диабет, болезнь Альцгеймера, грипп и пневмония, заболевания почек и сепсис. Стареющее население в сочетании с существующими факторами риска и достижения в области медицины, которые продлевают продолжительность, привели к выводу, что эти проблемы будут только усиливаться, если не эффективно решать их сейчас. [4]

1.1.2 Обзор статистики мировой смертности от НИЗ

Около 41 миллиона человек в год умирает от хронических заболеваний - это 71% всех случаев смерти в мире. По оценкам специалистов, в 2015 году на глобальном уровне произошло 40 миллионов смертей от неинфекционных заболеваний, что составляет 70% от общего числа смертей в мире. В 2016 году из 57 миллионов случаев смерти, более половины, были из-за хронических заболеваний. А в 2017 году умерло 42 миллиона людей от НИЗ. Статистика показывает, что самыми распространенными причина смерти в мире, являются заболевания группы НИЗ. На сайте Всемирной организации здравоохранения можно посмотреть статистику ведущих причин смерти с 2000 по 2016 гг.

Не смотря на то, что ученые и исследователи ведут отчаянную борьбу с самими НИЗ и с распространением этих заболеваний, за 16 лет верхушки ведущих причин смерти в мире почти не поменялись (Рисунок 1.2, 1.3).

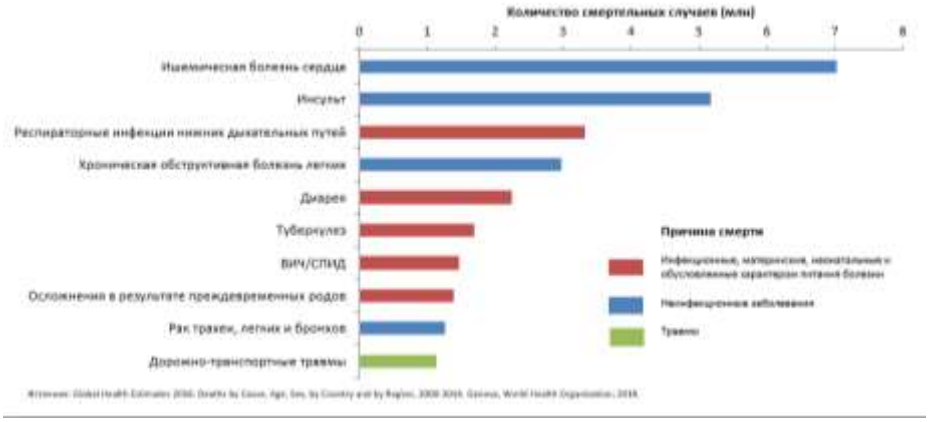

Рисунок 1.2 10 ведущих причин смерти в мире (2000 г.)

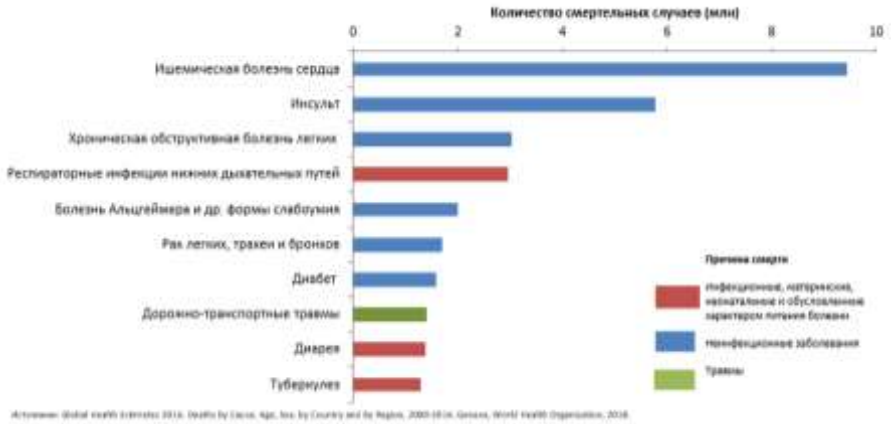

Рисунок 1.3 10 ведущих причин смерти в мире (2016 г.)

Шокирующим показателем является то, какое количество смертей можно было предотвратить, если бы люди более здраво и ответственно относились к своему здоровью. Мало что сделано для преобразования имеющихся знаний или фактических данных в эффективную политику и действия. За 2017 год умерло около 1,7 миллиона человек от заболеваний, связанных с диареей. Она входила в десятку первых причин смертности, а в некоторых странах являлась одной из главных убийц. Показатели смертей от неонатальных расстройств – смерть ребенка в первые 28 дней, значительно различаются между странами. Тот же 2017 год унес жизни более 1,8 миллиона новорожденных. В бедных странах мира 1 из 20 детей умирает в первые дни. В Японии же этот показатель менее 1 из 1000 детей. В следствие отсутствия лечения диабета каждые 30 секунд ампутируют нижнюю конечность человека. А за последние несколько лет процент заболеваемости раком в мире вырос до 33%. Если вдуматься в эти показатели становиться страшно. Люди не задумываются о своем здоровье до тех пор, пока не становится поздно. [5]

1.1.3 НИЗ в странах с высоким уровнем доходов и СНСД

Распространённость хронических заболеваний во многих странах с низким и средним уровнем дохода (СНСД) неуклонно растет (Рисунок 1.4).

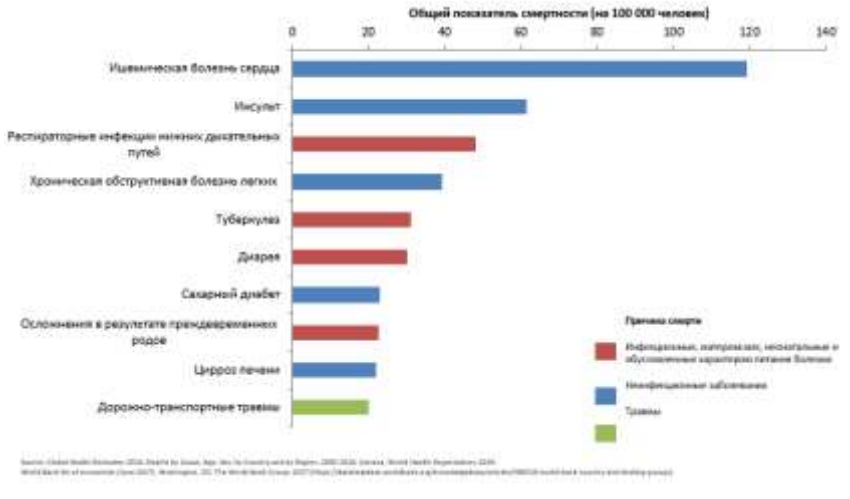

Рисунок 1.4 Ведущие причины смерти в СНСД в 2016 г.

В 2019 году этот показатель достиг 70% случаев смерти от неинфекционных заболеваний, которые происходят в таких странах. Как не

странно в странах с высоким уровнем доходов дела с хроническими заболеваниями обстоят также плохо, показатель достигает 88% случаев смерти от хронических заболеваний (Рисунок 1.5). НИЗ ответственны за 9 из 10 ведущих причин смерти в странах с высоким уровнем доходов. Все страны постоянно принимали и принимают основательные меры для борьбы с НИЗ.

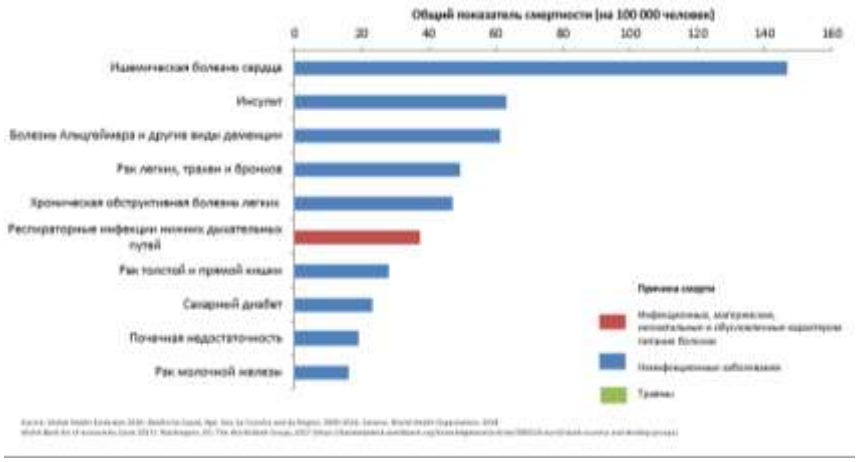

Рисунок 1.5 Ведущие причины смерти в странах с высоким уровнем доходов в 2016 г.

Приблизительно 1,8 миллиона новорожденных все еще умирают от осложнений при рождении. Очень низкие показатели неонатальной смертности в странах с высоким уровнем доходов и значительный прогресс в мире в последние десятилетия свидетельствуют о том, что мы знаем, как значительно сократить такие трагедии. Аналогичным образом, сокращены желудочно-кишечные заболевания, которые в 2017 году унесли 1,6 миллиона человек и являются одной из основных причин смерти детей в возрасте до 5 лет. Они также вылечены с помощью улучшенной воды, санитарии, гигиены и простой «оральной регидратационной соли» (ОРС). Малярия была успешно ликвидирована в некоторых регионах, и со временем должна быть возможность ее искоренить. Тем не менее, согласно исследованию IHME по глобальному бремени болезней (ГББ), в 2017 году от малярии все еще умерло около 620 000 человек.

В последние годы, в странах с высоким уровнем дохода стандартизованные по возрасту коэффициенты смертности от

сердечнососудистых заболеваний резко сократились, в то же время коэффициенты смертности от других основных НИЗ снижались более медленными темпами. При этом, несмотря на то, что стандартизованные по возрасту коэффициенты смертности от сердечнососудистых заболеваний и смертности от хронических респираторных заболеваний значительно снизились в странах с высоким уровнем дохода, в странах с низким и средним уровнем дохода, эти коэффициенты по-прежнему намного выше.

Мировой уровень потребления алкоголя в 2016 г. составил 6,4 литра чистого алкоголя на человека в возрасте 15 лет и старше, при этом отмечалась значительная разница в этом показателе между регионами ВОЗ. Имеющиеся данные указывают на то, что охват лечением лиц, злоупотребляющих алкоголем и наркотиками, недостаточен, хотя для улучшения измерения такого охвата необходима дальнейшая работа. В 2015 г. более 1,1 миллиарда человек курили табак, причем этот показатель гораздо выше у регулярно курящих мужчин, чем у регулярно курящих женщин.

1.1.4 НИЗ и COVID-19

Хронические заболевания приносят вред не только как самостоятельные заболевания, в тандеме с инфекционными болезнями они становятся бомбой и к большому сожалению даже не замедленного действия. Бич 2020 года – пандемия короновируса (COVID-19) этому доказательство. Минздрав, ВОЗ и многие иностранные сообщества и ученые доказали, что люди с НИЗ более подвержены заражение и плохо подлежат лечению от коронавируса. Эксперты общественного здравоохранения предупреждали о том, что пожилые пациенты с гипертонией, сахарным диабетом и другими хроническими заболеваниями могут быть особенно уязвимы к коронавирусу. Они более склонны испытывать серьезные симптомы COVID-19, респираторного заболевания, вызванного коронавирусом. Многие симптомы, связаны с постепенным ухудшением их иммунной системы, которые происходят из-за возраста (Рисунок 1.6). Поскольку, для хорошей работы организма, для борьбы с инфекциями все органы тела должны работать вместе, но когда один конец этой сложной

системы начинает напрягаться, то и у другой части системы возникают больше проблемы. Люди всех возрастов с общими хроническими заболеваниями, такими как астма и онкологические заболевания, также рискуют серьезно заболеть, если заражаются вирусом, распространяющимся по всему миру. [6]

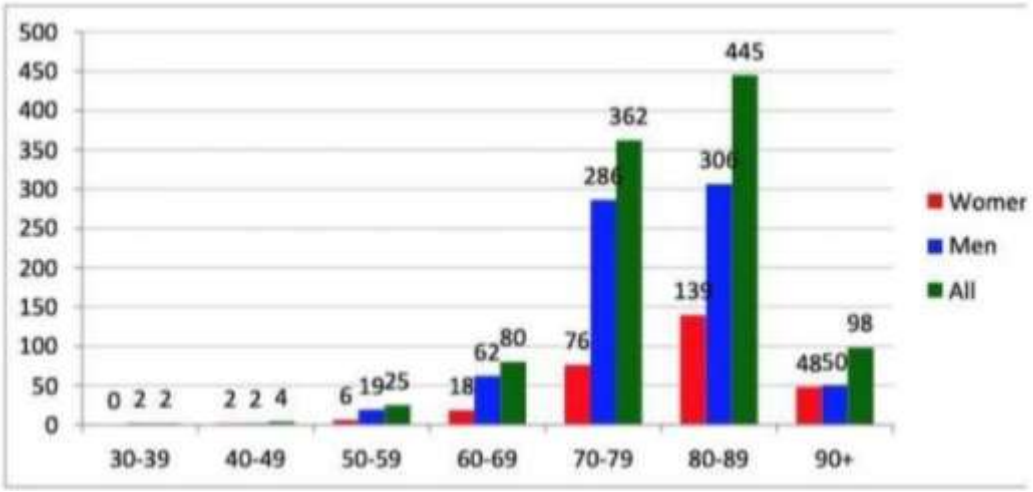

Рисунок 1.6 Смертность от коронавируса

Национальный центр CDC по профилактике хронических заболеваний и укреплению здоровья заявил, что данные на 31 марта 2020 года, показывают, что американцы с хроническими заболеваниями сталкиваются с повышенным риском тяжелой болезни от COVID-19, в соответствии с более ранними сообщениями из Китая и Италии. Исследователи изучили более 7000 случаев в США, в которых были доступны данные об основных состояниях здоровья и других потенциальных факторах риска. Они обнаружили, что среди людей, госпитализированных по поводу COVID-19, около 71% имели, по крайней мере одно основное заболевание, а также среди тех, кто поступил в реанимацию, около 78% имели одно хроническое заболевание. Всего 27 % людей с одним заболеванием не нуждались в госпитализации по поводу. Также наиболее часто сообщаемыми состояниями среди людей, больных COVID-19, были диабет, заболевание легких и болезни сердца.

1.2 Исследования в сфере хронических заболеваний

1.2.1 Обзор исследований в сфере хронических заболеваний

Хронические заболевания являются одними из самых распространенных и дорогостоящих заболеваний во всем мире. Неинфекционные заболевания,

включая рак, диабет, гипертонию, инсульт, болезни сердца, респираторные заболевания, артрит, ожирение и заболевания полости рта, могут привести к госпитализации, длительной не трудоспособности, к снижению качества жизни и смерти. Фактически, НИЗ являются основной причиной смерти и инвалидности в мире.

Во всем мире хронические заболевания влияют на здоровье и качество жизни многих граждан. Кроме того, хронические заболевания были основной движущее силой расходов на здравоохранение, а также влияли на структуру рабочей силы, включая, конечно, невыходы на работу. По данным организаций по контролю за заболеваниями, только в США хронические заболевания составляют почти 75% совокупных расходов на здравоохранение или примерно 5300долларов США на человека в год.

Сегодня перед системами здравоохранения стоит задача уменьшить показатели хронических заболеваний за счет проводимых исследований в области НИЗ, улучшения качества жизни, увеличения количества центров борьбы с хроническими заболеваниями, увеличения периода активной жизни, повышения и сохранения трудоспособность у пациентов с данными заболеваниями. НИЗ в значительной степени можно предотвратить, но мало что сделано для преобразования имеющихся знаний или фактических данных в эффективную политику и действия. Поэтому в первую очередь важно уделить внимание самым главным вопросам, касающимся НИЗ. Одним из таких глобальных вопросов на протяжении семи лет был план исследования в разработке профилактики и борьбе с неинфекционными заболеваниями. Главной целью плана на 2013-2020 год являлось содействие исследователям в приоритетных областях науки. Определение приоритетов важно, особенно в условиях ограниченных ресурсов, для наиболее эффективного использования ограниченных ресурсов.

Общественное здравоохранение может так же обратиться к традиционным подходам лечения хронических заболеваний, таким как йога, пилатес и натуротерапия. Ведь многие из этих подходов могут играть

первостепенную роль в профилактике и лечении заболеваний НИЗ. Но учитывая отсутствие убедительных данных о безопасности и эффективности таких альтернативных подходов лечения, эти области также должны быть изучены и исследованы. Основная проблема заключается в том, что в СНСД проводятся ограниченные исследования хронических неинфекционных заболеваний, и для таких исследований доступно мало ресурсов.

Для того что бы разобраться в эпидемиологии хронических заболеваний необходима система для мониторинга бремени НИЗ и изучения факторов риска. Поэтапный подход ВОЗ к эпидемиологическому надзору за факторами риска неинфекционных заболеваний направлен на сбор данных во всех регионах мира. Тем не менее, наука пошла дальше, после исследований факторов риска и распространённости заболеваний она перешла к исследованиям по выявлению лучших практик для снижения факторов риска. Большие, перспективные исследования необходимы, чтобы заполнить пробелы в эпидемиологии заболеваний, профилактике и контроле, прогрессировании и ответных реакциях на лечение.

1.2.2 Программа ВОЗ CINDI

Эксперты ВОЗ разработали программу для интегрированной медицинской профилактики основных неинфекционных заболеваний – CINDI. Целью этой программы является улучшение здоровья путем снижения смертности и заболеваемости от основных неинфекционных заболеваний (сердечнососудистых заболеваний, рака, травм, хронических респираторных заболеваний и других) посредством комплексных совместных мероприятий, которые предотвращают заболевания и укрепляют здоровье. Программа CINDI охватывает более 30 стран Европы и Канады, включая Болгарию. Для каждой страны программа имеет общегосударственное значение, так как реализуется в демонстрационных зонах, а за разработку отвечает соответствующее министерство здравоохранения. Целевой группой является население трудоспособного возраста (25-64 лет), включая группы с высоким риском

развития определенных заболеваний. Он также включает в себя детский компонент - ученики (14-18 лет), учителя и родители.

CINDI одобряет профилактику заболеваний через существующую структуру здравоохранения при активном участии общества и отдельных лиц. Существуют конкретные цели и задачи, основанные на точной эпидемиологической структуре, которая постоянно контролируется. За период 2000 - 2010 гг. Было проведено четыре мониторинга, оценивающих изменение поведения населения. Мониторинг показал положительные изменения на уровне населения с момента запуска программы. В 2004 году была введена детская составляющая программы - «Здоровые дети в здоровых семьях». Он был реализован в семи зонах и в настоящее время работает на местном уровне. На национальном уровне программа финансируется Министерством здравоохранения.

Целью CINDI является снижение риска неинфекционных заболеваний за счет уменьшения общих факторов риска, таких как курение, злоупотребление алкоголем, отсутствие физической активности и нездоровое питание. Основными стратегиями являются: санитарное просвещение населения для контроля основных факторов риска НИЗ и здоровья; наращивание потенциала среди медицинских специалистов и партнеров; участие учреждений в программных мероприятиях; разработка руководящих принципов надлежащей практики специалистов и партнеров, а также информационных материалов для населения. В программе участвует много партнеров Региональные инспекции здравоохранения, Региональный фонд медицинского страхования, больницы, медицинские и диагностические консультативные центры, СМИ, НПО, школы и детские сады, компании, союзы и клубы. [7]

1.2.3 Анализ Института Милкена

Во всем мире ведется борьба против распространённости хронических заболеваний. Так недавний анализ Института Милкена установил, что лечение семи наиболее распространенных хронических заболеваний в сочетании с потерей производительности будет стоить экономике США более 1 триллиона

долларов в год (Рисунок 1.7). Кроме того, по сравнению с другими развитыми странами, США плохо ранжируются по стоимости и результатам. Это происходит главным образом из-за нашей неспособности эффективно справляться с хроническими заболеваниями.

Рисунок 1.7 Общие прямые затраты на НИЗ в США в 2016 г. (млрд долл.)

И все же тот же анализ Милкена показывает, что медленное повышения здорового образа жизни может предотвратить или отсрочить 40 миллионов случаев хронических заболеваний в год. Если мы научимся эффективно управлять хроническими заболеваниями, избегая, таким образом, госпитализаций и серьезных осложнений, система здравоохранения сможет значительно улучшить качество жизни пациентов.

1.2.4 Общие данные об исследованиях в сфере НИЗ

Ученные со всего мира используют современную методологию для исследований по хроническим болезням и укреплению здоровья, включая иерархические модели, оценку малых территорий, а также географические информационные системы (ГИС) для связи данных о здоровье и окружающей среде. Успех работ по охране здоровья населения и лечению хронических заболеваний зависит от нескольких ключевых факторов: выявление пациентов, подверженных риску, получение доступа к нужным данным об этой группе населения, создание действенных представлений о пациентах и обучение их принятию более здоровых решений. Такие методы, как управляемая данными визуальная аналитика, помогают экспертам анализировать большие объемы

данных и получать информацию для принятия обоснованных решений в отношении хронических заболеваний.

Согласно данным Института медицины США и Национального исследования, видение здравоохранения 21-го века включает повышенное внимание к когнитивной поддержке при принятии решений. Это означает включение компьютерных инструментов и методов, которые помогают пониманию и познанию. Методы визуализации предлагают когнитивную поддержку, предлагая ментальные модели информации через визуальный интерфейс. Они сочетают статистические методы и модели с передовыми методами интерактивной визуализации, чтобы помочь скрыть сложность больших массивов, данных о состоянии здоровья и принять обоснованные решения.

1.3 Борьба с НИЗ

Неинфекционные заболевания занимают первые места среди причин смертности во всем мире. Страны с высоким, средним и низким уровнем доходов принимали основательные меры для борьбы с НИЗ. Сегодня во всех странах созданы и разрабатываются программы по боре с хроническими заболеваниями.

1.3.1 Обзор независимой комиссии ВОЗ по профилактике и лечению НИЗ

2017 год стал годом реновации в области профилактики и лечения хронических заболеваний, так как именно в этот год было проведено и создано огромное количество исследований и работ. Всемирная организация здравоохранения в тот же год организовала новую независимую комиссию высокого уровня по НИЗ. Комиссия представляет собой перечень требований и рекомендаций, благодаря которым многие страны могли бы уменьшить факторы риска и воздействие первой причины смертности в мире.

Инвестиции в связи с НИЗ должны быть сосредоточены на 16 ключевых вмешательствах, которые ВОЗ называет «Лучшими покупками». Это наиболее экономически эффективные вмешательства для предотвращения возникновения

НИЗ и преждевременной смерти людей с НИЗ, основанные на последних фактических данных. Важно отметить, что эти вмешательства осуществимы, доступны и приемлемы в большинстве случаев. В настоящее время все они реализованы на очень низких уровнях в СНСД.

1.3.2 Система профилактики НИЗ Дина Орниша

Население мира заинтересовано в том, чтобы будущее и нынешнее поколение было здорово. Поэтому проводятся ряды работ в борьбе против НИЗ. В США супружеская пара медиков Дин и Энн Орниши давно разработала уникальную систему профилактики хронических заболеваний. Они убеждены в том, что диета и правильный образ жизни влияет на человека и его организм. Хорошее здоровье, по их мнению, обуславливается здоровым выбором жизни. Супруги уверены, что успехи в медицине — это не только новые лекарства, лазеры, хирургические вмешательства с помощью робототехники, но и простые выборы, которые совершает человек в своей жизни каждый день – что мы едим, как мы реагируем на стресс, курим мы или нет, сколько мы тренируемся и качество наших отношений. [8]

На протяжении более 35 лет Дин Орниш, доктор медицинских наук и его коллеги из некоммерческого научно-исследовательского института профилактической медицины (PMRI) в сотрудничестве с Калифорнийским университетом, Сан-Франциско и другими ведущими научными учреждениями, провели серию исследований, показывающих, что изменения в рационе питания и образе жизни могут существенно повлиять на наше здоровье и благополучие. Программа доктора Орниша по борьбе с сердечными заболеваниями - это первая комплексная программа по борьбе с сердечными заболеваниями и другими хроническими заболеваниями.

Супругами были произведены исследования и приведены доказательства их медицинских успехов в области неинфекционных заболеваний. Они исследовали способность пациентов, участвующих в испытании сердца, в течение пяти лет выдерживать интенсивные изменения образа жизни, а также влияние этих изменений образа жизни на ишемическую болезнь сердца. Они

измерили склонность к изменениям образа жизни, изменениям в стенозе процентного диаметра коронарной артерии и сердечным осложнениям. Результаты в экспериментальной группе показали значительное улучшение по сравнению с начальными значениями. Кроме того, по сравнению с первым годом наблюдения, 5-летнее наблюдение показало большое улучшение по сравнению с начальными данными.

Также медики проводили работу с пациентами с диабетом и преддиабетом. Больные диабетом следовали программе Орнишей и продемонстрировали те же улучшения в отношении факторов риска коронарных заболеваний и качества жизни, что и пациенты без диабета. Пациенты показали статистически значимое снижение уровня HgbA1c - это форма гемоглобина, которая химически связана с сахаром. За все время работы, супруги провели более 12 доказательств своей эффективной работы, связанной с кровеносным давлением, депрессией, раком простаты, холестерином, ожирением, болезней сердца и др.

1.3.3 «Красная лента»

Бремя неинфекционных заболеваний растет и в Эфиопии. Рак уже уносит больше жизней, чем ВИЧ или туберкулез. Однако, хотя планы по расширению лечения рака в стране продолжаются, в настоящее время в Аддис-Абебе есть только один лечебный центр. Комплексный подход к расширению услуг имеет решающее значение для всех, кто нуждается в лечении.

Эфиопия запускает национальный стратегический план действий по неинфекционным заболеваниям и партнерство «Розовая лента» и «Красная лента». Страна представила свою национальную стратегию борьбы с неинфекционными заболеваниями и руководство по профилактике и борьбе с раком шейки матки в рамках празднования всемирного дня борьбы с раком.

ВОЗ выступает за обеспечение всех стран услугами здравоохранения и эффективной профилактики и борьбы с НИЗ. План по борьбе с раком шейки матки был запущен в то же время. Он представляет собой глобальное партнерство в области здравоохранения, основанное Институтом Джорджа

Буша-младшего при правительстве США в рамках Чрезвычайного плана Президента по оказанию помощи в связи со СПИДом (ПЕПФАР) Сьюзен Г. Комен и Объединенная программа Организации Объединенных Наций по ВИЧ / СПИДу (ЮНЭЙДС). Партнерство основывается на существующих программах здравоохранения в Эфиопии. Также оно добавляет эффективные меры по профилактике, скринингу и лечению рака шейки матки, который является основной причиной смертности от рака среди женщин в странах Африки к югу от Сахары, что усугубляется его связью с ВИЧ (Рисунок 1.8).

Рисунок 1.8 Отличие скрининга от ранней диагностики

1.3.4 Исследования в области легочной гипертензии

Борьбу с хроническими заболеваниями ведет весь мир. И Американская ассоциация кардиологов (АНА) не отстает. АНА – это один из ведущих спонсоров исследований в области сердечнососудистых заболеваний во всем мире. С 1996 года составляет ежегодный список основных достижений в области наук о болезнях сердца и инсультах. На 2019 год АНА добилась не малых успехов в области легочной гипертонии, повышенным холестерином, ожирением, повышенным артериальным давлением и т.д.

Новые исследования в области легочной гипертензии помогли выявить, как работает болезнь. Легочная гипертензия - это тип высокого кровяного давления, который возникает, когда сосуды, которые несут кровь от сердца к легким, становятся жесткими и узкими. Одно из таких исследований, обсуждалось в Американском журнале респираторной и реанимационной

медицины, где рассматривалась роль белка, известного как BMP9, который был низким у людей с определенным типом легочной гипертонии. В другом исследовании, опубликованном в журнале AHA Circulation, объясняется, как ген, известный как BOLA3, играет решающую роль в заболевании, открывая дверь для потенциальной терапии в будущем.

В декабре 2019 Управление по санитарному надзору за качеством пищевых продуктов и медикаментов одобрило использование икосапента этила, рецептурной формы омега-3 жирной кислоты EPA, в качестве дополнительной терапии для снижения риска сердечно - сосудистых событий среди взрослых с повышенным уровнем триглицеридов. Он продается под торговой маркой Vascepa. Одновременно с этим последовали исследования, опубликованные в Медицинском журнале Новой Англии. Исследования были о том, что производное рыбьего жира снижает риск инсульта или сердечной проблемы на 25% у людей с высоким уровнем триглицеридов, принимающих статины. Также, работы показали, что когда артерия, вызвавшая сердечный приступ, была открыта вместе с другими частично заблокированными сосудами, у пациентов был риск повторного сердечного приступа на 32% ниже по сравнению с очищением только артерии, вызвавшей приступ.

1.4 Данные для исследований

Данные для исследования были взяты с сайта центра по контролю и профилактики заболеваний (CDC). Они являются показателями хронических заболеваний США и связанные с ними демографические данные, привычки поведения, факторы здоровья полости рта и географические данные. Источником данных для этого исследования является не только центр по контролю и профилактике заболеваний, но и также совместная работа Советы государственных и территориальных эпидемиологов (CSTE) и Национальная ассоциация директоров по хроническим заболеваниям (NACDD). Отдел CDC предлагает комплексный набор из 815 000 рядов, 34 столбца и 124 показателей (Рисунок 1.9). [9]

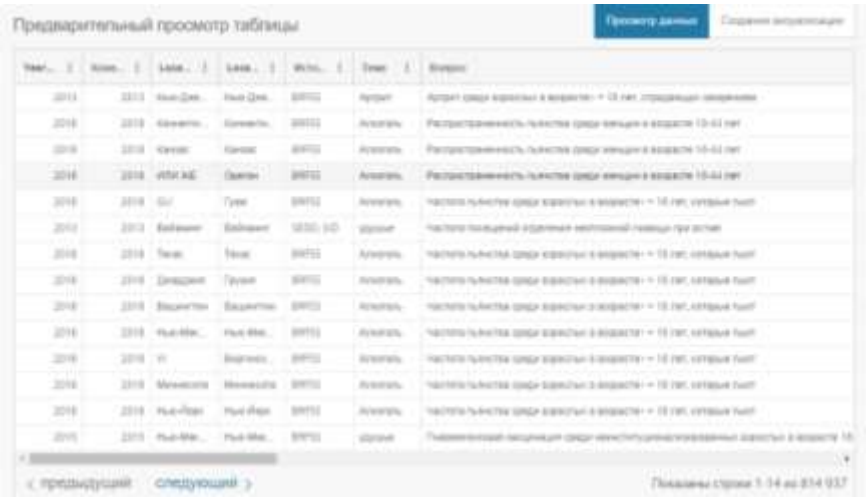

Рисунок 1.9 Предварительный просмотр данных

Эти индикаторы интегрированы из нескольких ресурсов с помощью веб-сайта Chronic Disease Indicator (CDI), который служит проводником для дополнительной информации и источников данных.

1.4.1 История составления показателей хронических заболеваний

Первоначальный показатели хронических заболеваний состояли из 73 показателей, принятых в 1998 году и затем дополненных в 2002 году. В 2012-13 годах CDC, CSTE и NACDD совместно провели серию обзоров, которые были основаны на предметном экспертном мнении для выработки рекомендаций по обновлению CDI. Цель этого обзора состояла в том, чтобы убедиться, что показатели хронических заболеваний реагируют на расширение приоритетов программ профилактики хронических заболеваний в государственных департаментах здравоохранения.

В результате CDI увеличился до 124 показателей в следующих 18 тематических группах:

- алкоголь;
- артрит;
- астма;
- рак;
- сердечно-сосудистые заболевания;
- хроническая болезнь почек;
- хроническое обструктивное заболевание легких;

- диабет;
- иммунизации;
- питание, физическая активность и весовой статус;
- здоровье ротовой полости;
- табачные изделия;
- общие условия;

и новые тематические области, которые включают:

- инвалидность;
- психическое здоровье;
- пожилые люди;
- репродуктивное здоровье;
- школьное здоровье.

Впервые индикаторы хронических заболеваний включают в себя 22 показателя систем и изменения окружающей среды.

1.4.2 CDC и CDI для исследования

CDI является примером сотрудничества между CDC и государственными департаментами здравоохранения в создании консенсусного набора показателей эпиднадзора, который следит за состоянием здоровья на уровне штатов. Это обновление реализовано для того, чтобы CDI оставался наиболее актуальной коллекцией данных эпиднадзора за хроническими заболеваниями для государственных эпидемиологов, должностных лиц программ по хроническим заболеваниям, а также должностных лиц по вопросам репродуктивного здоровья и охраны здоровья матери и ребенка. Стандартизированные определения показателей также будут способствовать согласованности эпиднадзора за хроническими заболеваниями на национальном, штатном и местном уровнях общественного здравоохранения.

Веб-сайт CDC позволяет создавать визуализацию, презентацию и многие другие операции с данными прямо на страницах сайта. С их помощью

возможно создавать гистограммы, столбчатые диаграммы, круговые диаграммы, графики, комбинированные и точечные диаграммы, а также просмотр данных по карте и календарю. Благодаря тому, что в данных присутствуют координаты мест сбора данных, то можно визуально, на карте видеть, просматривать и исследовать места сбора информации о хронических заболеваниях. Также сайт предоставляет возможность визуально просматривать данные по одному местоположению для всех показателей, по одному показателю для многих местоположений, определения индикаторов и сравнительный отчет (Рисунок 1.9). Такие функции позволяют анализировать и изучать нужную информацию в рамках одной страницы, а также говорят о том, что создатели продумали ходы, заинтересованных в данных сторон, и стараются для актуальности и популярности распространения борьбы с НИЗ.

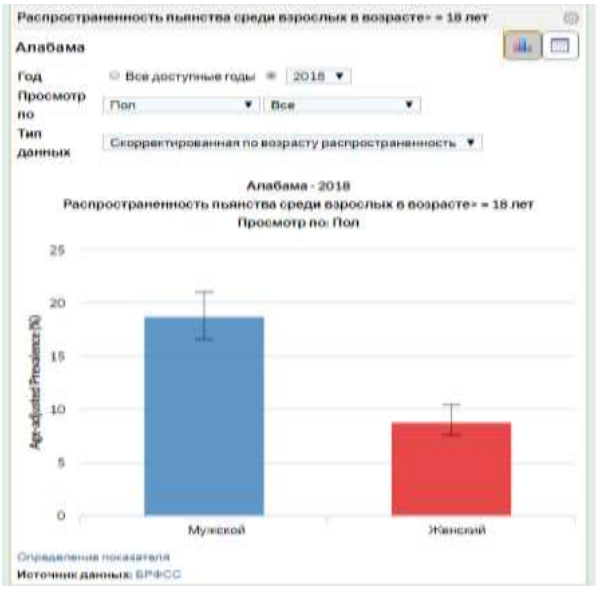

Рисунок 1.9 Визуальный просмотр данных на сайте CDC

Так же для исследования в качестве информации о показателях хронических заболеваний будут использованы «Рекомендации и отчеты» Отчета о заболеваемости и смертности (Morbidity and Mortality Weekly Report (MMWR)). [10]

В отчете представлены основные 18 группы показателей для наблюдения за хроническими заболеваниями Соединенных Штатов (Таблица 1.1)

Таблица 1.1

Сводка показателей хронических заболеваний по группам

Indicator group	Total	Population indicators	Systems and environmental indicators	Individual measures	Additional related measures under other topic groups
Alcohol	10	7	3	14	7
Arthritis	5	5	0	8	5
Asthma	6	6	0	12	4
Cancer	10	10	0	20	17
Cardiovascular disease	11	11	0	18	25
Chronic kidney disease	3	3	0	4	1
Chronic obstructive pulmonary disease	8	8	0	13	5
Diabetes	13	13	0	20	36
Disability	1	1	0	1	3
Immunization	1	1	0	1	16
Mental health	3	3	0	3	0
Nutrition, physical activity, and weight status	22	12	10	38	21
Older adults	4	4	0	5	33
Oral health	5	3	2	9	16
Overarching conditions	8	8	0	16	4
Reproductive health	3	3	0	3	36
School health	0	0	0	0	23
Tobacco	11	4	7	16	19
Total	124	102	22	201	—

В отчете содержится детально исследование каждого из 18 основных групп показателей и их подгруппы. Так, например, у группы Алкоголь существует 10 подгрупп, которые разделяются, по возрасту, по половому признаку и по другим критериям:

- Употребление алкоголя среди молодежи
- Употребление алкоголя до беременности
- Распространенность запоя среди молодежи
- Распространенность пьянства среди взрослых в возрасте ≥18 лет

- Распространенность пьянства среди женщин в возрасте 18-44 лет
- Частота запоя среди взрослых в возрасте ≥18 лет, употребляющих новый напиток
- Уровень запоя среди взрослых в возрасте ≥18 лет, которые пьют новый напиток
- Тяжелое пьянство среди взрослых в возрасте ≥18 лет
- Тяжелое пьянство среди женщин в возрасте 18-44 лет
- Хроническое заболевание печени
- Потребление алкоголя на душу населения среди лиц в возрасте ≥14 лет
- Размер акциза на алкоголь по видам напитков
- Коммерческий узел с ограниченной ответственностью
- Местные органы власти регулируют плотность производства алкоголя

1.4.3 Обзор данных

Данные, предоставляемые сайтом CDC в содружестве с Советом государственных и территориальных эпидемиологов (CSTE) и Национальной ассоциацией директоров по хроническим заболеваниям (NACDD) – это показатели хронических заболеваний (CDI) США. Данные предназначены для штатов, территорий и крупных городских районов США, включая 50 штатов и округ Колумбия, Гуам, Пуэрто-Рико и Виргинские острова США. Показатели хронических заболеваний записаны с мая 2016 по январь 2020 года, но сами значения данных идут с 2001 года, что говорит о большом количестве информации, за большой промежуток времени. Обновление происходило последний раз 28 января 2020 года, что свидетельствует о том, что информация из показателей является актуальной и обновленной.

Хронические заболевания разделяются по категориям и подкатегориям, а также имеют различные показатели (Таблица 1.2). Показатели для данной дипломной работы представляют собой набор индикаторов, показателей и

информации о НИЗ. Показатели и индикаторы были разработаны на основе консенсуса.

Таблица 1.2

Хронические заболевания и связанные с ними показатели

Категория	Подкатегория	Переменные (мера)	Определение
Хроническое состояние	Сахарный диабет	Сахарный диабет (%)	Распространенность диагностированного диабета среди взрослых в возрасте ≥18 лет - с 2018 по 2020 год
		Смертность от диабета (на 100 000)	Коэффициент смертности от диабета, указанный в качестве причины смерти, с 2018 по 2020 год
	Артрит	Артрит (%)	Распространенность артрита среди взрослых в возрасте ≥18 лет; 2018 - 2020 гг
	удушье	Астма (%)	Текущая распространенность астмы среди взрослых в возрасте ≥18 лет, в период с 2018 по 2020 год
		Смертность - астма (на 100 000 случаев)	Уровень смертности от астмы с 2018 по 2020 год
	Хроническая болезнь почек	Почки (%)	Распространенность хронического заболевания почек среди взрослых в возрасте ≥18 лет с 2018 по 2020 год
	Хроническое обструктивное заболевание легких	Легочный (%)	Распространенность хронической обструктивной болезни легких среди взрослых в возрасте ≥18 лет, в период с 2018 по 2020 год
		Смертность - легочная (на 100 000 случаев)	Смертность с хронической обструктивной болезнью легких как основной причиной среди взрослых в возрасте ≥45 лет, до 2014 и 2020 гг.
	Душевное здоровье	Умственный - женщины (%)	Общий показатель распространенности среди женщин в возрасте 18–44 лет, по крайней мере, в последние 14 дней в психически нездоровом состоянии, с с 2018 по 2020 год

Продолжение таблицы 1.2

	Послеродовой период (%)	Общий показатель распространенности послеродовых депрессивных симптомов в 2016 году
Алкоголь	Выпивка (%)	Распространенность пьянства среди взрослых в возрасте ≥18 лет, в период с 2018 по 2020 год
	Тяжелый напиток (%)	Пьянство среди взрослых в возрасте ≥18 лет, с 2018 по 2020 год
Питание, физическая активность и весовой статус	Физическая активность (%)	Нет физической активности в свободное время среди взрослых в возрасте ≥18 лет, с 2018 по 2020 год
	Табак (%)	Текущее курение среди взрослых в возрасте ≥18 лет, с 2018 по 2020 год
	Ожирение (%)	Ожирение среди взрослых в возрасте ≥18 лет, в период с 2018 по 2020 год
Пол	Пол (персонаж)	Мужской и женский
Этнос	Раса (персонаж)	Раса
Государственное местоположение	Расположение (персонаж)	50 штатов и округ Колумбия, Гуам, Пуэрто-Рико, Виргинские острова
Общие условия	Страхование (%)	Текущее отсутствие медицинского страхования среди взрослых в возрасте 18–64 лет, в период с 2018 по 2020 год
	Плохо - собственный уровень (%)	Удовлетворительная или плохая самооценка состояния здоровья среди взрослых в возрасте ≥18 лет
	Спать (%)	Распространенность достаточного количества сна среди взрослых в возрасте ≥18 лет

2 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

2.1 Технологии Big Data

Данная работа представляет собой исследование и уменьшение рисков показателей хронический заболеваний с помощью технологий больших данных (Big Data). Для того, чтобы провести это работу понадобятся методы, средства и технологии больших данных, самый популярный язык программирования – Python, а также пространство для работы – Jupyter Notebook. В данном исследовании будут анализироваться характеристики хронических заболеваний США, а также исследования по взаимосвязи между демографией, поведенческими привычками, полом, расовой принадлежностью и другими состояниями здоровья, и хроническими заболеваниями, а также будут учитываться рекомендации и отчеты MMWR. Таким образом, будет раскрыта основная задача данной дипломной работы и выполнены главные цели ВКР. Также в проведенной работе будет показана информация о здравоохранении и качестве жизни на уровне отдельных штатов и возможность использования данного исследования не только для индикаторов заболеваний соединённых штатов, но и для показателей других стан мира.

В этом исследовании, основанном на данных, используются технологии Big Data technologies. Данная технология – это методы, средства, подходы и инструменты для анализа и обработки информации из сложных или больших наборов данных. Данные могут быть структурированные и неструктурированные. Технология Big Data нужна для конкретных целей и задач, а также направлена на то, чтобы предоставить более эффективные способы понимания и анализа больших массивов данных. Нам нужны технологии обработки больших данных, чтобы анализировать огромные объемы данных в реальном времени и осуществлять выводы и прогнозы для снижения рисков в будущем. Технологии больших данных объединяет в себе аналитические возможности компьютера и способности специалистов-аналитиков, тем самым предлагая новые возможности контролирования

аналитических процессов и получения более конкретных и нужных результатов для дальнейшего эффективного использования.

Существует три принципа работы с большими данными: горизонтальная масштабируемость, отказоустойчивость, локальность данных.

Горизонтальная масштабируемость. Основной принцип обработки больших данных. С каждым днем большие данные становятся все более распространены, поэтому, нужно увеличить количество вычислительных узлов. Таким образом обработка будет происходить без вреда для производительности.

Отказоустойчивость. Это принцип вытекает из первого. Из-за того, что в секторе количество вычислительных узлов постоянно увеличивается, вероятность отказа машины также возрастает. Технологии больших данных должны предотвращать такие ситуации и предусматривать обязательные меры.

Локальность данных. Если данные находятся на одном сервере, а обрабатываются на другом, то соответственно, затраты на передачу данных будут огромными. Поэтому для того, чтобы затрат на передачу данных не было, желательно обрабатывать данные на том же компьютере, где они хранятся.

Все эти три принципа отличаются от привычных, традиционных централизованных вертикальных моделей, где производятся хранение хорошо структурированных данных. Именно поэтому для работы с большими данными, естественно, разрабатываются новые технологии и методы.

Big Data technologies применяет такие технологии, как инструменты бизнес-аналитики (BI), для объединения аналитических навыков человека с вычислительной мощностью ПК. Очевидно, что это исследование включает в себя, такие области, как визуализация, программирование, анализ данных, управление данными, объединение данных, эконометрик и статистика. Одним из ключевых понятий больших данных является то, что объединение этих разнообразных областей науки является самостоятельной научной дисциплиной.

В начале в Big Data включались средства массово-параллельной обработки структурированных данных, такие как СУБД NoSQL, алгоритмы MapReduce и Hadoop. Но с каждым годом число технологий увеличивается, они используются для агрегирования, манипулирования, управления и анализа больших массивов данных. Затем к этим технологиям стали относить и другие средства, похожие по характеристикам обработки больших массивов, данных и некоторые аппаратные средства. К таким новым добавлениям относятся, такие средства, как R — язык программирования для статистической обработки данных и работы с графикой, Аппаратные решения - корпорации Teradata, EMC и др. На данный момент можно насчитать около 24 средств технологий больших данных:

- **Big Table.**
- **Бизнес-аналитика (BI).**
- **Cassandra.** Система управления базами данных с открытым исходным кодом, предназначенная для обработки огромных объемов данных в распределенной среде. Эта система была первоначально разработана в Facebook и сейчас управляется как проект Apache Software foundation.
- **Облачные вычисления.** Вычислительная парадигма, в которой высоко масштабируемые вычисления, часто настроенные, как распределенная система, предоставляются как услуга через сеть.
- **Data mart.** Подмножество хранилища данных, обычно используемое для предоставления данных пользователям с помощью инструментов бизнес-аналитики.
- **Хранилище данных.** Специализированная база данных, оптимизированная для отчетности, часто используется для хранение больших объемов, структурированных данных.
- **Распределенная система.**
- **Dynamo.** Фирменная распределенная система хранения данных, разработанная компанией Amazon.

- **Извлечение, преобразование и загрузка (ETL).** Программные средства, используемые для извлечения данных из внешних источников, преобразует их в соответствии с оперативными потребностями и загружает в систему базы данных или хранилища данных.
- **Файловая Система Google.** Собственная распределенная файловая система, разработанная компанией Google.
- **Hadoop.** Программный фреймворк с открытым исходным кодом для обработки огромных массивов данных. Его разработка была совершена с помощью Google MapReduce и файловой системы Google. Первоначально он был разработан в компании Yahoo!, а теперь управляется как проект Apache Software Foundation.
- **HBase.** Открытая, распределенная, нереляционная база данных, смоделированная на основе Big Table Google. Первоначально он был разработан компанией Powerset, а теперь управляется как проект Apache Software foundation и как часть Hadoop.
- **MapReduce.** Программный фреймворк, представленный компанией Google для обработки огромных наборов данных**,** по определенным видам проблем в распределенной среде.
- **Mashup.** Приложение, которое использует и объединяет представленные данных или их функциональные возможности из двух или более источников для создания новых сервисов.
- **Метаданные.** Данные, описывающие содержание и контекст файлов данных, например, средства создания, цель, время и дата создания, а также автора.
- **Нереляционная база данных.** База данных, которая не хранит данные в таблицах. В отличие от реляционной базы данных.
- **R.** Язык программирования с открытым исходным кодом и программной средой для статистического вычисления и графика.

- **Реляционная база данных.** База данных, состоящая из набора таблиц, т. е., данные хранятся в строках и столбцах. Реляционные системы управления базами данных (СУБД) хранят тип структурированных данных. SQL является наиболее широко используемым языком для управление реляционными базами данных.

- **Полуструктурированные данные.** Данные, которые не соответствуют фиксированным полям, но содержат теги и другие маркеры для разделения элементов данных. Например, текст в формате XML или HTML.

- **SQL.** Первоначально аббревиатура для языка структурированных запросов, SQL - язык, предназначенный для управления данными в реляционных базах данных. Этот метод включает в себя возможность вставлять, запрашивать, обновлять и удалять данные, а также управлять ими.

- **Потоковая обработка.** Технологии, предназначенные для обработки больших потоков данных в реальном времени**.** Потоковая обработка позволяет использовать такие приложения, как алгоритмическая торговля финансовыми услугами, приложения для обработки событий RFID, обнаружение мошенничества, мониторинг технологических процессов и геолокационные услуги в области телекоммуникаций. Также известный как обработка потока событий.

- **Структурированные данные.** Данные, которые находятся в фиксированных полях. Примеры структурированных данных включают в себя реляционные базы данных или данные в электронных таблицах.

- **Неструктурированные данные.** Данные, которые не находятся в фиксированных полях.

- **Визуализация.** Технологии, используемые для создания изображений, диаграмм или анимаций, Представление информации таким образом, чтобы люди могли эффективно ее понимать.

2.1.1 Методы Big Data

Существует много методов и техник анализа больших данных, которые опираются на такие дисциплины, как статистика и информатика (особенно машинное обучение), которые могут быть использованы для анализа данных. Исследователи и программисты продолжают искать и развивать новые методы и совершенствовать существующие. Далеко не все эти методы строго требуют использования больших данных — некоторые из них могут быть эффективно применены к меньшим наборам данных (например, A/B тестирование, регрессионный анализ). Однако все методы, могут быть применены к большим данным и, как правило, более крупные и разнообразные наборы данных могут быть использованы для генерировать более многочисленные и точных результатов, чем меньшие, менее разнообразные.

Можно выделить 24 категорий методов, применимых в различных областях науки и промышленности:

- **А / Б тестирование.** Маркетинговая методика, использующаяся для оценки и управления эффективностью веб-страницы. Этот метод также известен как сплит-тестирование. Примером является определение того, что добавление текста, макетов, изображения, или цвета улучшат показатели конверсии на веб-сайте.

А / Б тестирование - это практика одновременного показа двух вариантов одной и той же веб-страницы разным сегментам посетителей и сравнения того, какой вариант вызывает больше конверсий. Как правило, тот, который дает более высокие конверсии, является выигрышным вариантом, который может помочь оптимизировать сайт для достижения лучших результатов.

- **Изучение правил ассоциации.** Эти методы состоят из множества алгоритмов для генерации и проверки возможных правил. Одним из применений является анализ рыночной корзины, в котором розничный торговец может определить, какие продукты часто покупаются вместе. А так

же использовать эту информацию для маркетинга. Например, многие покупатели супермаркетов, которые покупают подгузники тоже склонны покупать пиво. Используется для интеллектуального анализа данных.

- **Классификация.** Набор контролируемых, технологических средств, методов и подходов для определения категорий, в которых алгоритм учится на основе предоставленных ему данных, а затем использует это обучение для классификации новых наблюдений. Одним из применений является прогнозирование конкретного сегмента клиента. Например, решения о покупке, уровень оттока, уровень потребления, когда существует ясная гипотеза или объективный результат. Эти методы часто описываются как контролируемое обучение из-за наличия обучающего набора.
- **Ensemble обучения.** Это использование нескольких прогностических моделей, каждая из которых разработана с использованием статистики и (или) машинного обучения, для получения более высокой производительности. Это один из видов контролируемого обучения.
- **Генетический алгоритм.** Это метод, используемый для оптимизации. Основанием создания, послужил процесс естественной эволюции. Структура данных генетического алгоритма состоит из одной или большее количество хромосом (обычно из одной). Как правило, хромосома - это битовая строка, так что термин строка часто заменяет понятие "хромосома". Хромосомы могут объединяться и мутировать. Эти строки отбираются для выживания в моделируемой среде обитания, что определяет пригодность. Эти алгоритмы хорошо подходят для решение нелинейных задач. Например, улучшение планирования заданий в производстве и оптимизации показателей инвестиционного портфеля.
- **Методы класса Data Mining.** Это процесс поиска данных, аномалий, интеллектуальный анализ данных, анализ закономерностей и корреляций в больших наборах данных для прогнозирования результатов. Совокупность методов и технологий обнаружения в данных аномальных, ранее неизвестных полезных знаний, необходимых для принятия

решений и увеличения доходов, сокращения расходов, улучшения отношений с клиентами, снижения рисков и многого другого. К таким методам, в частности, относятся обучение ассоциативным правилам, классификация, кластерный анализ, регрессионный анализ, обнаружение и анализ отклонений и др.

- **Краудсорсинг.** Это практика привлечения «толпы» или группы для достижения общей цели, выполняющих эту работу без вступления в трудовые отношения.
- **Смешение и интеграция данных.** Комбинируя набор методов, которые анализируют и интегрируют данные из нескольких источников и решений, Аналитика становится более эффективной и потенциально более точной, если бы она разрабатывалась с использованием комбинированного набора методов
- **Машинное обучение.** Использование моделей, построенных на базе статистического анализа или машинного обучения для получения общего прогнозов.
- **Обработка естественного языка (NLP).** Набор техник, средств и методов по информатике. В рамках этой области, исторически называется "искусственным интеллектом". Метод создан для лингвистики, которая использует компьютерные алгоритмы для анализа человеческого языка. Многие техники NLP являются разновидностями машинного обучения. Одним из применений NLP является использование анализа настроений в социальных сетях для определения того, насколько перспективны клиенты, реагирующие на рекламную кампанию.
- **Искусственные нейронные сети.** Это сетевой анализ, который был придуман учеными, вдохновленными структурой и работами биологических нейронных сетей (то есть клеток и связей внутри мозга).
- **Распознавание образов.** Этот процесс может быть определен, как классификация данных на основе уже полученных знаний или статистической информации, извлеченной из образцов и / или их представления. Одним из важных аспектов распознавания образов является его осуществимость и

эффективность. Область распознавания образов связана с автоматическим обнаружением закономерностей в данных с помощью компьютерных алгоритмов и с помощью этих закономерностей для выполнения таких действий, как классификация данных по различным категориям.

- **Прогнозная аналитика.** Набор методов, с помощью которых создается или выбирается математическая модель, позволяющая наилучшим образом предсказать вероятность того или иного исхода.
- **Имитационное моделирование.** это важный метод анализа, позволяющий строить модели. В разных отраслях и дисциплинах имитационное моделирование предоставляет ценные решения, давая четкое представление о сложных системах. Имитационное моделирование можно рассматривать как разновидность экспериментальных испытаний.
- **Пространственный анализ.** Набор методов, некоторые из которых применяются из статистики, которые анализируют топологические, геометрические или географические свойства, закодированные в наборе данных. Часто данные для пространственного анализа поступают из геоинформационных систем (ГИС), которые включают информацию о местоположение.
- **Кластерный анализ.** Статистический метод классификации объектов, разбиение множества объектов на более мелкие группы сходные группы, или кластеры.
- **Регрессионный анализ.** На базовом уровне регрессионный анализ включает манипулирование некоторой независимой переменной, чтобы увидеть, как она влияет на зависимую переменную. Он описывает, как изменяется значение зависимой переменной при изменении независимой переменной. Лучше всего работает с непрерывными количественными данными, такими как вес, скорость или возраст.
- **Анализ временных рядов.** Набор методов для анализа последовательностей точек данных, представляющих значения в моменты времени для извлечения значимой информации и характеристик. Примером

анализа временных рядов, может служить количество диагностированных пациентов с определенным заболеванием каждый день.

Прогнозирование временных рядов - это использование модели, для прогнозирования будущих значений временного ряда, на основе известных значений из прошлого.

- **Статистика.** Этот метод работает для сбора, организации и интерпретации данных в рамках опросов и экспериментов.
- **Оптимизация.** Набор численных методов, используемых для получения лучшего варианта из всех возможных, для достижения определенной цели. Примеры применения включают в себя улучшение операционных процессов, таких как планирование, маршрутизация и планировка этажей, а также принятие стратегических решений.
- **Визуализация аналитических данных.** Это представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа. Очень важный этап анализа больших данных, позволяющий представить самые важные результаты анализа в наиболее удобном для восприятия виде.

Такой метод, как визуализация аналитических данных, помогает в анализе больших объемов данных и в получении информации для принятия обоснованных решений в отношении хронических заболеваний. Методы визуализации предлагают новый подход к старой аналитике. Он позволяет комбинировать свои базовые знания и креативность с огромными возможностями хранения и обработки современных компьютеров, чтобы получить представление о сложных проблемах. Аналитическая визуализация – это наука аналитического мышления, поддерживаемая интерактивными визуальными интерфейсами. Они сочетают статистические методы и модели с передовыми методами интерактивной визуализации, чтобы помочь скрыть сложность больших массивов данных, о состоянии здоровья и принять обоснованные решения.

- **Бесконтрольное обучение.** При неконтролируемом обучении модели глубокого обучения передается набор данных без четких инструкций о том, что с ним делать. Набор обучающих данных представляет собой набор примеров без определенного желаемого результата или правильного ответа. Затем нейронная сеть пытается автоматически найти структуру данных, извлекая полезные функции и анализируя ее структуру.
- **Контролируемое обучение.** Наличие полного набора помеченных данных при обучении алгоритму. Полностью помеченный означает, что каждый пример в наборе обучающих данных помечен ответом, который алгоритм должен придумать самостоятельно. Когда показано новое изображение, модель сравнивает его с примерами обучения, чтобы предсказать правильную метку.

2.1.2 Процессы технологии Data Mining

Data Mining – это интеллектуальный анализ данных, который анализирует наборы данных для выявления взаимосвязей, новых корреляций и тенденций, а также для извлечения полезных данных в форме шаблонов. Таким образом, этот процесс используется для определения допустимых, ценных и понятных форм данных.

Данные могут быть определены как любой факт, число или текст, который может быть обработан компьютером. Data Mining - это технология поиска информации и знаний. Эта технология возникла в 1950-х годах.

Все процессы интеллектуального анализа данных не могут быть завершены за один шаг. Другими словами, невозможно легко и быстро получить необходимую информацию из больших объемов данных. Все процессы Data Mining должны выполняться в указанном порядке.

Традиционный процесс Data Mining включает следующие этапы:

- Анализ предметной области.
- Постановка задачи.
- Подготовка данных.
- построение моделей;

- проверка и оценка моделей;
- выбор модели;
- применение модели;
- коррекция и обновление модели.

2.1.1.1 Анализ предметной области

Исследование — это процесс познания определенной предметной области, объекта или явления с определенной целью. [11]

Процесс исследования заключается в наблюдении свойств объектов с целью выявления и оценки важных, с точки зрения субъекта-исследователя, закономерных отношений между показателями данных свойств.

Предметная область — это мысленно ограниченная область реальной действительности, подлежащая описанию или моделированию и исследованию. [12]

Предметная область состоит из объектов, различаемых по свойствам и находящихся в определенных отношениях между собой или взаимодействующих каким-либо образом.

2.1.1.2 Постановка задачи

Постановка задачи Data Mining включает следующие формулировку задачи и формализацию задачи. Постановка задачи включает также описание статического и динамического поведения исследуемых объектов.

Описание статики подразумевает описание объектов и их свойств. При описании динамики описывается поведение объектов и те причины, которые влияют на их поведение. Динамика поведения объектов часто описывается вместе со статикой.

Технология Data Mining не может заменить аналитика и ответить на те вопросы, которые не были заданы. Поэтому постановка задачи является необходимым этапом процесса Data Mining. Иногда этапы анализа предметной области и постановки задачи объединяют в один этап. [13]

2.1.1.3 Подготовка данных

Цель этапа подготовки данных – разработка базы данных для Data Mining.

Подготовка данных является важнейшим этапом, от качества выполнения, которого зависит возможность получения качественных результатов всего процесса Data Mining. Кроме того, следует помнить, что на этап подготовки данных, по некоторым оценкам, может быть потрачено до 80% всего времени, отведенного на проект.

Сначала необходимо определить требования к данным. На этом этапе осуществляется так называемое моделирование данных, т. е. определение и анализ требований к данным, которые необходимы для осуществления Data Mining. При этом изучаются вопросы распределения пользователей (географическое, организационное, функциональное); вопросы доступа к данным, которые необходимы для анализа, необходимость во внешних и/или внутренних источниках данных; а также аналитические характеристики системы (измерения данных, основные виды выходных документов, последовательность преобразования информации и др.).

Затем происходит этап сбора данных. Наличие в организации хранилища данных делает анализ проще и эффективней, его использование, с точки зрения вложений, обходится дешевле, чем использование отдельных баз данных или витрин данных. Однако далеко не все предприятия оснащены хранилищами данных. В этом случае источником для исходных данных являются оперативные, справочные и архивные БД.

Если данные упорядочены, и мы имеем дело с временными рядами, желательно знать, включает ли такой набор данных сезонную/цикличную компоненту. В случае присутствия в наборе данных сезонной/цикличной компоненты, необходимо иметь данные как минимум за один сезон/цикл.

Если данные не упорядочены, то есть события из набора данных не связаны по времени, в ходе сбора данных следует соблюдать следующие правила.

Количество записей в наборе. Недостаточное количество записей в наборе данных может стать причиной построения некорректной модели. С точки зрения статистики, точность модели увеличивается с увеличением количества исследуемых данных. Возможно, некоторые данные являются устаревшими или описывают какую-то нетипичную ситуацию, и их нужно исключить из базы данных. Алгоритмы, используемые для построения моделей на сверхбольших базах данных, должны быть масштабируемыми.

Соотношение количества записей в наборе и количества входных переменных. При использовании многих алгоритмов необходимо определенное (желательное) соотношение входных переменных и количества наблюдений. Количество записей (примеров) в наборе данных должно быть значительно больше количества факторов (переменных).

Набор данных должен быть репрезентативным и представлять как можно больше возможных ситуаций. Однако, при наличии больших объемов данных следует сделать выборку, так как избыточное количество данных может оказать негативное влияние на восприятие общей картины. Пропорции представления различных примеров в наборе данных должны соответствовать реальной ситуации.

На следующем этапе происходит предварительная обработка данных. Анализировать можно как качественные, так и некачественные данные. Результат будет достигнут и в том, и в другом случае. Для обеспечения качественного анализа необходимо проведение предварительной обработки данных, которая является необходимым этапом процесса Data Mining.

Качество данных — это критерий, определяющий полноту, точность, своевременность и возможность интерпретации данных.

Данные могут быть высокого качества и низкого качества, последние — это так называемые «грязные данные». Данные высокого качества — это совершенные, истинные, полные, своевременные данные, которые поддаются интерпретации. Такие данные обеспечивают получение качественного

результата — знаний, которые смогут поддерживать процесс принятия решений.

Данные низкого качества, или грязные данные — это отсутствующие, неточные или бесполезные данные с точки зрения практического применения. [14]

Очевидно, что результаты Data Mining на основе грязных данных не могут считаться надежными и полезными. Однако наличие таких данных не обязательно означает необходимость их очистки или же предотвращения появления. Всегда должен быть разумный выбор между наличием грязных данных и стоимостью и/или временем, необходимым для их очистки.

Проблемы с качеством встречаются в отдельных наборах данных – таких как файлы и базы данных. Когда интеграции подлежит множество источников данных, необходимость в очистке данных существенно возрастает. Для обеспечения доступа к точным и согласованным данным необходима консолидация различных представлений данных и исключение дублирующийся информации. Специальные средства очистки обычно имеют дело с конкретными областями или же с исключением дубликатов. Преобразования обеспечиваются либо в форме библиотеки правил, либо пользователем в интерактивном режиме.

Преобразования данных могут быть автоматически получены с помощью средств согласования схемы.

Метод очистки данных должен удовлетворять ряду критериев.

1. Он должен выявлять и удалять все основные ошибки и несоответствия, как в отдельных источниках данных, так и при интеграции нескольких источников.

2. Метод должен поддерживаться определенными инструментами, чтобы сократить объемы ручной проверки и программирования, и быть гибким в плане работы с дополнительными источниками.

3. Очистка данных не должна производиться в отрыве от связанных со схемой преобразования данных, выполняемых на основе сложных метаданных.

4. Функции «маппирования» для очистки и других преобразований данных должны быть определены декларативным образом и подходить для использования в других источниках данных и в обработке запросов.

5. Инфраструктура технологического процесса должна особенно интенсивно поддерживаться для хранилищ данных, обеспечивая эффективное и надежное выполнение всех этапов преобразования для множества источников и больших наборов данных.

На сегодняшний день интерес к очистке данных возрастает. Целый ряд исследовательских групп занимается общими проблемами, связанными с очисткой данных, в том числе, со специфическими подходами к Data Mining и преобразованию данных на основании сопоставления схемы. В последнее время некоторые исследования коснулись единого, более сложного подхода к очистке данных, включающего ряд аспектов преобразования данных, специфических операторов и их реализации.

В целом очистка данных включает следующие этапы:

1. Анализ данных.

2. Определение порядка и правил преобразования данных.

3. Подтверждение.

4. Преобразования.

5. Противоток очищенных данных.

Такой процесс преобразования требует больших объемов метаданных (схем, характеристик данных уровня схемы, определений технологического процесса и др.). Для согласованности, гибкости и упрощения использования в других случаях, эти метаданные должны храниться в репозитории на основе СУБД. Для поддержки качества данных подробная информация о процессе преобразования должна записываться как в репозиторий, так и в трансформированные элементы данных, в особенности информация о полноте и свежести исходных данных и происхождения информации о первоисточнике трансформированных объектов и произведенных с ними изменениях. [15]

2.1.1.4 Построение модели

После окончания этапа подготовки данных можно переходить к построению модели. Для построения моделей потребуется набор данных для обучения алгоритма и еще набор для оценки модели, построенной алгоритмом, а также используются различные методы и алгоритмы Data Mining. Некоторые задачи могут быть решены при помощи моделей, построенных на основе различных методов.

Идеальной модели, которая бы позволила решать разнообразные задачи, не существует. Поэтому многие разработчики включают в инструменты Data Mining возможность построения различных моделей, многие также обеспечивают возможность расширяемости моделей. Некоторые инструменты Data Mining создаются специально для конкретных областей применения.

Среди большого разнообразия методов Data Mining должен быть выбран метод или же комбинация методов, при использовании которых построенная модель будет наилучшим образом описывать исследуемый объект. Иногда для выявления искомых закономерностей требуется использование нескольких методов и алгоритмов. В таком случае одни методы используются в начале моделирования, другие – на дальнейших этапах.

2.1.1.5 Проверка и оценка модели

Проверка модели подразумевает проверку ее достоверности или адекватности. Эта проверка заключается в определении степени соответствия модели реальности. Адекватность модели проверяется путем тестирования.

Адекватность модели – соответствие модели моделируемому объекту или процессу. [16]

Понятия достоверности и адекватности являются условными, поскольку мы не можем рассчитывать на полное соответствие модели реальному объекту, иначе это был бы сам объект, а не модель. Поэтому в процессе моделирования следует учитывать адекватность не модели вообще, а именно тех ее свойств, которые являются существенными с точки зрения проводимого исследования. В процессе проверки модели необходимо установить включение в модель всех

существенных факторов. Сложность решения этой проблемы зависит от сложности решаемой задачи.

Проверка модели также подразумевает определение той степени, в которой она действительно помогает менеджеру при принятии решений.

Оценка модели подразумевает проверку ее правильности. Оценка построенной модели осуществляется путем ее тестирования.

Тестирование модели заключается в «прогонке» построенной модели, заполненной данными, с целью определения ее характеристик, а также в проверке ее работоспособности. Тестирование модели включает в себя проведение множества экспериментов. На вход модели могут подаваться выборки различного объема. С точки зрения статистики, точность модели увеличивается с увеличением количества исследуемых данных. Алгоритмы, являющиеся основой для построения моделей на сверхбольших базах данных, должны обладать свойством масштабирования.

Для оценки результатов полученных моделей следует использовать знания специалистов предметной области. Если результаты полученной модели эксперт считает неудовлетворительными, следует вернуться на один из предыдущих шагов процесса Data Mining, а именно: подготовка данных, построение модели, выбор модели. Если же результаты моделирования эксперт считает приемлемыми, ее можно применять для решения реальных задач.

2.1.1.6 Выбор модели

Если в результате моделирования было построено несколько различных моделей, то на основании их оценки мы можем осуществить выбор лучшей из них. В ходе проверки и оценки различных моделей на основании их характеристик, а также с учетом мнения экспертов, следует выбор наилучшей.

В некоторых программных продуктах реализован ряд методов, разработанных для выбора модели. Многие из них основаны на так называемой «конкурентной оценке моделей», которая состоит в применении различных моделей к одному и тому же набору данных и последующем сравнении их характеристик.

2.1.1.7 Применение модели

После тестирования, оценки и выбора модели следует этап применения модели. На этом этапе выбранная модель используется применительно к новым данным с целью решения задач, поставленных в начале процесса Data Mining. Для классификационных и прогнозирующих моделей на этом этапе прогнозируется целевой (выходной) атрибут.

2.1.1.8 Коррекция и обновление модели

По прошествии определенного установленного промежутка времени с момента начала использования модели Data Mining следует проанализировать полученные результаты, определить, действительно ли она "успешна" или же возникли проблемы и сложности в ее использовании.

Существует много причин, требующих обучить модель заново, т.е. обновить ее, чтобы отразить определенные изменения. Основными причинами являются следующие:

- изменились входящие данные или их поведение;
- появились дополнительные данные для обучения;
- изменились требования к форме и количеству выходных данных;
- изменились цели бизнеса, которые повлияли на критерии принятия решений;
- изменилось внешнее окружение или среда (макроэкономика, политическая ситуация, научно-технический прогресс, появление новых конкурентов и товаров и т. д.).

2.1.3 Методы и задачи Data Mining

Существует много методов, используемых для интеллектуального анализа данных, но важнейшим шагом является выбор подходящего метода, в соответствии с формулировкой проблемы. Важно, что Data Mining, по сравнению с другими информационными технологиями, отличается нетривиальностью поиска данных. Все методы помогают предсказывать

будущее и затем принимать соответствующие решения. Они также помогают анализировать рыночные тенденции и увеличивать доходы компании.

Рассмотрим основные методы Data Mining и их приложения в конкретных системах.

1. Статистические методы. К базовым методам Data Mining традиционно причисляют все подходы, использующие элементы теории статистики. Последние версии почти всех известных статистических пакетов включают наряду с традиционными статистическими методами также элементы Data Mining. В качестве примеров наиболее мощных и распространенных статистических пакетов можно назвать SAS (компания SAS Institute), SPSS (компания SPSS), STATGRAPICS (компания Manugistics), STATISTICA, STADIA и другие.

2. Полный и ограниченный перебор. К базовым методам Data Mining принято относить также алгоритмы, основанные на переборе. Простой перебор всех исследуемых объектов требует $O(2^N)$ операций, где N – количество объектов. Следовательно, с увеличением количества данных объем вычислений растет экспоненциально, что при большом объеме делает решение любой задачи таким методом практически невозможным. Наиболее ярким современным представителем реализации этого подхода является система WizWhy компании WizSoft.

3. Нечеткая логика. Основным способом исследования задач анализа данных является их отображение на формализованный язык и последующий анализ полученной модели. Неопределенность по объему отсутствующей информации у системного аналитика можно разделить на три большие группы:

1) неизвестность,

2) неполнота (недостаточность, неадекватность),

3) недостоверность.

Недостоверность бывает физической (источником ее является внешняя среда) и лингвистической (возникает в результате словесного обобщения и обусловливается необходимостью описания бесконечного числа ситуаций

ограниченным числом слов за ограниченное время). Основной сферой применения нечеткой логики и во многом остается управление.

4. Генетические алгоритмы. Генетические алгоритмы относятся к числу универсальных методов оптимизации, позволяющих решать задачи различных типов (комбинаторные, общие задачи с ограничениями и без ограничений) и различной степени сложности. Одним из наиболее востребованных приложений генетического алгоритма в области Data Mining является поиск наиболее оптимальной модели (поиск алгоритма, соответствующего специфике конкретной области). Эти алгоритмы удобны тем, 2 что их легко распараллеливать. Пример системы, построенной на основе технологии Data Mining с применением генетических алгоритмов: система GeneHunter фирмы Ward Systems Group.

5. Нейронные сети. Нейронные сети – это класс моделей, основанных на биологической аналогии с мозгом человека и предназначенных для решения разнообразных задач анализа данных после прохождения этапа, так называемого обучения на имеющихся данных. При применении этого метода, прежде всего, встает вопрос выбора конкретной архитектуры сети (числа "слоев" и количества "нейронов" в каждом из них). Размер и структура сети должны соответствовать существу исследуемого явления. Примеры нейросетевых систем: BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic).

6. Деревья решений. Деревья решения являются одним из наиболее популярных подходов к решению задачи классификации. Они создают иерархическую структуру классифицирующих правил типа "если-то" (if-then), имеющую вид дерева. Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Большинство систем, построенных на основе технологии Data Mining, используют именно этот метод. Самыми известными являются See5/C5.0 (RuleQuest, Австралия), Clementine (Integral

Solutions, Великобритания), SIPINA (University of Lyon, Франция), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада).

Методы Data Mining помогают решить многие задачи, с которыми сталкивается аналитик. Из них основными являются: задача классификации и регрессии, задача поиска ассоциативных правил и задача кластеризации.

В основу технологии Data Mining положена концепция шаблонов, представляющих собой закономерности. В результате обнаружения этих, скрытых от невооруженного глаза закономерностей решаются задачи Data Mining. Различным типам закономерностей, которые могут быть выражены в форме, понятной человеку, соответствуют определенные задачи Data Mining.

Единого мнения относительно того, какие задачи следует относить к Data Mining, нет. Большинство авторитетных источников перечисляют следующие:

- классификация,
- кластеризация,
- прогнозирование,
- ассоциация,
- визуализация,
- анализ и обнаружение отклонений,
- оценивание,
- анализ связей,
- подведение итогов.

Классификация – наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; по этим признакам новый объект можно отнести к тому или иному классу.

Для решения задачи классификации могут использоваться методы:

- ближайшего соседа;

- k-ближайшего соседа;
- байесовские сети;
- индукция деревьев решений;
- нейронные сети.

Кластеризация является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на группы.

Ассоциация заключается в поиске ассоциативных правил, в ходе которого отыскиваются закономерности между связанными событиями в наборе данных.

Отличие ассоциации от двух предыдущих задач Data Mining: поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

Наиболее известный алгоритм решения задачи поиска ассоциативных правил – алгоритм «Apriori».

Последовательность позволяет найти временные закономерности между транзакциями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени (т.е. происходящими с некоторым определенным интервалом во времени). Другими словами, последовательность определяется высокой вероятностью цепочки связанных во времени событий. Фактически, ассоциация является частным случаем последовательности с временным лагом, равным нулю. Эту задачу Data Mining также называют задачей нахождения последовательных шаблонов.

В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей. Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

Краткое описание. Цель решения данной задачи определения отклонений или выбросов заключается в обнаружении и анализе данных, наиболее

отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

Задача оценивания сводится к предсказанию непрерывных значений признака.

Задача анализа связи – это задача нахождения зависимостей в наборе данных.

В результате визуализации создается графический образ анализируемых данных. Для решения задачи визуализации используются графические методы, показывающие наличие закономерностей в данных.

Подведение итогов – задача, цель которой – описание конкретных групп объектов из анализируемого набора данных.

2.1.4 Большие данные в промышленности

Большие данные стали ключевой основой конкуренции, которая служит основой для новых волн роста производительности, инноваций и потребительского излишка - при условии наличия правильной политики и инструментов.

Согласно отчету компании McKinsey «Global Institute, Big data: The next frontier for innovation, competition, and productivity», данные стали таким же важным фактором производства, как трудовые ресурсы и производственные активы [17]. За счет использования больших данных компании могут получать ощутимые конкурентные преимущества. Технологии Big Data могут быть полезными при решении следующих задач:

- прогнозирование рыночной ситуации
- маркетинг и оптимизация продаж
- совершенствование продукции
- принятие управленческих решений
- повышение производительности труда
- эффективная логистика
- мониторинг состояния основных фондов

На производственных предприятиях большие данные генерируются также вследствие внедрения технологий Промышленного интернета вещей. В ходе этого процесса основные узлы и детали станков и машин снабжаются датчиками, исполнительными устройствами, контроллерами и, иногда, недорогими процессорами, способными производить граничные (туманные) вычисления. В ходе производственного процесса осуществляется постоянный сбор данных и, возможно, их предварительная обработка (например, фильтрация). Аналитические платформы обрабатывают эти массивы информации в режиме реального времени, представляют результаты в наиболее удобном для восприятия виде и сохраняют для дальнейшего использования. На основе анализа полученных данных делаются выводы о состоянии оборудования, эффективности его работы, качестве выпускаемой продукции, необходимости внесения изменений в технологические процессы и т.д.

Благодаря мониторингу информации в режиме реального времени персонал предприятия может:

- сокращать количество простоев
- повышать производительность оборудования
- уменьшать расходы на эксплуатацию оборудования
- предотвращать несчастные случаи

Последний пункт особенно важен. Например, операторы, работающие на предприятиях нефтехимической промышленности, получают в среднем около 1500 аварийных сообщений в день, то есть более одного сообщения в минуту. Это приводит к повышенной усталости операторов, которым приходится постоянно принимать мгновенные решения о том, как реагировать на тот или иной сигнал. Но аналитическая платформа может отфильтровать второстепенную информацию, и тогда операторы получают возможность сосредоточиться в первую очередь на критических ситуациях. Это позволяет им более эффективно выявлять и предотвращать аварии и, возможно, несчастные случаи. В результате повышаются уровни надежности

производства, промышленной безопасности, готовности технологического оборудования, соответствия нормативным требованиям. [18]

Кроме того, по результатам анализа больших данных можно рассчитывать сроки окупаемости оборудования, перспективы изменения технологических режимов, сокращения или перераспределения обслуживающего персонала — т.е. принимать стратегические решения относительно дальнейшего развития предприятия.

2.2 Python

2.2.1 Анализ конкурентных языков программирования Python, R и Scala.

R, Python и Scala являются тремя основными языками для обработки данных и интеллектуального анализа данных. Важно и нужно разобраться со всеми их плюсами и минусами.

Организации и предприятия всех размеров могут анализировать большие запасы неструктурированных и структурированных данных, которые ежедневно пополняются, с целью выявления тенденций, моделей и корреляций. Такой анализ приведет к принятию более эффективных решений и получению дополнительных знаний.

Хранилище данных связано с аналитикой больших данных в том смысле, что оно также является важным фактором аналитики. В хранилище данных несколько источников корпоративных данных интегрированы в централизованное хранилище для отчетности, анализа и принятия решений.

Несмотря на то, что системы больших данных и системы хранилищ данных обычно различаются, некоторые хранилища данных SQL могут быть полезны для анализа больших данных, включая Cloudera Impala с открытым исходным кодом, Apache Hive и Apache Spark.

Самые популярные языки программирования для анализа больших данных:

R - это язык программирования, используется в основном для статистического анализа. Существует ряд пакетов для R, известных как

«Программирование с большими данными в R» (pbdR), которые облегчают анализ больших данных, распределенных по нескольким системам, с использованием кода R.

Гибкость R является сильной стороной, потому что вы можете работать практически на всех операционных системах. Кроме того, R обладает отличными графическими возможностями, которые могут оказаться полезными при попытке визуализировать шаблоны и ассоциации в системах больших данных.

Тем не менее, R не является языком общего назначения, поэтому разработчикам и ученым, работающим с данными, может быть трудно разобраться с ним по сравнению с более традиционным языком программирования.

Scala - это язык программирования общего назначения, созданный частично с целью решения некоторых основных проблем языка Java. Решение для кластерных вычислений Apache Spark написано на Scala, что объясняет популярность этого языка в науке о данных, особенно в анализе больших данных.

Scala раньше был обязательным для работы со Spark, но это было решено с помощью открытия конечных точек API, доступных на других языках. Scala имеет превосходную поддержку параллелизма, которая необходима для распараллеливания большого объема обработки, необходимой для больших наборов данных. Scala работает на виртуальной машине Java (JVM), что делает его идеальным для использования с такой средой, как Apache Hadoop.

Python - это скорее язык программирования общего назначения. Python также проще в освоении, и есть несколько отличных, полностью бесплатных онлайн-учебников, которые знакомят с основами. Python считается «связующим» языком, что означает, что он хорош, когда задачи анализа данных требуют интеграции с веб-приложениями.

Python - самый популярный язык, используемый исследователями данных для изучения больших данных, благодаря множеству полезных инструментов и

библиотек, таких как pandas и matplotlib. Python также обладает отличной производительностью и масштабируемостью для задач по обработке данных, и его можно использовать с быстрыми движками больших данных, такими как Apache Spark, через доступный API-интерфейс Python.

Было решено использовать Python, как самый популярный и удобный язык. Он используется с 1991 года для веб-разработок, разработок программного обеспечения, системных скриптов, а также для математики. Python может использоваться для:

- создания веб-приложений,
- обработки больших данных,
- выполнения сложной математики,
- разработки программного обеспечения,

а также он может подключаться к системам баз данных, читать файлы и изменять их. Этот язык программирования удобен в эксплуатации, так как работает на разных операционных системах, таких как Windows, Mac, Linux, Raspberry Pi и т.д. Python имеет простой синтаксис, похожий на английский язык, который позволяет разработчикам писать программы с меньшим количеством строк, чем некоторые другие языки программирования. Так же язык работает в системе интерпретатора, что означает, что команда может быть выполнена, как только написана. С Python можно обращаться объектно-ориентированным или функциональным образом.

В последнее время Python становится все более популярным инструментом для анализа данных. Язык занимается первоочередными задачами манипулирования, очистки и обработки данных в Python. Так же у Python есть ряд преимуществ:

- Можно осуществить не только обработку данных, но также их поиск и использование результата обработки в веб-приложении.
- Python 3 является одним из тех языков, которые отлично подходят на роль первого языка программирования.

- Прост и легок в изучении.
- Python 3 входит в пятерку самых популярных языков программирования.
- Постоянно развивается, обновляет и дополняется.
- В Python 3 встроен фреймворк для тестирования, входной барьер у которого очень низок.
- Фреймворк обеспечивает хорошее тестовое покрытие.

Таким образом, код будет надежен и удобен для многократного использования.

2.2.2 Средства Python для анализа данных

Python позволяет решить задачу автоматизации сбора данных, обработки данных, ускоряет анализ данных и позволяет реализовать на работе новые подходы к анализу, например, решать задачи с помощью обучения нейросетей**.**

2.2.2.1 Библиотеки Python и структуры данных

В Python 3 имеется четыре встроенных типа данных — *списки* (list), *кортежи* (turple), *словари* (dictionary) и *множества* (set).

Списки (list) - эти списки являются одной из самых универсальных структур данных в Python. Список можно просто определить, написав список значений через запятую в квадратных скобках. Списки могут содержать элементы разных типов, но обычно все элементы имеют одинаковый тип. Списки Python являются изменяемыми.

Кортежи (tuple) - кортежи представлены несколькими значениями, разделенными запятыми. Кортежи неизменны, а выходные данные заключены в круглые скобки, поэтому вложенные кортежи обрабатываются правильно. Хоть и кортежи являются неизменяемыми, они могут содержать изменяемые данные, если это необходимо.

Поскольку кортежи являются неизменяемыми и не могут изменяться, они быстрее обрабатываются по сравнению со списками.

Словари (dictionary) - представляет собой «записную книгу» — адрес человека зная только его имя: в словаре ключ (key) ассоциируется

со *значением* (value). Ключи должны быть уникальны. Пара скобок создает пустой словарь.

Множества (set) - Множества являются неупорядоченными коллекциями простых объектов, и используются в тех случаях, когда присутсвие объекта в коллекции важнее, чем порядок или количество вхождений этого элемента в одной коллекции. [19]

В отличие от некоторых других языков программирования, в Python, как правило, есть лучший способ сделать что-то. Тремя лучшими и наиболее важными библиотеками Python для науки о данных являются NumPy, Pandas и Matplotlib.

NumPy и Pandas отлично подходят для изучения и воспроизведения данных. Matplotlib - это библиотека визуализации данных, которая создает графики, которые вы можете найти в Excel или Google Sheets.

NumPy означает числовой Python. Самая мощная особенность NumPy - это n-мерный массив. Эта библиотека также содержит основные функции линейной алгебры, преобразования Фурье, расширенные возможности случайных чисел и инструменты для интеграции с другими языками низкого уровня, такими как Fortran, C и C ++.

Matplotlib для построения большого разнообразия графиков, начиная от гистограмм до линейных графиков. Вы можете использовать функцию Pylab в ipython, чтобы использовать эти встроенные функции построения графиков. Если вы игнорируете опцию inline, то pylab преобразует среду ipython в среду, очень похожую на Matlab. Вы также можете использовать команды Latex, чтобы добавить математику в свой график.

Pandas для структурированных операций с данными и манипуляций. Он широко используется для сбора и подготовки данных. Pandas был добавлен относительно недавно в Python и сыграл важную роль в расширении использования Python в сообществе специалистов по обработке данных.

2.2.2.2 Функции и методы Python для анализа данных

Функция plot() как часть класса DataFrame представляет не интерактивные визуализации. Библиотека Cufflinks связывает силу plotly с гибкостью pandas для лёгкого построения графиков.

%lsmagic – это набор удобных функций в Jupyter Notebook, которые предназначены для решения распространённых проблем анализа данных.

%pastebin загружает код в Pastebin — это сайт, где можуj сохранить обычный текст, например, фрагмент исходного кода, чтобы затем передать ссылку на него другим.

Команда %matplotlib inline используется для визуализации статических графиков matplotlib в Jupyter Notebook. Если заменить inline на notebook, то можно получить масштабируемые и изменяемые диаграммы.

Команда %run запускает скрипт внутри Jupyter Notebook.

%%writefile записывает содержимое ячейки в файл.

Функция %%latex отображает содержимое ячейки как LaTeX. Это полезно для написания математических формул и уравнений в ячейке.

Pandas.read_csv – функция библиотеки pandas, позволяющая загрузить файл с расширением csv.

Head() – функция, показывающая первые пять строк из какого-либо загруженного или созданного документа.

Import – позволяет загрузить файлы, библиотеки, дополнительные функции. Например, import pandas – загружает библиотеку pandas, давай доступ к ней из текущего документа Jupuyter Notebook.

Функция return позволяет вернуть те или иные значения для массива.

Большинство методов Python применимы только для одного типа значения. Например, .upper() работает со строками, но не работает с целыми числами. А .append() работает только со списками и не работает со строками, целыми числами или логическими значениями. Так что разбивать методы нужно по типу значения.

Строковые методы обычно используются на этапе очистки данных.

Функция a.lower() возвращает строчную версию строки.

Функция a.upper() противоположна lower().

Функция a.strip() используется, если строка имеет пробелы в начале или в конце, она удаляет их.

Функция a.replace('old', 'new') заменяет данную строку другой строкой.

Функция a.split('delimiter') разбивает строку в список.

Функция 'delimiter'.join(a) объединяет элементы списка в одну строку.

Функция a.count(arg) возвращает номер указанного значения в списке.

Функция a.clear() удаляет все элементы списка.

2.2.3 Jupyter Notebook

Jupyter Notebook— это веб-приложение с открытым исходным кодом, которое позволяет создавать и обмениваться документами, содержащими живой код, уравнения, визуализации и повествовательный текст. [20]

Проект Jupyter является преемником более раннего проекта IPython Notebook, который впервые был опубликован в качестве прототипа в 2010 году. Этот мощный инструмент, используется в науках о данных, для интерактивной разработки и представления проектов.

Интерфейс Notebook выглядит следующим образом (Рисунок 2.1)

Рисунок 2.1 Интерфейс Jupyter Notebook

У Jupyter Notebook есть существенные плюсы, которые облегчают и улучшают работу с ним:

- **Автозаполнение** помогает ускорить процесс работы. Эта функция позволяет быть Jupyter на ровне другими популярными IDE, такие как PyCharm.
- **Фрагменты.** Это расширение, позволяющее легко вставлять ячейки фрагмента другого кода в текущий Notebook.
- **Разделенные ячейки Notebook.** Это расширение разделяет ячейки рабочего интерфейса, что облегчает работу, а также создает аккуратный вид.

- **Содержание.** Это расширение позволяет собирать все заголовки и отображать их в плавающем окне, в виде боковой панели или с помощью меню навигации. Расширение также является перетаскиваемым, разборным и закрепляемым.
- Складные заголовки позволяют Notebook иметь разборные секции, разделенные заголовками. Поэтому, если в Notebook много грязного кода, вы можете просто свернуть его, чтобы избежать повторной прокрутки.
- **Autopep8** помогает переформатировать / предварительно подтвердить содержимое ячеек кода одним щелчком мыши.

3 ПРОЕКТНАЯ ЧАСТЬ

3.1 Подготовка и обзор данных

Для проведения данной работы будет использоваться 1 набор данных показателей хронический заболеваний США с 2001 – 2018 гг., взятый с сайта центра хронических заболеваний (CDC). Набор данных состоит из 815 000 рядов, 34 столбцов и 124 показателей. На основе этих показателей будет построена исходная модель, которая будет протестирована на части данных за 2018 год. После будет произведена корректировка модели.

Первый этап – это подготовка данных, поэтому для начала необходимо загрузить необходимые библиотеки, провести очистку данных и рассмотреть данные со всех возможных сторон.

Импортируем Pandas, Numpy для работы с нашими данными, Matplotlib для построения графиков и Seaborn (Рисунок 3.1).

Рисунок 3.1 Загрузка библиотек

Затем необходимо, с помощью *read_csv* загрузить исходный CVS файл с данными показателей хронических заболеваний США (Рисунок 3.2).

Рисунок 3.2 Загрузка данных

Данные состоят из 814937 строк, которые разбиты на 34 столбца:

- YearStart – начало исследования;
- YearEnd – конец исследования;
- LocationAbbr – наименование штата (абривиатура);
- LocationDesc – полное название штата;
- DataSource - источник данных;

- Topic - тема\название болезни;
- Question – проблема\описание болезни;
- Response – решение проблемы;
- DataValueUnit – единица измерения данных;
- DataValueType – тип данных
- DataValue – значение данных;
- DataValueAlt – значение данных;
- DataValueFootnoteSymbol – сноски о данных;
- DatavalueFootnote – примечания о данных;
- LowConfidenceLimit – нижний предел;
- HighConfidenceLimit – верхний предел;
- StratificationCategory1 – стратификация категория 1;
- Stratification1 – стратификация;
- StratificationCategory2 – стратификация категория 2;
- Stratification2 – стратификация 2;
- StratificationCategory3 – стратификация категория 3;
- Stratification3 – стратификация 3;
- GeoLocation – геолокация;
- ResponseID - ID решения проблемы;
- LocationID – ID названия штата;
- TopicID – ID названия болезни;
- QuestionID – ID проблемы;
- DataValueTypeID - ID типа данных;
- StratificationCategoryID1 – ID стратификации категории 1;
- StratificationID1 – ID стратификации;
- StratificationCategoryID2 – ID стратификации категории 2;
- StratificationID2 – ID стратификации 2;
- StratificationCategoryID3 – ID стратификации категории 3;
- StratificationID3 – ID стратификации 3;

Поскольку блокнот выдает предупреждение о то, что в 10 столбце имеются смешанные данные, нужно проверить тип данных каждого столбца с помощью команды *dtypes* (Рисунок 3.3).

Рисунок 3.3 Типы столбцов

Получаем список столбцов показателей НИЗ и список их типов данных: float64(13), int64(3), object(18). В нашем случае столбец 10 имеет тип данных object, что означает текс, но в столбце содержаться числа. Для удобства анализа и правильности решения с помощью команды .astype*()* можно поменять тип данных одного или нескольких столбцов, как и требовало предупреждение.

Но для начала необходимо отфильтровать данные. В используемой таблице слишком много столбцов, которые нам не понадобятся, а также те, в которых полностью отсутствует информация. С помощью метода *sns.heatmap* создадим тепловую карту, на которой можно легко увидеть, где есть данные и где их не хватает (Рисунок 3.4).

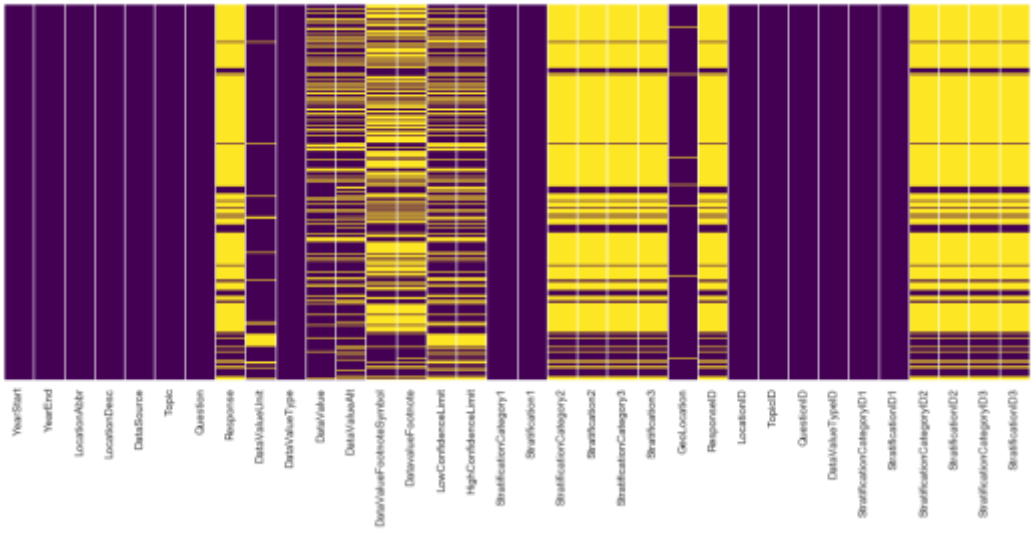

Рисунок 3.4 Тепловая карта с пропусками значений

Желтый цвет представляет недостающие данные. Каждый столбец является столбцом данных, а вертикальная ось - просто 814937 строк данных. Столбец Response и столбцы, относящиеся к StratificationCategory 2/3 и Stratification 2/3, содержат менее 20% данных. Хотя в StratificationCategory1 и Stratification1, по-видимому, имеются данные, которые потенциально полезны.

Следовательно, остальные колонки, где слишком много желтого цвета нужно удалить, для удобства работы с ними (Рисунок 3.5).

Рисунок 3.5 Удаление столбцов

Далее разберемся, о чем говорит каждая колонка. Хотя некоторые имена столбцов не требуют пояснений, я использовала set (dataframe ['ColumnName']), чтобы лучше понять уникальные категориальные данные. Вот некоторые примеры:

Topic. 814937 строк данных сгруппированы в следующие 18 категорий (Рисунок 3.6). Существует соответствующий столбец с именем TopicID, который просто дает сокращенную метку.

- Алкоголь;
- Артрит;
- Астма;
- Рак;
- Сердечно-сосудистые заболевания;
- Хроническая болезнь почек;
- Хроническая обструктивная болезнь легких;
- Диабет;
- Инвалидность;
- Иммунизация;
- Психическое здоровье;
- Питание, физическая активность и весовой статус;
- Пожилые люди;
- Здоровье полости рта;
- Общие условия;
- Репродуктивное здоровье;
- Табак.

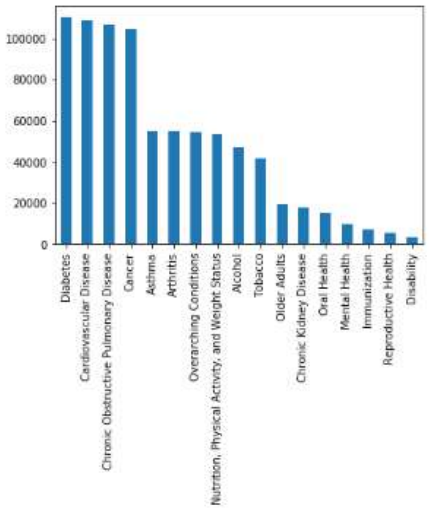

Рисунок 3.6 Распределение данных по темам

Из графика можно сделать вывод, что диабет, хроническая обструктивная болезнь легких, сердечно-сосудистые заболевания и рак являются 4 самыми распространенными болезнями.

Question. В каждой теме есть ряд вопросов. Существует так же столбец QuestionID, который так же дает сокращенную метку. Столбец вопросов - это 124 уникальных показателя (Рисунок 3.7).

```
Out[21]: Hospitalization for hip fracture among Medicare-eligible persons aged >= 65 years
         9984
         Hospitalization for chronic obstructive pulmonary disease as any diagnosis among Medicare-eligible persons aged >= 65 years
         9984
         Hospitalization for chronic obstructive pulmonary disease as first-listed diagnosis among Medicare-eligible persons aged >= 65
         years    9984
         Mortality due to diabetes reported as any listed cause of death
         9816
         Mortality with diabetic ketoacidosis reported as any listed cause of death
         9816

         ...
         Preventive dental care before pregnancy
         55
         Postpartum depressive symptoms
         55
         Prevalence of pre-pregnancy diabetes
         55
         Presence of regulations pertaining to avoiding sugar in early care and education settings
         54
         Prevalence of gestational diabetes
         52
         Name: Question, Length: 203, dtype: int64
```

Рисунок 3.7 Количество уникальных значений в столбце

DataSource. Учитывая, что у нас так много показателей, не удивительно, что есть 33 источника данных. Однако гистограмма показывает, что большая часть данных поступает из двух источников: BRFSS и NVSS (Рисунок 3.8).

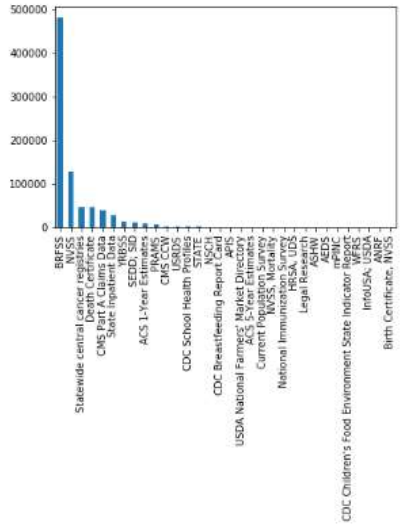

Рисунок 3.8 Распределение данных по источникам

DataValueUnit. значения в DataValue состоят из следующих единиц, включая проценты, суммы в долларах, годы и случаи на тысячи.

- $;
- %;

- Число;
- Годы;
- Количество случаев на 1000;
- Количество случаев на 1000000;
- Количество случаев на 10000;
- Количество случаев на 100000;
- Галлоны;
- продажи пачек на душу населения;
- на 100 000;
- на 100 000 жителей.

DataValue и DataValueAlt. DataValue представляет собой столбец с данными. Этот столбец состоит из числовых значений в виде строковых объектов, в то время как DataValueAlt является числовым float64.

Столбцы, связанные со стратификацией и категорией стратификации**.** Существует 12 колонок, связанных со стратификациями, которые являются подгруппами для каждого показателя. Например, пол, раса, возраст и т. д. В стратификационной категории1 есть пол, общая ситуация и раса. В Стратификации1 значения состоят из типов расы, типов пола и т.д.

3.2 Построение первоначальной модели

3.2.1 Анализ взаимосвязи между показателями хронического состояния здоровья населения

Используем сводную таблицу, как в Excel, для подведения итогов по Вопросу и связанным с ним значениям. Чтобы ее получить, создадим несколько индексов для каждого столбца: Topic, QuestionID, Question, DataValueUnit и DataValue, DataValueType (Рисунок 3.9).

Topic	QuestionID	Question	DataValueUnit	DataValueType	DataValueAlt
Alcohol	ALC1_1	Alcohol use among youth	%	Crude Prevalence	30.08
	ALC1_2	Alcohol use before pregnancy	%	Crude Prevalence	55.41
	ALC2_1	Binge drinking prevalence among youth	%	Crude Prevalence	16.57
	ALC2_2	Binge drinking prevalence among adults aged >= 18 years	%	Age-adjusted Prevalence	17.26
				Crude Prevalence	17.39
	ALC2_3	Binge drinking prevalence among women aged 18-44 years	%	Crude Prevalence	17.38
	ALC3_0	Binge drinking frequency among adults aged >= 18 years who binge drink	Number	Age-adjusted Mean	4.33
				Mean	4.18
	ALC4_0	Binge drinking intensity among adults aged >= 18 years who binge drink	Number	Age-adjusted Mean	7.08
				Mean	7.46
	ALC5_1	Heavy drinking among adults aged >= 18 years	%	Age-adjusted Prevalence	6.02
				Crude Prevalence	5.97
	ALC5_2	Heavy drinking among women aged 18-44 years	%	Crude Prevalence	6.17
	ALC6_0	Chronic liver disease mortality	cases per 100,000	Age-adjusted Rate	11.75
				Crude Rate	12.20
	ALC7_0	Per capita alcohol consumption among persons aged >= 14 years	gallons	Per capita alcohol consumption	2.44
	ALC8_0_1	Amount of alcohol excise tax by beverage type (beer)	$	US Dollars	0.30
	ALC8_0_2	Amount of alcohol excise tax by beverage type (wine)	$	US Dollars	0.78
	ALC8_0_3	Amount of alcohol excise tax by beverage type (distilled spirits)	$	US Dollars	4.38
Arthritis	ART1_1	Arthritis among adults aged >= 18 years	%	Age-adjusted Prevalence	24.70
				Crude Prevalence	24.19
	ART1_2	Arthritis among adults aged >= 18 years who are obese	%	Age-adjusted Prevalence	31.15

Рисунок 3.9 Сводная таблица, показывающая вопрос и значения

С помощью Pivot_table создали сводную таблицу и получили полную информацию по каждому вопросу и связанным с ним значениям. Например, у темы Алкоголь существует 14 вопросов:

- Смертность при хронических заболеваниях печени;
- Потребление на душу населения Алкоголь среди лиц в возрасте> = 14 лет;
- Интенсивность пьянства среди взрослых в возрасте> = 18 лет, которые пьют;
- Частота пьянства среди взрослых в возрасте> = 18 лет, которые пьют;
- Распространенность пьянства среди женщин в возрасте 18-44 лет;
- Распространенность пьянства среди взрослых в возрасте> = 18 лет;
- Пьянство среди взрослых в возрасте> = 18 лет;
- Пьянство среди женщин в возрасте 18-44 лет;
- Алкоголь употребляют до беременности;
- Законы об ответственности коммерческого хозяина (магазина драм);
- Сумма акцизного налога Алкоголь по типу напитка (пиво);

- Сумма акцизного налога Алкоголь по типу напитка (вино);
- Местные органы власти регулируют плотность на выходе Алкоголь;
- Сумма акцизного налога Алкоголь по типу напитка (крепкие спиртные напитки);
- Алкоголь используют среди молодежи;
- Распространенность пьянства среди молодежи.

Для каждого вопроса можно посмотреть его ID, тип значения данных, сами значения и единицу измерения значений. То есть, с помощью данной таблицы можно узнать абсолютно всю информацию, которая касается определенной болезни.

Далее создадим таблицу с информацией о стратификации (Рисунок 3.10) и таблицу с информацией о болезнях и вопросах, расположенных по штатам (Рисунок 3.11).

Topic	QuestionID	Question	StratificationID1	Stratification1	DataValueUnit	DataValueType	DataValueAlt
Alcohol	ALC1_1	Alcohol use among youth	OVR	Overall	%	Crude Prevalence	30.08
	ALC1_2	Alcohol use before pregnancy	OVR	Overall	%	Crude Prevalence	55.41
	ALC2_1	Binge drinking prevalence among youth	OVR	Overall	%	Crude Prevalence	16.57
	ALC2_2	Binge drinking prevalence among adults aged >= 18 years	BLK	Black, non-Hispanic	%	Age-adjusted Prevalence	13.49
						Crude Prevalence	13.82
			GENF	Female	%	Age-adjusted Prevalence	12.23

Рисунок 3.10 Суммирования информации о стратификации

				LocationAbbr	AK	AL	AR	AZ	CA	CO	CT	DC	DE	FL	...	TX
Topic	QuestionID	Question	DataValueUnit	DataValueType												
Alcohol	ALC1_1	Alcohol use among youth	%	Crude Prevalence	22.25	32.85	31.95	35.40	28.90	NaN	33.45	25.75	36.30	33.90	...	36.10
	ALC1_2	Alcohol use before pregnancy	%	Crude Prevalence	NaN	NaN	49.60	NaN	NaN	59.90	NaN	NaN	NaN	NaN	...	NaN
	ALC2_1	Binge drinking prevalence among youth	%	Crude Prevalence	12.65	17.75	19.60	19.55	15.10	NaN	17.00	10.30	20.40	15.95	...	21.00
	ALC2_2	Binge drinking prevalence among adults aged >= 18 years	%	Crude Prevalence	20.35	12.74	13.85	15.57	16.58	17.96	16.68	25.33	17.27	15.44	...	16.36
				Age-adjusted Prevalence	19.57	13.36	14.31	15.35	16.63	17.40	16.94	22.49	17.66	16.23	...	16.07
	ALC2_3	Binge drinking prevalence among	%	Crude Prevalence												

Рисунок 3.11 Сводная таблица вопросов и их местоположений

Теперь создадим визуализацию между всеми показателями и проверим есть ли корреляционная зависимость. При помощи таблиц мы увидели, что

представляют собой значения данных по вопросам, стратификации и местоположению. Нужно разобрать, что индикаторы вопросов говорят о себе и каковы отношения между ними.

Именно в таких вопросах помогает визуализация данных. Создадим новую таблицу с использованием предыдущей таблицы с индикаторами вопросов в качестве столбцов и сделаем тепловую карту (Рисунок 3.12)

Рисунок 3.12 Тепловая карта корреляции всех показателей

Даже с учетом того, что данные расположенные очень плотно можно рассмотреть положительную и отрицательную корреляцию. Оси тепловой карты показывают тему, QuestionID и сокращенный вопрос QuestionAbbr. Мы можем видеть некоторые области с более положительной корреляцией в розовом и более негативным в зеленом. Визуально, есть области сердечно-

сосудистой системы (CVD) и рака (CAN), которые, по-видимому, имеют более высокую корреляцию.

Сделаем еще более информативную таблицу, которая поможет не только увидеть зависимость, но и отобразит это в цифрах (Рисунок 3.13).

	QID1	QID2	DataValueAlt	Topic1	Topic2
4	CKD1_0	CVD1_4	0.999967	CKD	CVD
6	CKD1_0	CVD1_5	0.999958	CKD	CVD
12	CKD1_0	DIA1_1	0.999953	CKD	DIA
16	CVD1_5	OVC5_0	0.999941	CVD	OVC
17	CVD1_4	DIA1_1	0.999937	CVD	DIA
18	CVD1_1	OVC5_0	0.999934	CVD	OVC
20	CKD1_0	OVC5_0	0.999929	CKD	OVC
24	CVD1_4	OVC5_0	0.999926	CVD	OVC
28	CKD1_0	CVD1_1	0.999919	CKD	CVD
31	COPD1_2	CVD1_4	0.999917	COP	CVD
32	CVD1_2	OVC5_0	0.999915	CVD	OVC
34	COPD1_2	CVD1_5	0.999912	COP	CVD
38	CVD3_1	DIA9_0	0.999904	CVD	DIA
39	COPD1_2	OVC5_0	0.999901	COP	OVC
40	CKD1_0	CVD1_2	0.999899	CKD	CVD
41	CVD1_5	DIA1_1	0.999895	CVD	DIA
42	COPD1_2	DIA1_1	0.999892	COP	DIA
44	CKD1_0	COPD1_2	0.999890	CKD	COP
45	DIA1_1	OVC5_0	0.999889	DIA	OVC
47	COPD1_1	CVD1_5	0.999885	COP	CVD

Рисунок 3.13 Корреляция по темам по убыванию

На последней таблице в столбцах QID1 и QID2 можно заметить, что существуют зависимость между определенными темами. Например, (CKD) хроническое заболевание почек и (CVD) сердечнососудистые заболевания; CKD и диабет (DIA); CVD и DIA; общие условия (OVC) и CKD, DIA, CVD; и, наконец, пациенты с хронической обструктивной болезнью легких (COP) с DIA и CVD.

Сердечнососудистые заболевания (CVD). Факторы риска, которые увеличивают и поддерживают эту болезнь, включают диету, ожирение, диабет, чрезмерное употребление алкоголя и отсутствие физической активности.

Хроническое заболевание почек (CKD). Факторы риска, которые усиливают это заболевание, включают диабет, высокое кровяное давление,

курение, сердечнососудистые заболевания, пожилой возраст и некоторые этнические группы. По данным American Kidney Fund, когда почки не функционируют оптимально, это требует, чтобы сердце и сердечнососудистая система работали интенсивнее, что приводит ее перезагрузке и, следовательно, заболеванию.

Сахарный диабет (DIA). Бывает 1 и 2 типа, но причина типа 1 неизвестна. Диабет 2 типа имеет факторы риска, схожие с вышеуказанными заболеваниями, включая отсутствие физической активности, возраст, наследственность, высокое кровяное давление, вес, синдром поликистозных яичников и аномальный уровень холестерина / триглицеридов. Наличие диабета приводит к различным осложнениям, включая сердечно-сосудистые и хронические заболевания почек.

Общие условия (OVC). Это и есть факторы риска, к ним можно отнести диету, ожирение, отсутствие физической нагрузки, малоподвижный образ жизни, курение, алкоголь и т.д.

Следовательно, можно сделать вывод, что этот набор данных согласуется с научными работами в области НИЗ, и показывает, что существует высокая корреляция между сердечно - сосудистыми заболеваниями, хроническими заболеваниями почек и диабетом.

3.2.2 Стратификационный анализ преждевременной смертности по полу и расе среди взрослых

Поскольку индикатор OVC5_0 «Преждевременная смертность среди взрослых в возрасте 45–64 лет» появляется в ходе проделанной работы многократно, можно посмотреть, что этот индикатор может сказать нам путем расслоения по полу и расе. Нужно создать фрейм данных ind_OVC5_0_gender на основе ind1 (Рисунок 3.14).

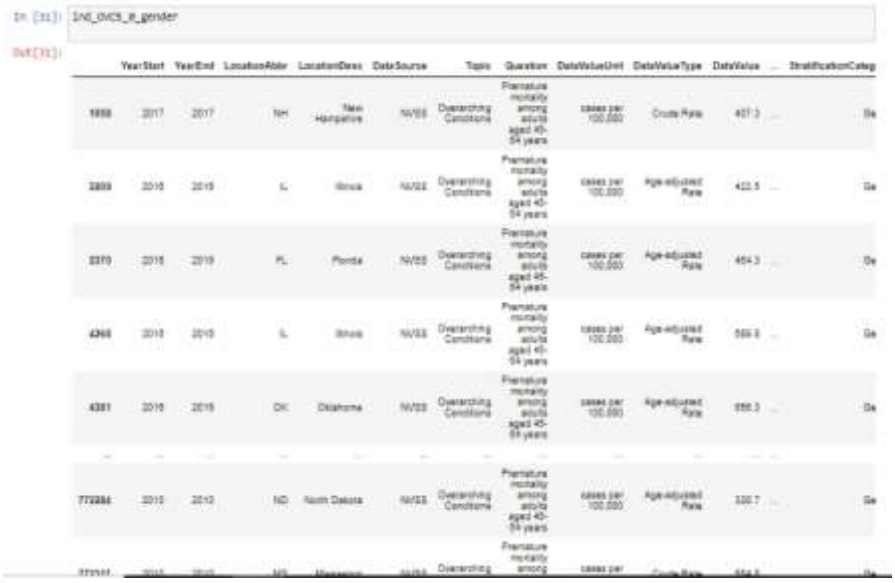

Рисунок 3.14 Стратификация по полу

Используя сводку по группам, мы видим, что доступные данные работают за 7 лет, а не за все 18 лет всего набора данных (Рисунок 3.15-3.16)

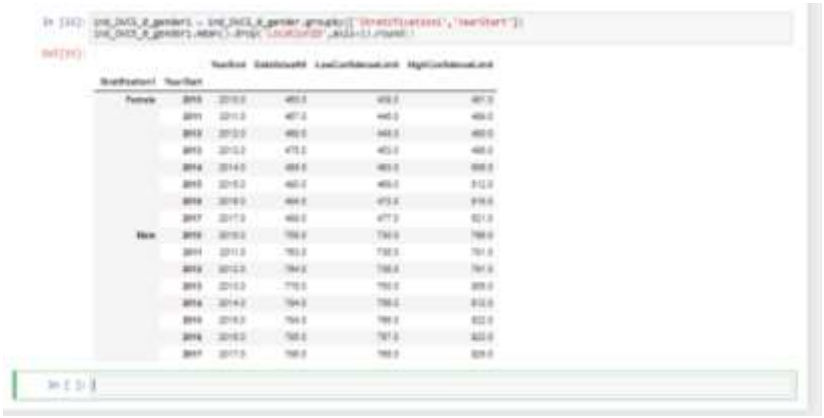

Рисунок 3.15 Использование Groupby по набору данных о стратификации полов

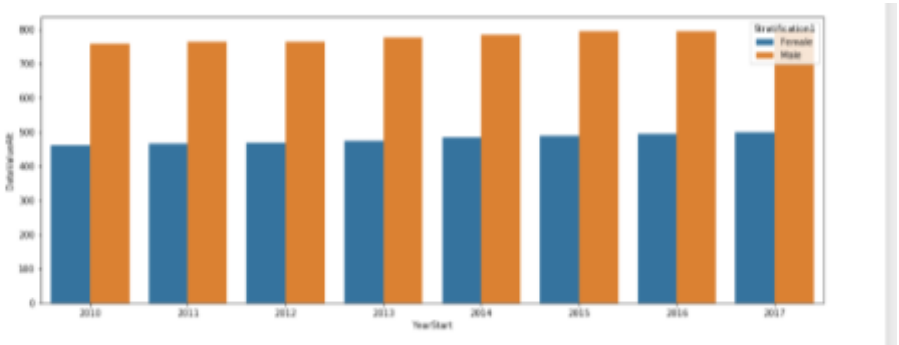

Рисунок 3.16 Тенденции преждевременной смертности в разбивке по полу в 2010–2017 гг. На 100 000 случаев

Гистограмма показывает, что в течении 2010-2017 годов преждевременная смертность населения в возрасте 45-64 лет медленно

увеличивается из года в год. Преждевременная смертность среди женского населения увеличивается на 5,4%, а мужского - на 3,4% Вертикальная ось соответствует DataValueAlt, а горизонтальная показывает года. Удивительно, что уровень преждевременной смертности растет быстрее среди мужчин, чем среди женщин в целом.

Подобно стратификации по полу, анализ стратификации по расе будет использовать следующие условия для фильтрации данных. Мы создадим новый фрейм данных ind_OVC5_0_race на основе OVC5_0, распространенности по возрасту с использованием «случаев на 100 000» и стратификационной категории 1 «Раса / Этнос» (Рисунок 3.17).

Stratification1	YearStart	
American Indian or Alaska Native	2010	788.0
	2011	787.0
	2012	772.0
	2013	786.0
	2014	819.0
	2015	852.0
	2016	844.0
	2017	843.0
Asian or Pacific Islander	2010	282.0
	2011	286.0
	2012	283.0
	2013	275.0
	2014	283.0
	2015	285.0
	2016	273.0
	2017	271.0
Black, non-Hispanic	2010	880.0
	2011	873.0
	2012	868.0
	2013	882.0
	2014	869.0
	2015	874.0
	2016	898.0
	2017	886.0
Hispanic	2010	348.0
	2011	354.0
	2012	355.0
	2013	364.0
	2014	383.0
	2015	388.0
	2016	383.0
	2017	375.0
White, non-Hispanic	2010	585.0
	2011	594.0
	2012	596.0
	2013	606.0
	2014	618.0
	2015	626.0
	2016	626.0
	2017	630.0

Рисунок 3.17 Расслоение по расам со средними значениями DataValueAlt

Далее мы будем использовать метод .groupby () для суммирования подфрейма данных ind_OVC5_0_race в таблицу на основе расы, года и DataValueAlt. Категории для расы в Стратификации1 дают нам 5 категорий:

- Индейцы или уроженцы Аляски;
- Выходец из Азии или Тихого океана;

- Черный, не латино;
- Испанец;
- Белый, не латино.

Для YearStart у нас есть данные за 7 лет, так, как только за эти года были доступны данные о стратификации. DataValueAlt - это среднее значение. Чтобы сделать гистограмму (Рисунка 3.17), нужно отобразить все расы и значения DataValueAlt. Параметры для гистограммы включают в себя ось x = YearStart, ось y = DataValueAlt (Рисунок 3.18).

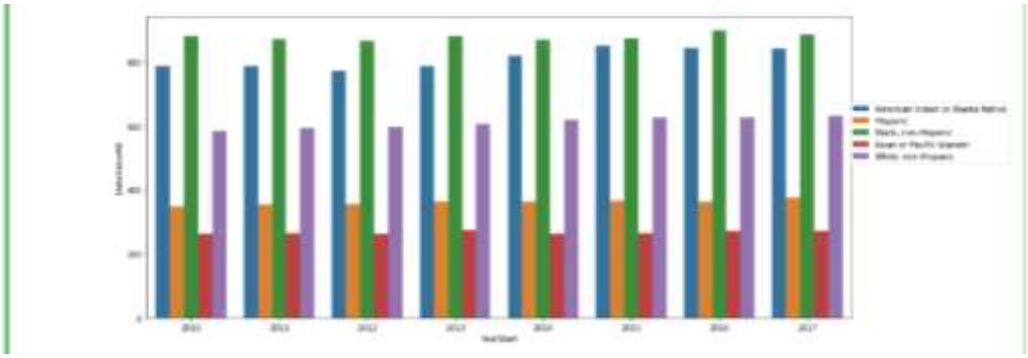

Рисунок 3.18 Тенденции преждевременной смертности в 2010–2017 гг., в зависимости от расы

Данные с таблицы и гистограммы показывают, что преждевременная смертность несколько увеличивается из года в год, с постепенным увеличением. Преждевременная смертность среди взрослых в возрасте 45–64 лет наиболее сильно влияет на группы чернокожих, не латиноамериканцев и коренных американцев или жителей Аляски. По другому спектру мы видим, что популяции азиатских или тихоокеанских островов и латиноамериканцев являются наименее тяжелыми. Для сравнения, мы видим, что показатели преждевременной смертности для чернокожих, не латиноамериканцев более чем вдвое превышают показатели латиноамериканцев.

Можно предположить, что корреляции с числами преждевременной смертности для каждой стратификации, выходит за рамки привязки описательного элемента к причинно-следственной связи. Именно (Рисунок 3.18) дает общее понимание серьезности факторов, влияющих на все население.

3.2.3 Анализ тем по годам

Далее будем использовать визуализацию для анализа данных по годам, поэтому необходимо провести построение диаграмм, показывающих отношение тем (болезней) к годам. Но для начала проверим, все ли года имели хорошие показатели (Рисунок 3.19).

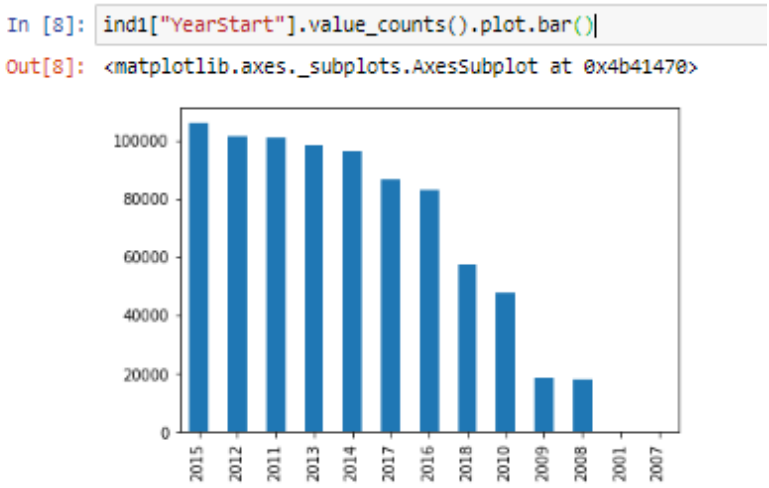

Рисунок 3.19 Количество показателей НИЗ по годам

По графику видно, что данные ведутся с 2001 по 2018 год. За 2001, 2007, 2008 2009 гг. данные очень малы, их нужно удалить, так как они не несут для нас никакой ценной информации, а данных за 2002, 2003, 2004, 2005, 2006 совсем нет. Проводим удаление не нужных строк и получаем следующее:

- 2015 - 105825
- 2012 - 101548
- 2011 - 100821
- 2013 - 98562
- 2014 - 96542
- 2017 - 86652
- 2016 - 82868
- 2018 - 57295
- 2010 – 47833

Построим диаграммы за 2010, 2011, 2012 год (Рисунок 3.20), 2013, 2014, 2015 (Рисунок 3.21), 2016, 2017, 2018 (Рисунок 3.22).

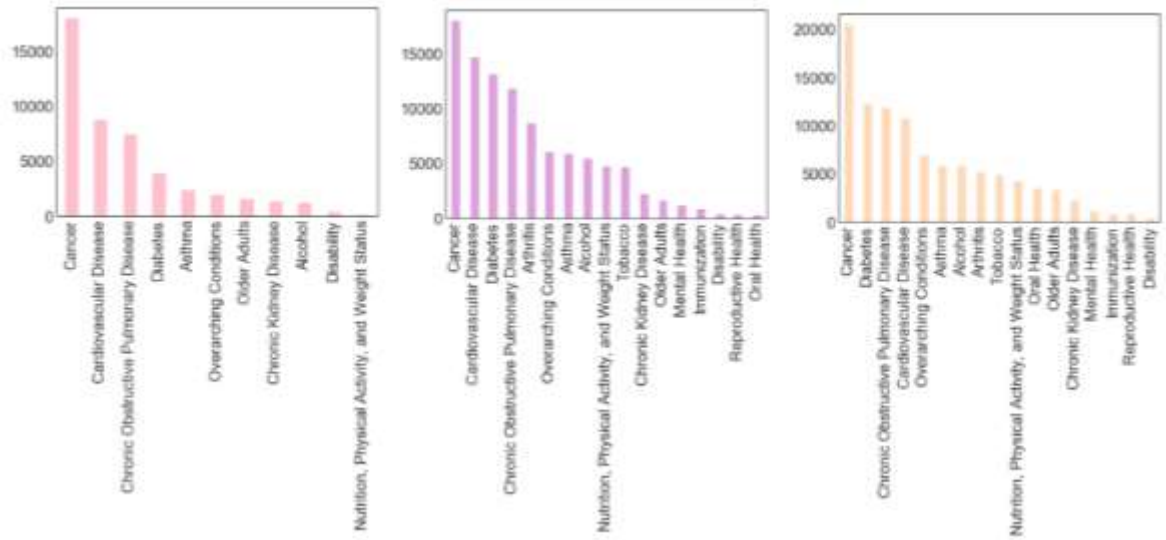

Рисунок 3.20 Расслоение тем (болезней) по годам (2010, 2011, 2012)

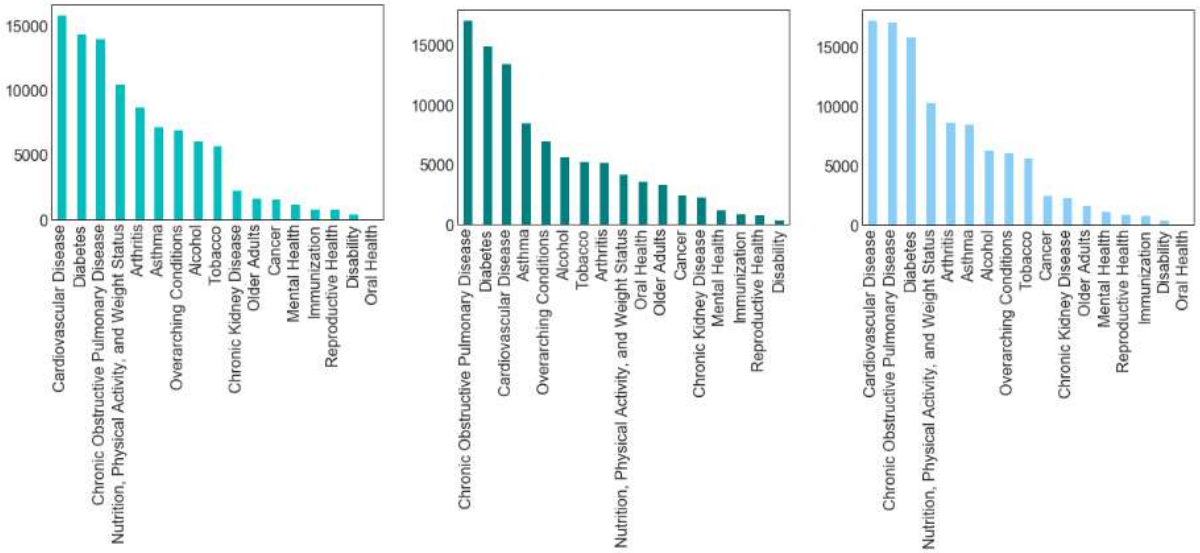

Рисунок 3.21 Расслоение тем (болезней) по годам (2013, 2014, 2015)

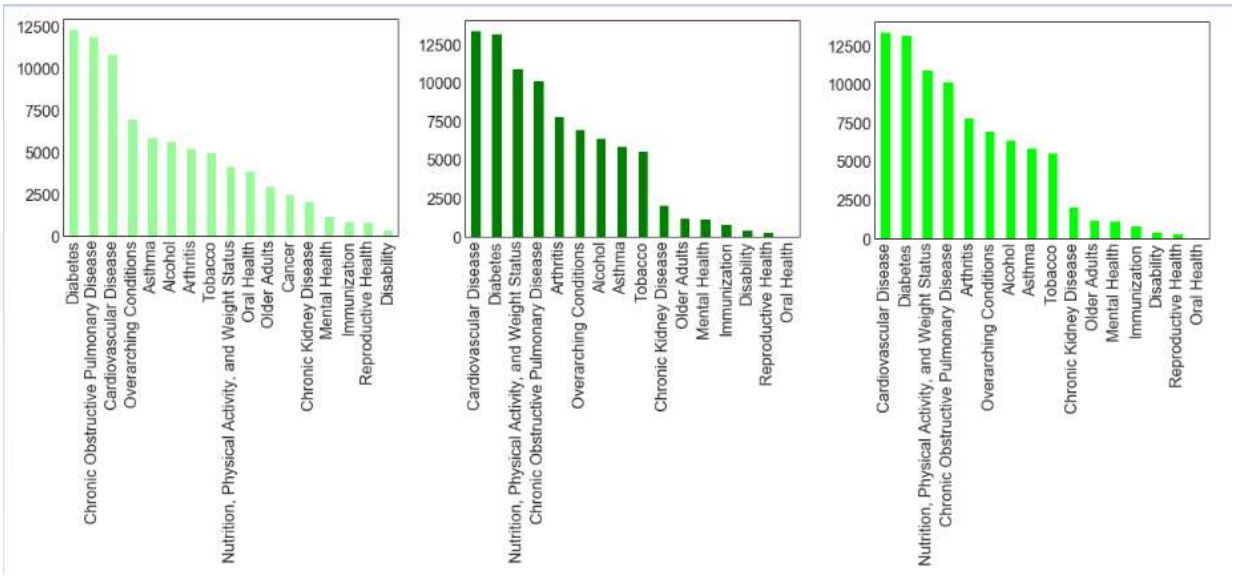

Рисунок 3.22 Расслоение тем (болезней) по годам (2016, 2017, 2018)

Просмотрев данные диаграмм, можно заметить, что увеличилось количество показателей заболеваний в целом, а также определенные болезни, которые актуальны сейчас или были актуальны несколько лет назад (Таблица 3.1).

Таким образом, можно сделать вывод, что с 2010 по 2012 год проблемой здравоохранительных органов в основном были онкологические заболевания, а в последние годы это стали сердечно-сосудистые заболевания и диабет. Следовательно, можно предположить, что возможна аналогия со смертностью в эти года от этих болезней. Статистки мировой смертности доказывает мое предположение (Рисунок 1.2-1.3).

Таблица 3.1

Количество показателей в год по алфавиту

Название темы	Количество								
	Года								
	2010	2011	2012	2013	2014	2015	2016	2017	2018
Alcohol	1282	5467	5839	6124	5718	6362	5677	6399	4130
Arthritis		8700	5220	8700	5220	8700	5220	7830	5220
Asthma	2442	5902	5902	7171	8494	8494	5902	5902	4675
Cancer	18096	18096	20591	1625	2495	2495	2495		2495
Cardiovascular Disease	8847	14725	10863	15880	13461	17266	12349	13420	3480
Chronic Kidney Disease	1435	2305	2305	2305	2305	2305	2097	2097	870
Chronic Obstructive Pulmonary Disease	7518	11910	11910	14016	17103	17103	11910	10170	5220
Diabetes	7518	13219	12349	14431	14947	15817	12349	13219	9895
Disability	424	424	424	424	424	424	424	424	
Immunization		870	870	870	870	870	870	870	870
Mental Health		1195	1195	1195	1195	1195	1195	1195	1195
Nutrition, Physical Activity, and Weight Status	164	4785	424	10528	4245	10363	4188	10968	4025
Older Adults	1680	1680	3420	1680	3420	1680	2988	1248	1740
Oral Health		330	3590	55	3590	55	3865	55	3480
Overarching Conditions	1975	6105	6975	6975	6975	6105	6975	6975	5000
Reproductive Health		380	855	869	827	911	855	325	325
Tobacco		4730	4995	5714	5253	5680	4995	5555	4675

3.2.4 Анализ тем по штатам

Далее провели построение диаграмм, показывающих зависимость каждой из темы «Topic» от местоположения «LocationDesc», т.е. в каких штатах распространена определенная болезнь (Рисунок 3.23).

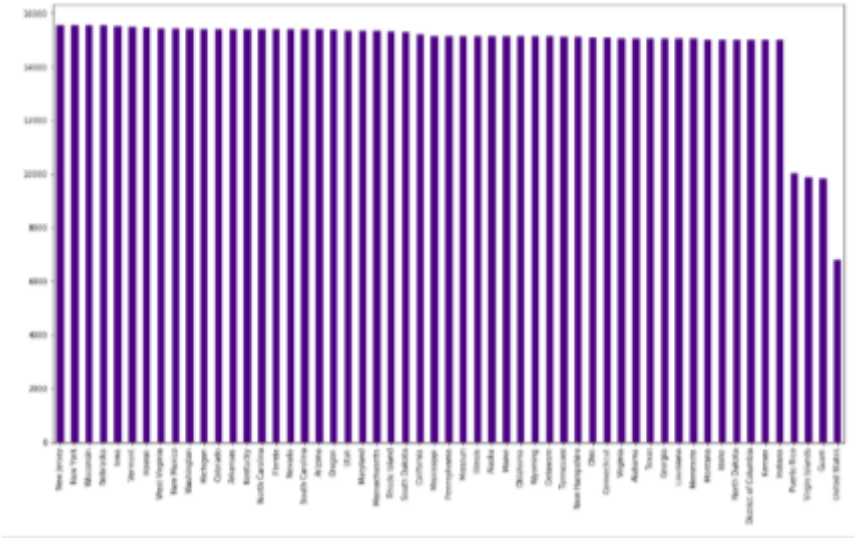

Рисунок 3.23 Анализ по штатам

По гистограмме можно заметить, что данных о болезнях прмерно одинаковое количество, но можно выделить пять штатов с самым высоким количество данных и 4 штата с наименьшим (Таблица 3.2).

Таблица 3.2

Штаты с max\min количеством данных

Назавние штата	Количество данных
Max	
New Jersey	15548
New York	15548
Wisconsin	15548
Nebraska	15548
Iowa	15520
Min	
Puerto Rico	10043
Virgin Islands	9892
Guam	9843
United States	6796

Предыдущие анализы выявили самые проблемные темы, поэтому создадим диаграммы и проведем анализ только пяти тем: рак (Рисунок 3.24),

сердечно-сосудистые заболевания (Рисунок 3.25), хроническая обструктивная болезнь легких (Рисунок 3.26), диабет (Рисунок 3.27) и хроническая болезнь почек (Рисунок 3.28).

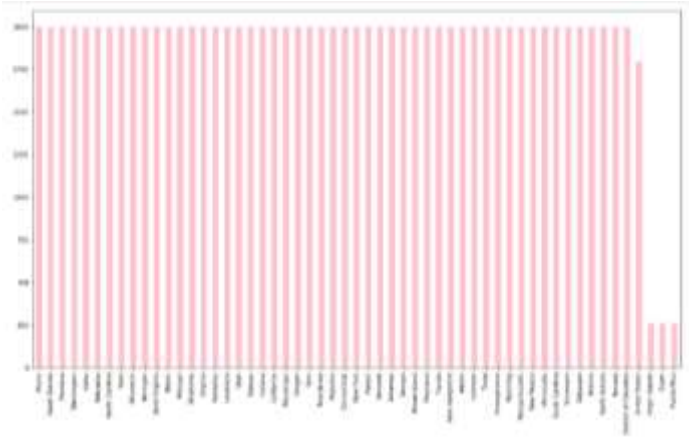

Рисунок 3.22 Распространённость рака по штатам

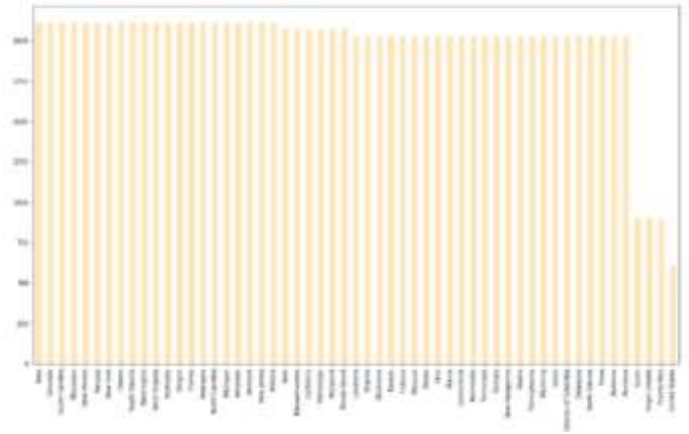

Рисунок 3.25 Распространённость сердечнососудистых заболеваний по штатам

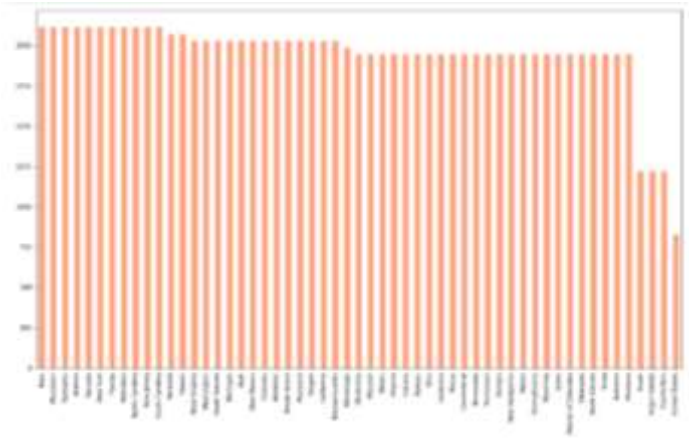

Рисунок 3.26 Распространённость хронической обструктивной болезни легких по штатам

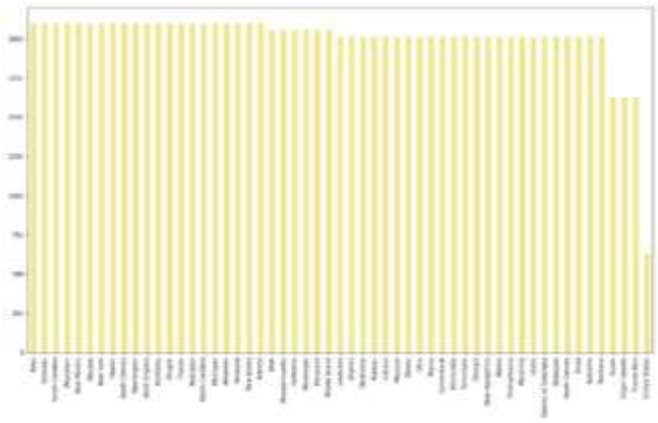

Рисунок 3.27 Распространённость диабета по штатам

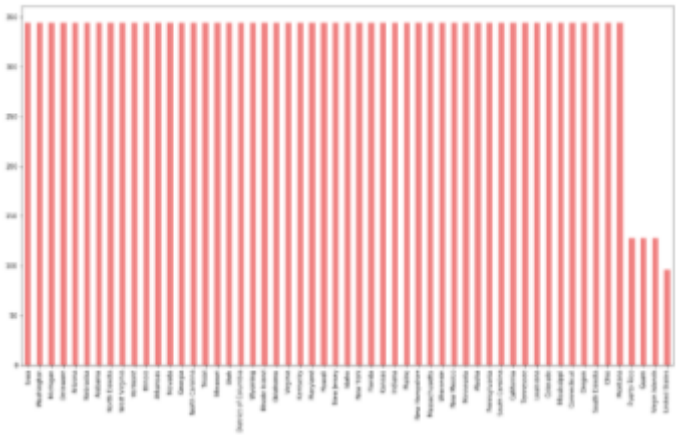

Рисунок 3.28 Распространённость хронической болезни почек по штатам

Для удобства оценивая была построена таблица со всеми пятью темами болезни (Таблица 3.3).

Таблица 3.3

Количество данных о болезнях по штатам

Название «Topic»	Количество данных по штатам	
	Max	Min
Cancer	Maine	Puerto Rico
Cardiovascular Disease	Iowa	Puerto Rico
Chronic Obstructive Pulmonary Disease	Iowa	Puerto Rico
Diabetes	Iowa	Puerto Rico
Chronic Kidney Disease	Iowa	Puerto Rico

Исходя из данных таблицы, можно сделать вывод, что штаты Iowa и Maine имеют большое количество данных и информации по основным болезням. Так же по таблице можно заметить, что Puerto Rico имеет очень мало данных об этих заболеваниях.

Исследуя данные штаты, пришли к выводу, что Puerto Rico не является настоящим штатом США. Это государство имеет неофициальное название 51 штата США и не позднее 1 января 2021 года должен стать официальным. Puerto Rico — свободно ассоциированное государство на острове в Карибском море. Именно поэтому, данный массив данных о НИЗ имеет очень мало показателей из данного государства. Так же Puerto Rico был передан Соединенным Штатам и попал под суверенитет этой страны в соответствии с Парижским договором, положившим конец испано-американской войне 1898 года. [21]

Штаты Iowa и Maine являются штатами со среднем уровнем доходов по всем категориям включая доходы домохозяйства и доходы на душу населения. Следовательно, можно сказать, что высокий объём данных о распространении хронических заболеваний из этих штатов является понятным и не удивительным.

3.2.5 Анализ источников данных

В наборе данных 33 источника данных. И большая часть данных поступает из двух источников: BRFSS и NVSS (Рисунок 3.8). Для более подробной информации по источникам данных составим диаграммы источников по годам (Рисунок 3.29) и (Рисунок 3.30).

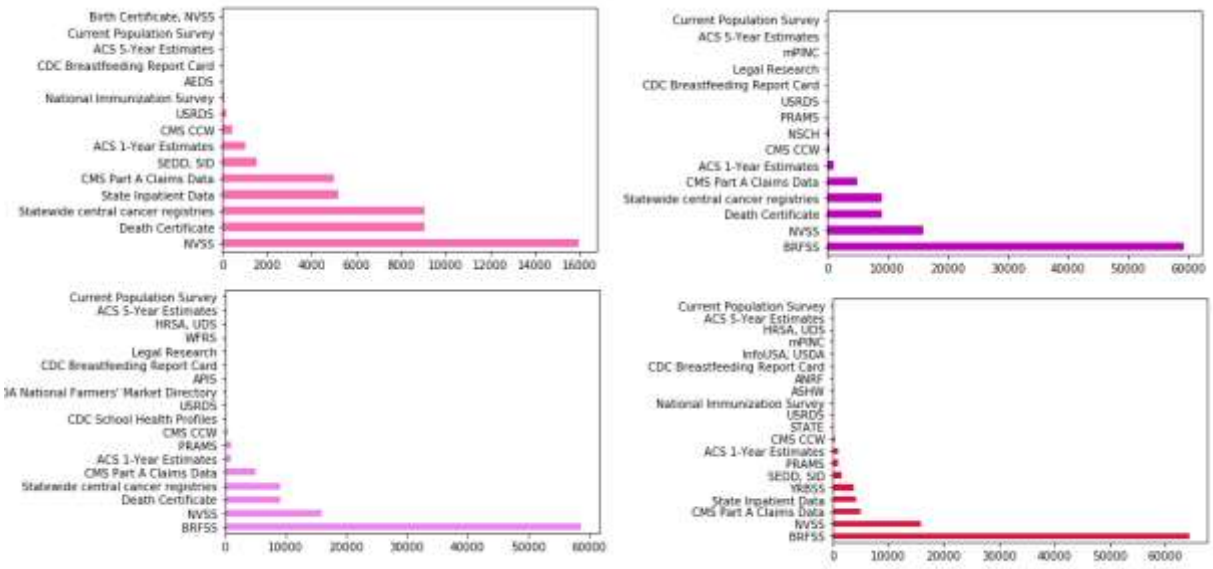

Рисунок 3.29 Источники данных по 2010-2013 гг.

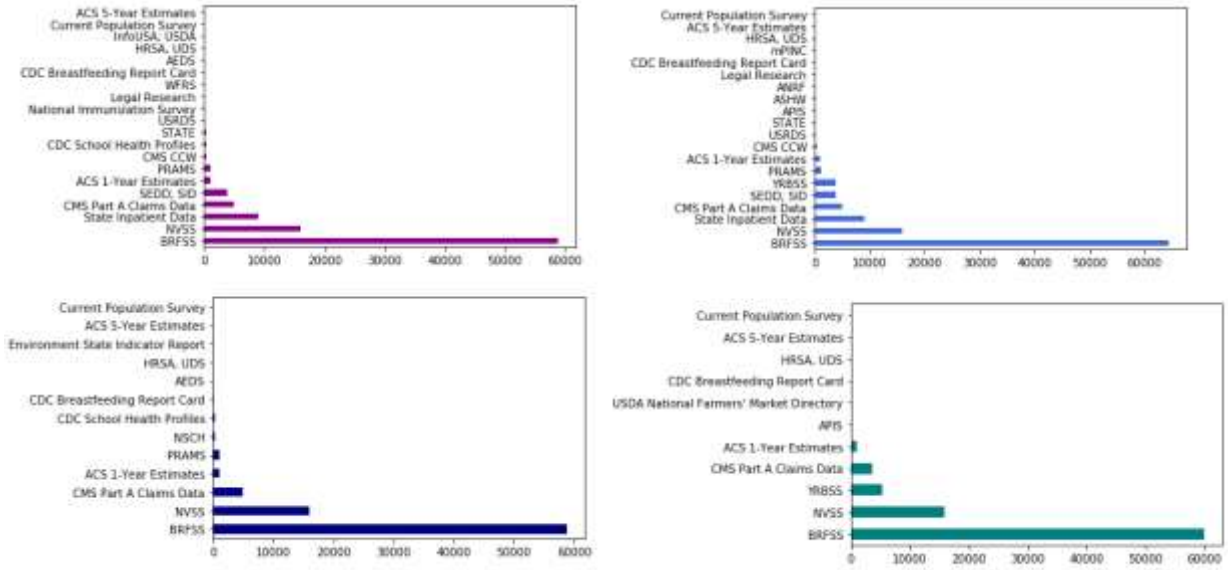

Рисунок 3.30 Источники данных по 2014-2017 гг.

Система эпиднадзора за поведенческими факторами риска (BRFSS) - это ведущая национальная система телефонных опросов, связанных со здоровьем, которая собирает данные о жителях США, касающиеся их поведения в отношении здоровья, хронических заболеваний и использования профилактических услуг. Основанная в 1984 году в 15 штатах, BRFSS собирает данные во всех 50 штатах, а также в округе Колумбия и трех территориях США.

Национальная система статистики естественного движения населения (NVSS) предоставляет наиболее полные данные о рождении и смерти в Соединенных Штатах.

Исходя из гистограмм, можно сделать вывод о том, что начиная с 2010 года поступление информации из источников изменились. С 2011 года BRFSS являлся главным источником, а в 2010 это был NVSS. Поэтому проведем анализ по штатам, только двух главных источников данных (Рисунок 3.31).

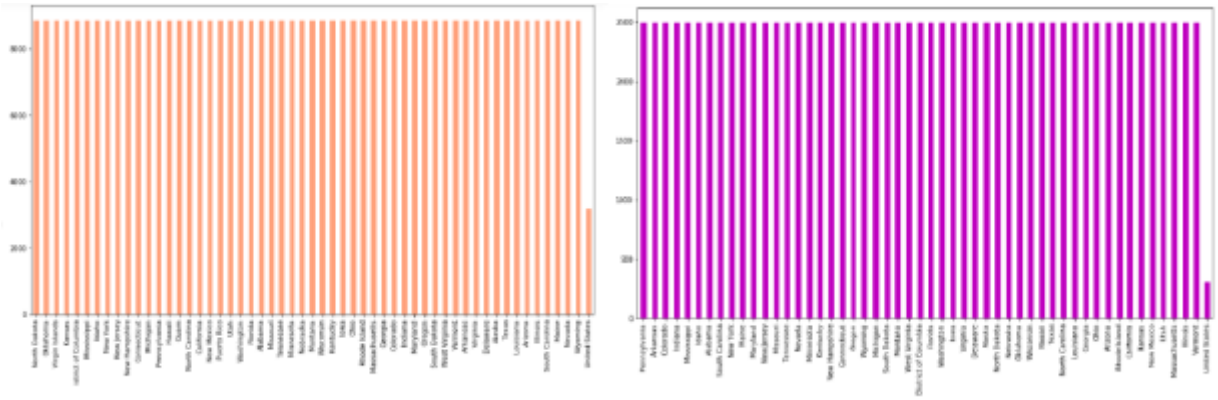

Рисунок 3.31 Распределение BRFSS и NVSS по штатам

Распределение BRFSS по штатам равномерно, так же, как и для NVSS по штатам, но у первого показатели намного выше.

3.2.6 Кросс-факторный анализ

Из-за того, что текстовая информация почти всегда воспринимается глазом хуже, чем визуальная, следует визуализировать все полученные данные. Проведем кросс-факторные анализ, с использованием библиотеки seaborn и функции heatmap(), которая служит для создания тепловых карт. Создадим визуализацию типов болезней по годам (Рисунок 3.32), болезней по штатам (Рисунок 3.33) и болезней по стратификации (Рисунок 3.34).

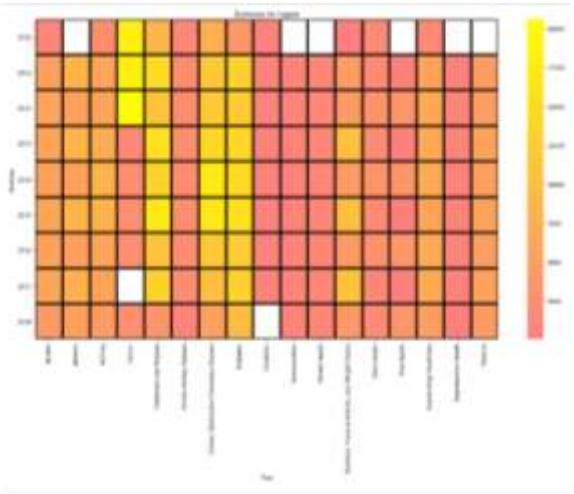

Рисунок 3.32 Тепловая карта распространенности болезни по годам

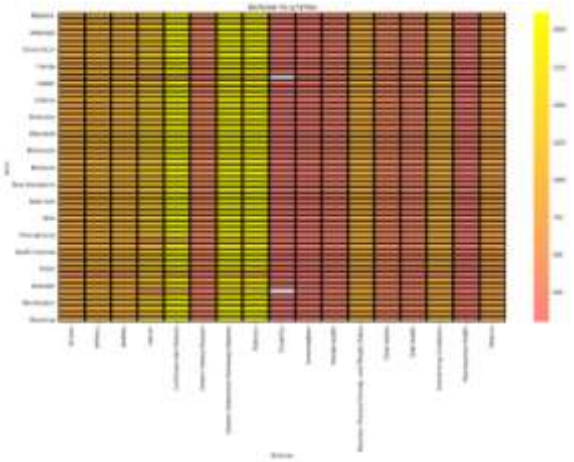

Рисунок 3.33 Тепловая карта распространенности болезни по штатам

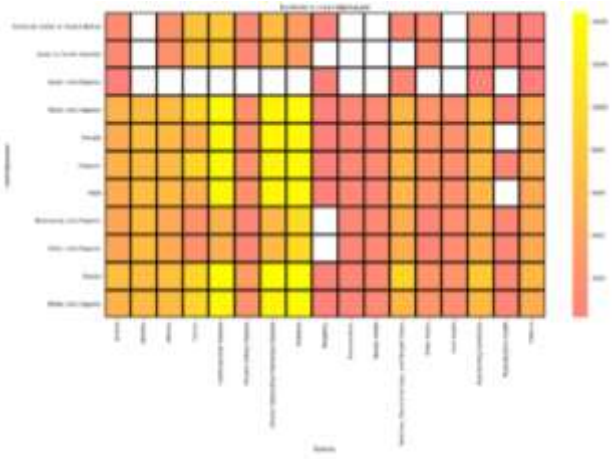

Рисунок 3.34 Тепловая карта распространенности болезни по стратификации

Так же, для более качественного анализа разобьем распространение болезней по штатам еще и на 2010, 2011, 2012 году (Рисунок 3.35), 2013, 2014, 2015 (Рисунок 3.36) и на 2016, 2017, 2018 (Рисунок 3.37).

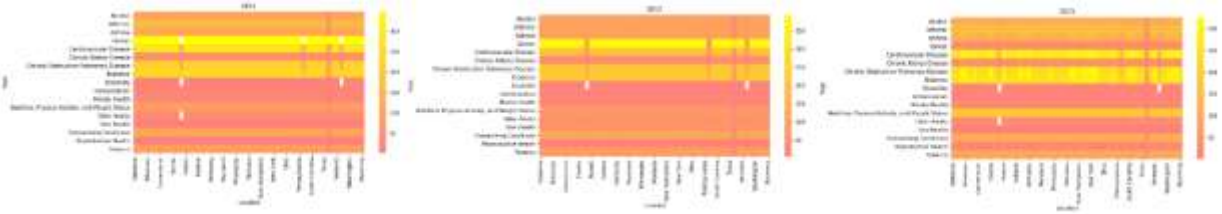

Рисунок 3.35 Распространение болезней по штатам в 2010, 2011 и 2012 гг.

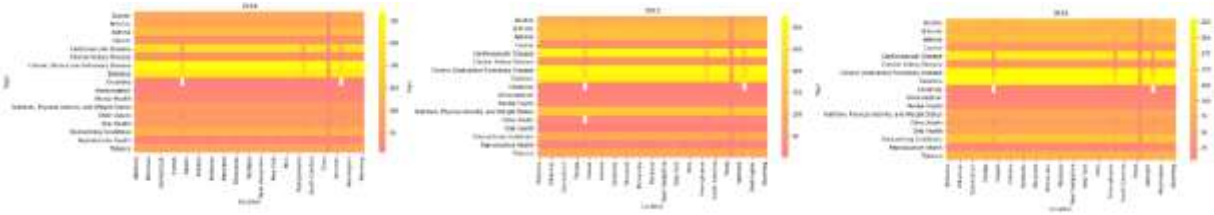

Рисунок 3.36 Распространение болезней по штатам в 2013, 2014 и 2015 гг.

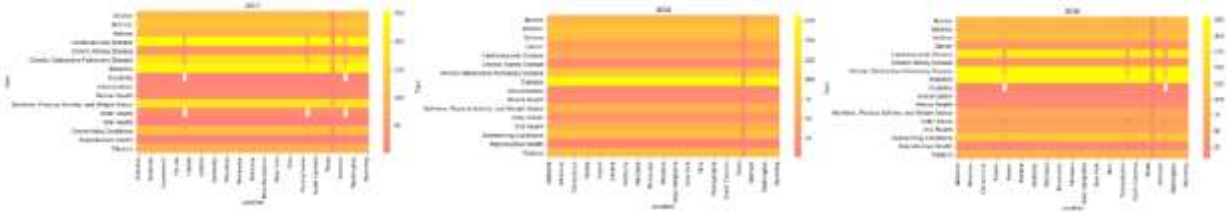

Рисунок 3.37 Распространение болезней по штатам в 2016, 2017 и 2018 гг.

Рассмотрев тепловые карты можно сделать вывод, что розовый цвет означает наименьшее количество данных, желтый наибольшее, отсутствие данных обозначается белым. На карте распространенности болезней по годам видно, что с 2010 по 2012 год проблемой здравоохранительных органов в основном были онкологические заболевания, а в последние годы это стали сердечнососудистые заболевания и диабет.

На карте распространённости болезней по штатам видно, что штаты Iowa и Maine имеют большое количество данных и информации по основным болезням. Так же по таблице можно заметить, что Puerto Rico имеет очень мало данных об этих заболеваниях. Так как это государство еще не является штатом США.

По данным карты распространенности болезней по стратификации видно, что диабет, рак, сердечнососудистые заболевания и болезнь легких являются самыми распространенными проблемами. Так же, исходя из данной карты, можно сделать вывод о предрасположенности людей определенной расы и пола к определенному заболеванию. Так, например, белокожие, темнокожие и латиноамериканцы имеют не сильную, но существенную предрасположенность к раку. Абсолютно все расы и полы имеют предрасположенность к заболеванию сердечнососудистой системы, заболеванию легких и диабету. Научное подтверждение этих выводов – это всемирная статистика смертности, в которой первые причины смертности, как раз являются эти заболевания.

3.2.7 Анализ по широте и долготе

В данном наборе показателей хронических заболеваний присутствует колонка с геолокацией. Для того, чтобы разобраться к какому местоположению относятся данные координаты, как проанализировать их, нужно разбить столбец GeoLocation на 2 колонки.

С помощью str.split и str.get разбиваем нашу колонку на две новых, обозначающих долготу и ширину. После разбиения столбца проверяем тип данных, так как в случае неправильно выбранного типа данных, могут возникнуть ошибки в процессе работы. Данные в столбце долгота и широта имеют тип данных object, это означает, что нам понадобиться функция .astype('float') для того, что бы поменять тип данных на float64.

Когда данные пришли в подходящий для их анализа вид, преобразуем их в карту (Рисунок 3.38).

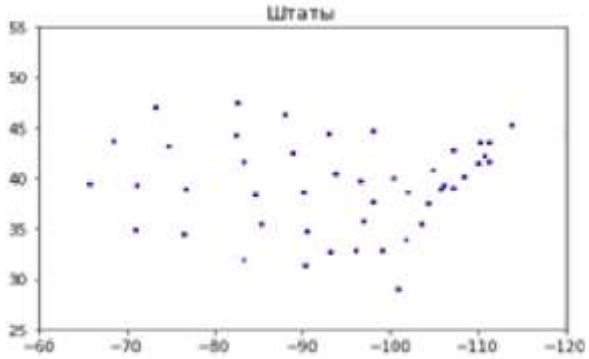

Рисунок 3.38 Карта геолокаций за весь период

По данной карте можно заметить существенную схожесть точек с координатами центров штатов США. Для доказательства данного предположения приведу карту Соединенных Штатов (Рисунок 3.39).

Рисунок 3.39 Карта США

Действительно, составленная мной карта по долготе и широте из данных о показателях хронических заболеваний является координатами штатов США. Для того, чтобы понять какие штаты несли основной вклад в составление показателей НИЗ составим карту за 2011, 2012, 2013, 2014 гг. (Рисунок 3.40), 2015, 2016, 2017, 2018 (Рисунок 3.41).

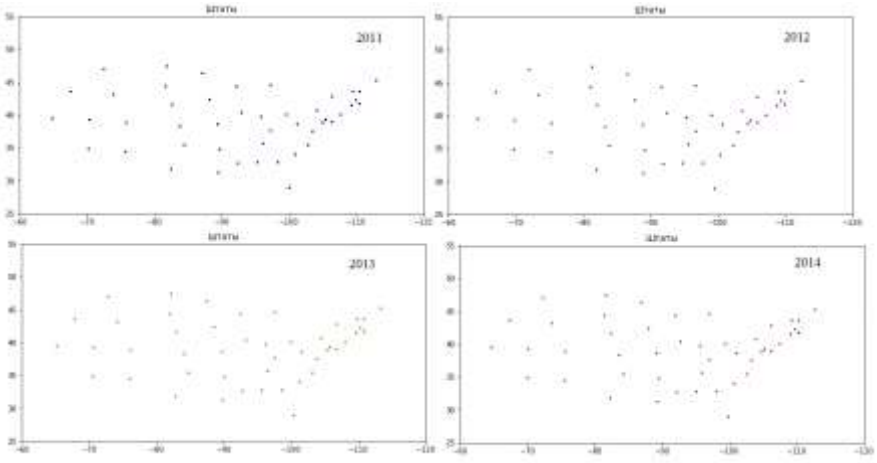

Рисунок 3.40 Карты штатов за 2011-2014 гг.

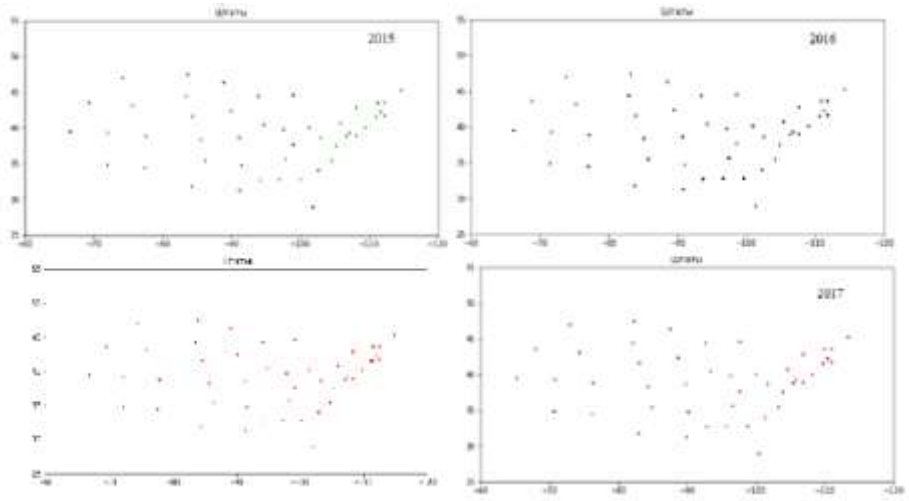

Рисунок 3.41 Карты штатов за 2015-2018 гг.

Исходя из карт за 2011-2018 гг. можно сделать вывод, что данные, которые находились в колонке GeoLocation – это координаты штатов Соединённых Штатов Америки. С 2011 года поступление информации о заболеваниях и о показателях НИЗ мало поменялись.

3.2.8 Визуализация данных с помощью ресурсов CDC

На сайте представителя набора данных CDC, можно онлайн провести визуализацию данных о НИЗ. Рассмотрим координаты точек в столбце геоданные (Рисунок 3.42).

Рисунок 3.42 Визуализация координат точек геоданных на веб-сайте

Исходя из данной карты, можно подтвердить предположение о том, что координаты действительно являются координатами штатов США. Далее рассмотрим увеличенную версию данной карты (Рисунок 3.43).

Рисунок 3.43 Увеличенная карта США с количеством данных

При увеличении карты, можно рассмотреть количество данных поступающих в набор показателей НИЗ из определенных штатов.

3.2.9 Исходная модель

После проведения данной работы, можно сделать вывод о том, что (CKD) хроническое заболевание почек и (CVD) сердечнососудистые заболевания; CKD и диабет (DIA); CVD и DIA; общие условия (OVC) и CKD, DIA, CVD; и, наконец, пациенты с хронической обструктивной болезнью легких (COP) с DIA и CVD имеют высокую корреляцию.

Преждевременная смертность среди взрослых в возрасте 45–64 лет наиболее сильно влияет на группы чернокожих, не латиноамериканцев и коренных американцев или жителей Аляски. Так же этот показатель имеет наименьшую активность на популяции азиатских или тихоокеанских островов и латиноамериканцев.

С 2010 по 2012 год проблемой здравоохранительных органов в основном были онкологические заболевания, а в последние годы это стали сердечнососудистые заболевания и диабет. Следовательно, можно предположить, что возможна аналогия со смертностью в эти года от этих болезней. Статистки мировой смертности доказывает наше предположение.

Так же из модели мы получаем, что штаты Iowa и Maine имеют большое количество данных и информации по основным болезням. Они являются штатами со среднем уровнем доходов по всем категориям включая доходы

домохозяйства и доходы на душу населения. Следовательно, можно сказать, что высокий объём данных о распространении хронических заболеваний из этих штатов является понятным и не удивительным.

В наборе данных присутствуют 33 источника данных, но большая часть данных поступает из двух источников: BRFSS, которая представляет собой систему наблюдения за поведенческим фактором риска CDC, и NVSS, которая является Национальной системой статистики естественного движения населения.

С помощью построение карт по точкам и визуализации координат стало понятно, что данные из колонки Geolocation – это координаты центров штатов. Так же на сайте CDC можно построить визуализацию геоданных и просмотреть количество показателей из определенного штата.

3.3 Итоговая модель

Учитывая, что в 2020 году произошла пандемия короновируса COVID-19, что немного затрудняет исследования в области хронических заболеваний, можно сделать несколько предположений касающихся этого и 2021 гг.

Статистика за 2020 год может отличатся от всех, что были раньше и будут потом, так как из-за всемирной пандемии коронавируса все показатели смертности и здоровья могут сбиться. Смертность в этом году возрастет до небывалых размеров. COVID-19 уже убил около 300 тысяч человек, и, к сожалению, это не предел. Пандемия «заставляет» сидеть дома и боятся заражения и смерти. Если придерживаться сведениями об этом заболевании, то можно сделать вывод, что не сам вирус убивает людей, а паника и неосведомленность. Из-за паники, постоянного стресса и отсутствия физической активности могут развиться огромное количество хронических заболеваний. Больше всего заболеванию подвержены пожилые люди с хроническими заболеваниями. Так как в силу возраста и хронического заболевания иммунитет ослаблен, возможность заболеть увеличивается в несколько раз.

Поскольку COVID-19 поражает легкие, можно предположить, что в 2020 году хроническая обструктивная болезнь легких займет первое место по распространенности НИЗ, а также по всемирной статистики смертности. Сердечнососудистые заболевания займут второе место. Это можно объяснить тем, что нервная система любого человек, в сложный для всего мира период, подвержена стрессу, депрессии, паники и чувству тревоги, которые являются одними из факторов риска этого заболевания. Уже сейчас многие пожилые люди не выдерживают происходящее в мире, и у них случается инсульт и инфаркт.

Душевное здоровье вместе с онкологическими заболеваниями, так же поднимутся по шкале смертности и будут занимать эти места еще не один год. Поскольку, одно зависит от другого, и любая болезнь имеет свойство накапливаться, т.е. стресс, испытанный человеком сегодня, может показать свое действие через несколько лет, в виде больного сердца, легких или онкологических заболеваний.

Если рассматривать данную ситуацию с другой стороны, то можно спрогнозировать хорошие показатели на 2021 год. Все человечество начало активную борьбу за свое здоровье. После окончания карантина и полной победы над вирусом многие начнут заниматься спортом, вести правильный образ жизни, откажутся от вредных привычек, начнут соблюдать диету и т.д. Эта тенденция продлится не долго, так как многие люди имеют импульсивное отношение к таким вещам. Если выражаться грубо, то как быстро они поддались панике и начали активно вносить в свою жизнь спорт и правильное питание, так же быстро они откажутся и забудут все это. Но не стоит обобщать всех людей, те, кто практиковали правильный образ жизни, продолжат в том же направлении, те, кому не хватало мотивации, получили ее и будут двигаться дальше.

Итоговая модель была построена на показателях хронических заболеваний Соединенных Штатов Америки, и для дальнейшего исследования

и анализа проблемы распространённости НИЗ нужно провести сравнение экономических, экологических и социальных сфер США и России.

3.3.1 Экономическая, экологическая и социальная сферы

3.3.1.1 Экономическая сфера

Россия классифицируется как экономика с высоким уровнем дохода Всемирным банком и является членом БРИКС (Бразилия, Россия, Индия, Китай и Южная Африка). БРИКС как группа наций, рекламируемая Всемирным банком и международным валютным фондом (МВФ), является следующей группой глобальных сверхдержав, способных превзойти существующих экономических лидеров. Это 8-е место по величине ВВП и 6-е место по паритету покупательной способности (ППС).

Соединенные Штаты являются крупнейшей экономикой в мире, с самым большим ВВП и самым высоким ППС. Хотя экономика США в настоящее время сталкивается с многочисленными препятствиями, причем некоторые прогнозы говорят о том, что Китай в конечном итоге превзойдет экономическое превосходство бывшего, нет никаких сомнений в том, что экономическая мощь страны по-прежнему остается самой доминирующей в мире.

3.3.1.2 Экологическая сфера

Россия по экологическим показателям страдает от многих причин:

- загрязнение воздуха из-за тяжелой промышленности;
- промышленное, муниципальное и сельскохозяйственное загрязнение внутренних водных путей и побережий;
- вырубка леса;
- эрозия почвы;
- загрязнение почвы от неправильного применения сельскохозяйственных химикатов;
- загрязнение подземных вод токсичными отходами.

Такие же показатели можно привести и для США:

- загрязнение воздуха, приводящее к кислотным дождям в США и Канаде;
- США являются крупнейшим источником выбросов углекислого газа при сжигании ископаемого топлива;
- загрязнение воды от слива пестицидов и удобрений;
- ограниченные природные ресурсы пресной воды;
- опустынивание.

3.3.1.3 Социальная сфера

Россия и США вместе составляют 6% населения мира. США насчитывают 325,7 млн. Человек на своих территориях, в то время как в России их менее половины, а их 137,7 млн. Наиболее густонаселенным регионом в США является восточное побережье, в то время как в России большая часть населения проживает в западной части. Обе страны сильно урбанизированы: 81% населения США проживает в городских районах, а в России - 74%

Приведем сравнительную статистику США и России по экономической социальной и экологической сферам (Таблица 3.4).

Таблица 3.4

Сравнение статистики России и США

Название показателя	Показатели	
	Экономическая сфера	
	Россия	США
Величина внутреннего валового продукта	1,28 трлн $	18,57 трлн $
Величина ВВП на душу населения	26 490 $	57 436 $
Индекс человеческого развития	0,804 (49-я позиция в рейтинге)	0,920 (10-е место в мировом рейтинге)
Процентное отношение государственного долга к ВВП	17%	105%
Размер международных резервов	385 млрд $ (6-е место среди всех стран мира)	118 млрд $ (11-е место в мире)
МРОТ – минимальный размер оплаты труда	160$ в месяц	7,25 $ в час
	Экологическая сфера	
Скорректированный чистый национальный доход	774,44 млрд. долларов США	$ 11,13 трлн. Занимает 2-е место. В 14 раз больше чем России
Выбросы CO2 на 1000	10,65 - на 19 месте.	19,86 - на 4 месте. На 86% больше чем в России

Продолжение таблицы 3.4

Выбросы CO2 от производства электроэнергии и тепла, всего	1 000,18 - 4-м месте.	2 478,03 - 2-е место. В 2 раза больше чем России
Экологический след	5.36 - 31-е место	12.22 - 2 место. В 2 раза больше чем России
Площадь леса на душу населения	56,51 км² на 1000 человек - 9 место. В 6 раз больше, чем в США	10,22 км² на 1000 человек
	Социальная сфера	
Средний возраст	42,38 лет	44,38 лет
Население в возрасте 0-14 лет	16,99%	16,71%
Население в возрасте 15-24 лет	11,68%	11,22%
Уровень рождаемости	12,11 рождений/1000 населения	13,66 рождений/1000 населения
Смертность	13,97 смертей/1000 населения	8,39 смертей/1000 населения
	Медицинская сфера	
Ожидаемая продолжительность жизни при рождении	66,29 лет	78,37 лет
Врачи	4,25 на 1000 человек	2,3 на 1000 человек
Заболеваемость туберкулезом (на 100 000 человек)	91	3,6
Рак. Коэффициент смертности от рака (на 100 000 населения)	142	133
Средний ИМТ человека	23,25	27,82
Причина смерти от неинфекционных заболеваний	82,46%	86,57%
Диабет	9,74%	9,35%

После приведенного сравнения, можно сделать вывод, что показатели статистики экономической сферы США и России различны, но заметны значительные сходства в социальной и медицинских сферах. Так как значительной разницы между условиями жизни среднего класса нет, а разница в статистике других сфер жизни человека компенсирует друг друга, то данная модель, построенная на основе медицинских данных США, может использоваться и для медицинских данных России.

3.3.4 Уменьшение рисков распространенности НИЗ

Приведем выводы, которые получили в ходе проделанной работы.

- (CKD) хроническое заболевание почек и (CVD) сердечно-сосудистые заболевания; CKD и диабет (DIA); CVD и DIA; общие условия (OVC) и CKD, DIA, CVD; и, наконец, пациенты с хронической обструктивной болезнью легких (COP) с DIA и CVD имеют высокую корреляцию.

К сожалению зависимость одной болезни от другой исправить нельзя. Для хорошей работы организма, необходима отличная работоспособность каждого отдельного органа, тогда организм будет иметь силы на защиту себя от патогенных инфекций, а не тратить энергию на спасение определенного органа. Этот вывод сам по себе является мотиватором к заботе о здоровье.

Для уменьшения риска распространенности данного факта, нужно провести несколько операций:

1. Уведомление всех людей, о величине проблемы распространенности хронических заболеваний.
2. Проводить, как можно больше сборов, уроков, лекций и других мероприятий, посвященных проблеме распространенности хронических заболеваний.
3. Сделать информацию о НИЗ более доступной, а борьбу против НИЗ привлекательнее, за счет рекламы и дополнительных приложений.
4. Уведомление людей, имеющих одно или несколько хронических заболеваний о данном факте.

- Преждевременная смертность среди взрослых в возрасте 45–64 лет.

Наиболее распространенными причинами болезней и преждевременной смерти во всем мире являются курение, высокое кровяное давление, высокий уровень холестерина в крови, ожирение, чрезмерное употребление алкоголя и отсутствие физической активности. Так же в силу возраста иммунитет ослаблен и возможность заболеть увеличивается в несколько раз.

Можно значительно снизить риск ранней смерти, сделав несколько простых изменений в образе жизни, а также провести некоторые мероприятия:

1. Не курить.
2. Соблюдая здоровую диету.
3. Регулярно заниматься физическими упражнениями.
4. Ограничить употребление алкоголя или уменьшить его количество.

- Белокожие, темнокожие и латиноамериканцы имеют не сильную, но существенную предрасположенность к раку.

Исследования показывают, что до 50% случаев заболевания раком и около 50% случаев смерти от рака можно предотвратить с помощью изменения образа жизни и отношения к своему здоровью.

Для уменьшения риска заболевания онкологическими заболеваниями и для профилактики данного заболевания нужно:

1. Не курить.
2. Соблюдать здоровую диету.
3. Регулярно заниматься физическими упражнениями.
4. Защищаться от солнца.
5. Сделать прививку от Гепатита В и вируса папилломы человека (ВПЧ).
6. Избегать инфекций, передающихся половым путем и через кровь.
7. Наблюдаться у врача.

- Абсолютно все расы и полы имеют предрасположенность к заболеванию сердечнососудистой системы, заболеванию легких и диабету 2 типа.

Диабет 2 типа, обструктивное заболевание легких и сердечнососудистые заболевания имеют несколько причин, но самые главные это:

1. Генетическая предрасположенность.
2. Образ жизни.

3.4 Практическая значимость модели

Данная модель может быть использована в медицинской сфере услуг для осведомления людей об их предрасположенности к определенному хроническому заболеванию, а также для анкетирования или тестирования людей по желанию.

Предполагается, что итоговая модель будет основой для дальнейшей работы создания теста на определение предрасположенности к НИЗ, а также рекомендации по правильному образу жизни.

В наше время у организации здравоохранения и почти у всех медицинских учреждений существует веб-сайты, на которых можно совершить запись на прием ко врачу в режиме онлайн. Так же веб-сайты оборудованы информационными и контактными страницами, личным кабинетом, а некоторые уже имеет в своем арсенала опросы и тесты, касающиеся той или иной болезни.

Обычно, с возникновение определенной медицинской проблемы увеличивается и спрос на информацию о ней. Так, например, с возникновение коронавируса чисто онлайн-тестов и опросов на определение данного заболевания, увеличилось на 55%. Естественно, что это связанно с тем, что люди начали массово искать информацию о симптоматике COVID-19. Такие организации, как Минздрав России и портал Госуслуг уже создали приложения и онлайн-тесты на определение коронавируса и для поддержания здоровья человека.

Таким образом, если рассматривать хронические неинфекционные заболевания, как глобальную проблему для всего мирового сообщества, имеющую не только медицинское, но и огромное социально-экономическое значение, а также учитывать то, что они занимают первые строки статистики

мировой смертности уже многие годы, то можно сделать вывод о том, что такой онлайн-тест необходим и актуален.

3.5 Структура модели

В связи с тем, что Госуслуги, Минздрав Росии, ВОЗ и другие организации имеют веб-сайты и приложения, которыми управляют штаты программистов, разумно будет привлечь к этапу разработки одного специалиста отдела программирования. Поскольку создание онлайн-теста не является сложной работой, тот же программист справится с дизайном и оформлением.

Стратегия заключается в добавлении новой ссылки в меню веб-сайта или приложения организации, которая будет переносить пользователя на онлайн-тест. Тест будет показывать предрасположенность человека к определенным хроническим заболеваниям, а также рекомендации по правильному образу жизни. Предрасположенность и рекомендации будут строиться исходя из заполненных пользователем ответов на характерные вопросы о человеке, его генетике, образу жизни, полу, расе и т.д. Также в построение рекомендаций будут использоваться исторические данные, которые будут, обновляется и корректироваться по мере необходимости. Пользователи будут получать информацию о своем здоровье и рекомендации по профилактике заболевания, к которому предрасположены, а также информацию по оздоровлению организма в целом после прохождения теста.

Ценность данного нововведения заключается в том, что российские организации здравоохранения смогут обеспечить информацией о хронических неинфекционных заболеваниях огромное количество людей, о предрасположенности людей к определенному заболеванию и смотивировать на переход к правильному образу жизни.

Таким образом, данная опция будет интегрирована в приложение или веб-сайт для людей, как дополнительная информация о своем здоровье, профилактика болезней и борьба с НИЗ.

ЗАКЛЮЧЕНИЕ

Данная работа включает в себя три части: (1) Аналитическую часть: анализ информации и исследований в области хронических неинфекционных заболеваний; (2) Теоретическую часть: ознакомление с технологией Data Mining, выбор основного метода и инструмента; (3) Проектную часть: интеллектуальный анализ больших данных показателей хронических заболеваний США (CDI), составление итоговой модели, подсчет практической значимости и анализ рисков распространенности НИЗ.

В Аналитической части была рассмотрена информация и исследования в области НИЗ. Приведены примеры борьбы с хроническими заболеваниями и мировая статистика смертности, где первые строчки занимали такие заболевания, как сердечнососудистые заболевания, рак, обструктивная болезнь легких и диабет. Также были выявлены основные тенденции и актуальные вопросы, касающиеся данной темы: влияние наличия хронических заболеваний на заражение коронавирусом COVID-19.

В Теоретической части был произведен детальный обзор технологий и методов Big Data. Как основной метод реализации работы была выбрана, и технология Data Mining, а также произведен полный ее разбор. Были выявлены основные процессы, задачи и технологии интеллектуального анализа данных. К основным методам Data Mining относятся: линейная регрессия, нейронные сети, визуализация, деревья решений, полиномиальные нейронные сети, метод k-ближайшего соседа. По результатам проведения сравнительного анализа всех перечисленных методов, был выбран наиболее подходящий под специфику задачи – метод визуализации. Помимо этого, с помощью сравнительного анализа был выбран основной инструмент для проведения интеллектуального анализа больших данных – язык программирования Python 3. В качестве платформы был выбран Jupyter Notebook, а для хороших графиков и диаграмм загрузили библиотеки Pandas, Numpy, Matplotlib и Seaborn.

В Проектной части были использованы данные центра хронических заболеваний США (CDC), в количестве 815 тысяч показателей и индикаторов хронических заболеваний. Работа началась с подготовки и очистки данных, был произведен первичный анализ, построение исходной модели, построение итоговой модели, написание практической значимости и структуры проекта. По результатам анализа было выявлено, что (CKD) хроническое заболевание почек и (CVD) сердечнососудистые заболевания; CKD и диабет (DIA); CVD и DIA; общие условия (OVC) и CKD, DIA, CVD; и, наконец, пациенты с хронической обструктивной болезнью легких (COP) с DIA и CVD имеют высокую корреляцию. Преждевременная смертность среди взрослых в возрасте 45–64 лет наиболее сильно влияет на группы чернокожих, не латиноамериканцев и коренных американцев или жителей Аляски. В последние годы проблемой здравоохранительных органов стали сердечнососудистые заболевания и диабет. Так же из модели мы получили, что штаты Iowa и Maine имеют большое количество данных и информации по основным болезням.

В ходе построения тепловой карты были получены выводы касательно того, что в наборе данных присутствуют 33 источника данных, но большая часть данных поступает из двух источников: BRFSS и NVSS.

Также, по результатам прогнозных значений 2020 год будет отличатся от всех, что были раньше и будут потом, так как из-за всемирной пандемии коронавируса все показатели смертности и здоровья могут сбиться. Душевное здоровье вместе с онкологическими заболеваниями, так же поднимутся по шкале смертности и будут занимать эти места еще не один год. А вот в 2021 году намечаются хорошие показатели, так как весь мир начнет активную борьбу за свое здоровье.

Люди склонны принимать смерти от НИЗ как неизбежность. Бремя болезней слишком велико и сложно, чтобы с ним справиться. Тем не менее, данная модель новый стратегический ответ, основанный на последних данных и надежном анализе, позволяет предположить, что пора изменить взгляд на хронические заболевания. Интегрируя эффективные меры в глобальном

масштабе, можно спасти миллионы жизней, получить экономические выгоды и добиться значительного прогресса в достижении целей в экономической, социальной и медицинской сферах.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

1. ВОЗ – глобальный веб-сайт Всемирного здравоохранения [Электронный ресурс]. Режим доступа: https://www.who.int/chp/ru/, свободный - (дата обращения: 10.04.2020).

2. CDC – американский портал центра хронических заболеваний [Электронный ресурс]. Режим доступа: https://www.cdc.gov/chronicdisease/about/costs/index.htm, свободный – (дата обращения: 10.04.2020).

3. Википедия — свободная энциклопедия [Электронный ресурс]. Режим доступа: https://en.wikipedia.org/wiki/2015#Deaths, свободный – (дата обращения 11.04.2020).

4. ВОЗ – глобальный веб-сайт Всемирного здравоохранения [Электронный ресурс]. Режим доступа: https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death, свободный - (дата обращения: 10.04.2020).

5. ВОЗ – глобальный веб-сайт Всемирного здравоохранения [Электронный ресурс]. Режим доступа: https://www.who.int/ru/news-room/fact-sheets/detail/newborns-reducing-mortality, свободный - (дата обращения: 13.04.2020).

6. НВ – Украинский общественно-политический журнал [Электронный ресурс]. Режим доступа: https://nv.ua/health/medicine/koronavirus-i-vozrast-kto-bolshe-vsego-riskuet-umeret-ot-covid-19-50076696.html, свободный - (дата обращения: 13.04.2020).

7. Chrodis – открытый веб-сайт о внедрении передовых опытов для НИЗ [Электронный ресурс]. Режим доступа: http://chrodis.eu/good-practice/cindi-countrywide-integrated-non-communicable-disease-intervention-bulgaria/, свободный - (дата обращения: 14.04.2020).

8. Ornish lifestyle medicine – открытый веб-сайт по борьбе с НИЗ [Электронный ресурс]. Режим доступа: https://www.ornish.com/proven-program/the-research/, свободный - (дата обращения: 14.04.2020).

9. CDC – американский портал центра хронических заболеваний [Электронный ресурс]. Режим доступа: https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-/g4ie-h725, свободный - (дата обращения: 14.04.2020).

10. CDC – американский портал центра хронических заболеваний [Электронный ресурс]. Режим доступа: https://www.cdc.gov/mmwr/pdf/rr/rr6401.pdf, свободный - (дата обращения: 17.04.2020).

11. Википедия — свободная энциклопедия [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/wiki/исследование, свободный - (дата обращения: 01.05.2020).

12. Википедия — свободная энциклопедия [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/wiki/предметная_область, свободный - (дата обращения: 01.05.2020).

13. Bibs-science — учебное пособие по основам эконометрического анализа [Электронный ресурс]. Режим доступа: http://bibs-science.ru/archive/sbornik_11/gaisyonok.pdf, свободный - (дата обращения: 02.05.2020).

14. Sas — веб-сайт в формате электронной библиотеки с элементами новостного сайта [Электронный ресурс]. Режим доступа https://www.sas.com/ru_ru/insights/analytics/data-mining.html, свободный - (дата обращения: 02.05.2020).

15. Encyclopedia — электронная энциклопедия [Электронный ресурс]. Режим доступа: http://www.encyclopedia.ru/data_mining, свободный - (дата обращения: 11.05.2020).

16. Intuit – национальный открытый университет [Электронный ресурс]. Режим доступа:

https://www.intuit.ru/studies/courses/6/6/lecture/196?page=4, свободный - (дата обращения: 11.05.2020).

17. McKinsey – веб-сайт международной консалтинговой компании [Электронный ресурс]. Режим доступа: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation, свободный - (дата обращения: 11.05.2020).

18. CRN/RE - издание, посвященное бизнесу в сфере информационных технологий [Электронный ресурс]. Режим доступа: https://www.crn.ru/news/detail.php?ID=117807], свободный - (дата обращения: 12.05.2020).

19. Бонцанини М. — Анализ социальных медиа на Python. Извлекайте и анализируйте данные из всех уголков социальной паутины на Python - Издательство "ДМК Пресс" - 2018 - ISBN: 978-5-97060-574-5 - Текст электронный // ЭБС Лань - URL: https://e.lanbook.com/book/108129

20. Википедия — свободная энциклопедия [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/wiki/jupuyer_notebook, свободный - (дата обращения: 13.05.2020).

21. Википедия — свободная энциклопедия [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/wiki/%D0%9F%D1%83%D1%8D%D1%80%D1%82%D0%BE-%D0%A0%D0%B8%D0%BA%D0%BE, свободный - (дата обращения: 13.05.2020).

ПРИЛОЖЕНИЕ

Листинг кода

Подключаем необходимые библиотеки:

```
import collections
import seaborn as sb
import numpy as np
import pandas as pd
import matplotlib
%pylab inline
```

Загружаем файл CSV с данными:

```
ind=pd.read_csv('C:\\Users\\acep\\Desktop\\U.S._Chronic_Disease_Indicators__CDI_.csv')
```

Вывод загруженных данных:

```
print (ind)
```

Строим тепловую карту, для обнаружения пустых значений:

```
sb.heatmap (ind.isnull (), yticklabels = False, cbar = False, cmap = 'viridis')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)
```

Вывод информации о данных:

```
ind.info()
```

Удаление ненужных столбцов:

```
df=ind.drop(ind.columns[[1, 2, 7, 8, 12, 13, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]], axis='columns')
```

Вывод оставшихся столбцов:

```
print (df)
```

Вывод индексов колонок:

```
df.columns
```

Гистограмма распределение данных по темам:

```
df["Topic"].value_counts().plot.bar()
```

Гистограмма распределение данных по источникам:

df["DataSource"].value_counts().plot.bar()

Составление сводной таблицы, показывающий вопрос и значения:

table=df.pivot_table(index=['Topic','QuestionID','Question','DataValueUnit','DataValueType'],columns=None,dropna=True)

Составление таблицы суммирования информации о стратификации:

table1=df.groupby(['Topic','QuestionID','Question','StratificationID1','Stratification1','DataValueUnit','DataValueType']).mean().round(2)

Составление сводной таблицы вопросов и их местоположений:

table2=df.pivot_table(values='DataValueAlt',index=['Topic','QuestionID','Question','DataValueUnit','DataValueType'], columns='LocationAbbr',aggfunc='mean',dropna=True).round(2)

Составление карт корреляций:

table_new= df.groupby(['Topic','QuestionID','Question']).mean().round(2)

table3 = table_new.transpose().corr()

mask = np.zeros_like(table_new, dtype=np.bool)

mask[np.triu_indices_from(mask)] = True

f, ax = plt.subplots(figsize=(150, 150))

f.suptitle('Карта корреляции всех хронических показателей', x=0.4,y=0.85,fontsize=150)

ax.tick_params(labelsize=50)

cmap = sns.diverging_palette(190, 10, as_cmap=True)

sns.heatmap(table_new, mask=mask, cmap=cmap, vmax=.3, center=0,square=True, linewidths=2, cbar_kws={"shrink": .3})

Составляем корреляция по темам по убыванию:

top_corr=(table_new*np.tril(np.ones(table_new.shape),1)).stack().sort_values(by=['DataValueAlt'],ascending=False)

Создаем новую группировку:

ind_OVC5_0_gender = ind1[(ind1['QuestionID'] == 'OVC5_0') & (ind1['StratificationCategory1'] == 'Gender') &

(ind1['DataValueUnit'] == 'cases per 100,000')]

ind_OVC5_0_gender .info()

Группируем и создаем столбчатую гистограмму по полу со значения, в разбивке по годам:

ind_OVC5_0_gender1=ind_OVC5_0_gender.groupby(['Stratification1','YearStart'])

ind_OVC5_0_gender1.mean().drop('LocationID',axis=1).round()

plt.figure(figsize=(16, 6))

sns.barplot(x='YearStart',y='DataValueAlt',data=ind_OVC5_0_gender,hue='Stratification1',ci=None,saturation=0.7)

Удаляем ненужные столбцы:

ind_gender2=ind1.drop(columns=['Question','Topic','DataValueUnit','DataValueType','StratificationCategory1','TopicID','DataValueTypeID','StratificationCategoryID1','StratificationID1','QuestionAbbr'])

Создаем таблицу:

ind_gender2_loc_qid=ind_gender2.pivot_table(values='DataValueAlt',index=['LocationID','Stratification1'],columns=['QuestionID'],aggfunc=np.mean)

ind_gender2_loc_qid.reset_index(level='Stratification1',inplace=True)

ind_gender2_loc_qid.fillna(ind_gender2_loc_qid.mean(),inplace=True)

ind_gender2_loc_qid.head()

ind_OVC5_0_race1=ind_OVC5_0_race.groupby(['Stratification1','YearStart']).mean().round(0)

plt.figure(figsize=(16, 7))

sns.barplot(x='YearStart',y='DataValueAlt',data=ind_OVC5_0_race,hue='Stratification1',ci=None,saturation=0.7)

plt.legend(loc='best',bbox_to_anchor=(0.5, 0,.73, .73))

Удаление не нужных строк:

ind = ind.loc[ind['YearStart'] != 2001]

ind = ind.loc[ind['YearStart'] != 2007]

ind = ind.loc[ind['YearStart'] != 2008]

```
ind = ind.loc[ind['YearStart'] != 2009]
```

болезни по годам:

```
df = ind.loc[ind['YearStart'] == 2010]
df["Topic"].value_counts().plot.bar(color = 'pink')
df = ind.loc[ind['YearStart'] == 2011]
df["Topic"].value_counts().plot.bar(color = 'plum')
df = ind.loc[ind['YearStart'] == 2012]
df["Topic"].value_counts().plot.bar(color = 'peachpuff')
df = ind.loc[ind['YearStart'] == 2013]
df["Topic"].value_counts().plot.bar(color = 'c')
df = ind.loc[ind['YearStart'] == 2014]
df["Topic"].value_counts().plot.bar(color = 'teal')
df = ind.loc[ind['YearStart'] == 2015]
df["Topic"].value_counts().plot.bar(color = 'lightskyblue')
df = ind.loc[ind['YearStart'] == 2016]
df["Topic"].value_counts().plot.bar(color = 'palegreen')
df = ind.loc[ind['YearStart'] == 2017]
df["Topic"].value_counts().plot.bar(color = 'g')
df = ind.loc[ind['YearStart'] == 2018]
df["Topic"].value_counts().plot.bar(color = 'lime')
```

Болезни по штатам:

```
ind1["LocationDesc"].value_counts().plot.bar(color = 'indigo')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(18.5, 10.5)
fig.savefig('test2png.png', dpi=100)
```

Онкологические заболевания по штатам:

```
df = ind.loc[ind.Topic == 'Cancer']
df["LocationDesc"].value_counts().plot.bar(color = 'pink')
```

```
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(18.5, 10.5)
fig.savefig('test2png.png', dpi=100)
```

Сердечно-сосудистые заболевания по штатам:

```
df = ind.loc[ind.Topic == 'Cardiovascular Disease']
df["LocationDesc"].value_counts().plot.bar(color = 'moccasin')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(18.5, 10.5)
fig.savefig('test2png.png', dpi=100)
```

Обструктивная болезнь легких по штатам:

```
df = ind.loc[ind.Topic == 'Chronic Obstructive Pulmonary Disease']
df["LocationDesc"].value_counts().plot.bar(color = 'lightsalmon')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(18.5, 10.5)
fig.savefig('test2png.png', dpi=100)
```

Диабет по штатам:

```
df = ind.loc[ind.Topic == 'Diabetes']
df["LocationDesc"].value_counts().plot.bar(color = 'khaki')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(18.5, 10.5)
fig.savefig('test2png.png', dpi=100)
```

Хроническое заболевание почек по штатам:

```
df = ind.loc[ind.Topic == 'Chronic Kidney Disease']
df["LocationDesc"].value_counts().plot.bar(color = 'lightcoral')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(18.5, 10.5)
```

```
fig.savefig('test2png.png', dpi=100)
```

Анализ источников данных:

```
ind["DataSource"].value_counts().plot.barh(color = 'teal')
```

Анализ источника BRFSS:

```
df = ind.loc[ind.DataSource == 'BRFSS']
df["LocationDesc"].value_counts().plot.bar(color = 'lightsalmon')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(15, 8)
fig.savefig('test2png.png', dpi=100)
```

Анализ источника NVSS:

```
df = ind.loc[ind.DataSource == 'NVSS']
df["LocationDesc"].value_counts().plot.bar(color = 'm')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(15, 8)
fig.savefig('test2png.png', dpi=100)
```

Анализ источников данных по годам:

```
df = ind.loc[ind['YearStart'] == 2010]
df["DataSource"].value_counts().plot.barh(color = 'hotpink')
df = ind.loc[ind['YearStart'] == 2011]
df["DataSource"].value_counts().plot.barh(color = 'm')
df = ind.loc[ind['YearStart'] == 2012]
df["DataSource"].value_counts().plot.barh(color = 'violet')
df = ind.loc[ind['YearStart'] == 2013]
df["DataSource"].value_counts().plot.barh(color = 'crimson')
df = ind.loc[ind['YearStart'] == 2014]
df["DataSource"].value_counts().plot.barh(color = 'darkmagenta')
```

```
df = ind.loc[ind['YearStart'] == 2015]
df["DataSource"].value_counts().plot.barh(color = 'royalblue')
df = ind.loc[ind['YearStart'] == 2016]
df["DataSource"].value_counts().plot.barh(color = 'navy')
df = ind.loc[ind['YearStart'] == 2017]
df["DataSource"].value_counts().plot.barh(color = 'teal')
```

Кросс-факторный анализ

Болезни по годам:

```
by_cross=ind.groupby(["YearStart", "Topic"]).size().unstack()
sb.heatmap(by_cross, center= 0, cmap= 'spring', linewidths=3, linecolor='black')
plt.title('Болезни по годам')
xlabel('Год')
ylabel('болезнь')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(18.5, 10.5)
fig.savefig('test2png.png', dpi=100)
```

Бользнь по штатам:

```
by_cross1=ind.groupby(["LocationDesc", "Topic"]).size().unstack()
sb.heatmap(by_cross1, center= 0, cmap= 'spring', linewidths=3, linecolor='black')
plt.title('Болезни по штатам')
xlabel('болезнь')
ylabel('штат')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(18.5, 10.5)
fig.savefig('test2png.png', dpi=100)
```

Болезнь по стратификации:

```
by_cross2=ind.groupby(["Stratification1", "Topic"]).size().unstack()
sb.heatmap(by_cross2, center= 0, cmap= 'spring', linewidths=3, linecolor='black')
plt.title('Болезни и стратификация')
xlabel('болезнь')
ylabel('стратификация')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(18.5, 10.5)
fig.savefig('test2png.png', dpi=100)
```

Болезни по штатам по годам:

```
df2011 = df.loc[df['YearStart'] == 2011]
Topic_by_location = df2011.groupby(["Topic","LocationDesc"]).size().unstack()
sb.heatmap(Topic_by_location, center= 0, cmap= 'spring')
plt.title('2011')
xlabel('Location')
ylabel('Topic')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)

df2012 = df.loc[df['YearStart'] == 2012]
Topic_by_location = df2012.groupby(["Topic","LocationDesc"]).size().unstack()
sb.heatmap(Topic_by_location, center= 0, cmap= 'spring')
plt.title('2012')
xlabel('Location')
ylabel('Topic')
fig = matplotlib.pyplot.gcf()
```

```
        fig.set_size_inches(10, 5)
        fig.savefig('test2png.png', dpi=100)

        df2013 = df.loc[df['YearStart'] == 2013]
        Topic_by_location =
df2013.groupby(["Topic","LocationDesc"]).size().unstack()
        sb.heatmap(Topic_by_location, center= 0, cmap= 'spring')
        plt.title('2013')
        xlabel('Location')
        ylabel('Topic')
        fig = matplotlib.pyplot.gcf()
        fig.set_size_inches(10, 5)
        fig.savefig('test2png.png', dpi=100)

        df2014 = df.loc[df['YearStart'] == 2014]
        Topic_by_location =
df2014.groupby(["Topic","LocationDesc"]).size().unstack()
        sb.heatmap(Topic_by_location, center= 0, cmap= 'spring')
        plt.title('2014')
        xlabel('Location')
        ylabel('Topic')
        fig = matplotlib.pyplot.gcf()
        fig.set_size_inches(10, 5)
        fig.savefig('test2png.png', dpi=100)

        df2015 = df.loc[df['YearStart'] == 2015]
        Topic_by_location =
df2015.groupby(["Topic","LocationDesc"]).size().unstack()
        sb.heatmap(Topic_by_location, center= 0, cmap= 'spring')
        plt.title('2015')
```

```
        xlabel('Location')
        ylabel('Topic')
        fig = matplotlib.pyplot.gcf()
        fig.set_size_inches(10, 5)
        fig.savefig('test2png.png', dpi=100)

        df2016 = df.loc[df['YearStart'] == 2016]
        Topic_by_location =
df2016.groupby(["Topic","LocationDesc"]).size().unstack()
        sb.heatmap(Topic_by_location, center= 0, cmap= 'spring')
        plt.title('2016')
        xlabel('Location')
        ylabel('Topic')
        fig = matplotlib.pyplot.gcf()
        fig.set_size_inches(10, 5)
        fig.savefig('test2png.png', dpi=100)

        df2017 = df.loc[df['YearStart'] == 2017]
        Topic_by_location =
df2017.groupby(["Topic","LocationDesc"]).size().unstack()
        sb.heatmap(Topic_by_location, center= 0, cmap= 'spring')
        plt.title('2017')
        xlabel('Location')
        ylabel('Topic')
        fig = matplotlib.pyplot.gcf()
        fig.set_size_inches(10, 5)
        fig.savefig('test2png.png', dpi=100)

        df2018 = df.loc[df['YearStart'] == 2018]
        Topic_by_location=
```

```
df2018.groupby(["Topic","LocationDesc"]).size().unstack()
sb.heatmap(Topic_by_location, center= 0, cmap= 'spring')
plt.title('2018')
xlabel('Location')
ylabel('Topic')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)
```

Геоданные

Разбиваем один столбец на два:

```
df['shirota']=df['GeoLocation'].str.split(',').str.get(0)
df['dolgota']=df['GeoLocation'].str.split(',').str.get(1)
```

Карта штатов:

```
plot(df['dolgota1'], df['shirota2'],'.',ms=0.9, color='DarkBlue', alpha=1)
xlim(-65,-115)
ylim(25,50)
plt.title('Штаты')
```

Карта штатов по годам:

```
df11 = df.loc[df['YearStart'] == 2011]
plot(df11['dolgota1'], df11['shirota2'],'.',ms=1, color = 'DarkBlue', alpha=7)
xlim(-60,-120)
ylim(25,55)
plt.title('Штаты')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)
```

```
df12 = df.loc[df['YearStart'] == 2012]
plot(df12['dolgota1'], df12['shirota2'],'.',ms=1, color = 'DarkMagenta', alpha=7)
xlim(-60,-120)
ylim(25,55)
plt.title('Штаты')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)

df13 = df.loc[df['YearStart'] == 2013]
plot(df13['dolgota1'], df13['shirota2'],'.',ms=1, color = 'Olive', alpha=7)
xlim(-60,-120)
ylim(25,55)
plt.title('Штаты')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)

df14 = df.loc[df['YearStart'] == 2014]
plot(df14['dolgota1'], df14['shirota2'],'.',ms=1, color = 'brown', alpha=7)
xlim(-60,-120)
ylim(25,55)
plt.title('Штаты')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)

df15 = df.loc[df['YearStart'] == 2015]
plot(df15['dolgota1'], df15['shirota2'],'.',ms=1, color = 'green', alpha=7)
```

```
xlim(-60,-120)
ylim(25,55)
plt.title('Штаты')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)

df16 = df.loc[df['YearStart'] == 2016]
plot(df16['dolgota1'], df16['shirota2'],'.',ms=1, color = 'black', alpha=7)
xlim(-60,-120)
ylim(25,55)
plt.title('Штаты')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)

df17 = df.loc[df['YearStart'] == 2017]
plot(df17['dolgota1'], df17['shirota2'],'.',ms=1, color = 'brown', alpha=7)
xlim(-60,-120)
ylim(25,55)
plt.title('Штаты')
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)

df18 = df.loc[df['YearStart'] == 2018]
plot(df18['dolgota1'], df18['shirota2'],'.',ms=1, color = 'red', alpha=7)
xlim(-60,-120)
ylim(25,55)
plt.title('Штаты')
```

```
fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 5)
fig.savefig('test2png.png', dpi=100)
```