МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО БРАЗОВАНИЯ
«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 данных мероприятий Министерства культуры РФ с использованием технологий Big Data»

Студентка _____ Елизавета Александровна Фролова

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

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 Фролова Елизавета Александровна **группы** 161-381

1. **Тема**: «Анализ данных мероприятий министерства культуры РФ с использованием технологий Big Data»
2. **Утверждена** приказом ФГБОУ ВО «Московский политехнический университет» от 22.05.2020 № 301-с
3. **Исходные данные к работе**: открытые данные Министерства за 2015-2020 год
4. **Содержание ВКР (**перечень подлежащих разработке вопросов**)**

№ п/п	Наименование раздела	Содержание раздела
1	Аналитическая часть	1.1 Основные цели и задачи Министерства культуры
		1.2 Главные проекты Министерства культуры
		1.3 Реконструкция и реставрация
		1.4 Развитие театров
		1.5 Развитие музеев
		1.6 Развитие кино
		1.7 Информационные технологии культуры
		1.8 Влияние COVID-19 на проведение мероприятий
		1.9 Обзор данных
2	Теоретическая часть	2.1 Технология Data Mining
		2.2 Процесс внедрения Data Mining
		2.3 Задачи Data Mining
		2.4 Методы Data Mining
		2.5 Языки программирования для работы с Big Data
		2.6 Преимущества языка программирования Python 3
3	Проектная часть	3.1 Подготовка данных
		3.2 Построение первоначальной модели
		3.3 Проверка исходной модели на данных
		3.4 Рекомендации по дальнейшему развитию

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 20.06.2020**

Задание выдал 20.04.20 Задание получил 20.04.2020

Руководитель Студентка

Наталья Ивановна Царькова Елизавета Александровна Фролова

АННОТАЦИЯ

Тема выпускной квалификационной работы: «Анализ данных мероприятий министерства культуры РФ с использованием технологий Big Data».

Работа содержит 92 страницы, 55 рисунков, 7 таблиц и 25 источников.

Цель ВКР: анализ данных о мероприятиях, проводимых Министерством культуры РФ для ознакомления с развитием Министерства культуры в данной сфере, а также выявление сильных и слабых сторон и выводы о дальнейшем развитии и улучшениях, применив к имеющимся данным технологии больших данных.

Данная работа состоит из трех частей:

В Аналитической части были рассмотрены основные действия Министерства культуры по развитию различных сфер культуры и приобщению граждан к мировому культурному и природному наследию.

В Теоретической части был произведен обзор технологий и методов Big Data. Были сгруппированы задачи интеллектуального анализа данных, а также были вывалены методы анализа данных и основной инструмент для анализа (Python 3).

В Проектной части были использованы данные Министерства культуры по проведенным мероприятиям за 2015-2020 года. Данные были подготовлены и преобразованы для анализа, далее была построена исходная и итоговая модели, после чего была произведена их проверка, и сформулированы выводы о дальнейшем развитии и улучшениях.

Ключевые слова: BIG DATA (большие данные), DATA MINING (Интеллектуальный анализ данных), PYTHON 3, МЕРОПРИЯТИЯ В СФЕРЕ КУЛЬТУРЫ, МЕТОД ВИЗУАЛИЗАЦИИ ДАННЫХ, МИНИСТЕРСТВО КУЛЬТУРЫ РФ.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Министерство культуры Российской Федерации (Минкультуры России) является федеральным органом исполнительной власти, осуществляющим функции по выработке и реализации государственной политики и нормативно-правовому регулированию в сфере культуры, искусства, культурного наследия (в том числе археологического наследия), кинематографии, туристской деятельности, авторского права и смежных прав и функции по управлению государственным имуществом и оказанию государственных услуг в сфере культуры и кинематографии, а также по охране культурного наследия, авторского права и смежных прав, по контролю и надзору в указанной сфере деятельности.

В сфере туристской деятельности Минкультуры России осуществляет координацию и контроль деятельности подведомственного ему Федерального агентства по туризму. В сфере международной деятельности и продвижения российской культуры за рубежом Минкультуры России взаимодействует с Министерством иностранных дел Российской Федерации.

При выработке государственной политики и нормативно-правовом регулировании в сфере образования в области культуры и искусства Министерство культуры Российской Федерации осуществляет свою деятельность во взаимодействии с другими федеральными органами исполнительной власти, в том числе с Министерством образования и науки Российской Федерации. Минкультуры России осуществляет государственный контроль и надзор за соблюдением требований законодательства Российской Федерации в сфере защиты детей от информации, причиняющей вред их здоровью и (или) развитию, к обороту информационной продукции, относящейся к аудиовизуальной продукции, на любых видах носителей.

Начиная с 2014 года, Министерство культуры Российской Федерации начало публиковать данные о проведенных мероприятиях в открытом доступе.

Минкультуры России разделило первое место с Минфином России в интегральном рейтинге открытых данных, что свидетельствует как о высоком качестве раскрываемых данных, так и об их востребованности.

Интегральный рейтинг публикации информации в формате открытых данных учитывает все показатели публикационной активности, востребованности, качества опубликованных наборов, а также степень выполнения требований законодательства.

До настоящего момента цели и задачи Минкультуры России в рамках реализации государственной политики в закрепленной сфере в наибольшей степени определялись содержанием Государственной программы «Развитие культуры и туризма» на 2013-2020 годы, а также Федеральными целевыми программами, в выполнении которых оно принимало участие. После принятия Стратегии государственной культурной политики возникла необходимость гармонизации целей и задач Государственной и федеральных целевых программ с положениями и требованиями утвержденной Стратегии [1].

Довольно ценной особенностью интеллектуального анализа данных является возможность получение ответов на широкий спектр вопросов. Ведь с его помощью можно не только определить степень развития в данной сфере и количество мероприятий за определенный промежуток времени, но и понять наиболее популярные направления, в зависимости от множества факторов. Более того, с помощью такого анализа можно определить направление дальнейшего развития.

Следовательно, интеллектуальный анализ больших данных в сфере мероприятий министерства культуры РФ будет также практически применим и будет иметь цену для самого Минкультуры РФ.

Цель работы – анализ данных мероприятий министерства культуры РФ с использованием технологий Big Data.

Для достижения поставленной цели были определены следующие задачи:

1. Анализ деятельности Минкультуры РФ.
2. Подготовка данных для анализа.
3. Построение исходной модели.
4. Проверка корректности модели на имеющихся данных.
5. Произведение корректировки модели.
6. Формирование рекомендаций и итоговой модели.

Объектом работы является Министерство культуры Российской Федерации.

Предметом исследования являются данные Министерства культуры Российской Федерации о проведенных мероприятиях по всей России, содержащие в себе более 550 тысяч записей за последние 6 лет. Данные включают в себя более 120 параметров, включая: начало мероприятия, его название, краткое описание, место проведения и т.д.

В рамках данной работы использовались такие техники, как интеллектуальный анализ больших данных (Big Data Mining), визуализация данных, кросс-факторный анализ данных.

Основными источниками данных для анализа, являются портал открытых данных Министерства культуры Российской Федерации и официальный сайт Министерства культуры.

1 АНАЛИТИЧЕСКАЯ ЧАСТЬ

Культура России, формировавшаяся на протяжении столетий, является гордостью и главным богатством многонационального российского народа, одним из ключевых факторов его единства. Важнейшая роль культуры в жизни российского общества закреплена в официальных документах — Основах государственной культурной политики и Стратегии государственной культурной политики РФ на период до 2030 года.

В 2018 году по поручению Президента РФ дан старт подготовке нового всеобъемлющего закона о культуре, который позволит отразить особенности и специфику отрасли, а также закрепить конкретные экономические механизмы её поддержки.

1.1 Основные цели и задачи Министерства культуры

С 2012 года произошло глобальное переосмысление приоритетов деятельности Минкультуры. Особое внимание ведомства сосредоточено на решении проблем регионов. Значительная часть ежегодно выделяемых на культуру средств направляется в субъекты, преимущественно в малые города и сёла. Министерство культуры Российской Федерации осуществляет множество функций в сфере культуры, культурного и археологического наследия, искусства, кинематографии и т.д.

Все цели и задачи Министерства культуры РФ в наибольшей степени определяется содержанием Государственной программы «Развитие культуры и туризма» на 2013-2020 годы. Также Минкультуры приняло участие в выполнении Федеральных целевых программ.

После того, как была принята Стратегия государственной культурной политики появилась необходимость в том, чтобы гармонизировать все цели, а также и задачи Государственной и федеральных целевых программ с учетом всех положений, и требований утвержденной Стратегии.

В задачи данной Государственной программы входят:

1. Сохранение культурного и исторического наследия народа, обеспечение доступа граждан к культурным ценностям и участию в культурной жизни, реализация творческого потенциала нации;
2. Повышение качества и доступности услуг в сфере внутреннего и международного туризма;
3. Создание благоприятных условий для устойчивого развития сфер культуры и туризма.

1.2 Главные проекты Министерства культуры

В связи с указами Президента РФ, в России были проведены определенные мероприятия и проекты, с целью популяризации искусства, литературы и их сохранения.

1.2.1 Год культуры

2014 год официально был объявлен годом культуры в России. Президент Владимир Путин подписал соответствующий указ с целью: «"…привлечения внимания общества к вопросам развития культуры, сохранения культурно-исторического наследия и роли российской культуры во всем мире» [2].

Этот год стал определенным стимулом для множества субъектов РФ, после чего были приняты определенные меры по развитию культуры. Основой для изменений стала разработка целевых программ для регионов страны.

Это был своего рода сигнал, благодаря которому начали происходить существенные изменения: развитие государственно-частного партнёрства, развитие благотворительности, повышение престижа профессии работника культуры, а также у жителей РФ начало появляться больше возможностей для саморазвития и отдыха, связанного с культурной деятельностью.

В основной программе Года культуры была поддержка музеев, различных учреждений культуры, коллективов народного творчества, театров. Были выделены средства на поддержку проектов, связанных с сохранением исторического облика малых городов России, а также на строительство многофункциональных культурных центров. Были проведены различные

мероприятия для поддержки библиотек, отечественного кино и российского фольклора.

Согласно созданному указу, правительство РФ создало специальный организационный комитет, который провел Год культуры и обеспечил разработку плана по проведению всех основных мероприятий этого года.

1.2.2 Год литературы

2015 год в России был официально объявлен Годом литературы. В данный проект было включено более ста мероприятий, связанных с различными ярмарками, фестивалями и выставками, проводимых по всей стране.

Год литературы начался с церемонии, прошедшей 28 января с трансляции в прямом эфире в МХТ имени Чехова.

"Цель самого проведения Года литературы – напомнить об исключительной ее – литературы – значимости и ее особой миссии, – такие слова были произнесены Президентом РФ на самой церемонии. – Рассчитываю, что Год литературы, действительно, пройдет широко, и в столицах, и во всех российских регионах, поможет вернуть в нашу жизнь, жизнь молодежи понимание хорошей литературы и, конечно, самого слова, всех удивительных возможностей нашего родного языка, который, по праву входит в число самых выразительных и образных языков мира" [3].

1.2.3 Год российского кино

2016 год был официально объявлен в России Годом Кино. Основной целью являлась популяризация именно отечественного кинематографа, а также увеличение количества выпускаемых кинофильмов и продвижение киноискусства в различные регионы нашей большой страны.

Были проведены модернизация и реорганизация «Ленфильма» и Киностудии имени Горького. Также Музей кино был перенесен в один из павильонов на ВДНХ и стал так называемым «центром пропаганды

российского киноискусства». «Союзмультфильм» был перенесен в свое собственное здание, находящееся на улице Академика Королева.

Был создан и запущен проект по увеличению кинозалов в средних, а также малых городов России, благодаря которому открылось более 150 новых залов, для показа кинофильмов.

В вечернем эфире телеканала «Россия К», начиная с 28 августа, была показана премьера авторского цикла Александра Казакевича «Библиотека приключений» вместе с лучшими приключенческими фильмами отечественного и мирового кинематографа.

Можно с уверенностью заявить, что после распада СССР кино пережило огромное количество действительно трудных моментов. Несмотря на это, появилось множество талантливых людей, которые были готовы к разного рода трудностям и посвятили свою жизнь творчеству, новым идеям и развитию культуры страны.

1.2.4 Год театра

Прошедший 2019 год был официально объявлен годом театра в России. В Москве 28 апреля был подписан соответствующий указ.

До 1 июня 2018 года был сформирован специальный комитет для организации и проведения Года театра. Был полностью утвержден состав и был разработан план всех проводимых мероприятий.

Была проведена встреча между Президентом Владимиром Путиным и Александром Калягиным, председателем Союза театральных деятелей (СТД), а также руководителем московского театра Et Cetera.

«Помимо тех проблем, которые мы бы решили в театральный год, в Год театра, мы наметили бы какой-то путь и составили стратегический план до 2030 года», – отметил при встрече Александр Калягин.

Также председатель СТД заявил, что большинство театров требуют капитального ремонта в связи со своим плачевным состоянием. Президент, в свою очередь, согласился на рассмотрение мер, предложенных Калягиным.

1.3 Реконструкция и реставрация

Реконструкция и реставрация различных зданий, музеев и помещений является одной из главных задач для Министерства культуры РФ, так как именно благодаря этому увеличивается количество мест для проведения различного рода выставок, концертов и других мероприятий.

Всего с 2012 по 2019 годы было построено, отреставрировано и реконструировано более 900 объектов, включая:

- 585 сельских домов культуры
- 39 центров инновационного и культурного развития
- 34 театра
- 10 цирков
- Более 220 объектов культурного наследия
- Более 40 объектов федеральных учреждений культуры

Начиная с 2017 года Министерство культуры РФ создало программу по строительству и реконструкции сельских домов культуры, на что правительством ежегодно выделяется около 2,9 млрд рублей.

1.4 Развитие театров

Начиная с 2012 года, были предприняты определенные меры для развития мероприятий в данной области. В 2014 году Министерство культуры проявило инициативу и возродило гастрольную программу для всех театров России, благодаря чему данная область начала свое развитие (Рисунок 1.1)

Рисунок 1.1 Результаты гастрольной программы.

Также был принят в силу проект «Культура малой Родины», главная задача которого заключается в повышении качества работы учреждений культуры, а также возможность обеспечения доступа к участию культурной жизни страны для всех граждан РФ. Немаловажно была и возможность для получения дополнительного образования, и разнообразия досуга.

На 2019 год по данной программе были выплачены субсидии в размере: 900 млн рублей 154 детским театрам, 700 млн рублей 152 театрам малых городов, 1,4 млрд рублей 5,9 тысячам домам культуры.

1.5 Развитие музеев

Стратегия развития деятельности музеев в РФ была сформулирована и запланирована на период до 2030 года. Были приняты меры по реконструированную и строительству музеев, а также разработана специальная программа по мероприятиям, проходимым в самих музеях, для увеличения посещаемости и культурного развития граждан РФ (Рисунок 1.2).

Рисунок 1.2 Результаты стратегии развития деятельности музеев

1.6 Развитие кино

Были предприняты меры по развитию кино, для увеличения количества зрителей кинозалов и роста сборов, что к 2018 увеличило долю российского кино по кассовым сборам в 1,7 раз (Рисунок 1.3).

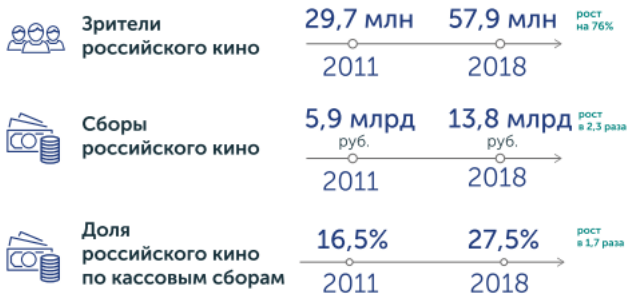


Рисунок 1.3 Результаты стратегии развития деятельности кино.

Также были проведены программы по реконструкции и строительству кинозалов в городах и других населенных пунктах, где численность населения не превышает 500 тыс. чел.

В результате программ с 2015 по 2019 года были поддержаны заявки более 1000 кинозалов в 878 населенных пунктах, а посещаемость открытых кинозалов в 2019 составила 10,9 млн чел, в то время, как кассовые сборы превысили 2 млрд рублей.

1.7 Информационные технологии культуры

В 2013 году появился интернет портал «Культура.РФ», который был посвящен культуре России. Данный портал предоставляет возможность узнать о значимых событиях и заниматься саморазвитиям в совершенно различных сферах.

На платформе представлены фильмы, спектакли, лекции, статьи, книги и многое другое. Более того, тут проводятся прямые трансляции, создаются мультимедийные проекты, и проводятся виртуальные туры по разным музеям и красивым местам. На платформе находится огромное количество информации как для самообразования и развития, так и для интересного проведения досуга, что не мало важно, в связи с нынешней ситуацией, когда нет возможности организовать какое-либо мероприятие из-за ограничений и мер предосторожности.

Данный проект – прекрасный вариант для организации онлайн-мероприятий, связанных с развитием и продвижением культуры России в массы, благодаря своей доступности и наличием огромного количества полезной и интересной информации.

1.8 Влияние COVID-19 на проведение мероприятий

В 2020 году произошла пандемия коронавируса, в результате в России были объявлены дни самоизоляции. В Москве нерабочие дни были объявлены с 30 марта до 3 апреля, для предотвращения распространения вируса, позднее они были продлены до 14 апреля, а после и до 31 мая.

В эти даты обычным гражданам можно покидать квартиру только при необходимости: оказание медицинской помощи, выгул питомца, поход в магазин и аптеку, поездка на работу, а также вынос мусора.

Безусловно, в такой ситуации не идет и речи о посещениях разного рода развлекательных программ, так как это способствует скоплению большого количества людей, также данные события по факту не несут никакой необходимости или значимости.

В данной ситуации количество проводимых мероприятий должно значительно снизиться, однако возможен вариант переноса определенных событий в формат онлайн. Например, возможны проведения лекций, показы фильмов, прямые эфиры, виртуальные экскурсии и многое другое.

В связи с данной ситуацией стало понятно, что введение определенных мер необходимо, иначе вся работа, проводимая Министерством культуры в формате мероприятий будет невозможна, что означает огромное количество потраченных средств и невыполненные планы и стратегии, чего допускать нельзя.

1.9 Обзор данных

Все данные были взяты с официального портала открытых данных Министерства культуры. В них хранится информация о всех мероприятиях, проводимых Министерством культуры РФ, начиная с 2014 года, по сегодняшний день. Информация периодически обновляется и перезаписывается, данные, предоставленные в работе были загружены 18 мая 2020 года.

Исходя из целей, поставленных Министерством культуры Российской Федерации, а также деятельности, осуществляемой им, можно с уверенностью сказать, что анализ просто необходим для выявления наиболее благоприятной и эффективной стратегии действий.

Так как большое количество мероприятий в сфере культуры проводятся на бесплатной основе, за счет выделенного государством бюджета, то особенно необходимо понять предпочтения людей, а также актуальность того или иного вида мероприятия в соответствии с временем и датой проведения. Более того, с помощью анализа можно выявить наиболее сильные и слабые сферы, в которых происходят мероприятия, что может помочь определить путь дальнейшего развития и возможное перераспределение бюджета.

Также можно проверить, насколько Министерство культуры Российской Федерации готово к ведению мероприятий онлайн, в связи с карантином. Можно отследить количество таких проводимых и запланированных мероприятий.

2 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

2.1 Технология Data Mining

Интеллектуальный анализ данных, в области компьютерных наук – процесс обнаружения интересных и полезных моделей и взаимосвязей в больших объемах данных. Эта область объединяет инструменты из статистики и искусственного интеллекта (такие как нейронные сети и машинное обучение) с управлением базой данных для анализа больших цифровых коллекций, известных как наборы данных. Интеллектуальный анализ данных широко используется в бизнесе (страхование, банковское дело, розничная торговля), научных исследованиях (астрономия, медицина) и государственной безопасности (обнаружение преступников и террористов).

Распространение многочисленных крупных, а иногда и связанных, правительственных и частных баз данных привело к принятию правил, обеспечивающих точность и безопасность отдельных записей от несанкционированного просмотра или взлома. Большинство типов интеллектуального анализа данных нацелены на выяснение общих знаний о группе.

Data Mining – это процесс, который используется различными компаниями для преобразования необработанных данных в полезную и нужную информацию. При использовании ПО для поиска шаблонов в больших пакетах данных компании могут больше узнавать о своих клиентах, разрабатывать более эффективные маркетинговые стратегии, увеличивать продажи и снижать затраты. Интеллектуальный анализ данных зависит от нескольких факторов: от того, насколько эффективен сбор данных, от складирования, а также от компьютерной обработки.

Программы интеллектуального анализа данных анализируют отношения и закономерности в данных, основываясь на запросах пользователей. Например, компания может использовать программное

обеспечение для интеллектуального анализа данных для создания классов информации. Например, в нашем случае, мы хотим использовать интеллектуальный анализ данных, чтобы определить, какое время является наиболее популярным и благоприятным для проведения мероприятия, а именно: время года, месяц, день. Он просматривает собранную информацию и создает классы на основе того, когда люди посещают мероприятия, а также тип самого мероприятия.

Складирование является важным аспектом интеллектуального анализа данных. Складирование – централизация данных компании в одной базе данных или программе. С помощью хранилища данных организация может выделять сегменты данных для анализа и использования конкретными пользователями.

Однако в других случаях аналитики могут начать с нужных им данных и создать хранилище данных на основе этих спецификаций. Независимо от того, как предприятия и другие организации организуют свои данные, они используют их для поддержки процессов принятия решений руководством.

2.2 Процесс внедрения Data Mining

Для того, чтобы воспользоваться анализом данных с помощью Data Mining, необходимо выявить все его процессы и следовать определенной структуре. Давайте подробно изучим процесс внедрения Data Mining.

CRISP-DM – это надежная модель интеллектуального анализа данных, состоящая из шести этапов. Это циклический процесс, который обеспечивает структурированный подход к процессу интеллектуального анализа данных. Шесть этапов могут быть реализованы в любом порядке, но иногда потребуется откат к предыдущим шагам и повторение действий.

Этапы внедрения Data Mining:

- Анализ предметной области
- Проверка данных
- Подготовка данных

- Построение моделей
- Проверка и оценка моделей
- Представление полученных знаний

2.2.1 Анализ предметной области

Исследование: в предельно широком смысле – поиск новых знаний или систематическое расследование с целью установления фактов; в более узком смысле исследование – научный метод изучения чего-либо; результат такого действия, научный труд, документ с описанием изученного объекта или чего-либо. [12]

Решение абсолютно любой задачи в сфере разработки программного обеспечения начинается именно с изучения предметной области.

Предметная область – множество объектов, рассматриваемых в пределах отдельного рассуждения, научной теории. П. о. включает прежде всего индивиды, т. е. элементарные объекты, изучаемые теорией, а также свойства, отношения и функции, рассматриваемые в теории. [8]

Знания из различных источников должны быть форматизированы при помощи каких-либо средств. Это могут быть текстовые описания предметной области или специализированные графические нотации. Модель предметной области описывает процессы, происходящие в предметной области, и данные, которые в этих процессах используются. [10]

Таким образом, на этапе анализа предметной области необходимо:

1. Четко понять цели и выяснить, каковы потребности от самого анализа.
2. Оценить текущую ситуацию, найти ресурсы, предположения, ограничения и другие важные факторы, которые следует учитывать.
3. Исходя из бизнес-целей и текущих ситуаций, создать цели интеллектуального анализа данных для достижения поставленных целей в текущей ситуации.

2.2.2 Проверка данных

В самом начале необходимо выполнить проверку работоспособности данных, чтобы убедиться в том, что они подходят для поставленных нами целей анализа. В противном случае от них нельзя будет получить никакой полезной информации, либо мы получим некорректные и ложные сведения.

Существуют такие проблемы, как сопоставление объектов и интеграция схем, которые могут возникнуть в процессе интеграции данных. В начале, при необходимости, данные собираются из нескольких источников, доступных в организации. Эти источники могут включать в себя несколько баз данных, плоский файл или кубы данных.

В нашем случае данные были взяты из плоского файла, что означает собрание записей в определенном формате одна за другой, т. е. список. [13]

Далее необходимо выполнить поиск свойств, полученных данных. Хороший способ исследовать данные – это ответить на вопросы интеллектуального анализа, используя инструменты запросов, отчетов и визуализации. На основании результатов запроса должно быть установлено качество данных, также можно произвести проверку на их отсутствие для дальнейшей корректировки и заполнения пустых значений при необходимости.

Уверенность в том, что процесс сбора не приводит к неточностям, поможет обеспечить общее качество последующего анализа.

2.2.3 Подготовка данных

Процесс подготовки данных может занимать значительное время от выполнения всего проекта.

Интеграция данных и их очистка просто необходимы, когда речь идет об анализе и построении моделей. Под этими терминами понимается переход от необработанных, «грязных данных» к статистическому анализу.

Грязные данные являются неточными, неполными или противоречивыми данными. [20] Они ежегодно обходятся компаниям в

миллионы долларов. В частности, ошибки и упущения вызывают дорогостоящие перерывы в работе.

Существует огромное количество типов грязных данных, рассмотрим самые популярные из них:

1. Неполные данные

Важные поля в записях основных данных, полезные для решений определенных целей и задач, часто остаются пустыми.

2. Дублирующиеся данные

Большинство компаний имеют дело с дублирующимися записями клиентов. Это может дорого стоить компании из-за избытка запасов и неоптимальных решений.

3. Неверные данные

Неверные данные могут возникать, когда значения полей создаются вне допустимого диапазона значений. Например, значение в поле месяца должно находится в промежутке от 1 до 12.

4. Неточные данные

Бывают случаи, когда данные являются технически правильными, но неточными, учитывая бизнес-контекст. Например, незначительные ошибки в адресах клиентов могут привести к доставке в неправильные местоположения, даже если адреса являются фактическими адресами.

В нашем случае может стоять некорректная цена посещения мероприятия.

5. Нарушения бизнес-правил

Часто существуют большие коллекции плохо документированных бизнес-правил, связанных с основными данными, которые специфичны для отрасли или бизнес-контекста. Например, дата вступления в силу всегда должна предшествовать дате истечения срока действия.

6. Несовместимые данные

Избыточность данных, т.е. одинаковые значения полей, хранящиеся в разных местах, часто приводят к несоответствиям. Например, большинство компаний хранят информацию о клиентах в нескольких системах, и данные часто не синхронизируются.

При работе с набором данных необходимо учитывать перечисленные выше проблемы, чтобы в дальнейшем правильно откорректировать значения и выполнить анализ без ошибок.

После ввода почти всегда необходимо преобразовать необработанные данные в переменные, которые можно использовать при анализе. Существует множество различных преобразований, которые можно, а иногда просто необходимо выполнить.

Рассмотрим наиболее распространенные из них:

1. Недостающие значения

Многие программы анализа автоматически обрабатывают пустые значения как пропущенные. В других случаях необходимо указать конкретные значения для их представления. Например, можно заполнить их цифрами или текстом, который будет показывать, что данные в определенном месте не были заполнены.

2. Сторнирование предметов

В шкалах и опросах иногда используются элементы реверсирования, чтобы уменьшить вероятность набора ответов. Когда происходит анализ данных, следует сделать так, чтобы все баллы для элементов шкалы были в одном направлении, где высокие и низкие баллы означают одно и то же. В этих случаях необходимо отменить оценки для некоторых из пунктов шкалы.

3. Очистка данных

В некоторых случаях могут быть выбросы данных, значения, сильно отличающиеся от остальных. Например, могут встречаться аномальные значения в столбце с ценой, что говорит об опечатке или некорректном

заполнении, такие значения необходимо будет убрать из таблицы. Также эту проблему можно будет убрать с помощью нормализации, а значит поставить ограничение по максимуму или минимуму.

4. Преобразование значений

В нашем случае имеется столбец с датой и временем. Изначально тип данного столбца String, однако лучше перевести данный столбец в тип Date для удобства дальнейших вычислений и анализа.

В результате будет получен окончательный набор данных, который уже можно будет использовать при моделировании без последующих неудобств или ошибок.

2.2.4 Построение моделей

На этом этапе математические модели используются для определения структуры данных.

При построении аналитической модели важно помнить, что она должна давать значимые и интерпретируемые результаты реальных ситуаций.

Различные алгоритмы применяются к данным для получения полезной и нужной информации из них. Алгоритмы могут быть классифицированы как описательные, прогнозирующие или предписывающие.

Каждая модель машинного обучения должна давать значимые и интерпретируемые результаты реальных ситуаций. Прогностическая модель должна быть проверена на соответствие действительности, чтобы считаться значимой и полезной. Поэтому человеческий вклад и опыт всегда необходимы и полезны для осмысления результатов, полученных с помощью алгоритмов.

Квалифицированный специалист по данным должен быть в состоянии продемонстрировать доказательства успешного завершения реального проекта по науке о данных, который включает в себя все этапы рабочего процесса в области и машинного обучения, такие как формирование проблем, сбор и анализ данных, построение, тестирование моделей, ее оценка и развертывание.

2.2.5 Проверка и оценка моделей

На этапе оценки результаты модели должны оцениваться в контексте поставленных на первом этапе целей. На этом этапе новые требования могут быть повышены из-за новых шаблонов, обнаруженных в результатах модели, или из-за других факторов.

Проверка модели определяется в нормативном руководстве как «набор процессов и действий, предназначенных для проверки того, что модели работают так, как ожидается, в соответствии с их целями проектирования и бизнес-применениями». Оно также определяет «потенциальные ограничения и предположения и оценивает их возможное влияние».

Как правило, действия по проверке выполняются лицами, независимыми от разработки или использования модели. Следовательно, модели не должны проверяться их владельцами, поскольку они могут быть очень техническими, и некоторые учреждения могут столкнуться с трудностями при создании группы по риску для моделей, которая обладает достаточным функциональным и техническим опытом для проведения независимой проверки. Столкнувшись с этим препятствием, учреждения часто передают задачу проверки третьим сторонам.

В статистике проверка модели является задачей подтверждения того, что результаты статистической модели приемлемы по отношению к реальному процессу генерирования данных. Другими словами, проверка модели – это задача подтверждения того, что выходные данные статистической модели имеют достаточную точность по отношению к выходным данным процесса генерирования данных, чтобы можно было достичь целей исследования.

Проверка модели состоит из четырех важных элементов, которые следует учитывать:

1. Концептуальный дизайн

Основой любой валидации модели является ее концептуальный дизайн, который требует документированной оценки покрытия, которая поддерживает

способность модели удовлетворять потребности бизнеса и регулятора, а также уникальные риски. Дизайн и возможности модели могут оказать глубокое влияние на общую эффективность.

Проверка должна независимо оспаривать базовый концептуальный дизайн и обеспечивать соответствие документации для поддержки логики модели и способности модели достигать желаемых нормативных и бизнес-результатов, для которых она предназначена.

2. Проверка системы

Все технологии и автоматизированные системы, реализованные для поддержки моделей, имеют ограничения.

Эффективная проверка включает в себя: во-первых, оценку процессов, используемых для интеграции концептуального дизайна и функциональности модели в бизнес-среду организации; и, во-вторых, изучение процессов, реализованных для выполнения общего дизайна модели. Там, где наблюдаются пробелы или ограничения, следует оценить элементы управления, чтобы модель могла эффективно функционировать.

3. Проверка данных и оценка качества

Ошибки или неточности в данных ухудшают результаты и могут привести к тому, что организация не сможет определить риски и ответить на них. Передовой опыт показывает, что учреждения должны применять проверку данных на основе рисков, что позволяет рецензенту учитывать риски, уникальные для организации и модели.

Чтобы установить надежную структуру для проверки данных, руководство указывает, что точность исходных данных должна быть оценена. Это очень важный шаг, поскольку данные могут быть получены из различных источников, в некоторых из которых может отсутствовать контроль целостности данных, поэтому данные могут быть неполными или неточными.

4. Проверка процесса

Чтобы убедиться, что модель работает эффективно, важно доказать, что установленные процессы для текущего администрирования модели, включая политики и процедуры управления, поддерживают устойчивость модели. Обзор процессов также определяет, дают ли модели точные, эффективные результаты и подлежат ли соответствующему контролю.

Если все сделано эффективно, проверка модели позволит быть уверенным в точности своих различных моделей, а также привести их в соответствие с деловыми и нормативными ожиданиями.

2.2.6 Представление полученных знаний

Знания или информация, полученные в процессе анализа данных, должны быть представлены таким образом, чтобы заинтересованные стороны могли использовать их в любое время. Исходя из прописанных целей и требований, данный этап может быть таким же простым, как создание отчета, или таким сложным, как повторяющийся процесс анализа данных во всей организации.

На этапе представления полученных знаний должны быть созданы планы развертывания, обслуживания и мониторинга для реализации, а также поддержки в будущем: формируется стратегия мониторинга и поддержки результатов модели интеллектуального анализа данных для проверки ее полезности, составляются окончательные отчеты и проводится проверка всего процесса, чтобы проверить любую ошибку и посмотреть, повторяется ли какой-либо шаг.

2.3 Задачи Data Mining

Интеллектуальный анализ данных может быть использован для решения сотен бизнес-задач. Исходя из характера этих проблем, мы можем сгруппировать их в следующие задачи интеллектуального анализа данных.

1. Классификация

Классификация является одной из самых популярных задач интеллектуального анализа данных. Бизнес-проблемы, такие как анализ

оттока, управление рисками и таргетинг объявлений, обычно включают классификацию.

Классификация относится к распределению случаев по категориям на основе предсказуемого атрибута. Каждый случай содержит набор атрибутов, одним из которых является атрибут класса (прогнозируемый атрибут). Задача требует поиска модели, которая описывает атрибут класса как функцию входных атрибутов.

Для обучения модели классификации необходимо знать значение класса входных случаев в наборе данных обучения, которые обычно представляют собой исторические данные.

Алгоритмы интеллектуального анализа данных, для которых требуется обучение, рассматриваются как контролируемые алгоритмы. Типичные алгоритмы классификации включают деревья решений, нейронную сеть и наивный байесовский алгоритм.

2. Кластеризация

Кластеризация также называется сегментацией. Она используется для идентификации естественных группировок на основе набора атрибутов. Случаи в одной группе имеют более или менее сходные значения атрибутов.

Алгоритм кластеризации группирует набор данных в три сегмента на основе этих двух атрибутов (Рисунок 2.1). Кластер 1 содержит более молодое население с низким уровнем дохода. Кластер 2 содержит клиентов среднего возраста с более высокими доходами. Кластер 3 – это группа пожилых людей с относительно низким доходом. Кластеризация – это задача неуправляемого анализа данных. Ни один атрибут не используется для управления процессом обучения. Все входные атрибуты обрабатываются одинаково. Большинство алгоритмов кластеризации строят модель через ряд итераций и останавливаются, когда модель сходится, то есть когда границы этих сегментов стабилизируются.

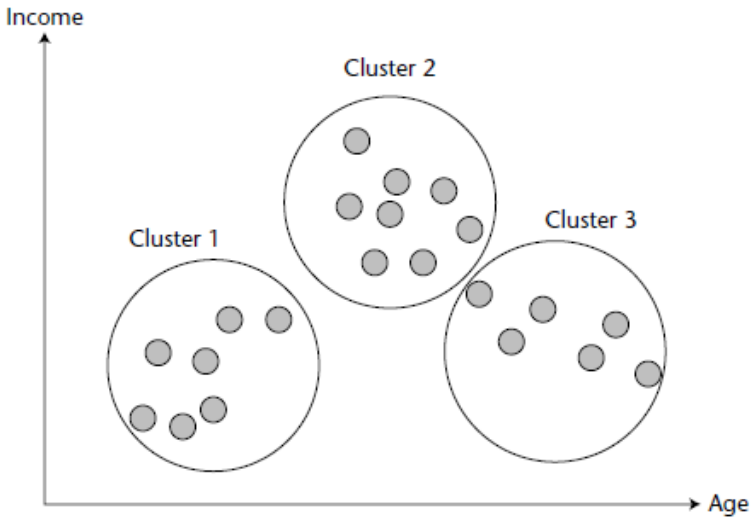


Рисунок 2.1 Распределение кластеров

3. Ассоциация

Ассоциация является еще одной популярной задачей интеллектуального анализа данных. Типичная бизнес-проблема ассоциации – анализ таблицы транзакций продаж и определение продуктов, которые часто продаются в одной и той же корзине покупок. Обычно используется ассоциация для определения общих наборов товаров (частые наборы товаров) и правил для перекрестных продаж.

Вероятность также упоминается как достоверность в литературе интеллектуального анализа данных. Вероятность – это пороговое значение, которое пользователь должен указать перед обучением модели ассоциации.

4. Регрессия

Задача регрессии аналогична классификации. Основное отличие состоит в том, что прогнозируемый атрибут представляет собой непрерывное число.

Методы регрессии веками широко изучались в области статистики. Линейная регрессия и логистическая регрессия являются наиболее популярными методами регрессии. Другие методы регрессии включают деревья регрессии и нейронные сети.

5. Прогнозирование

Прогнозирование является еще одной важной задачей интеллектуального анализа данных.

Обычно метод принимает в качестве входного набора данных временного ряда, например, последовательность чисел с атрибутом, представляющим время. Данные временного ряда обычно содержат смежные наблюдения, которые зависят от порядка. Методы прогнозирования имеют дело с общими тенденциями, периодичностью и шумовой фильтрацией.

6. Анализ последовательности

Анализ последовательности используется для поиска шаблонов в дискретном ряду. Последовательность состоит из ряда дискретных значений (или состояний). Например, последовательность ДНК представляет собой длинный ряд, состоящий из четырех различных состояний: A, G, C и T. Последовательность щелчков в Интернете содержит ряд URL-адресов.

Данные как последовательности, так и временного ряда содержат смежные наблюдения, которые являются зависимыми. Разница заключается в том, что ряд последовательности содержит дискретные состояния, а временной ряд содержит непрерывные числа. Данные последовательности и ассоциации похожи в том смысле, что каждый отдельный случай содержит набор элементов или состояний.

Различие между моделями последовательности и ассоциации заключается в том, что модели последовательности анализируют переходы состояний, в то время как модель ассоциации рассматривает каждый элемент в корзине покупок как равный и независимый.

7. Анализ отклонений

Это также называется обнаружением выброса, которое относится к обнаружению значительных изменений от ранее наблюдаемого поведения.

Анализ отклонений может использоваться во многих приложениях. Наиболее распространенным является обнаружение мошенничества с кредитными картами. Выявить необычные случаи из миллионов транзакций –

очень сложная задача. Другие приложения включают обнаружение вторжений в сеть, анализ производственных ошибок и так далее. Не существует стандартной методики для анализа отклонений.

2.4 Методы Data Mining

2.4.1 Нейронные сети

Нейронные сети представляют собой метафору мозга для обработки информации, так как именно нейронные связи в мозге были вдохновением для создание этого типа моделей. Нейронные сети оказались очень перспективными системами во многих приложениях для прогнозирования и классификации бизнеса благодаря их способности «извлекать уроки» из данных, их непараметрическому характеру (т. е. без жестких допущений) и способностью обобщать.

Нейронные вычисления относятся к методологии распознавания образов для машинного обучения. Полученная модель из нейронного вычисления часто называют искусственной нейронной сетью (artificial neural network или ANN) или просто нейронной сетью.

Метод нейронной сети используется для классификации, кластеризации, анализа возможностей, прогнозирования и распознавания образов. Он имитирует структуру нейронов животных, базируется на модели M-P и правилах обучения Хебба, поэтому по сути это структура распределенных матриц. С помощью интеллектуального анализа данных метод нейронной сети постепенно вычисляет (включая повторную итерацию или совокупный расчет) веса подключенной нейронной сети.

2.4.2 Линейная регрессия

Регрессия – это метод интеллектуального анализа данных, используемый для прогнозирования диапазона числовых значений (также называемых непрерывными значениями) для конкретного набора данных.

Регрессия используется во многих отраслях для планирования бизнеса и маркетинга, финансового прогнозирования, моделирования окружающей среды и анализа тенденций.

Самая простая форма регрессии – это линейная регрессия, используемая для оценки взаимосвязи между двумя переменными. Этот метод использует математическую формулу прямой линии (y = mx + b). Проще говоря, это просто означает, что, учитывая график с осью Y и X, связь между X и Y является прямой линией с небольшим количеством выбросов.

Продвинутые методы, такие как множественная регрессия, предсказывают взаимосвязь между несколькими переменными. Добавление большего количества переменных значительно увеличивает сложность прогноза. Существует несколько типов методов множественной регрессии, включая стандартные, иерархические, пошаговые и пошаговые, каждый со своим приложением.

2.4.3 Автокорреляционная функция

Автокорреляция относится к степени корреляции между значениями одних и тех же переменных в разных наблюдениях в данных. Концепция автокорреляции чаще всего обсуждается в контексте данных временных рядов, в которых наблюдения происходят в разные моменты времени. Например, можно ожидать, что мероприятий будет больше ближе к началу праздников. Если значения количества мероприятий, которые произошли ближе друг к другу во времени, на самом деле более похожи, чем значения, которые произошли дальше друг от друга во времени, данные будут автоматически коррелированы.

Когда автокорреляция используется для обнаружения неслучайности, обычно представляет интерес только первая (лаг 1) автокорреляция. Когда автокорреляция используется для определения подходящей модели временных рядов, автокорреляции обычно строятся для многих лагов.

2.4.4 Деревья решений

Дерево решений – это контролируемый подход к обучению, мы обучаем имеющиеся данные, уже зная, какова целевая переменная на самом деле. Как следует из названия, этот алгоритм имеет древовидную структуру.

В Дереве решений алгоритм разбивает набор данных на подмножества на основе наиболее важного или значимого атрибута. Наиболее значимый атрибут обозначен в корневом узле, где и происходит разделение всего набора данных, присутствующего в корневом узле. Такое разделение известно, как узлы принятия решений. Если разделение больше невозможно, этот узел называется конечным узлом.

Чтобы остановить алгоритм, для достижения определенной стадии, используется критерий остановки. Одним из критериев остановки является минимальное количество наблюдений в узле до разделения. Применяя дерево решений для разделения набора данных, нужно быть осторожным, так как многие узлы могут просто содержать зашумленные данные.

2.4.5 Полиномиальная нейронная сеть

Полиномиальная нейронная сеть известна также, как групповой метод обработки данных (Group Method of Data Handling – GMDH). Он применялся в самых разных областях для глубокого обучения и обнаружения знаний, прогнозирования и анализа данных, оптимизации и распознавания образов. Индуктивные алгоритмы GMDH позволяют автоматически находить взаимосвязи в данных, выбирать оптимальную структуру модели или сети и повышать точность существующих алгоритмов.

Этот оригинальный самоорганизующийся подход отличается от дедуктивных методов, используемых для моделирования. Он имеет индуктивный характер – он находит лучшее решение путем сортировки возможных вариантов.

При сортировке различных решений сети GMDH направлены на минимизацию влияния автора на результаты моделирования. Компьютер сам

находит структуру оптимальной модели или законы, которые действуют в системе.

2.4.6 Метод k-ближайших соседей

Алгоритм k-ближайших соседей (KNN) представляет собой простой, легко реализуемый контролируемый алгоритм машинного обучения, который можно использовать для решения как задач классификации, так и регрессии.

K-Nearest Neighbours является одним из самых основных, но важных алгоритмов классификации в машинном обучении. Он принадлежит к контролируемой области обучения и находит широкое применение в распознавании образов, интеллектуальном анализе данных и обнаружении вторжений.

Он широко доступен в реальных сценариях, поскольку он непараметрический, то есть он не делает каких-либо базовых предположений о распределении данных.

Работу K-NN можно объяснить на основе приведенного ниже алгоритма:

Шаг 1: Выбрать число K соседей

Шаг 2: Рассчитать евклидово расстояние K числа соседей

Шаг 3: Взять K ближайших соседей согласно расчетному евклидову расстоянию.

Шаг 4: Среди этих k соседей подсчитать количество точек данных в каждой категории.

Шаг 5: Назначить новые точки данных той категории, для которой число соседей является максимальным.

2.4.7 Методы визуализации

Методы визуализации данных относятся к созданию графического представления информации. Визуализация играет важную роль в анализе данных и помогает интерпретировать большие данные в структуре реального времени, используя сложные наборы числовых или фактических данных.

С кажущимися бесконечными потоками данных, легко доступных для современных предприятий в разных отраслях, проблема заключается в интерпретации данных, которая является наиболее ценным понятием для отдельной организации, а также ее целей и задач. В данном случае и помогает визуализация данных.

Благодаря тому, как человеческий мозг обрабатывает информацию, представление диаграмм или графиков для визуализации значительных объемов сложных данных более доступно, чем использование электронных таблиц или отчетов.

Визуализации предлагают быстрый, интуитивно понятный и простой способ универсальной передачи критических концепций. Также присутствует возможность экспериментов с различными сценариями, остается только внести небольшие изменения.

Поэтому визуализация данных имеет решающее значение для устойчивого успеха бизнеса и помогает извлечь максимальную выгоду из этого проверенного средства анализа и представления важной информации.

2.5 Языки программирования для работы с Big Data

Наука о данных – это концепция объединения статистики, анализа данных и связанных с ними стратегий. Она включает в себя теории и методы, взятые из различных областей статистики, математики и информатики.

С развитием машинного обучения наука о данных приобретает все большую популярность. Чтобы понять и стать специалистом по данным, необходимо выучить хотя бы один язык программирования, хотя знание более чем одного полезно для соискателей.

Наука о данных – это увлекательная область для работы, сочетающая количественные навыки и передовые статистические данные с реальными возможностями программирования.

Для того, чтобы решить определенные задачи с большими данными, необходимо выбрать наиболее подходящий для поставленных задач язык программирования. Рассмотрим наиболее популярные из них.

2.5.1 Python

Python – чрезвычайно популярный динамический язык общего назначения, широко используемый в сообществе специалистов по данным. Обычно его называют самым простым языком программирования для чтения и изучения. Поскольку он сочетает в себе быстрое улучшение с возможностью взаимодействия с высокопроизводительными алгоритмами, написанными на языке Fortran или C, он стал ведущим языком программирования для науки об открытых данных.

С развитием технологий, таких как искусственный интеллект, машинное обучение и прогнозная аналитика, спрос на экспертов с навыками Python значительно возрастает. Он широко используется в веб-разработке, научных вычислениях, интеллектуальном анализе данных и т.д.

2.5.2 R

Это один из наиболее часто используемых инструментов. R – это язык с открытым исходным кодом и программная среда для статистических вычислений и графики для статистических вычислений. Этот набор навыков востребован среди рекрутеров в области машинного обучения и науки о данных.

R предоставляет множество статистических моделей, и многие аналитики создали свои приложения в R. Это верхний предел открытого статистического анализа, и существует четкое внимание к статистическим моделям, которые были составлены с использованием R. Публичный архив пакетов R содержит более 8000 сетей. Microsoft, RStudio и различные организации оказывают бизнес-поддержку R-вычислениям.

2.5.3 Java

Java - чрезвычайно популярный язык общего назначения, работающий на виртуальной машине Java (JVM). Многие организации, в частности организации MNC, используют этот язык для создания серверных систем и настольных / мобильных / веб-приложений. Это поддерживаемая Oracle уникальная вычислительная система, которая обеспечивает переносимость между платформами.

Из-за того, что спрос на навыки Java растет, навыки Java были востребованы, особенно для архитекторов программного обеспечения, разработчиков программного обеспечения и инженеров DevOps.

2.5.4 SQL

SQL (язык структурированных запросов) является одним из самых популярных в области науки о данных. Он хорошо используется для запроса и редактирования информации, хранящейся в реляционной базе данных. А также, для хранения и извлечения данных в течение десятилетий. Он используется для управления особенно большими базами данных, сокращая время обработки онлайн-запросов за счет быстрого времени обработки. Владение навыками SQL может быть самым большим преимуществом для специалистов в области машинного обучения и обработки данных, поскольку SQL является наиболее предпочтительным набором навыков для всех организаций.

2.5.5 Julia

Julia – это язык динамического программирования высокого уровня, разработанный для удовлетворения потребностей высокопроизводительного численного анализа, и научные вычисления быстро завоевывают популярность среди ученых, занимающихся данными. Это более новый язык, также способный к программированию общего назначения, и он существует не так давно, как R или Python.

Благодаря быстрому выполнению Julia стала идеальным выбором для работы со сложными проектами, содержащими большие объемы данных. Многие базовые тесты выполняются в 30 раз быстрее, чем Python, и регулярно выполняются несколько быстрее, чем C-код.

2.5.6 Scala

Scala (масштабируемый язык) – один из самых известных языков с одной из самых больших пользовательских баз. Это универсальный язык программирования с открытым исходным кодом, который работает на JVM.

Scala – это идеальный выбор языка для тех, кто работает с массивами больших объемов данных, и имеет полную поддержку функционального программирования и мощную систему статических типов.

Поскольку он был разработан для работы на JVM, он обеспечивает совместимость с самой Java, что делает Scala отличным языком общего назначения, а также идеальным вариантом для науки о данных.

Платформа кластерных вычислений Apache Spark написана на Scala. Если стоит задача манипулировать данными в кластере из тысячи процессоров и иметь кучу унаследованного кода Java, то Scala – это невероятное решение с открытым исходным кодом.

2.5.7 MATLAB

Он разработан и лицензирован MathWorks. Это быстрый, стабильный язык, который гарантирует надежные алгоритмы для числовых вычислительных языков, используемых всей академией и промышленностью. Считается подходящим языком для математиков и ученых, занимающихся сложными математическими потребностями, такими как преобразования Фурье, обработка сигналов, обработка изображений и матричная алгебра.

MATLAB, широко используемый в статистическом анализе, таком как приложения или повседневная роль, требует интенсивной, расширенной функциональности в математике, что делает его хорошим выбором для работы с данными.

2.5.8 TensorFlow

TensorFlow – превосходная библиотека программного обеспечения с открытым исходным кодом для численных расчетов. Это структура машинного обучения, подходящая для крупномасштабных данных. Работает по основной концепции. Например, если необходимо выполнить график вычислений в Python, после того как он был определен, TensorFlow запустит его, используя набор настроенного кода C ++.

Одним из наиболее значительных преимуществ TensorFlow является то, что график может быть разбит на множество кусков, которые могут работать параллельно на разных GPU или CPU. А также поддерживает распределенные вычисления.

2.6 Преимущества языка программирования Python 3

Python был признан самым быстрорастущим языком программирования согласно Stack Overflow Trends. Согласно исследованию Stack Overflow Developers' 2019, Python является вторым «самым любимым» языком, и 73% разработчиков выбирают его среди других языков, преобладающих на рынке.

Python и Big Data – это новая комбинация, завоевывающая рыночное пространство. Python пользуется большим спросом среди компаний Big Data. Разберем основные преимущества использования Python, а также почему Python стал предпочтительным выбором для бизнеса в наши дни.

1. Простое кодирование

Программирование на Python требует меньше строк кода по сравнению с другими языками, доступными для программирования. Более того, Python автоматически предлагает помощь для идентификации и связывания типов данных.

Джек Янсен – сопровождающий MacPython и автор многих специальных модулей для Macintosh в Python говорит: «Python – действительно замечательный язык. Когда кто-то приходит с хорошей идеей, требуется около

1 минуты и пяти строк, чтобы запрограммировать то, что почти делает то, что вы хотите». [22]

Программирование на Python следует структуре вложенности на основе отступов. Язык может обрабатывать длительные задачи за короткий промежуток времени. Поскольку нет никаких ограничений на обработку данных, то есть возможность вычислять данные на обычных компьютерах, ноутбуках, облаке и настольных компьютерах.

2. Открытый исходный код

Разработанный с помощью модели сообщества, Python является языком программирования с открытым исходным кодом. Будучи языком с открытым исходным кодом, Python поддерживает несколько платформ. Кроме того, он может быть запущен в различных средах, таких как Windows и Linux.

«Мой любимый язык для сопровождения – это Python. Он имеет простой, чистый синтаксис, инкапсуляцию объектов, хорошую поддержку библиотек и необязательные именованные параметры», – сказал Брэм Коэн, американский программист, автор протокола BitTorrent и программы BitTorrent. [24]

3. Поддержка библиотек

Программирование на Python предлагает использование нескольких библиотек. Поскольку Big Data включает в себя большой объем анализа данных и научных вычислений, Python и Big Data являются отличными компаньонами.

Python предлагает ряд хорошо протестированных аналитических библиотек. Эти библиотеки состоят из пакетов, таких как: численные вычисления, анализ данных, статистический анализ, визуализация, машинное обучение.

4. Скорость

Python считается одним из самых популярных языков для разработки программного обеспечения из-за его высокой скорости и производительности.

Программирование на Python поддерживает идеи создания прототипов, которые помогают сделать код быстрым. Более того, Python также поддерживает прозрачность между кодом и процессом.

Ранее Python считался более медленным языком по сравнению с некоторыми из его аналогов, такими как Java и Scala, но сценарий изменился. Появление платформы Anaconda предложило языку большую скорость. Вот почему Python для больших данных стал одним из самых популярных вариантов в отрасли.

5. Объем

Python позволяет пользователям упростить операции с данными. Поскольку Python является объектно-ориентированным языком, он поддерживает сложные структуры данных. Некоторые структуры данных, которыми управляет Python, включают списки, наборы, словари и многое другое.

Помимо этого, Python помогает в поддержке научных вычислительных операций, таких как матричные операции, фреймы данных и т.д. Эти невероятные возможности Python помогают расширить область применения языка, позволяя тем самым ускорить операции с данными. Это то, что делает Python и Big Data превосходной комбинацией.

6. Поддержка обработки данных

Python имеет встроенную функцию поддержки обработки данных. По этой причине компании, работающие с большими данными, предпочитают выбирать именно этот язык программирования.

Python обеспечивает расширенную поддержку изображений и голосовых данных благодаря встроенным функциям поддержки обработки данных для неструктурированных и нетрадиционных данных, что может быть необходимо при анализе.

Таким образом, есть четкое представление о том, почему Python для больших данных считается наиболее подходящим вариантом.

Python – это простой язык с открытым исходным кодом, обладающий высокой скоростью и надежной поддержкой библиотеки.

С использованием технологии больших данных, распространяющейся по всему миру, удовлетворение требований в отрасли, безусловно, является непростой задачей. Но благодаря своим невероятным преимуществам Python стал подходящим выбором для больших данных.

2.6.1 Встроенные функции и методы

Справедливо сказать, что использование функций – самое большое преимущество Python. Они довольно часто используются в различных проектах Data Science.

Функции Python работают очень просто. Необходимо лишь вызывать функцию и указать необходимые аргументы, затем она возвращает результаты. Тип аргумента (например, строка, список, целое число, логическое значение и т.д.) Может быть ограничен (например, в некоторых случаях он должен быть целым числом), но в большинстве случаев это может быть несколько типов значений.

Давайте рассмотрим наиболее важные встроенные функции и методы:

print() – вывод на экран

abs() – возвращает абсолютное значение (например, целое число или число с плавающей запятой)

round() – возвращает округленное значение числа

min() – возвращает наименьший элемент списка или введенных аргументов, также это может быть строка

max() – возвращает наибольший элемент списка или введенных аргументов, также это может быть строка

sorted() – сортирует список по возрастанию, список может содержать строки или цифры

sum() – суммирует список, список может иметь все типы числовых значений

type() – возвращает тип переменной

open() – открывает указанный файл

index() – ищет заданный элемент с начала списка и возвращает самый низкий индекс, где элемент появляется

len() – функция возвращает количество элементов (длину) в объекте

filter() – метод создает итератор из элементов итерируемого, для которого функция возвращает true, проще говоря, метод filter () фильтрует заданную итерацию с помощью функции, которая проверяет, является ли каждый элемент в итерируемом истинным или нет

append() – метод, который добавляет один элемент в существующий список, при этом не возвращает новый список элементов, но изменяет исходный, добавляя элемент в конец списка, после выполнения метода, добавляемого в список, размер списка увеличивается на единицу

Это встроенные, самые простые функции и методы Python, которые используются при анализе довольно регулярно.

2.6.2 Модуль PyLab

PyLab – это удобный модуль, который массово импортирует matplotlib.pyplot (для построения графиков) и NumPy (для математики и работы с массивами) в одном пространстве имен.

Основные функции модуля:

%pylab inline – функция, с помощью которой интерпретатор IPython импортирует модули matplotlib и NumPy

import – с помощью данной функции можно загрузить различные файлы, библиотеки или дополнительные функции, например, мы можем загрузить библиотеку pandas (import pandas)

pandas.read_csv – функция считывает файл значений, разделенных запятыми (csv), в DataFrame, также поддерживает опциональную итерацию или разбиение файла на куски.

value_counts() – функция возвращает объект, содержащий количество уникальных значений, полученные объекты будут в порядке убывания, так что первый элемент является наиболее часто встречающимся элементом, исключает значения NA по умолчанию

groupby() – функция используется для разделения данных на группы по некоторым критериям, объекты pandas можно разделить на любую из их осей, абстрактное определение группировки состоит в том, чтобы обеспечить отображение меток на имена групп

get_group() – чтобы выбрать группу, мы можем использовать данную функцию

normalize() – относится к изменению масштаба числовых атрибутов с действительными значениями в диапазоне 0 и 1, полезно масштабировать входные атрибуты для модели, которая опирается на величину значений, таких как меры расстояния, используемые в k-ближайших соседях, и при подготовке коэффициентов в регрессия.

map() – функция возвращает объект карты (который является итератором) результатов после применения данной функции к каждому элементу данной итерируемой, например списка

drop() – функция используется для удаления указанных меток из строк или столбцов, удаление строки или столбца происходит, когда указаны имена меток и соответствующие оси или указаны имена индексов или столбцов, при использовании многоиндексных меток на разных уровнях можно удалить, указав уровень

dropna() – функция используется для удаления пропущенных значений, 0 или «index»: удалить строки, содержащие пропущенные значения. 1 или «columns»: удаление столбцов, которые содержат пропущенное значение.

fillna() – с помощью этого метода заполняются значения NA / NaN

2.6.3 Matplotlib для визуализации в Python

Теперь рассмотрим пакет Matplotlib для визуализации в Python.

Matplotlib – это двухмерная библиотека для построения графиков, которая помогает визуализировать фигуры. Matplotlib эмулирует Matlab как графики и визуализации. Matlab не бесплатен, его сложно масштабировать, а язык программирования довольно сложен в изучении. Поэтому мы используем matplotlib в Python как надежную, бесплатную и простую библиотеку для визуализации данных.

Фигура содержит общее окно, в котором происходит построение графика, а внутри фигуры – то, где строятся фактические графики. Каждый график имеет оси X и Y для построения. Внутри осей содержатся названия и метки, связанные с каждой осью (Рисунок 2.2). Важной фигурой matplotlib является то, что мы можем построить и показать больше, чем просто оси на фигуре, которая помогает в построении нескольких графиков. В matplotlib pyplot используется для создания фигур и изменения их характеристик.

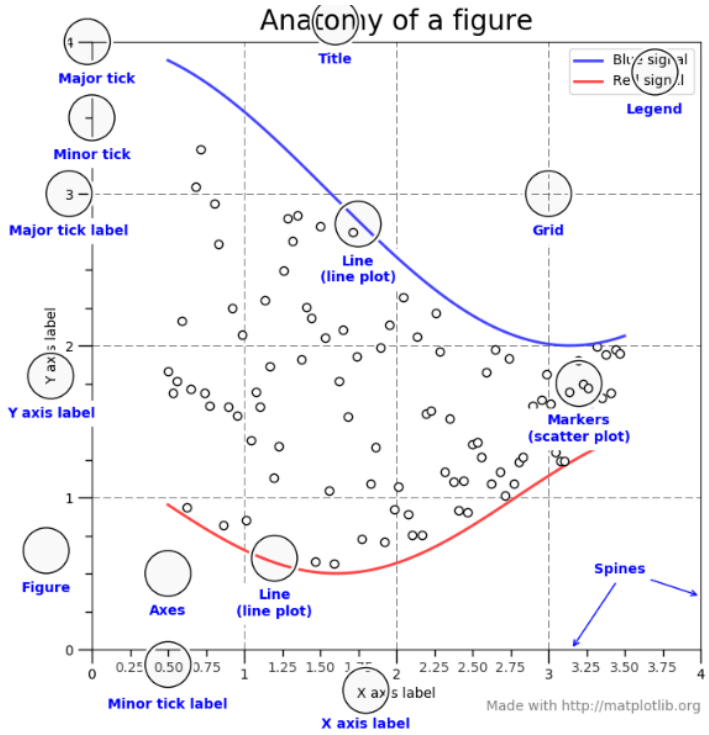


Рисунок 2.2 График в Matplotlib

Построения в Matplotlib довольно просты. Как правило, при построении графика выполняются одни и те же шаги. Matplotlib имеет модуль под

названием pyplot, который помогает в построении фигуры. Блокнот Jupyter используется для запуска графиков.

Основные функции для визуализации:

import matplotlib.pyplot as plt – импорт pyplot для вызова модуля package

plt.plot() – функция для построения линейного графика аналогично вместо графика используются другие функции для построения

plot.xlabel, plt.ylabel – используются для маркировки осей x и y соответственно

plt.xticks, plt.yticks – используются для маркировки точек отметок наблюдения x и y соответственно

plt.legend () – функция обозначения переменных наблюдения

plt.title () – функция установки заголовка

plot.show () – функция отображения графика

plt.hist() – функция для построения гистограммы

plt.pie() – построение круговой диаграммы

plt.bar() – функция для построения гистограммы, показывает распределение данных по нескольким группам

plt.heatmap() – построение тепловой карты для визуализации скалярных функций двух переменных

У каждого графика есть определенные параметры, которые можно менять вне зависимости от данных: цвет, отображение надписей (заголовка, значений, процентного соотношения и т.д.), расположение столбцов, наложение нескольких диаграмм на одну ось и т.п.

Таким образом, можно, как строить графики визуализации данных, используя одну числовую переменную, так и несколько переменных, при этом настраивать визуализацию так, как нам удобно. Теперь можно легко строить графики для интуитивного понимания наших данных с помощью визуализаций.

3 ПРОЕКТНАЯ ЧАСТЬ

3.1 Подготовка данных

В исходной работе используется набор данных, в которых описаны мероприятия министерства культуры РФ, взятые с официального сайта министерства. Набор дынных содержит все мероприятия, проходящие на территории России, начиная с 2015 года. На основе этих данных будет построена исходная модель, которая будет протестирована на части данных за 2020 год. После будет произведена корректировка модели.

На первом этапе идет подготовка данных: загрузка необходимых для работы библиотек Pandas, Seaborn и Matplotlib (Рисунок 3.1).

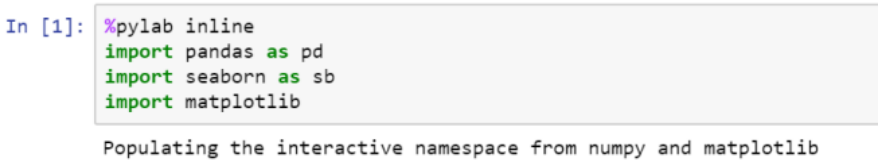


Рисунок 3.1 Загрузка библиотек

После мы загружаем исходный CVS файл с данными о мероприятиях в сфере культуры (Рисунок 3.2) произошедшие с 2015 года на территории России.

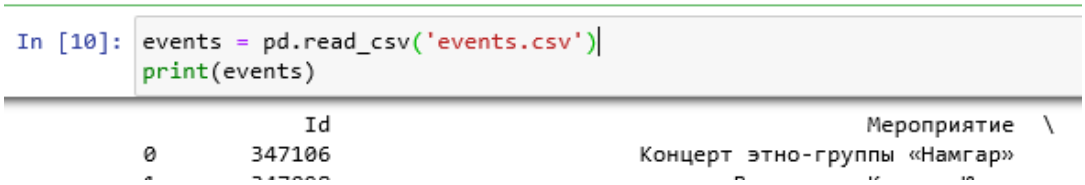


Рисунок 3.2 Загрузка данных

В наборе огромное количество данных: 582032 строки и 122 колонки. Можно сделать вывод, что нам не понадобятся большинство столбцов, так как в них хранится дополнительная, не нужная нам информация.

Так же стоит отметить, что большое количество записей, хранимых в документе, не имеют какой-либо ценности.

Для большей наглядности выведем всю информацию, не распределяя ее в таблицу (Рисунок 3.3) и просмотрим все столбцы.

```
In [*]: f = open('events.csv', encoding='utf-8')
        for line in f:
            print(line)
```

```
Id,Мероприятие,имя организатора мероприятия,краткое описание события,Полное описание,возрастное ограничение,Бесплатное (1-да),
стоимость посещения,максимальная стоимость посещения,Статус мероприятия,SaleLink,NeedMedia,"Вместительность, чел.",Начало мер
оприятия,Окончание мероприятия,Id,Изображение,Список изображений,Место проведеня мероприятия,Hosting,Тип категории,Категория,
Категория (лат),Id,Id организации ЕИПСК,Организация,ИНН организации,Тип организации,наименование улицы,комментарии,Идентифика
тор дома в ФИАС,Идентификатор улицы в ФИАС,Полный адрес,Items,Место проведения,Часовой пояс,Место проведения (лат),Id,Items,М
есто проведения,Часовой пояс,Место проведения (лат),Id,Network,Место проведения ,Часовой пояс,Место проведения (лат),Id,Id,Ме
сто проведения,Описание,Статус,Название улицы,Комментарии,Идентификатор дома в ФИАС,Идентификатор улицы в ФИАС,Полный адрес,т
ип категории,название категории,системное имя категории,Id,сайт,электронная почта,номер телефона,ссылка,EipskId,Culturarf,Gos
catalogId,Statistic,ArtType,AudienceType,Language,ProfessionalLevel,VirtualTour,Items,Items,Изображение,Items,Месторасположен
ие,Часовой пояс,Месторасположение (лат),идентификатор локали,Id,название организации,ИНН,тип организации,Улица,Комментарии,Ид
ентификатор дома в ФИАС,Идентификатор улицы в ФИАС,полный адрес,Items,Месторасположение,Часовой пояс,системное имя локали,иде
нтификатор локали,Items,название локали,Часовой пояс,системное имя локали,идентификатор локали,Network,Месторасположение,Часо
вой пояс,Месторасположение (лат),идентификатор локали,Месторасположение,дата начала мероприятия,Id,Hosting,0,1,2,3,4,5,6,Id,д
ата начала мероприятия,Категория,Дата создания,Дата обновления
```

Рисунок 3.3 Загрузка данных

Среди столбцов можно увидеть «Вместительность чел.», что может быть довольно полезной информацией, в нашем случае. Ведь появляется возможность просмотреть, на какое количество людей было рассчитано каждое мероприятие, и проследить увеличение или уменьшение интереса к тому или иному событию.

Однако после проверки выяснилось, что данный столбец заполнен лишь пустыми значениями. Следовательно, необходимо отфильтровать подобные данные, для большего удобства работы.

Это мы сделаем, воспользовавшись функцией Pandas dataframe.filter() (Рисунок 3.4), которая используется для подмножества строк или столбцов данных в соответствии с метками в указанном индексе.

```
In [19]: events = events.filter(['Начало мероприятия','Мероприятие','Место проведения','Месторасположение','Категория','возрастное огранич
         print(events)
```

```
             Начало мероприятия  \
0       2019-02-02T14:00:00.000Z
1       2019-01-24T06:00:00.000Z
2       2019-01-22T07:00:00.000Z
3       2019-02-02T07:00:00.000Z
4       2019-02-02T07:00:00.000Z
...                          ...
```

Рисунок 3.4 Фильтр данных

Далее необходимо проверить данные на наличие нулевых значений и удалить строки, в которых они совершенно недопустимы, а именно в колонках: категория, место проведения, мероприятие, начало мероприятия (Рисунок 3.5).

```
In [25]: events.dropna(subset=['Категория', 'Место проведения', 'Мероприятие', 'Начало мероприятия'], inplace=True)
         print(events)

                Начало мероприятия  \
0       2019-02-02T14:00:00.000Z
1       2019-01-24T06:00:00.000Z
2       2019-01-22T07:00:00.000Z
3       2019-02-02T07:00:00.000Z
4       2019-02-02T07:00:00.000Z
```

Рисунок 3.5 Удаление пустых строк

Используемый нами объем данных довольно велик, поэтому самостоятельно проверить наличие пустых строк крайне сложно. Проверим правильность работы кода и посмотрим, не осталось ли в данных строках недопустимых значений NaN, проверку выполним для каждого столбца, после чего будет получено значение «False», если ячейка не пустая и «True», если ячейка пустая (Рисунок 3.6).

```
In [27]: None in events["Место проведения"]

Out[27]: False
```

Рисунок 3.6 Проверка на пустые значения

Ячеек с пустыми данными в нашей таблице нет. Следовательно, можно сделать вывод, что код был выполнен корректно, и в дальнейшем мы не получим ошибок и проблем при работе с данными.

Далее рассмотрим колонки, в которых значение NaN можно заменить на 0, а именно: стоимость посещения и возрастное ограничение.

Действительно, в данных столбцах значение 0 допустимо, так как ограничение по возрасту может просто не ставиться, а вход или участие в мероприятии может быть бесплатным. Заменим все оставшиеся пустые ячейки на 0 (Рисунок 3.7).

```
In [10]: events = events.fillna(0)
         print(events)
```

Рисунок 3.7 Замена пустых значений

Также необходимо проверить данные на наличие некорректных записей или выбросов, которые также следует убрать. Например, в столбце «стоимость посещения» можно увидеть значения: 30030033350, 800800800, 50505050 и

т.д. Можно с уверенностью сказать, что это ошибка или опечатка, а это значит, что следует убрать эти строки, так как они являются выбросами.

Поставим фильтр и удалим ненужные и выделяющиеся данные из нашей таблицы, чтобы они не повлияли на наш конечный результат (Рисунок 3.8).

```
In [12]: events.drop(events[events["стоимость посещения"] > 130000].index, inplace=True)
         print(events)

                 Начало мероприятия  \
0       2019-02-02T14:00:00.000Z
1       2019-01-24T06:00:00.000Z
2       2019-01-22T07:00:00.000Z
3       2019-02-02T07:00:00.000Z
4       2019-02-02T07:00:00.000Z
```

Рисунок 3.8 Удаление выбросов

В первой колонке: «Начало мероприятия», хранится информация о дате и времени проведения мероприятия. Тип данного столбца String, однако мы можем перевести данный столбец в тип Date для удобства дальнейших вычислений и анализа (Рисунок 3.9).

```
In [32]: events['Начало мероприятия'] = events['Начало мероприятия'].map(pd.to_datetime)
         print(events)

                 Начало мероприятия  \
0       2019-02-02 14:00:00+00:00
1       2019-01-24 06:00:00+00:00
2       2019-01-22 07:00:00+00:00
3       2019-02-02 07:00:00+00:00
4       2019-02-02 07:00:00+00:00
...                           ...
540655 2020-03-30 16:00:00+00:00
540656 2020-04-24 07:00:00+00:00
540657 2020-04-01 15:00:00+00:00
540658 2020-03-30 13:00:00+00:00
540659 2020-02-24 04:00:00+00:00
```

Рисунок 3.9 Изменение формата данных

Как мы можем видеть, формат данных изменился, теперь он имеет тип Date, и мы четко можем видеть дату и время каждого мероприятия.

Кроме этого можно удалить информацию о времени, так как нам важна лишь дата проведения мероприятия (Рисунок 3.10). Следовательно, мы можем просто обнулить значение времени с помощью normalize(), который в datetime преобразует параметр time в полночь, то есть 00:00:00. Это полезно в нашем случае, так как для нас время не имеет значения.

```
In [36]: from datetime import datetime as dt
         events['Начало мероприятия'] = events['Начало мероприятия'].dt.normalize()
         print(events)

                 Начало мероприятия  \
0       2019-02-02 00:00:00+00:00
1       2019-01-24 00:00:00+00:00
2       2019-01-22 00:00:00+00:00
3       2019-02-02 00:00:00+00:00
4       2019-02-02 00:00:00+00:00
...                           ...
```

Рисунок 3.10 Изменение времени

Далее мы можем взять информацию о дате проведения мероприятия создать новые столбцы, где будут храниться число, месяц и год по отдельности, для удобства дальнейшей работы над данными (Рисунок 3.11).

```
In [17]: def get_dom(dt) :
             return dt.day
         def get_month(dt) :
             return dt.month
         def get_year(dt) :
             return dt.year
         events ['День'] = events['Начало мероприятия'].map(get_dom)
         events ['Месяц'] = events['Начало мероприятия'].map(get_month)
         events ['Год'] = events['Начало мероприятия'].map(get_year)
         print(events)
```

Рисунок 3.11 Создание отдельных столбцов

Необходимо проверить наличие некорректных записей, а именно: проверить наличие неотображаемых символов в записях (Рисунок 3.12). Это немаловажная процедура, так как далее для работы нам потребуются данные, которые выводятся и отображаются без ошибок и нечитаемых символов. Также, если пропустить данный этап, то возможны ошибки при поиске в наборе данных, чего допускать не следует.

```
In [27]: import string
         printable_chars = set(string.printable + "абвгдеёжзийклмнопрстуфхцчшщъыьэюяАБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ«»–—…")

         events[events.applymap(lambda x: set(str(x)).issubset(printable_chars)).all(1)]
```

Out[27]:

	Начало мероприятия	Мероприятие	Место проведения	Категория	возрастное ограничение	стоимость посещения	День	Месяц	Год
0	2019-02-02 00:00:00+00:00	Концерт этно-группы «Намгар»	Тольятти	Концерты	12.0	350.0	2	2	2019
1	2019-01-24 00:00:00+00:00	Выставка «Краски Юга»	Ростовская область	Выставки	0.0	20.0	24	1	2019
2	2019-01-22 00:00:00+00:00	Выставка «Секреты павловопосадских узоров»	Ростовская область	Выставки	0.0	20.0	22	1	2019
3	2019-02-02 00:00:00+00:00	Книжно-иллюстративная выставка «Он поучает муд...	Белгородская область	Выставки	12.0	0.0	2	2	2019
4	2019-02-02 00:00:00+00:00	Книжно-иллюстративная выставка «Сокровища родн...	Белгородская область	Выставки	12.0	0.0	2	2	2019
...	...	...	...	...	...	...	...	...	...

Рисунок 3.12 Проверка наличия некорректных записей

Такая проверка происходит с помощью фильтра, далее результат выводится, но не перезаписывается, чтобы суметь проанализировать отфильтрованные данные и сделать вывод о корректности самой фильтрации.

В данном случае видно, что после такой фильтрации было удалено более 10 000 записей, что довольно много, а это означает, что фильтр подобран скорее всего неправильно. В таком случае необходимо вывести отфильтрованные результаты, чтобы проверить и исправить при надобности шаблон, используемый в фильтре (Рисунок 3.13).

```
In [28]: events[events.applymap(lambda x: not set(str(x)).issubset(printable_chars))].dropna(how="all")
```

Out[28]:

	Начало мероприятия	Мероприятие	Место проведения	Категория	возрастное ограничение	стоимость посещения	День	Месяц	Год
27	NaT	Лекция «Я непременно устою в водоворотах всяко...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
46	NaT	Лекция «Никогда я не был на Босфоре…»	NaN	NaN	NaN	NaN	NaN	NaN	NaN
49	NaT	Литературно-музыкальный вечер «Кот Бегемот, ил...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
92	NaT	Выставка «Закрыв глаза, я слушаю Стамбул…»	NaN	NaN	NaN	NaN	NaN	NaN	NaN
100	NaT	Литературный вечер «Я храню твоё объятье…»	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...

Рисунок 3.13 Проверка наличия некорректных записей

Как мы можем увидеть, в некоторых записях используются знак троеточия, а значит его необходимо добавить, как и другие отображаемые символы. В итоге было отфильтровано около 2 000 некорректно отображаемых записей.

На данном этапе мы проверили наши данные на наличие некорректных значений и убрали их, заполнили пустые строки, а также отфильтровали столбцы, выбрав только интересующие нас значения, по которым и будет проводиться дальнейший анализ. Более того, мы обнулили время для всех мероприятий, так как для нас имеет значение только параметр даты и распределили число, месяц и год по отдельным столбцам.

После корректировки каждый набор данных содержит в себе следующие столбцы: Начало мероприятия, Мероприятие, Место проведения, Категория, возрастное ограничение, стоимость посещения, День, Месяц, Год

3.2 Построение первоначальной модели

3.2.1 Мероприятия за каждый год и месяц

Теперь посчитаем количество мероприятий за каждый год. Воспользуемся функцией value_counts(). Полученный объект будет в порядке убывания, поэтому первый элемент является наиболее часто встречающимся элементом (Рисунок 3.14).

```
In [10]: count = events['Год'].value_counts()
         print(count)

         2019    188864
         2018    131005
         2017     92117
         2020     89831
         2016     44296
         2015     12431
         2014        44
         2021         4
         2025         1
         2005         1
         Name: Год, dtype: int64
```

Рисунок 3.14 Подсчет с помощью value_counts()

Как мы можем видеть, в таблице находятся данные 2005, 2014, 2021, 2025 годов. Эти значения следует удалить из таблицы, так как они не имеют для нас никакой информационной ценности. Данных за 2005 и 2014 год слишком мало, очевидно, что там нет большинства мероприятий, произошедших в эти годы. 2021 и 2025 года мы не рассматриваем в нашем наборе данных, поэтому мы их тоже удалим.

В начале мы находим индексы каждого, ненужного нам мероприятия по значению года: 2005, 2014, 2021, 2025. Записываем получившиеся результаты в список indexName (Рисунок 3.15).

```
In [24]: indexNames = []
         indexNames.append(events[(events['Год'] == 2005)].index)
         indexNames.append(events[(events['Год'] == 2025)].index)
         indexNames.append(events[(events['Год'] == 2014)].index)
         indexNames.append(events[(events['Год'] == 2021)].index)
         print(indexNames)

         [Int64Index([284559], dtype='int64'), Int64Index([328750], dtype='int64'), Int64Index([232836, 232864, 249069, 255605, 258471,
         269034, 276961, 278092,
                     283023, 283025, 283037, 283039, 283138, 283140, 283141, 283144,
                     283545, 283551, 283587, 283635, 283637, 283640, 284125, 284196,
                     284198, 284532, 284616, 284641, 284642, 284644, 284648, 284649,
                     284650, 284652, 284675, 284699, 287017, 287529, 287530, 287531,
                     287566, 287737, 287962, 287995],
                    dtype='int64'), Int64Index([424404, 487926, 516607, 526349], dtype='int64')]
```

Рисунок 3.15 Заполнение списка индексами

Создадим небольшой цикл, который проходится по значениям списка indexName, и удаляет строки в наборе данных. Сразу же проверим количество мероприятий за каждый год снова, чтобы убедиться в том, что код сработал корректно и мы избавились от ненужных нам значений (Рисунок 3.16).

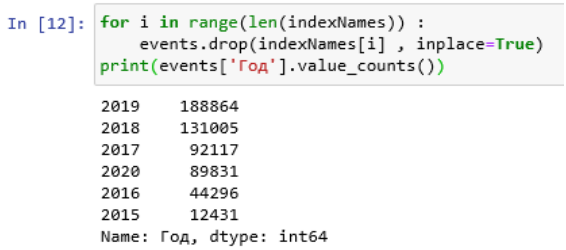


Рисунок 3.16 Удаление строк

Как мы можем видеть, все сработало и строки были удалены, теперь коррекция данных выполнена окончательно, и мы можем начать наш анализ.

Количество данных:

- 2015 год – 12431
- 2016 год – 44296
- 2017 год – 92117
- 2018 год – 131005
- 2019 год – 188864
- 2020 год – 89831

Посчитаем так же количество данных за каждый месяц и распределим значения по годам с помощью операции groupby() (Рисунок 3.17). Данная операция включает некоторую комбинацию разделения объекта, применения функции и объединения результатов. Поэтому мы используем данный метод для группировки больших объемов данных и вычисления операций в нужных нам группах.

```
In [13]: month = events.groupby(["Год", "Месяц"]).size().reset_index(name="Количество")
         grouped_df = month.groupby('Год')
         for key, item in grouped_df:
             print(grouped_df.get_group(key), "\n\n")
```

```
     Год  Месяц  Количество
0   2015      1          57
1   2015      2         128
2   2015      3         436
3   2015      4         707
4   2015      5        1406
5   2015      6         462
6   2015      7         391
7   2015      8         538
8   2015      9        1223
9   2015     10        1833
10  2015     11        2874
11  2015     12        2376


     Год  Месяц  Количество
12  2016      1        1922
13  2016      2        2651
```

Рисунок 3.17 Использование функции groupby()

Занесем полученные данные по количеству мероприятий за каждый год и месяц в таблицу, для наглядности (Таблица 3.1).

Таблица 3.1

Количество мероприятий за каждый год и месяц

Месяц\Год	Количество					
	2015	2016	2017	2018	2019	2020
1	57	1922	4641	7164	10254	16071
2	128	2651	6759	10501	14271	21969
3	436	3002	7315	11461	17495	22998
4	707	4467	9714	13119	17499	14764
5	1406	3470	7835	12329	15317	12722
6	462	2554	7243	10541	14522	884
7	391	1518	4748	7215	10366	81
8	538	2485	4988	7579	11563	65
9	1223	3890	6942	9233	13633	92
10	1833	4457	8849	11752	17573	78
11	2874	7296	12751	15541	24366	77
12	2376	6584	10332	14570	22005	30

Исходя из количества мероприятий 2020 года, начиная с июня, можно сделать вывод, что данные еще не заполнены и многие мероприятия не записаны в базу, а значит мы можем их удалить (Рисунок 3.18).

```
In [24]: indexNames = []
         indexNames.append(events.query("`Год` == 2020 & `Месяц` > 5").index)
         print(indexNames)

[Int64Index([355864, 374185, 383471, 390749, 395432, 395447, 395468, 395484,
             395486, 401425,
             ...
             558512, 558513, 558528, 558549, 558575, 558576, 558590, 558597,
             558598, 558603],
            dtype='int64', length=1307)]
```

Рисунок 3.18 Заполнение списка новыми индексами

На данном этапе 2020 год рассматриваться не будет, однако предварительная очистка была необходима для будущей корректной проверки данных на полученной модели.

После того, как данные были окончательно очищены, можно приступать к визуализации данных с помощью построения графиков.

Построим диаграммы по количеству проведенных мероприятий в месяц. На графиках указано количество мероприятий, проведенных за последние 5 лет: 2015 год (Рисунок 3.19), 2016 год (Рисунок 3.20), 2017 год (Рисунок 3.21), 2018 год (Рисунок 3.22), 2019 год (Рисунок 3.23).

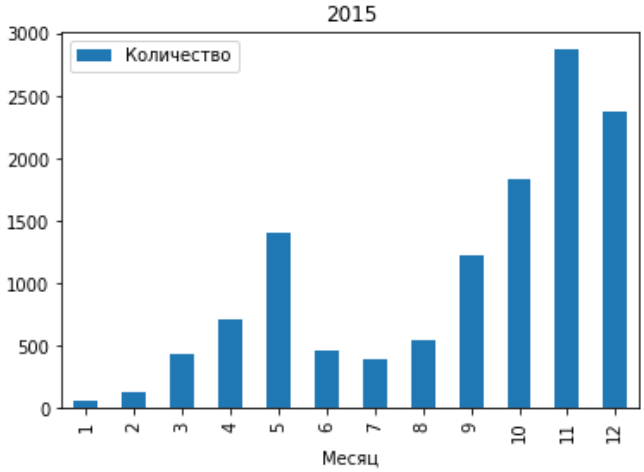


Рисунок 3.19 Количество мероприятий за 2015 год по месяцам

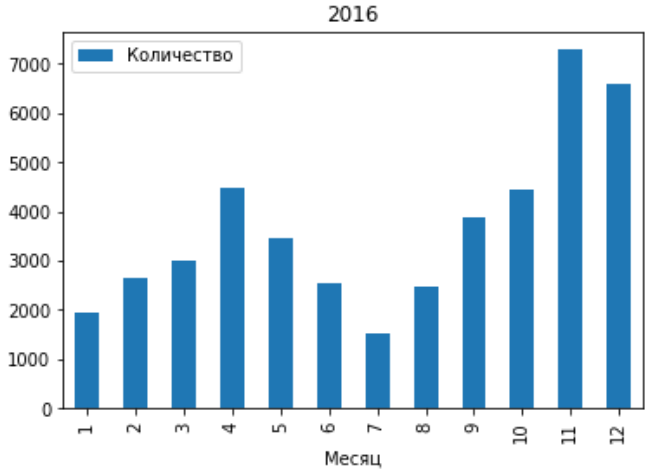


Рисунок 3.20 Количество мероприятий за 2016 год по месяцам

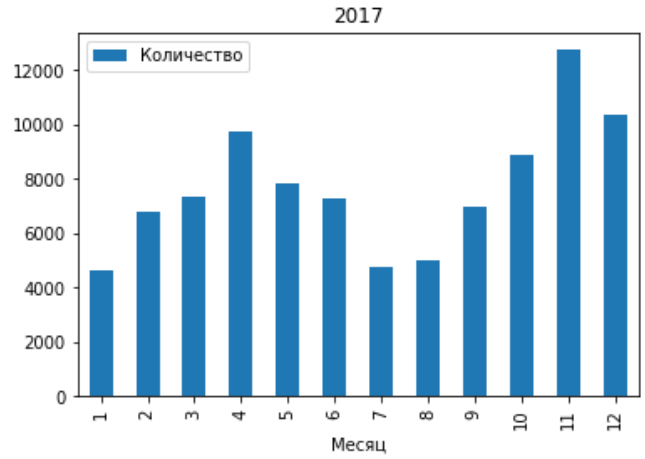


Рисунок 3.21 Количество мероприятий за 2017 год по месяцам

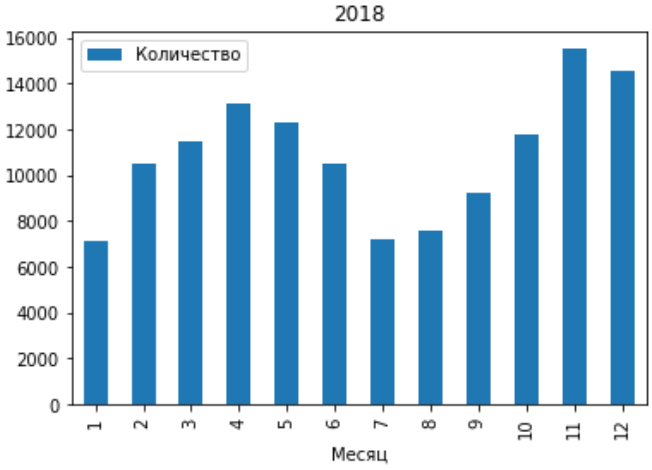


Рисунок 3.22 Количество мероприятий за 2018 год по месяцам

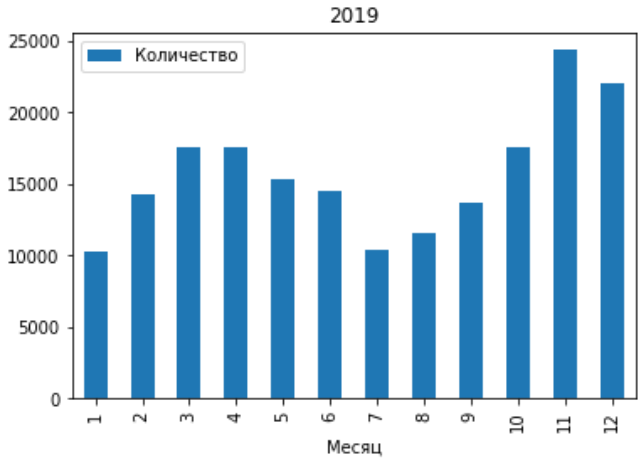


Рисунок 3.23 Количество мероприятий за 2019 год по месяцам

Просмотрев все графики, можно с уверенностью сказать, что присутствует определенная цикличность количества проведенных мероприятий, предположительно зависящая от месяца года.

Построим диаграмму, содержащую в себе информацию за все 6 лет на одной плоскости: с января по июнь (Рисунок 3.24) и с июля по декабрь (Рисунок 3.25). Синий – 2015 год, оранжевый – 2016 год, зеленый – 2017 год, красный – 2018 год, фиолетовый – 2019 год.

Можно увидеть, что количество мероприятий увеличивается с каждым годом. Также видно, что наименее загруженными месяцами являются январь и июль, после них количество мероприятий постепенно увеличивается. Наиболее загруженными являются месяца апрель и ноябрь, после этих месяцев количество мероприятий постепенно идет на спад.

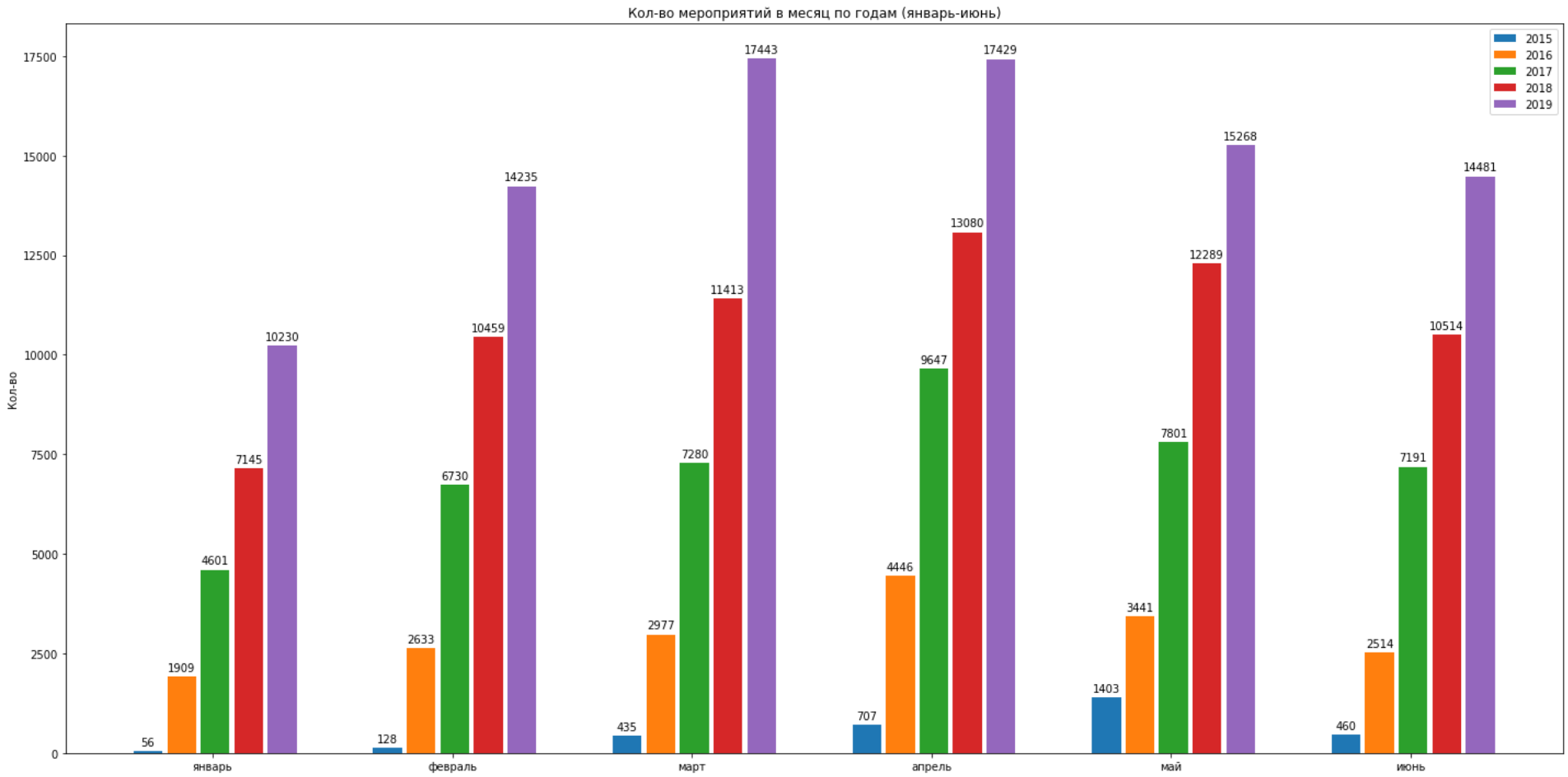


Рисунок 3.24 Количество мероприятий в месяц по годам (январь – июнь)

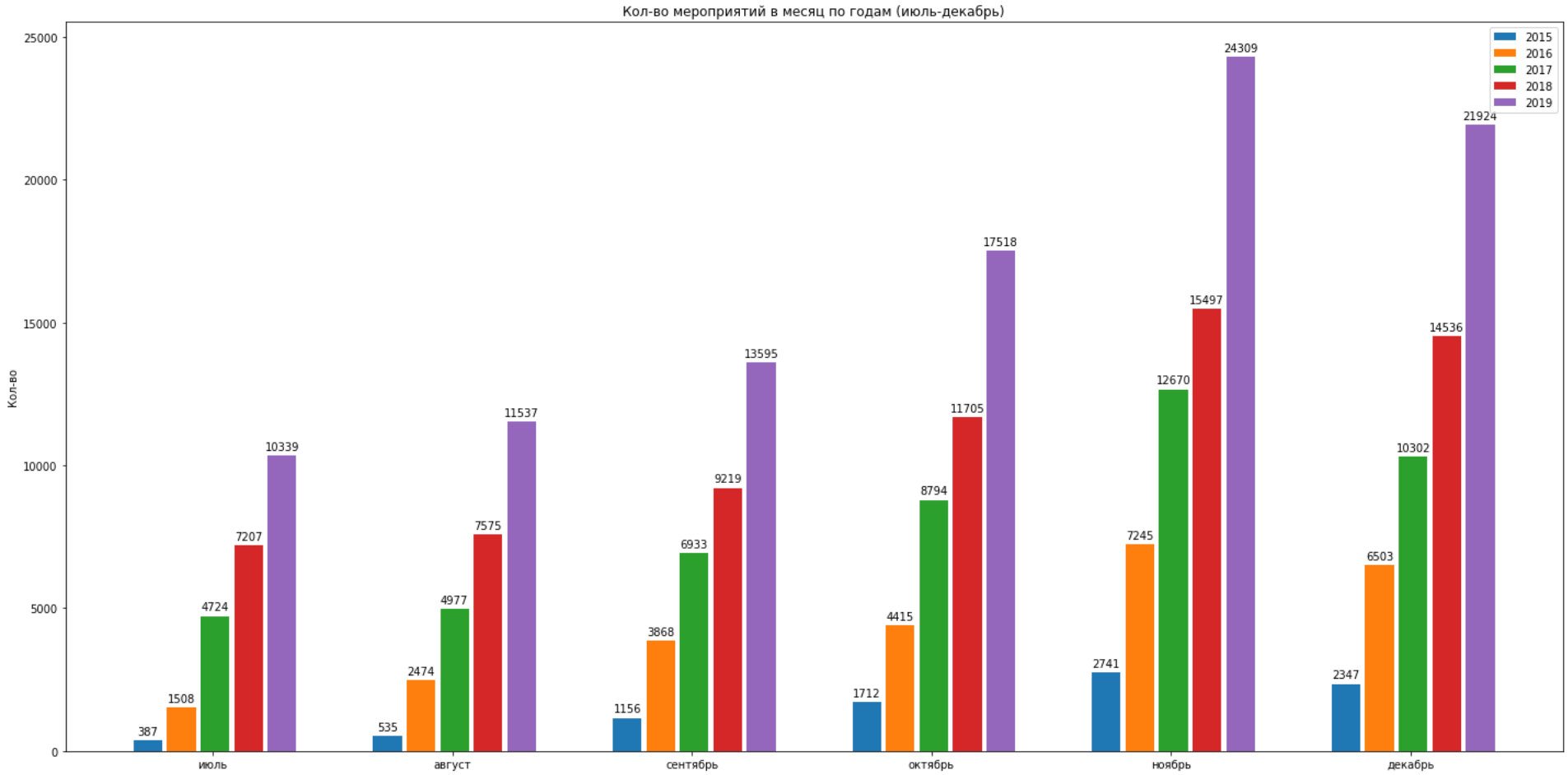


Рисунок 3.25 Количество мероприятий в месяц по годам (июль – декабрь)

3.2.2 Автокорреляционная функция

Для того, чтобы окончательно сделать вывод о наличии сезонных колебаний, будет построена автокорреляционная функция.

Для начала необходимо сгруппировать все данные по месяцам и годами распределить их в отдельную таблицу по времени (Таблица 3.2).

Таблица 3.2

Данные распределенные по месяцам

МесяцГод	Кол-во	МесяцГод	Кол-во	МесяцГод	Кол-во	МесяцГод	Кол-во
2015.01	56	2016.04	4446	2017.07	4724	2018.10	11705
2015.02	128	2016.05	3441	2017.08	4977	2018.11	15497
2015.03	435	2016.06	2514	2017.09	6933	2018.12	14536
2015.04	707	2016.07	1508	2017.10	8794	2019.01	10230
2015.05	1403	2016.08	2474	2017.11	12670	2019.02	14235
2015.06	460	2016.09	3868	2017.12	10302	2019.03	17443
2015.07	387	2016.10	4415	2018.01	7145	2019.04	17429
2015.08	535	2016.11	7245	2018.02	10459	2019.05	15268
2015.09	1156	2016.12	6503	2018.03	11413	2019.06	14481
2015.10	1712	2017.01	4601	2018.04	13080	2019.07	10339
2015.11	2741	2017.02	6730	2018.05	12289	2019.08	11537
2015.12	2347	2017.03	7280	2018.06	10514	2019.09	13595
2016.01	1909	2017.04	9647	2018.07	7207	2019.10	17518
2016.02	2633	2017.05	7801	2018.08	7575	2019.11	24309
2016.03	2977	2017.06	7191	2018.09	9219	2019.12	21924

Теперь необходимо построить общий график, показывающий количество мероприятий за каждый месяц до конца 2019 года (Рисунок 3.26).

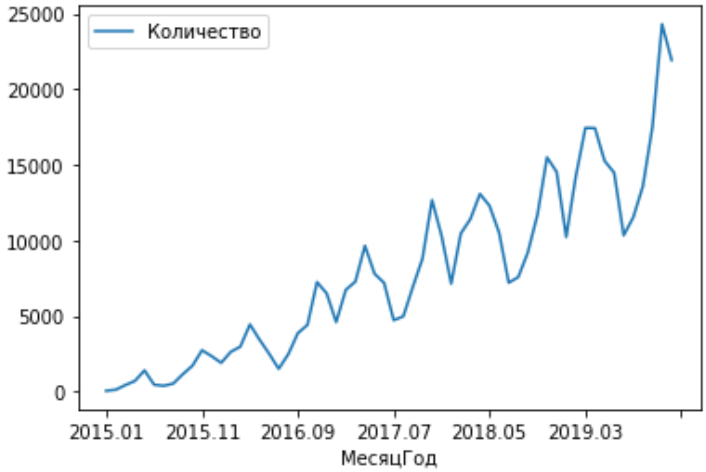


Рисунок 3.26 График мероприятий по месяцам за 2015-2019 года

Как и предполагалось ранее, данный ряд не является стационарным, так как имеется наличие тренда.

Найдем коэффициенты корреляции между рядами, для проверки тесноты связи между ними. Так как в нашем наборе данных 60 значений, и мы будем искать коэффициент 1-го порядка, то расчёт будет проводится по 59 парам, где в одном из исходных рядов будет сдвиг на 1 уровень.

Все вычисления производятся в Python, что значительно упрощает работу, благодаря встроенным функциям и методам. В итоге было получено значение коэффициента равное 0,929. Следовательно, между рядами связь тесная и прямая.

Однако необходимо проверить значимость получившегося коэффициента корреляции, для того, чтобы окончательно принять верность наших утверждений.

Проверка происходила по таблице Стьюдента при степени свободы равным 57, а уровне значимости равным 0,05. В ходе вычислений значение критерия Стьюдента получилось равным 26,42, что больше значения из таблицы. Следовательно, можно с уверенностью сказать, что коэффициент корреляции статистически значим и в наших данных имеется наличие сезонных колебаний.

Таким образом были вычислены коэффициенты корреляции и построена коррелограмма для большей наглядности полученных при решении значений (Рисунок 3.27).

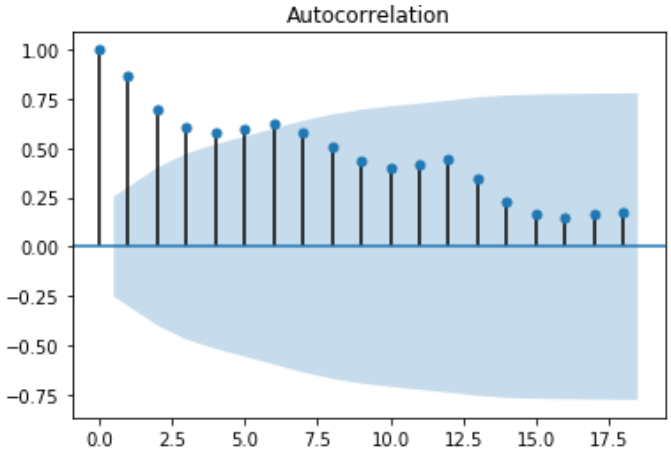


Рисунок 3.27 Коррелограмма

По коррелограмме также видно, что данные не хаотичны и взаимосвязаны.

3.2.3 Анализ категорий

Теперь посчитаем количество мероприятий за каждый год по распределенным категориям. Необходимо сразу сгруппировать данные по категориям и по годам (Рисунок 3.28).

```
In [38]: category = events.groupby(["Год", "Категория"]).size().reset_index(name="Количество")
grouped_df = category.groupby('Год')
for key, item in grouped_df:
    print(grouped_df.get_group(key), "\n\n")
```

	Год	Категория	Количество
0	2015	Встречи	4057
1	2015	Выставки	2407
2	2015	Концерты	1935
3	2015	Обучение	174
4	2015	Праздники	1507
5	2015	Прочие	867
6	2015	Спектакли	1120

Рисунок 3.28 Использование функции groupby()

Занесем полученные данные по количеству мероприятий за каждый год по распределенным категориям в таблицу, для наглядности (Таблица 3.3).

Таблица 3.3

Количество мероприятий за каждый год по распределенным категориям

Месяц\Год	Количество				
	2015	2016	2017	2018	2019
Встречи	4057	17021	38903	60874	88768

Выставки	2407	7997	13103	17491	23432
Концерты	1935	8777	16848	22308	27992
Обучение	174	965	2761	4613	10036
Праздники	1507	4629	8588	14566	19497
Прочие	867	1002	3174	4534	10668
Спектакли	1120	3541	8273	6253	7915

Построим графики по количеству проведенных мероприятий исходя из категорий. Чтобы было намного легче оценивать соотношение количества мероприятий друг с другом, было решено представить данные в виде круговых диаграмм.

На диаграммах указано количество мероприятий, распределенных по категориям, проведенных за последние 6 лет: 2015 год (Рисунок 3.29), 2016 год (Рисунок 3.30), 2017 год (Рисунок 3.31), 2018 год (Рисунок 3.32), 2019 год (Рисунок 3.33).

Рисунок 3.29 Процентное соотношение категорий за 2015 год

Рисунок 3.30 Процентное соотношение категорий за 2016 год

Рисунок 3.31 Процентное соотношение категорий за 2017 год

Рисунок 3.32 Процентное соотношение категорий за 2018 год

Рисунок 3.33 Процентное соотношение категорий за 2019 год

Можно сделать вывод, что основной упор идет в развитие такого направления, как «Встречи». Мероприятия такого рода подразумевают лекции, связанные с историей, искусством, литературой, природоведением и т.д. Также в таких мероприятиях проводятся различные конкурсы, соревнования, игры и т.д. Такие лекции захватывают практически все сферы культуры, а также есть возможность переноса лекций в онлайн пространство,

что может быть актуально во время карантина. Это позволит продолжать деятельность Министерства культуры вне зависимости от введения карантина.

Мероприятия в категории «Обучение» постепенно увеличивается в количественном и процентном соотношении, можно с уверенностью сказать, что данная сфера постепенно развивается. Однако количество спектаклей, концертов, выставок и мероприятий, связанных с праздниками в 2018 и 2019 годах, уменьшаются в процентном соотношении. Можно предположить, что основные средства уходят на развитие категории «Встречи».

Составим диаграмму за период с 2015 по 2020 год и посмотрим количество мероприятий, проведенных за это время, распределим их также по категориям (Рисунок 3.34). Синий – 2015 год, оранжевый – 2016 год, зеленый – 2017 год, красный – 2018 год, фиолетовый – 2019 год, коричневый – 2020 год.

На графике видно, что количество мероприятий постепенно увеличивается с каждым годом, единственным исключением является категория «Спектакли», в которой количество резко уменьшилось в 2018 году, а в следующем году снова увеличилось, возможно благодаря тому, что 2019 год был объявлен Годом театра, что способствовало росту событий.

Если просмотреть данные за 2020 год, то можно предположить, что рост количества мероприятий в разных сферах также продолжится, так как данные были заполнены лишь за первые 5 месяцев этого года, однако в некоторых категориях их количество уже довольно велико. Например, в категории «Прочие», в которой находится информация по различным акциям, кинопоказам, торжествам и мероприятиям, не попавших в другие категории, можно увидеть, что их количество уже превышает показатели предыдущего 2019 года.

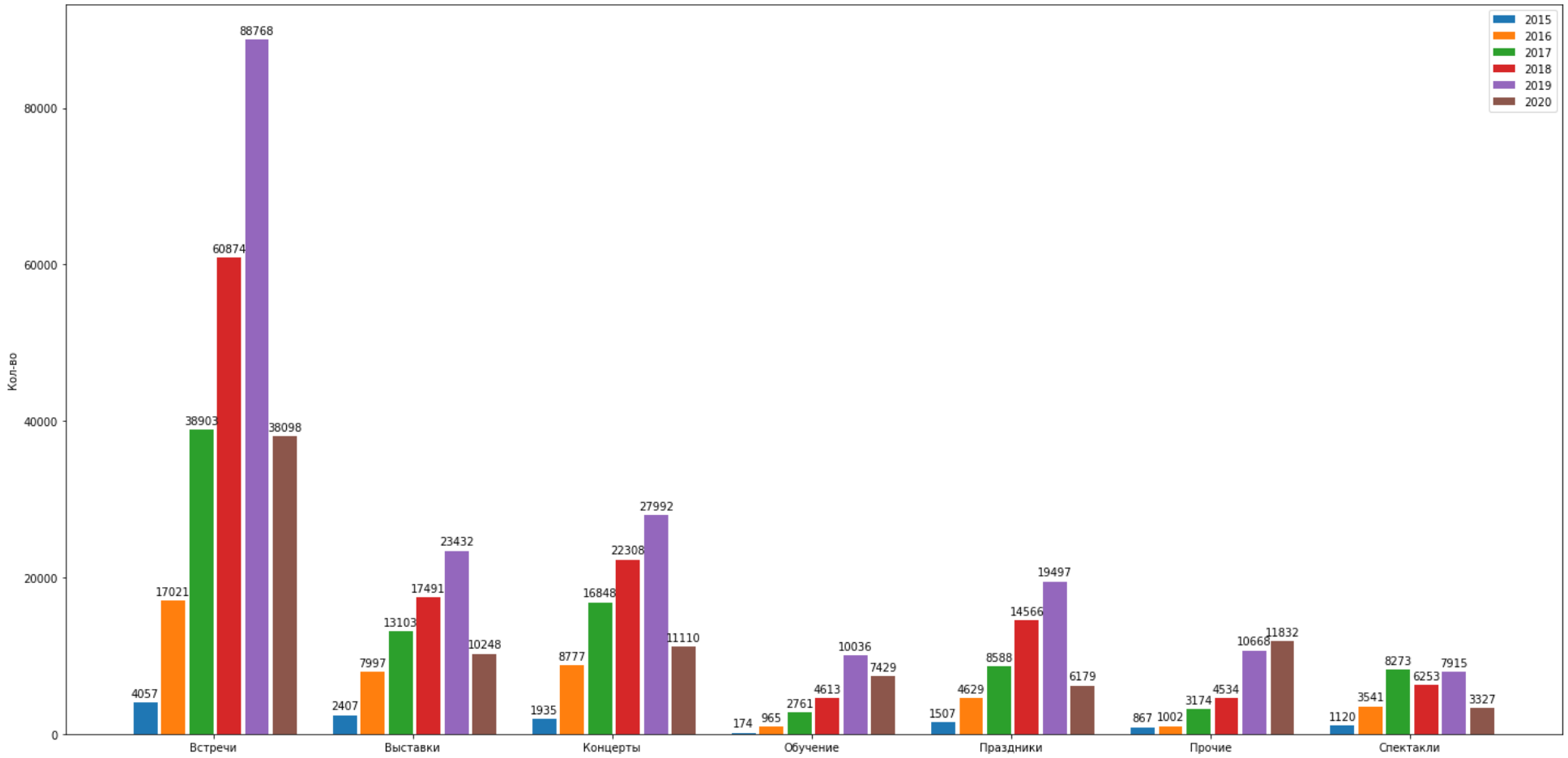


Рисунок 3.34 Количество мероприятий по категориям (2015 – 2020 года)

3.2.4 Кросс-факторный анализ

Чтобы выявить наиболее загруженные дни и времена года по количеству мероприятий, был произведен кросс-факторный анализ по месяцам и дням.

Чтобы информация была наглядна, все данные, полученные из таблицы, были перенесены на тепловую карту (heatmap). Данные были взяты за 2015 (Рисунок 3.35), 2016 (Рисунок 3.36), 2017 (Рисунок 3.37), 2018 (Рисунок 3.38) и 2019 года (Рисунок 3.39).

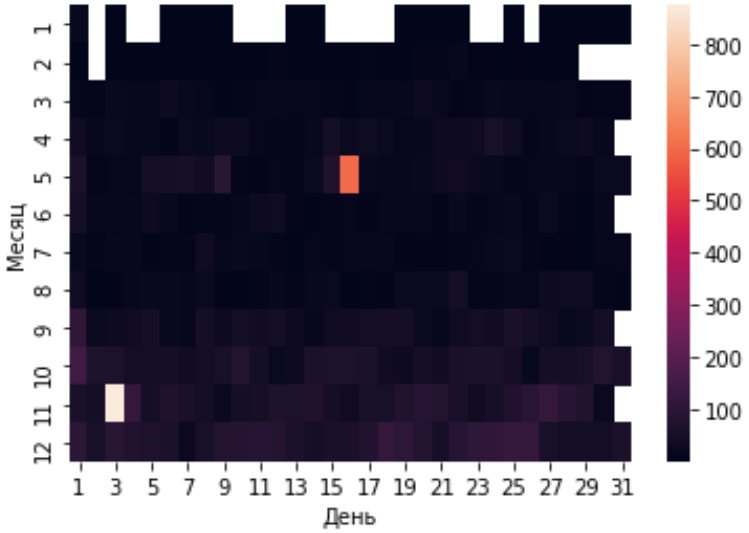


Рисунок 3.35 Тепловая карта мероприятий за 2015 год

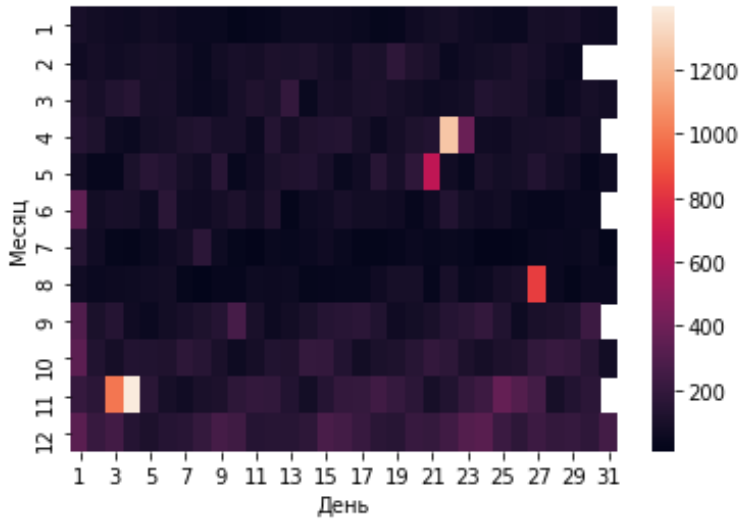


Рисунок 3.36 Тепловая карта мероприятий за 2016 год

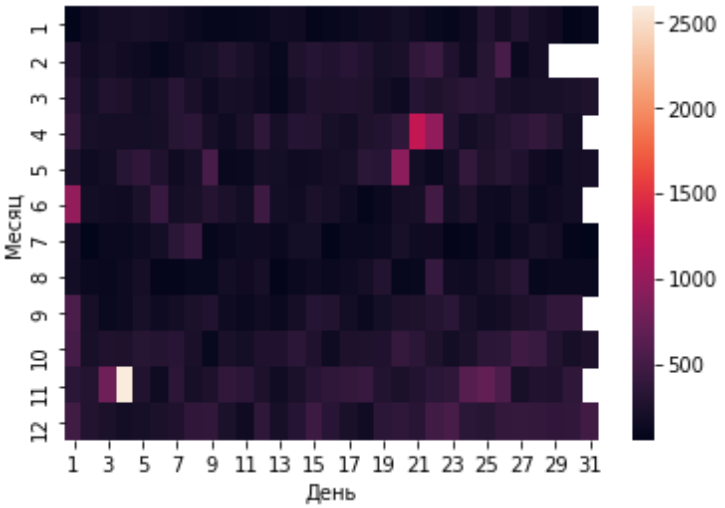


Рисунок 3.37 Тепловая карта мероприятий за 2017 год

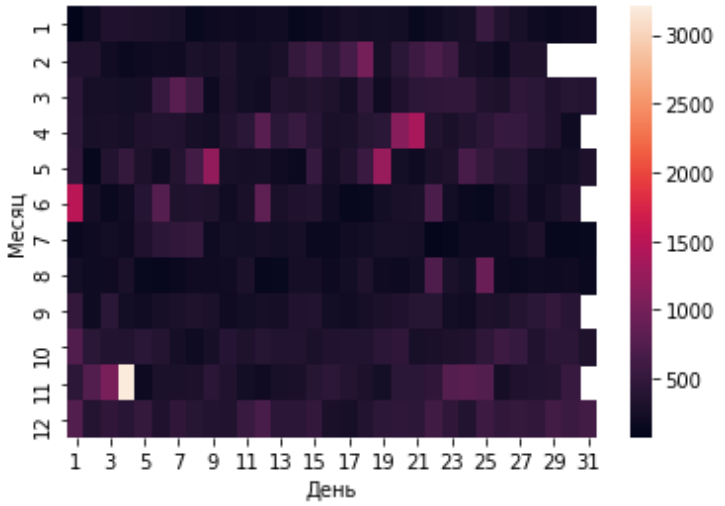


Рисунок 3.38 Тепловая карта мероприятий за 2018 год

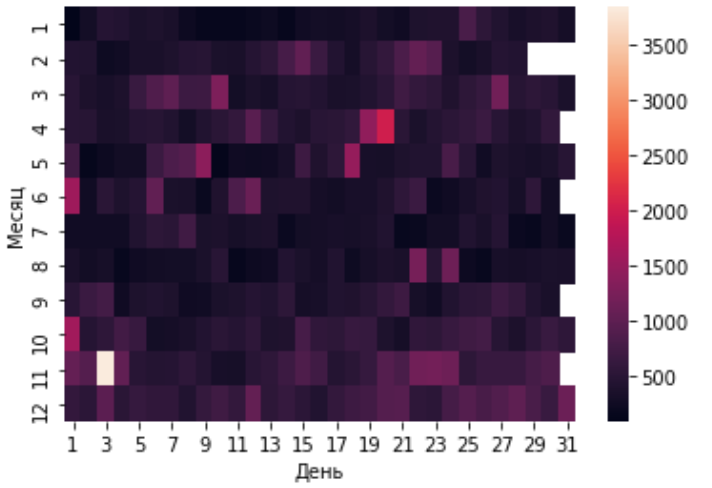


Рисунок 3.39 Тепловая карта мероприятий за 2019 год

Исходя из данных тепловых карт, можно сделать вывод о том, что некоторые месяцы наиболее загружены, по сравнению с другими. Наименьшее количество мероприятий проходит в январе, феврале, июле и августе, дни данных месяцев практически полностью закрашены темно-фиолетовым цветом, что означает низкие показатели. Наиболее «популярными» месяцами оказались декабрь, ноябрь и апрель, что подтверждают диаграммы, составленные по месяцам, которые были рассмотрены ранее.

Можно заметить, что каждый год, в определенные даты количество мероприятий принимает наибольшее значение. Если мы посмотрим на все года, то самое большое значение принимает 3 и 4 ноября, они чередуются в некоторых годах. Вполне возможно, что это связано с Днем народного единства, который проходит 3 ноября. Большое количество записей в данную дату связаны именно с мероприятиями, посвящёнными единству Родины.

Также загруженным днем является 1 июня. В этот день происходит большое количество мероприятий для детей: концерты, спектакли и развлекательные программы. Это можно понять, просмотрев ограничения по возрасту, которые установлены либо от 6 лет, либо они вообще сняты (0+). Можно предположить, что это связанно с Всемирным днем ребёнка.

В апреле в начале 20 чисел резко повышается количество мероприятий, просмотрев записи за эти дни, можно сделать вывод, что это связано с Днем книги, так как проводится большое количество мероприятий, связанных именно с ним: библионочи, библиосумерки, выставки, посвященные книгам.

3.2.5 Виртуальные и онлайн мероприятия

Так как у Министерства культуры есть интернет портал «Культура.РФ», посвященный различным проектам и информированию в сфере культуры, то можно предположить, что деятельность также будет развиваться и в онлайн формате.

Построим графики по количеству проведенных онлайн-мероприятий в месяц. На графиках указано количество онлайн-мероприятий, проведенных за

последние 5 лет: 2015 год (Рисунок 3.40), 2016 год (Рисунок 3.41), 2017 год (Рисунок 3.42), 2018 год (Рисунок 3.43), 2019 год (Рисунок 3.44).

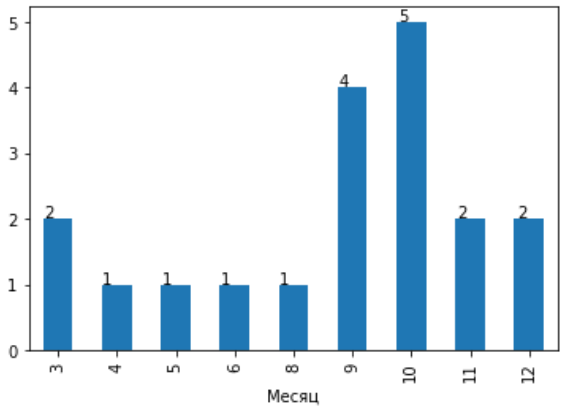


Рисунок 3.40 Количество онлайн-мероприятий за 2015 год по месяцам

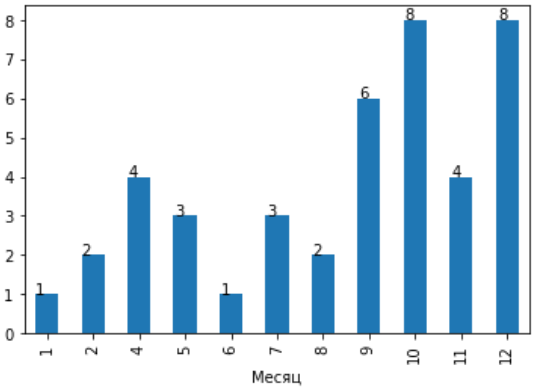


Рисунок 3.41 Количество онлайн-мероприятий за 2016 год по месяцам

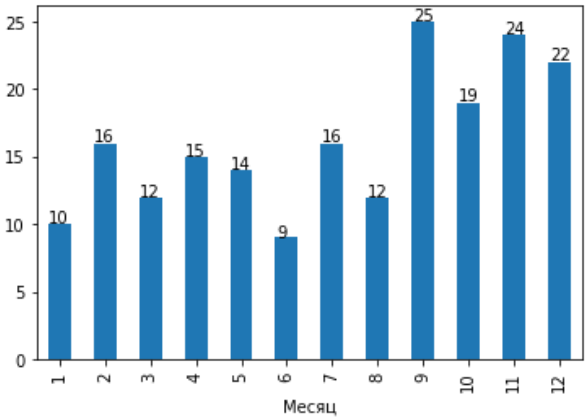


Рисунок 3.42 Количество онлайн-мероприятий за 2017 год по месяцам

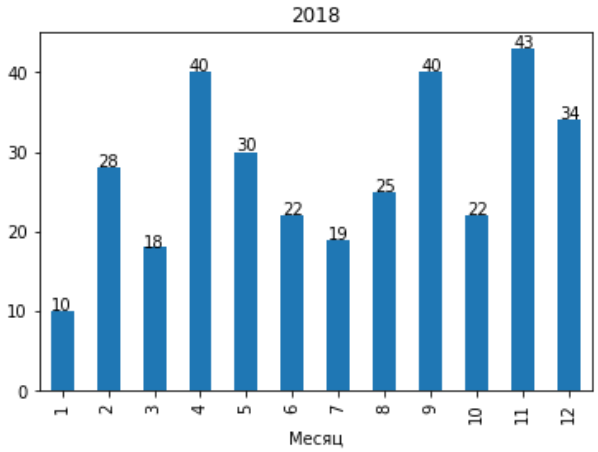


Рисунок 3.43 Количество онлайн-мероприятий за 2018 год по месяцам

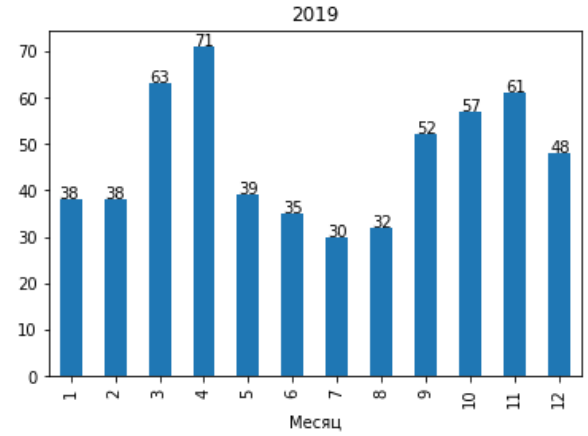


Рисунок 3.44 Количество онлайн-мероприятий за 2019 год по месяцам

Просмотрев данные графики, можно с уверенностью сказать, что постепенно количество проводимых онлайн-мероприятий увеличивается. Стоит заметить, что количество виртуальных событий крайне мало по сравнению ко всем другим, как и их рост.

3.2.6 Стоимость посещения

Просмотрим количество мероприятий за каждый год, за посещение которых необходимо внести плату.

Занесем полученные данные по количеству платных мероприятий за каждый год по распределенным категориям в таблицу, для наглядности (Таблица 3.4).

Таблица 3.4

Количество мероприятий за каждый год по распределенным категориям

Месяц\Год	Количество				
	2015	2016	2017	2018	2019
Встречи	791	2906	3849	4891	6621
Выставки	1156	3727	4507	6006	6818
Концерты	1025	5225	7853	9499	11717
Обучение	66	382	944	1597	2707
Праздники	361	967	1199	1860	2363
Прочие	141	231	858	1714	2841
Спектакли	1015	3306	7718	5446	6331

Если сравнить эти данные с таблицей 3.3, то можно сделать вывод, что в основном платными мероприятиями являются мероприятия из категорий: спектакли, концерты. Примерно половина платных событий приходится на категории выставки. Большое количество бесплатных событий приходится на категории: обучение, встречи, праздники, прочие.

Просмотрим среднее арифметическое цен за каждую категорию, данные были представлены в виде графика, для большей наглядности (Рисунок 3.45). Синий – 2015 год, оранжевый – 2016 год, зеленый – 2017 год, красный – 2018 год, фиолетовый – 2019 год.

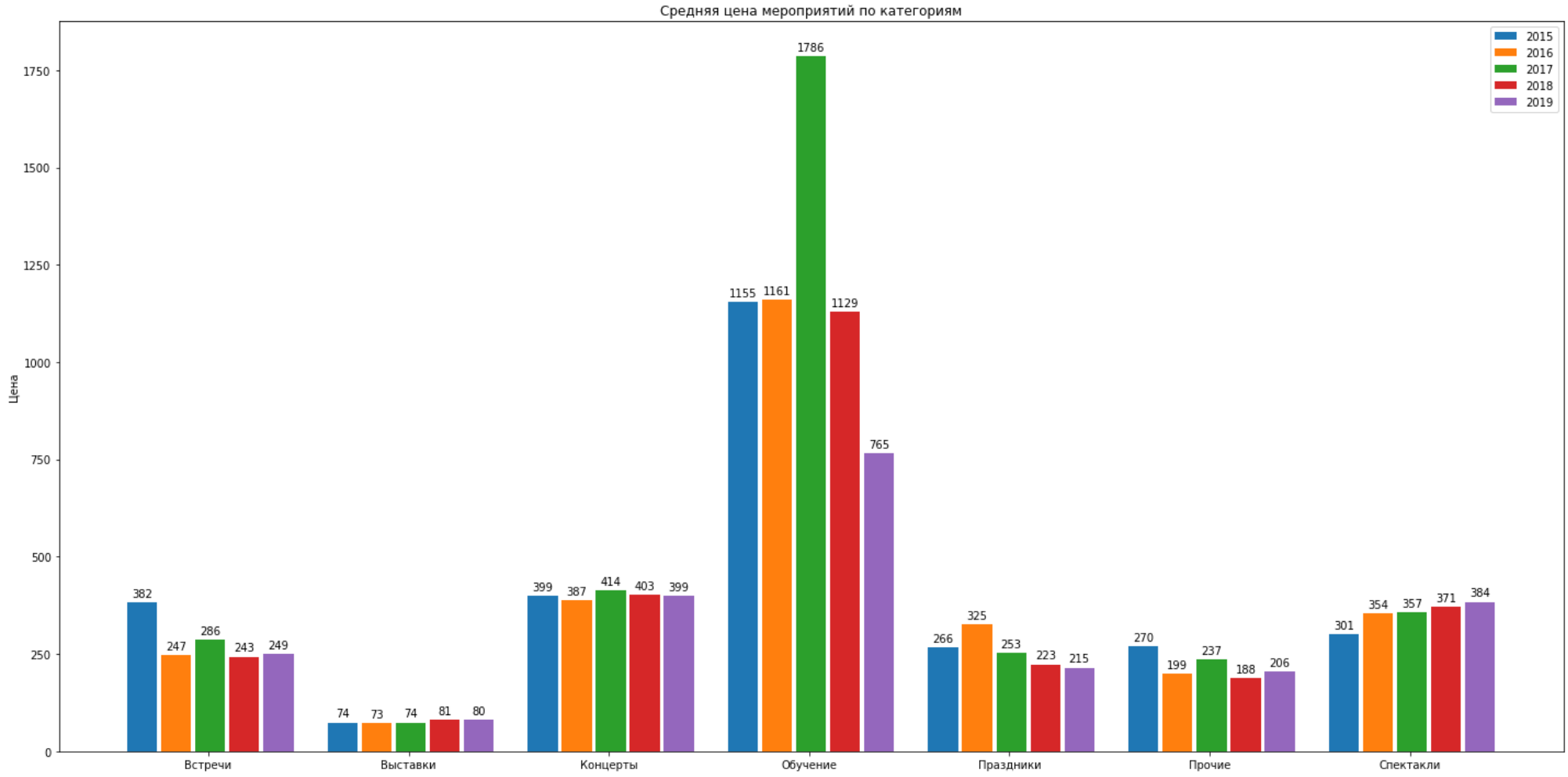


Рисунок 3.45 Стоимость мероприятий по категориям в рублях (2015 – 2019 года)

Исходя из данных, представленных на диаграмме можно сделать вывод, что цена за категории «Спектакли» постепенно возрастает, следовательно, ожидается, что цена и в 2020 году тоже увеличится.

По другим категориям нельзя сделать точного вывода, так как данные изменялись довольно хаотично. Однако можно все-таки предположить, что средняя цена за посещения мероприятий по категориям «Прочие», «Встречи», «Праздники» и «Обучение» уменьшится, так как в основном, в данных по годам видно постепенное понижение в цене.

3.2.7 Исходная модель

Проведя анализ данных, можно предположить, что в 2020 количество мероприятий за каждый месяц только увеличится, так как с каждым годом идет развитие в сфере мероприятий. Однако если учесть ситуацию с самоизоляцией, то результатом скорее будет уменьшение количества мероприятий за апрель.

Внимательно изучив тепловые карты за предыдущие года, исходя из того, что модель мы проверяем на данных за 2020 год за первые пять месяцев, можно предположить, что наиболее загруженными днями будут 20 числа февраля и начало марта, в связи с возможным резким сокращением количества проводимых мероприятий в апреле.

Если говорить о развитии конкретных категорий, то можно рассчитывать на продолжение развития «Образования», и на то, что наиболее популярной будет категория «Встречи». Более того, из-за режима самоизоляции ожидается резкое увеличение количества онлайн мероприятий и виртуальных экскурсий в апреле и мае.

3.3 Проверка исходной модели на данных

В ходе работы была получена модель, которую мы протестируем на данных за 2020 год, данные были взяты за январь, февраль, март, апрель и май.

3.3.1 Мероприятия за каждый месяц

Была построена диаграмма, показывающая количество проведенных мероприятий в 2020 году за первые 5 месяцев (Рисунок 3.46).

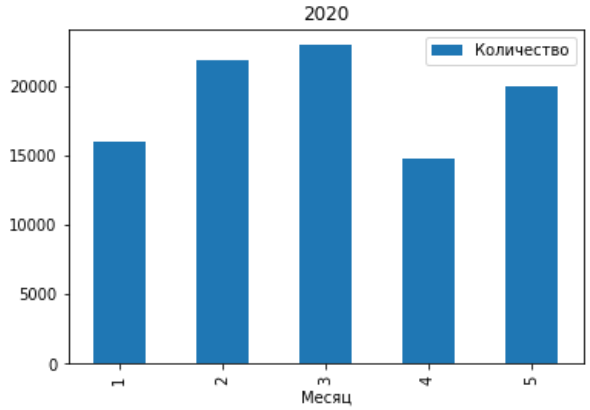


Рисунок 3.46 Количество мероприятий за 2020 год по месяцам

Как можно увидеть на диаграмме, количество мероприятий, начиная с января, постепенно увеличивается, доходит до максимального значения в марте, далее идет на спад. Более того, если сравнить количество мероприятий за каждый месяц с предыдущими годами, то мы увидим, что произошел рост за январь, февраль, март и май (Таблица 3.5).

Таблица 3.5

Количество мероприятий за каждый год и месяц

Месяц\Год	Количество					
	2015	2016	2017	2018	2019	2020
1	57	1922	4641	7164	10254	16015
2	128	2651	6759	10501	14271	21891
3	436	3002	7315	11461	17495	22934
4	707	4467	9714	13119	17499	14742
5	1406	3470	7835	12329	15317	20007
6	462	2554	7243	10541	14522	-
7	391	1518	4748	7215	10366	-
8	538	2485	4988	7579	11563	-
9	1223	3890	6942	9233	13633	-
10	1833	4457	8849	11752	17573	-
11	2874	7296	12751	15541	24366	-
12	2376	6584	10332	14570	22005	-

За апрель видно резкое уменьшение количества проводимых мероприятий, как и ожидалось. Это произошло из-за введения режима самоизоляции, что означает отмену всех событий, которых невозможно провести в режиме онлайн.

3.3.2 Распределение мероприятий по категориям

Была построена круговая диаграмма, показывающая количество проведенных мероприятий в 2020 году за первые 5 месяцев в процентном соотношении (Рисунок 3.47).

Рисунок 3.47 Процентное соотношение категорий за 2020 год

По количеству все так же лидирует категория «Встречи», при этом процент мероприятий, вошедших в категории «Обучение» увеличился. Просмотрим таблицу, где собраны мероприятия за все 6 лет, распределив их по категориям. (Таблица 3.6).

Таблица 3.6

Количество мероприятий за 6 лет по категориям

Месяц\Год	Количество					
	2015	2016	2017	2018	2019	2020
Встречи	4057	17021	38903	60874	88768	40151
Выставки	2407	7997	13103	17491	23432	11142
Концерты	1935	8777	16848	22308	27992	11705
Обучение	174	965	2761	4613	10036	8656
Праздники	1507	4629	8588	14566	19497	6375
Прочие	867	1002	3174	4534	10668	13965
Спектакли	1120	3541	8273	6253	7915	3595

По таблице можно увидеть, что количество мероприятий по некоторым категориям превышает половину, а именно: обучение, прочие. Причем категория «Прочие» превышает количество мероприятий, проведенных в этой же категории в 2019 году.

Другие же категории по количеству прошедших мероприятий составляют немного больше 45% от проведенных мероприятий прошлого 2019 года. Ожидалось, что количество событий будет выше, однако уменьшение произошло именно из-за самоизоляции в апреле и мае.

3.3.3 Распределение мероприятий по тепловой карте

Была построена тепловая карта за пять месяцев 2020 года, для выявления наиболее загруженных дней по мероприятиям (Рисунок 3.48).

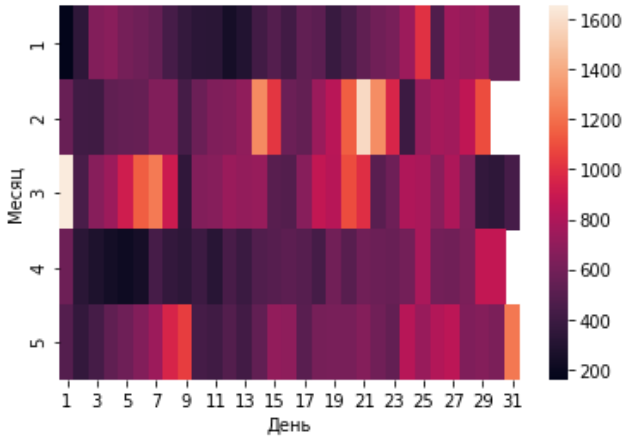


Рисунок 3.48 Тепловая карта мероприятий за 2020 год

Как и предполагалось, наиболее загруженными месяцами оказались февраль и март. Причем наибольшее количество мероприятий пришлось на 20 числа февраля (с 20 по 24) и на первые числа мая (3 число, числа с 5 по 9). Так как с 30 марта были официально объявлены нерабочие дни, то и количество мероприятий заметно уменьшилось в этот период.

При этом количество мероприятий на конец мая сильно увеличилось, чего не было в предыдущие года, что связанно с введением обязательной самоизоляцией.

3.3.4 Онлайн мероприятия в 2020 году

Была построена диаграмма по количеству онлайн-мероприятий и виртуальных экскурсий за первые пять месяцев 2020 года (Рисунок 3.49).

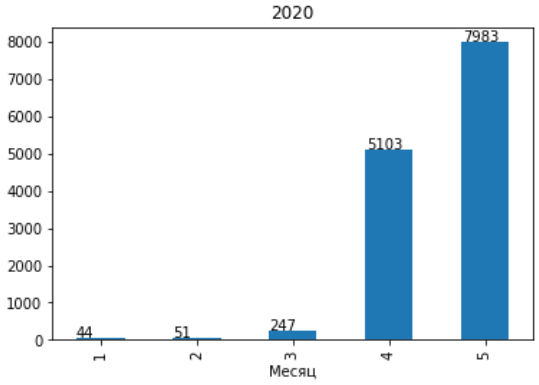


Рисунок 3.49 Количество онлайн-мероприятий за 2020 год по месяцам

Как и ожидалось, произошло резкое увеличение количества проводимых мероприятий в онлайн режиме за апрель и май, в связи с объявленной самоизоляцией. Развитие данного направления может помочь быстро и легко переносить определенные мероприятия в режим онлайн при необходимости, что и произошло в апреле и мае.

3.3.5 Стоимость посещения в 2020 году

Была построена диаграмма по средней арифметической цене за мероприятия, проводимые по категориям за первые 5 месяцев 2020 года, также в диаграмме показана средняя арифметическая цена за категории по предыдущим годам (Рисунок 3.50). Синий – 2015 год, оранжевый – 2016 год, зеленый – 2017 год, красный – 2018 год, фиолетовый – 2019 год, коричневый – 2020 год.

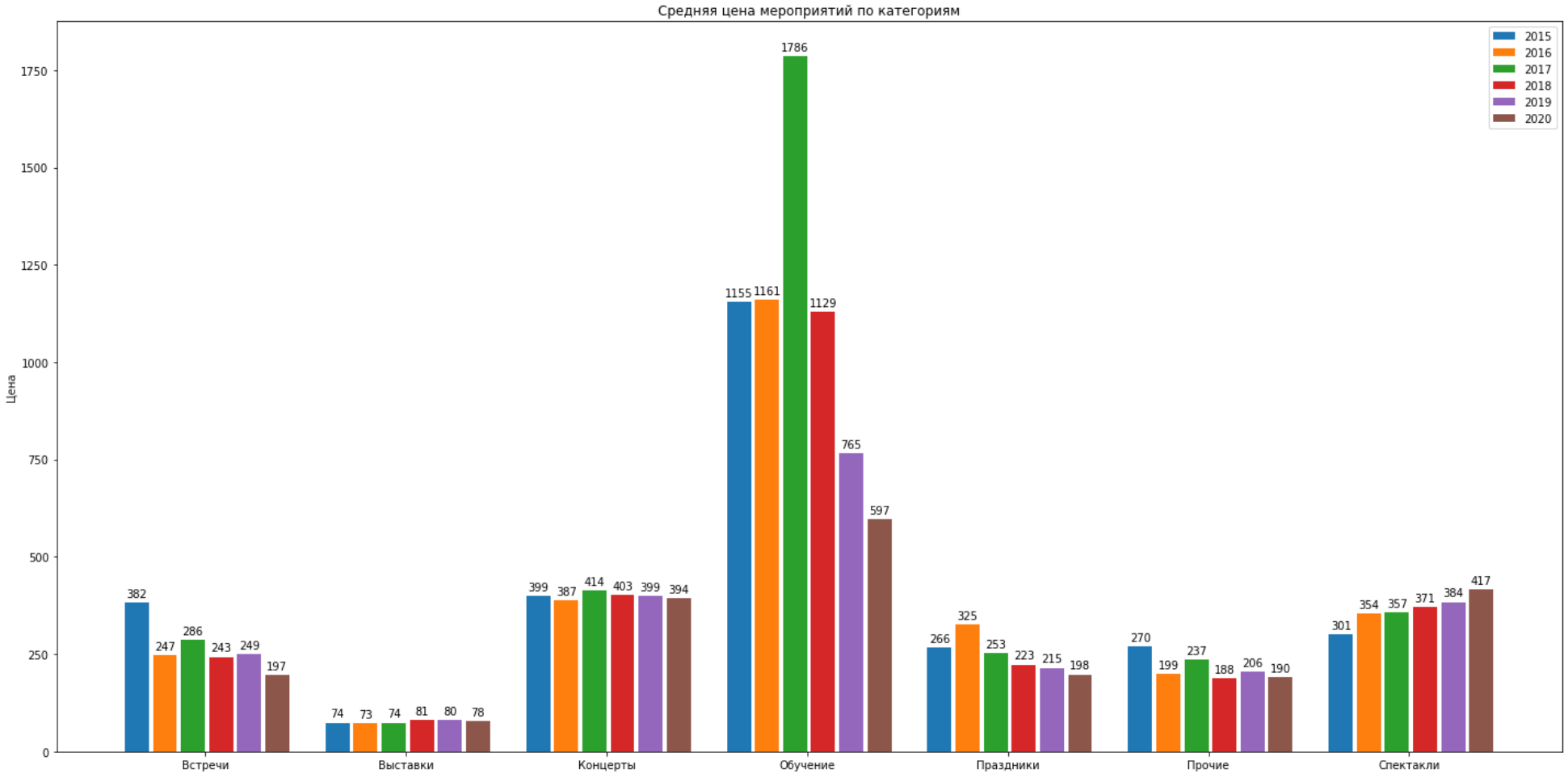


Рисунок 3.50 Стоимость мероприятий по категориям в рублях (2015 – 2020 года)

Среднее арифметическое цены за мероприятия в категории «Спектакли» было увеличено, а за категории «Прочие», «Встречи», «Праздники» уменьшилось. Более того, если посмотреть на среднюю цену по годам за категорию «Обучение», то можно увидеть, что она снова уменьшилась. Также можно увидеть, что среднее арифметическое цены за «Выставки» и «Концерты» уменьшилось, причем разница в цене довольно мала.

Просмотрим количество мероприятий за каждый год, за посещение которых необходимо внести плату и занесем их по распределенным категориям в таблицу, для наглядности (Таблица 3.7).

Таблица 3.7

Количество платных мероприятий за каждый год по категориям

Месяц\Год	Количество					
	2015	2016	2017	2018	2019	2020
Встречи	791	2906	3849	4891	6621	1870
Выставки	1156	3727	4507	6006	6818	1807
Концерты	1025	5225	7853	9499	11717	4410
Обучение	66	382	944	1597	2707	888
Праздники	361	967	1199	1860	2363	732
Прочие	141	231	858	1714	2841	1281
Спектакли	1015	3306	7718	5446	6331	1791

В сравнении данной таблицы с таблицей 3.6 можно заметить, что количество платных мероприятий в определенных категориях намного меньше от общего количества проводимых событий, а именно: встречи, праздники, прочие, выставки, обучение. Однако остаются и категории, в которых доля платных мероприятий намного больше: спектакли, концерты.

3.3.6 Итоговая модель

Проведя анализ данных можно с уверенностью сказать, что количество проводимых мероприятий зависит от времени года и от дат проведения.

Наиболее популярным месяцем является ноябрь, после него количество мероприятий постепенно уменьшается, доходя до минимума в январе. После

января количество проводимых событий снова постепенно увеличивается и доходит до второго максимума в апреле, далее постепенно уменьшаясь и доходя до второго минимума в июле.

В каждый год есть числа, когда количество мероприятий принимает наибольшее значение: 3 и 4 ноября, которые чередуются в некоторых годах; 20 числа апреля и 1 июня. Можно сделать определенный вывод, что это связанно с Днем народного единства 3 ноября, Всемирным днем ребёнка 1 июня и Днем книги 23 апреля.

Основной упор идет в развитие категории «Встречи». Мероприятия в категории «Обучение» постепенно увеличивается в процентном соотношении. Однако количество спектаклей, концертов, выставок и мероприятий уменьшаются в процентном соотношении.

В среднем цена за определенные категории постепенно понижается, а именно: обучение, встречи, праздники, прочие. «Спектакли» напротив постепенно растут в цене, а цена по категориям «Концерты» и «Выставки» колеблется с каждым годом то возрастая, то понижаясь.

Готовность к переходу в режим онлайн сильно повлияла на мероприятия в апреле 2020 года, так как меры по самоизоляции были введены довольно быстро, и не все события были перенесены в онлайн, что сильно сказалось на статистике. Количество мероприятий в апреле резко уменьшилось и достигло минимума в 2020 году, что несвойственно для предыдущих лет.

3.4 Рекомендации по дальнейшему развитию

Данная модель была разработана с целью ознакомления с развитием Министерства культуры РФ в сфере проведения мероприятий. Основная ее функция – понять сильные и слабые стороны, сделать выводы о дальнейшем развитии и улучшениях.

Получив данные после анализа, можно с уверенностью сказать, что до весны 2020 года возможность проведения виртуальных мероприятий и улучшение онлайн портала в этом направлении совершенно не

рассматривались, однако ситуация с COVID-19 вынудила перенаправить основные силы на данную сферу. Так как дальнейшая ситуация с коронавирусной инфекцией все еще остается неизвестной, и вполне возможны дальнейшие вспышки данного заболевания, то стоит быть готовыми к переносу мероприятий в режим онлайн.

Если говорить в общем о развитии различных категорий, то стоит продолжать развивать «Встречи», так как данные события представляют из себя различного рода лекции, связанные с историей, искусством, литературой, природоведением и т.д. Более того, эти мероприятия носят и развлекательный характер: соревнования, игры и прочие подобные события. Такие виды мероприятий помогают привлечь больший интерес, при этом охватывая практически все сферы культуры. Более того, данные лекции легко перенести в онлайн пространство.

Однако следует обратить больше внимания на категории спектаклей, концертов, выставок и мероприятий, связанных с праздниками, так как их количество в процентном соотношении уменьшается по сравнению с другими категориями.

ЗАКЛЮЧЕНИЕ

Данная работа в итоге включает в себя три части: (1) Аналитическую часть: основные цели и задачи Министерства культуры, проекты и развитие; (2) Теоретическую часть: ознакомление с технологией Data Mining, выбор основного метода и инструмента; (3) Проектную часть: интеллектуальный анализ больших данных мероприятий Министерства культуры Российской Федерации, составление итоговой модели, практическая значимость полученной модели, анализ проведенных мероприятий.

В Аналитической части были рассмотрены основные действия Министерства культуры по развитию различных сфер культуры и приобщению граждан к мировому культурному и природному наследию.

В Теоретической части был произведен обзор технологий и методов Big Data. Были сгруппированы задачи интеллектуального анализа данных, а именно: классификация, кластеризация, ассоциация, регрессия, прогнозирование, анализ последовательности, анализ отклонений. Среди методов были выявлены: нейронные сети, линейная регрессия, деревья решений, полиномиальная нейронная сеть, метод k-ближайшего соседа, методы визуализации.

В ходе работы было принято решение использовать визуализацию, так как она играет важную роль в анализе данных и помогает интерпретировать большие данные в структуре реального времени.

Основным инструментом для проведения анализа был язык программирования Python 3, а в качестве платформы – Jupyter Notebook. Так как основным методом выбрана визуализация данных, то было принято решение использовать библиотеки Pandas, Numpy, Matplotlib и Seaborn для построения графиков и диаграм.

В заключающей Проектной части были использованы данные Министерства культуры по проведенным мероприятиям за 2015-2020 года. Данные были подготовлены и преобразованы для анализа. Далее была

построена исходная и итоговая модели, затем произошла ее проверка на данных за первые пять месяцев 2020 года.

В результате было выявлено, что количество проводимых мероприятий зависит от времени года. Наиболее популярным месяцем для проведения мероприятий является ноябрь, после него количество мероприятий постепенно уменьшается, доходя до минимума в январе. После января количество проводимых событий снова постепенно увеличивается и доходит до второго максимума в апреле, далее постепенно уменьшаясь и доходя до второго минимума в июле. Большое количество мероприятий приходится на праздники: День народного единства, Всемирный день ребенка, День книги.

Также были выведены наиболее популярные категории и категории, продолжающие развиваться: встречи, обучение. Выяснилось, что количество мероприятий постепенно увеличивается с каждым годом, единственным исключением является категория «Спектакли», в которой количество мероприятий уменьшилось в 2018 году, однако 2019 год был объявлен Годом театра, что вновь поспособствовало росту.

Цена за определенные категории постепенно понижается: обучение, встречи, праздники, прочие. «Спектакли» напротив растут в цене, а стоимость по категориям «Концерты» и «Выставки» колеблется с каждым годом то возрастая, то понижаясь.

Также стоит отметить, что готовность к переходу в режим онлайн сильно повлияла на деятельность Минкультуры в апреле 2020 года, так как не все события были перенесены в онлайн, что сильно сказалось на статистике. Количество мероприятий в апреле резко уменьшилось и достигло минимума в этом году, что несвойственно для предыдущих лет. В связи с данной ситуацией, данные за 2020 год сильно отличаются от предыдущих лет, однако в 2021 году должны произойти значительные улучшения и готовность к определенным трудностям, связанных с введением карантина и самоизоляцией.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

1. Mkrf.ru – Министерство культуры РФ [Электронный ресурс]. Режим доступа: https://www.mkrf.ru/documents/plan-deyatelnosti-ministerstva-kultury-rossiyskoy-federatsii-na-2016-2021-gody/, свободный — (дата обращения: 18.05.2020)
2. Kremlin.ru – Администрация Президента России [Электронный ресурс]. Режим доступа: http://kremlin.ru/acts/news/17944, свободный – (дата обращения: 18.05.2020)
3. Onf.ru – Общероссийский народный фронт [Электронный ресурс]. Режим доступа: https://onf.ru/ 2015/01/29/putin-cel-provedeniya-goda-literatury-napomnit-ob-isklyuchitelnoy -znachimosti-literatury/, свободный — (дата обращения: 18.05.2020)
4. Mkrf.ru – Министерство культуры РФ [Электронный ресурс]. Режим доступа: https://www.mkrf.ru/upload/iblock/37a/37afb75287077b40ee57a017a17d6f9.pdf, свободный — (дата обращения: 18.05.2020)
5. Mkrf.ru – Министерство культуры РФ [Электронный ресурс]. Режим доступа: https://www.mkrf.ru/activities/plan/, свободный — (дата обращения: 18.05.2020)
6. Rbc.ru – РБК [Электронный ресурс]. Режим доступа: https://www.rbc.ru/society/11/05/2020/5eb997529a79478657f8a6b0, свободный – (дата обращения: 18.05.2020)
7. Medium – платформа для социальной журналистики [Электронный ресурс]. Режим доступа: https://medium.com/@murielporter3/6-stages-of-data-mining-process-in-wisdom-of-business-71bfe62423ea, свободный — (дата обращения: 18.05.2020)
8. Medium https://medium.com/yogesh-khuranas-blogs/difference-between-model-validation-and-model-evaluation-1a931d908240

9. BarnRaisers – агентство полного цифрового маркетинга [Электронный ресурс]. Режим доступа: https://barnraisersllc.com/2018/10/01/data-mining-process-essential-steps/, свободный — (дата обращения: 18.05.2020).
10. Википедия — свободная энциклопедия [Электронный ресурс]. Режим доступа: https://ru.wikipedia.org/wiki/ecofriendly, свободный - (дата обращения: 18.05.2020).
11. Dic.academic.ru – Академик словарь [Электронный ресурс]. Режим доступа: https://dic.academic.ru/dic.nsf/logic/291, свободный - (дата обращения: 18.05.2020).
12. Research Methods Knowledge Base – Методы исследования База знаний [Электронный ресурс]. Режим доступа: https://conjointly.com/kb/data-preparation/, свободный — (дата обращения: 18.05.2020).
13. SQL-soft.ru [Электронный ресурс]. Режим доступа: http://www.sql-soft.ru/glava-01-ploskie-fajly.htm, свободный — (дата обращения: 18.05.2020).
14. Software Testing Help – Блог о тестировании программного обеспечения [Электронный ресурс]. Режим доступа: https://www.softwaretestinghelp.com/data-mining-process/, свободный — (дата обращения: 19.05.2020).
15. Wisdom IT Services – IT Сервис [Электронный ресурс]. Режим доступа: https://www.wisdomjobs.com/e-university/data-mining-tutorial-199/data-mining-tasks-1871.html, свободный — (дата обращения: 19.05.2020).
16. National Institute of Strategic Studies – Национальный институт стратегических исследований [Электронный ресурс]. Режим доступа: http://www.gmdh.net/, свободный — (дата обращения: 20.05.2020).
17. Международный журнал по электронике и вычислительной технике [Электронный ресурс]. Режим доступа: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.640.136&rep=rep1&type=pdf, свободный — (дата обращения: 20.05.2020).

18. Big Data Make Simple – ведущий технологический портал в Big Data [Электронный ресурс]. Режим доступа: https://bigdata-madesimple.com/top-8-programming-languages-every-data-scientist-should-master-in-2019/, свободный — (дата обращения: 22.05.2020).
19. Википедия — свободная энциклопедия [Электронный ресурс]. Режим доступа: https://en.wikipedia.org/wiki/Dirty_data, свободный — (дата обращения: 22.05.2020).
20. Российская государственная библиотека [Электронный ресурс]. Режим доступа: https://www.rsl.ru/ru/editions/bibliography-editions/, свободный — (дата обращения: 25.05.2020).
21. Srinimf.com – Srinimf сайт для разработчиков программного обеспечения [Электронный ресурс]. Режим доступа: https://srinimf.com/2017/02/12/top-python-benefits-really-better-option-for-analytics/, свободный — (дата обращения: 25.05.2020).
22. Rcfoundation.ru – Российский фонд культуры [Электронный ресурс]. Режим доступа: https://rcfoundation.ru/about.html, свободный — (дата обращения: 25.05.2020).
23. Brainyquote.com – [Электронный ресурс]. Режим доступа: https://www.brainyquote.com/quotes/bram_cohen_219754, свободный — (дата обращения: 25.05.2020).
24. Oreily.com – O'Reilly издательская компания [Электронный ресурс]. Режим доступа: https://www.oreilly.com/library/view/python-data-science/9781491912126/ ch04.html, свободный — (дата обращения: 25.05.2020).
25. Opendata.mkrf.ru – Открытые данные Минкультуры [Электронный ресурс]. Режим доступа: https://opendata.mkrf.ru/opendata/7705851331-events#a:eyJ0YWIiOiJidWlsZF90YWJsZSJ9, свободный — (дата обращения: 18.05.2020)

ПРИЛОЖЕНИЕ

Листинг кода

Загрузка библиотек:

```
%pylab inline
import pandas as pd
import seaborn as sb
import matplotlib
```

Загрузка данных и их фильтрация:

```
events = pd.read_csv('events.csv')
events = events.filter(['Начало мероприятия','Мероприятие','Место
проведения','Категория','возрастное ограничение','стоимость посещения'])
events.dropna(subset=['Категория', 'Место проведения', 'Мероприятие', 'Начало
мероприятия'], inplace=True)
events = events.fillna(0)
events.drop(events[events["стоимость посещения"] > 130000].index,
inplace=True)
import string
printable_chars = set(string.printable + "
абвгдеёжзийклмнопрстуфхцчшщъыьэюяАБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ«»—–—…é-
№\uc2a0")
events = events[events.applymap(lambda x:
set(str(x)).issubset(printable_chars)).all(1)]
```

Преобразование столбца «Начало мероприятия»:

```
events['Начало мероприятия'] = events['Начало
мероприятия'].map(pd.to_datetime)
from datetime import datetime as dt
events['Начало мероприятия'] = events['Начало мероприятия'].dt.normalize()
def get_dom(dt) :
    return dt.day
def get_month(dt) :
    return dt.month
def get_year(dt) :
    return dt.year
events ['День'] = events['Начало мероприятия'].map(get_dom)
events ['Месяц'] = events['Начало мероприятия'].map(get_month)
events ['Год'] = events['Начало мероприятия'].map(get_year)
```

Удаление ненужных значений:

```
indexNames = []
indexNames.append(events[(events['Год'] == 2005)].index)
indexNames.append(events[(events['Год'] == 2025)].index)
indexNames.append(events[(events['Год'] == 2014)].index)
indexNames.append(events[(events['Год'] == 2021)].index)
for i in range(len(indexNames)) :
    events.drop(indexNames[i] , inplace=True)
indexNames = events.query('Год' == 2020 & `Месяц` > 6").index
for i in range(len(indexNames)) :
    events.drop(indexNames[i] , inplace=True)
```

Построение диаграмм по количеству мероприятий за каждый месяц:

```
year = 2014
grouped_df = month.groupby('Год')
for key, item in grouped_df:
    group = grouped_df.get_group(key)
    year += 1
    group = group.filter(['Месяц','Количество'])
    group.plot.bar(x='Месяц', title = year)
```

Построение диаграмм по количеству мероприятий за все года:

```
import collections
months_names = ["январь", "февраль", "март", "апрель", "май", "июнь", "июль",
"август", "сентябрь", "октябрь", "ноябрь", "декабрь"]
months = np.arange(1, 12 + 1)
years = np.arange(2015, 2019 + 1)
def plot_groups_of_bars(title, ylabel, xlabels, groups_of_bars):
    center_x = np.arange(len(xlabels))
    width = 0.12
    offset = 0.02
    fig, ax = plt.subplots(figsize=(20, 10))
    rects_groups = []
    groups_count = len(groups_of_bars)
    for index, (name, values) in enumerate(groups_of_bars.items()):
        left_x = center_x - (groups_count - 1) * width / 2 - (groups_count -
2) * offset / 2
        rects = ax.bar(
            left_x + index * (width + offset),
            values,
            width,
            label=name)
        rects_groups.append(rects)
    ax.set_ylabel(ylabel)
    ax.set_title(title)
    ax.set_xticks(center_x)
    ax.set_xticklabels(xlabels)
    ax.legend()
    for rects in rects_groups:
        for rect in rects:
            height = rect.get_height()
            ax.annotate(
                '{}'.format(height),
                xy=(rect.get_x() + rect.get_width() / 2, height),
                xytext=(0, 3),
                textcoords="offset points",
                ha="center", va='bottom'
            )
    fig.tight_layout()
    plt.show()
sorted_events = events.groupby(["Месяц",
"Год"]).size().reset_index(name="Количество")
events_by_years = sorted_events.groupby(["Год"])
counts_by_years = {year: events["Количество"].tolist() for year, events in
events_by_years}
del counts_by_years[2020]
counts_by_years_1 = {year: counts[:6] for year, counts in
counts_by_years.items()}
counts_by_years_2 = {year: counts[6:] for year, counts in
counts_by_years.items()}
print(counts_by_years_1, counts_by_years_2)
```

Построение автокорреляционной функции:

```
from math import sqrt
from numpy.random import seed
from numpy.random import randn
from numpy import mean
from scipy.stats import sem
from scipy.stats import t
import statsmodels.api as sm
from statsmodels.iolib.table import SimpleTable
from sklearn.metrics import r2_score
import ml_metrics as metrics
from pandas.plotting import lag_plot
from statsmodels.graphics import tsaplots
def independent_ttest(data1, data2, alpha):
    mean1, mean2 = mean(data1), mean(data2)
    se1, se2 = sem(data1), sem(data2)
    sed = sqrt(se1**2.0 + se2**2.0)
    t_stat = (mean1 - mean2) / sed
    df = len(data1) + len(data2) - 2
    cv = t.ppf(1.0 - alpha, df)
    p = (1.0 - t.cdf(abs(t_stat), df)) * 2.0
    return t_stat, df, cv, p
events['МесяцГод'] = events.apply(lambda row:
f'{row["Год"]}.{row["Месяц"]:02}', axis = 1)
series = events[['МесяцГод',
'День']].groupby("МесяцГод").count().reset_index().rename(columns =
{"День":"Количество"})
series2019 = series.head(-6)
series.head(-6).plot(x = 'МесяцГод')
series.plot(x = 'МесяцГод')
fig = tsaplots.plot_acf(series.head(-6)['Количество'])
seed(1)
data1 = 5 * randn(100) + 50
data2 = 5 * randn(100) + 51
alpha = 0.05
t_stat, df, cv, p = independent_ttest(data1, data2, alpha)
print('t=%.3f, df=%d, cv=%.3f, p=%.3f' % (t_stat, df, cv, p))
```

Построение круговых диаграмм по категориям:

```
year = 2014
grouped_df = category.groupby('Год')
for key, item in grouped_df:
    group = grouped_df.get_group(key)
    year += 1
    group = group.filter(['Категория','Количество'])
    fig1, ax1 = plt.subplots()
    explode = (0.05,0.05,0.05,0.05,0.05,0.05,0.05)
    sizes = group['Количество']
    labels = group['Категория']
    ax1.pie(sizes, labels=labels, explode = explode, autopct='%1.0f%%',
startangle=90)
    ax1.axis('equal')
    plt.show()
```

Построение диаграмм онлайн мероприятий:

```
online = events[events["Мероприятие"].str.match('онлайн|виртуал', case =
False)]
online = online.groupby(["Год",
"Месяц"]).size().reset_index(name="Количество")
year = 2014
grouped_df = online.groupby('Год')
for key, item in grouped_df:
    group = grouped_df.get_group(key)
    year += 1
    group = group.filter(['Месяц','Количество'])
    ax = group.plot.bar(x='Месяц', title = year, legend = None)
    for p in ax.patches:
        ax.annotate(str(p.get_height()), (p.get_x() * 1.005, p.get_height() *
1.005))
```

Пример построения тепловых карт:

```
group2 = grouped_df.get_group(2020)
group2 = group2.filter(['Количество', 'Месяц', 'День'])
heatmap2_data = pd.pivot_table(group2, values='Количество', index=['Месяц'],
columns='День')
sb.heatmap(heatmap2_data)
```

Расчет средней стоимости посещения:

```
category_money = events.drop(events[events["стоимость посещения"] ==
0].index)
category_money = category_money.groupby(["Год", "Категория"]).agg({"стоимость
посещения":'mean'}).round().astype(int)
grouped_df = category_money.groupby('Год')
```

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

Факультет информационных технологий

Кафедра «Прикладная информатика»

Форма обучения: очная

ИЛЛЮСТРАТИВНЫЙ МАТЕРИАЛ К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

по направлению 01.03.02 «Прикладная математика и информатика»

на тему «Анализ данных мероприятий Министерства культуры РФ с использованием технологий Big Data»

Студентка _____ Елизавета Александровна Фролова

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

ДОПУСКАЕТСЯ К ЗАЩИТЕ

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

Москва 2020

Анализ данных мероприятий Министерства культуры РФ с использованием технологий Big Data

Студентка

Е. А. Фролова
Группа 161-381

Руководитель,
к.п.н. доцент

Н. И. Царькова

Цели и задачи работы

Цель работы – анализ данных мероприятий министерства культуры РФ с использованием технологий Big Data.

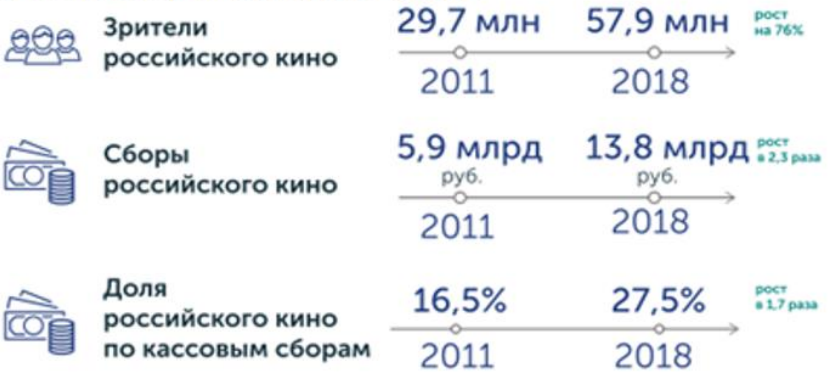
Для достижения поставленной цели были определены следующие задачи:

1. Анализ деятельности Минкультуры РФ.
2. Подготовка данных для анализа с использованием технологий Big Data.
3. Построение исходной модели.
4. Проверка корректности модели на имеющихся данных.
5. Произведение корректировки модели.
6. Формирование рекомендаций и итоговой модели.

Основные проекты

Развития деятельности театров:

Развитие российского кино:



Развития деятельности музеев:

Основные преимущества Python 3

Python и Big Data – это новая комбинация, завоевавшая рыночное пространство. Python пользуется большим спросом среди компаний Big Data.

Преимущества:

1. Простое кодирование
2. Открытый исходный код
3. Поддержка аналитических библиотек
4. Высокая скорость
5. Поддержка обработки данных

Так как в качестве анализа была выбрана визуализация, то была использована бесплатная библиотека matplotlib для визуализации данных.

Исходные данные и их преобразование

Количество данных: 582032 строки и 122 колонки.

Удаление пустых строк:

```
events.dropna(subset=['Категория', 'Место проведения', 'Мероприятие', 'Начало мероприятия'], inplace=True)
```

Фильтрация данных:

```
events = events.filter(['Начало мероприятия','Мероприятие','Место проведения','Категория','возрастное ограничение','стоимость по
print(events)
```

Преобразование String в формат Date:

```
events['Начало мероприятия'] = events['Начало мероприятия'].map(pd.to_datetime)
print(events)
```

```
             Начало мероприятия  \
0      2019-02-02 14:00:00+00:00
1      2019-01-24 06:00:00+00:00
2      2019-01-22 07:00:00+00:00
3      2019-02-02 07:00:00+00:00
4      2019-02-02 07:00:00+00:00
```

Замена пустых значений:

```
In [10]: events = events.fillna(0)
         print(events)
```

Исходные данные и их преобразование

Удаление некорректных записей:

```
import string
printable_chars = set(string.printable + "абвгдеёжзийклмнопрстуфхцчшщъыьэюяАБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ«»–—…é-№\uc2a0")

events = events[events.applymap(lambda x: set(str(x)).issubset(printable_chars)).all(1)]
```

Добавление новых столбцов:

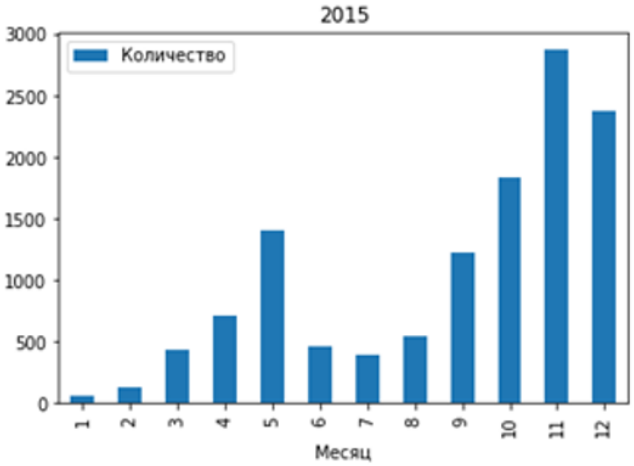
```
def get_dom(dt) :
    return dt.day
def get_month(dt) :
    return dt.month
def get_year(dt) :
    return dt.year
events ['День'] = events['Начало мероприятия'].map(get_dom)
events ['Месяц'] = events['Начало мероприятия'].map(get_month)
events ['Год'] = events['Начало мероприятия'].map(get_year)
```

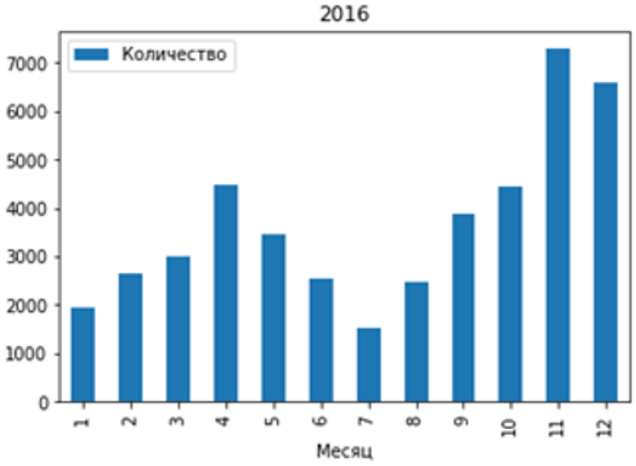
Удаление выбросов:

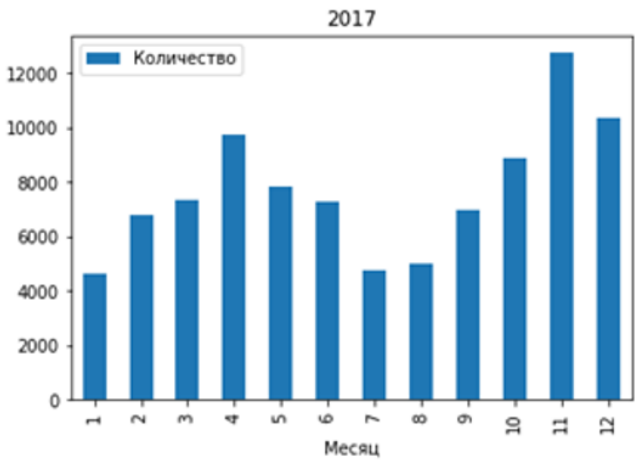
```
events = events.fillna(0)
events.drop(events[events["стоимость посещения"] > 130000].index, inplace=True)
```

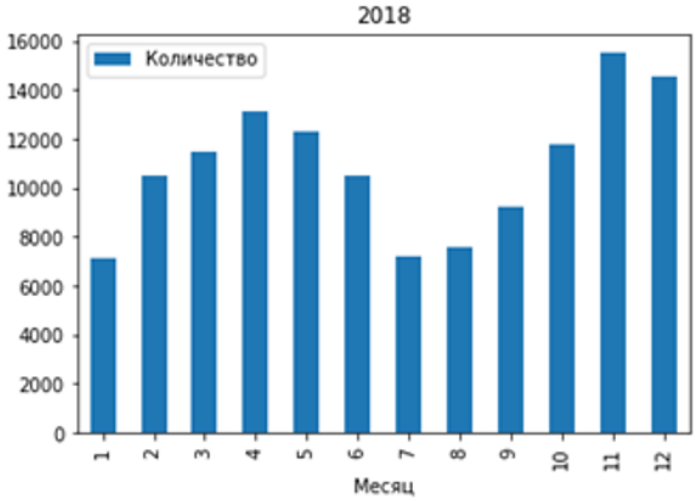
После корректировки набор данных содержит 579500 строк и столбцы:
Начало мероприятия, Мероприятие, Категория, возрастное ограничение, стоимость посещения, День, Месяц, Год

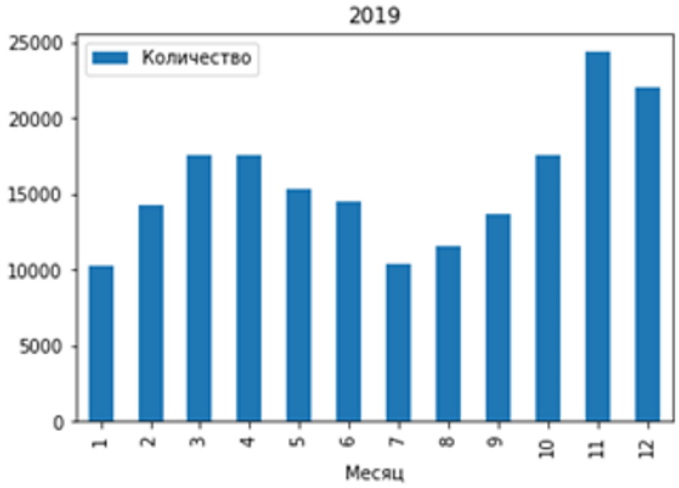
Анализ по количеству мероприятий в месяц







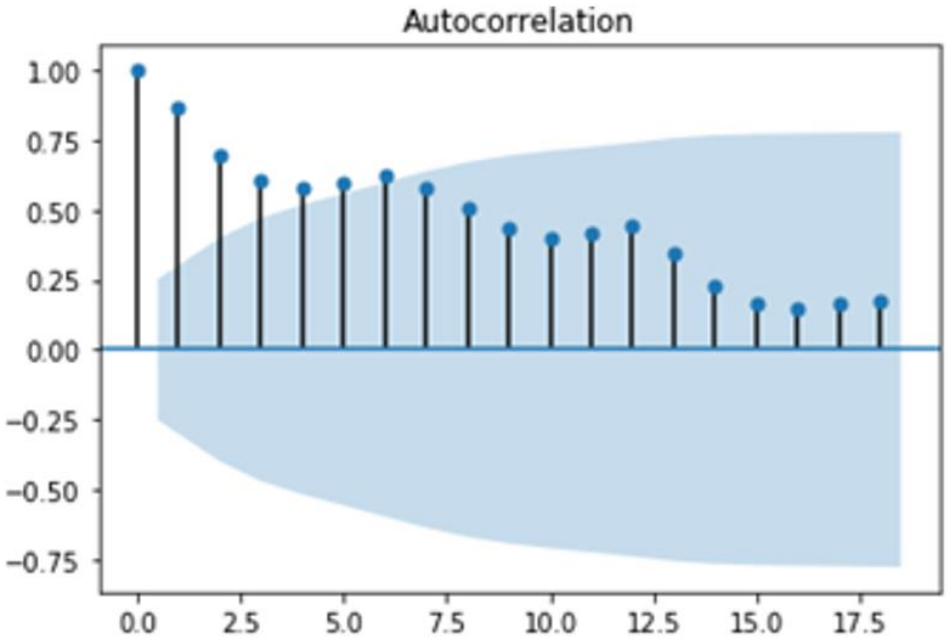




Автокорреляционная функция

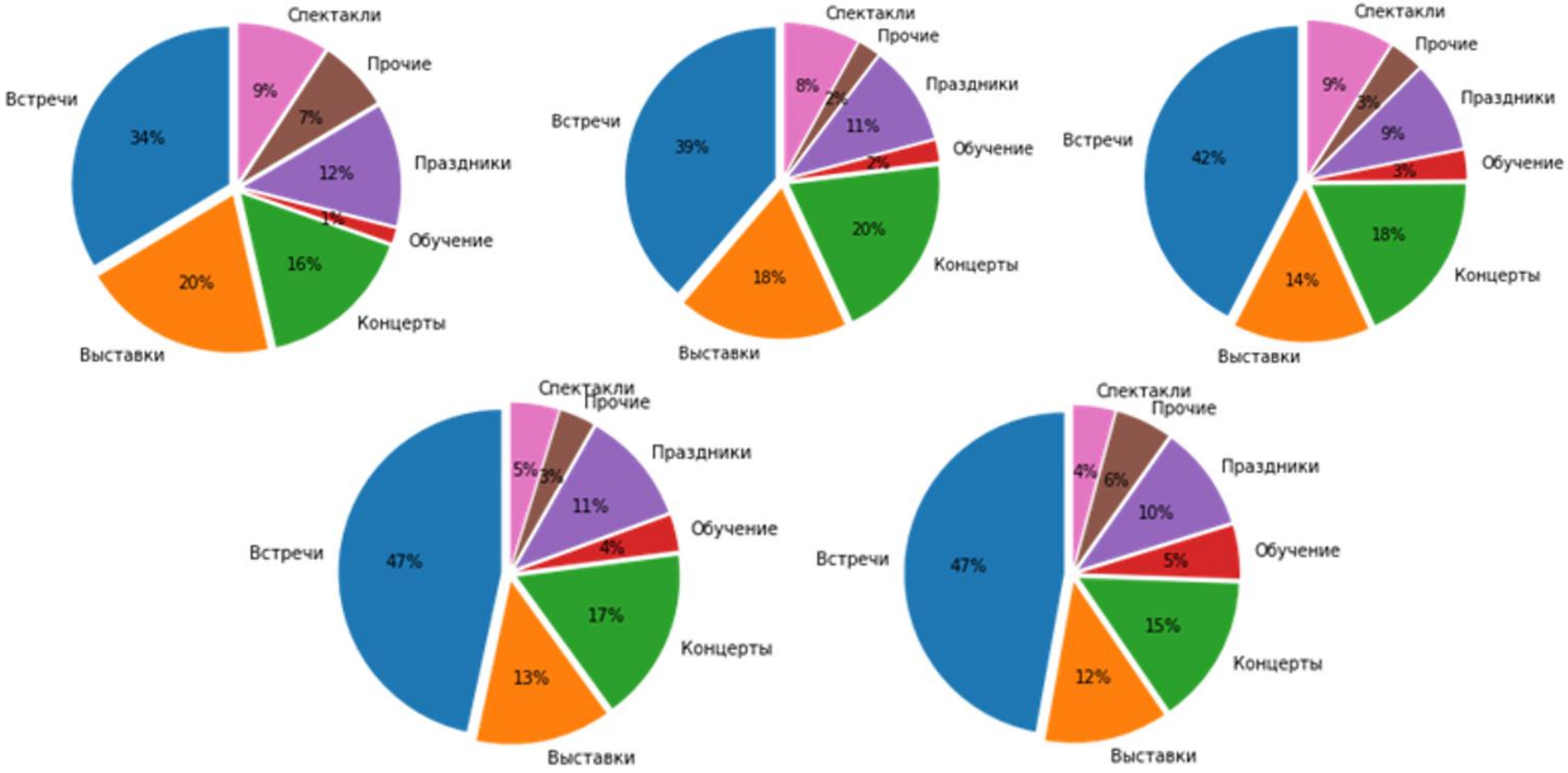
Было получено значение коэффициента корреляции 1-го порядка равное 0,929

Далее была вычислена последовательность коэффициентов автокорреляции уровней первого, второго и т. д. порядков, и построена автокорреляционная функция.



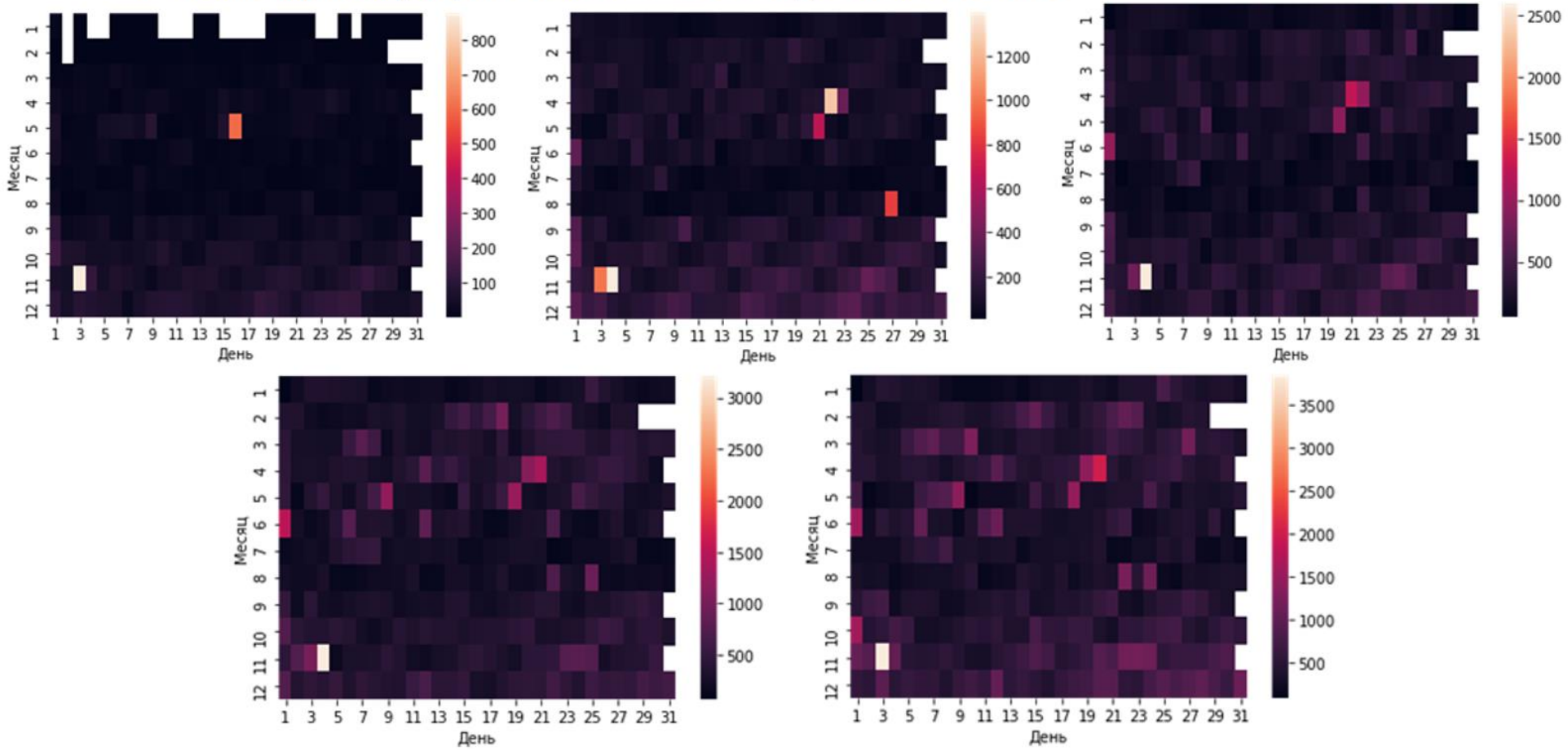
Мероприятия по категориям

Соотношение категорий за 2015-2019 года соответственно:

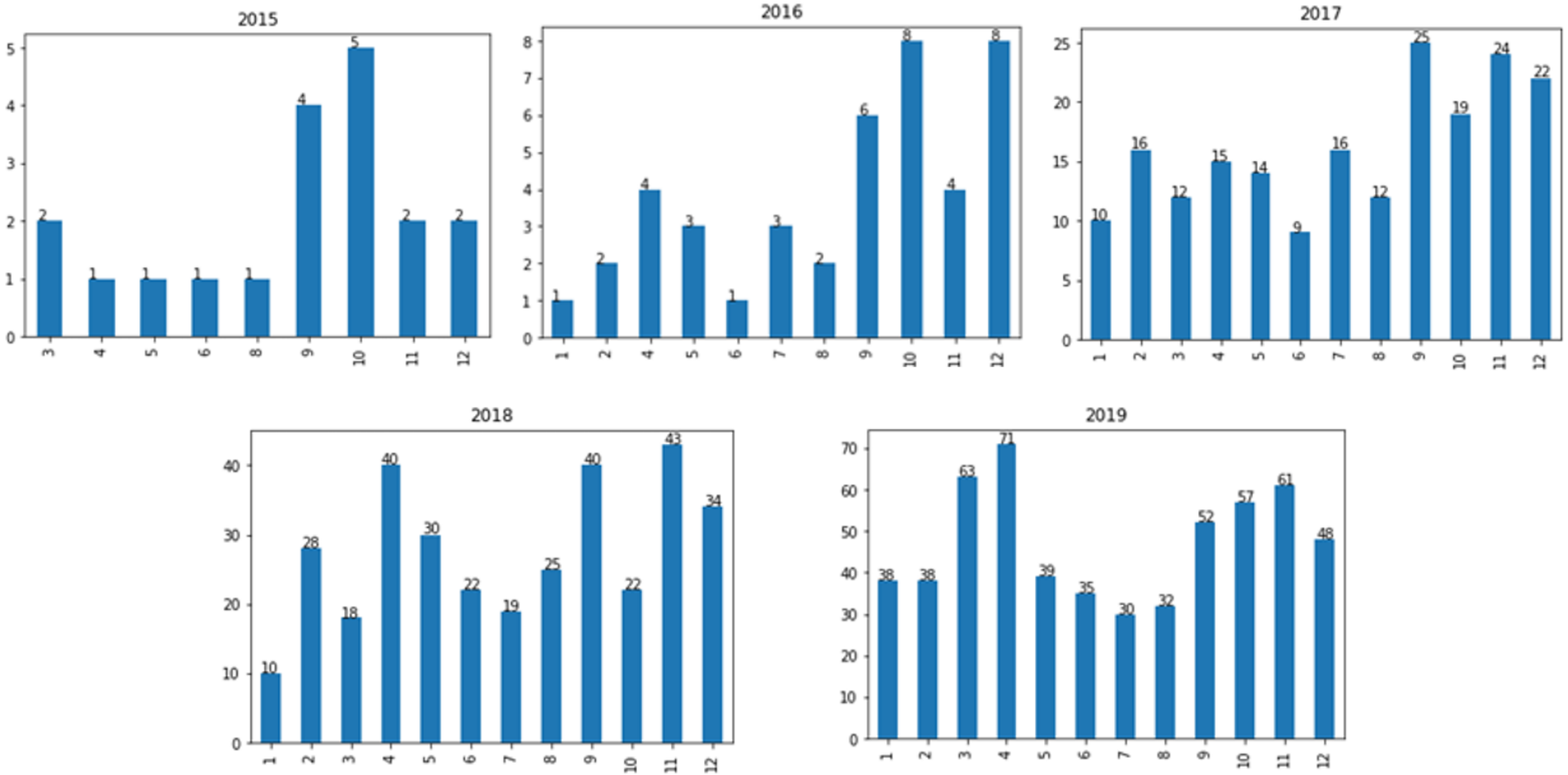


Количество мероприятий по дням и месяцам

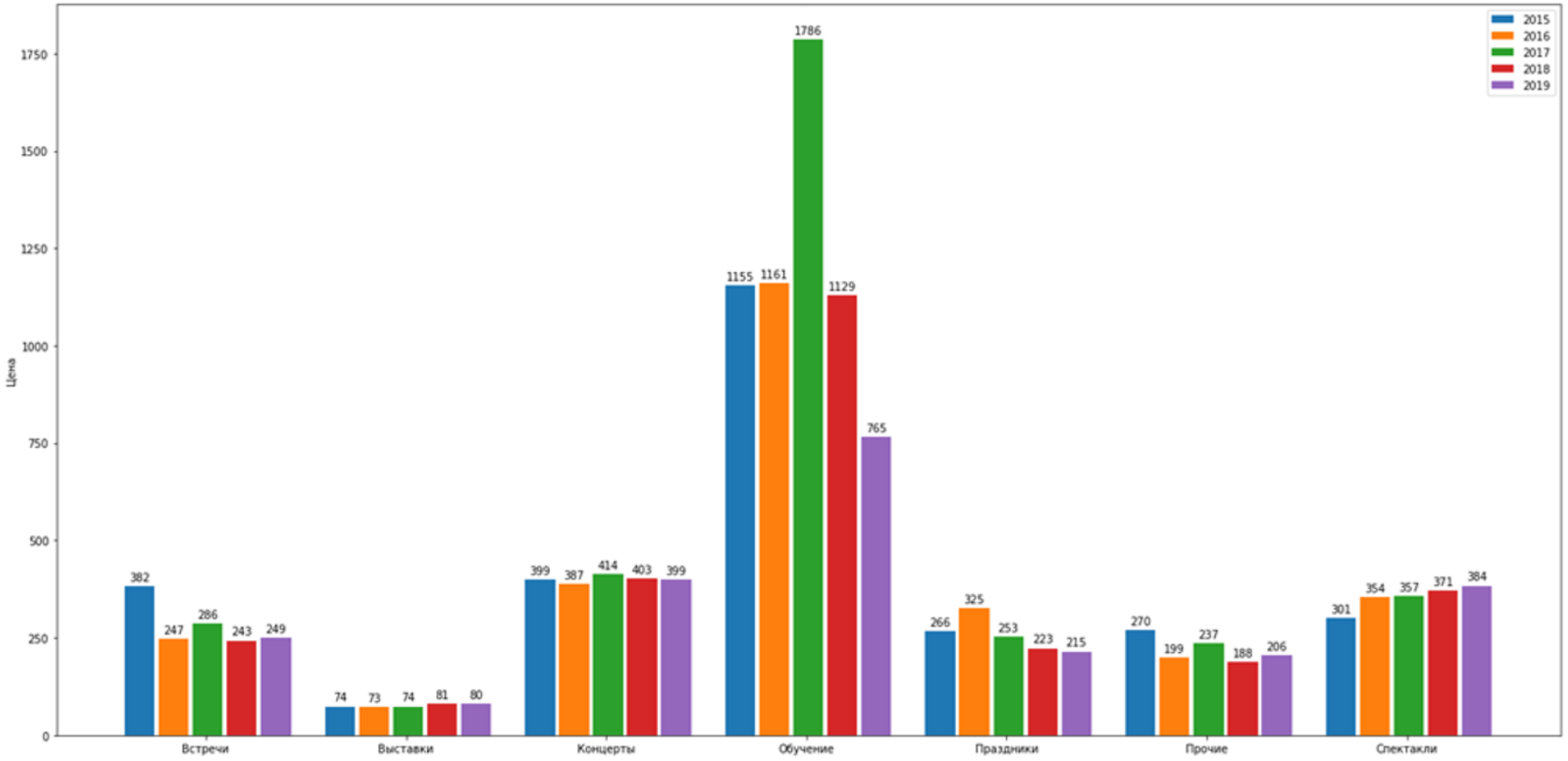
Тепловые карты мероприятий за 2015-2019 года соответственно:



Количество онлайн мероприятий



Стоимость по категориям в рублях



Словесное описание исходной модели на 2020 год

Наиболее загруженные месяца: ноябрь и апрель

Наименее загруженные месяца: январь, июль

Самые загруженные числа: 20 числа февраля и начало марта

Наиболее развивающиеся категории: встречи, обучение

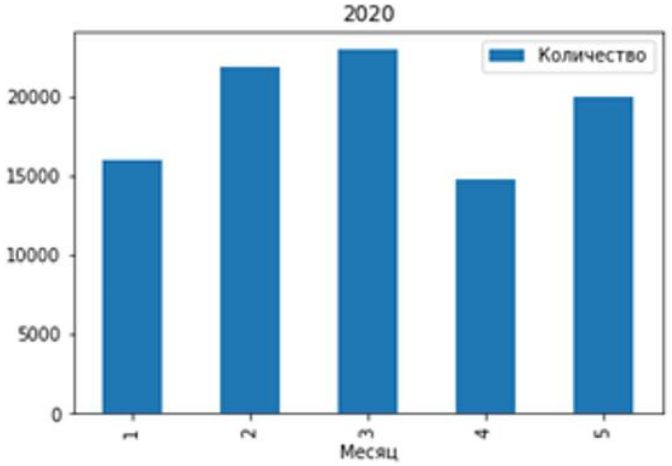
Снижение развития категорий: спектакли, концерты, выставки, праздники

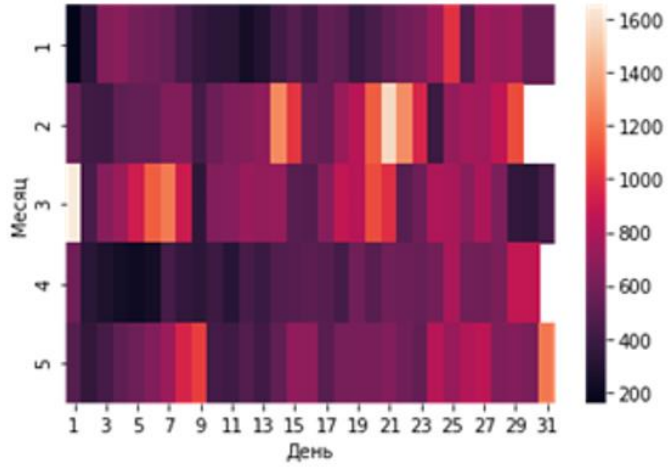
Повышение в цене: спектакли

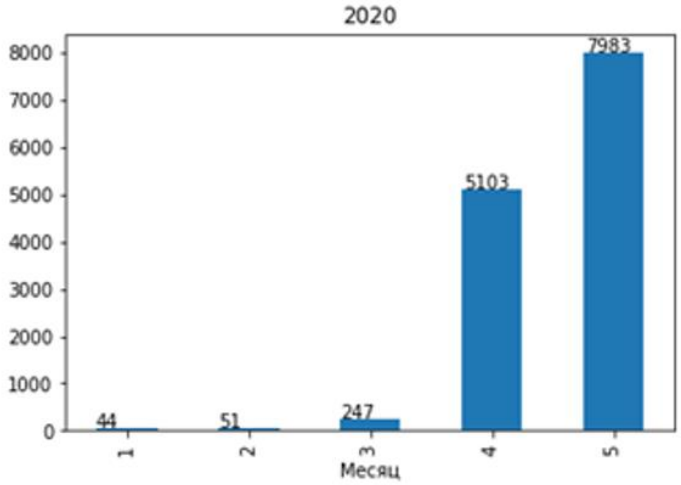
Постепенное понижение в цене: обучение, встречи, праздники, прочие

Ожидается резкое увеличение количества онлайн мероприятий в апреле

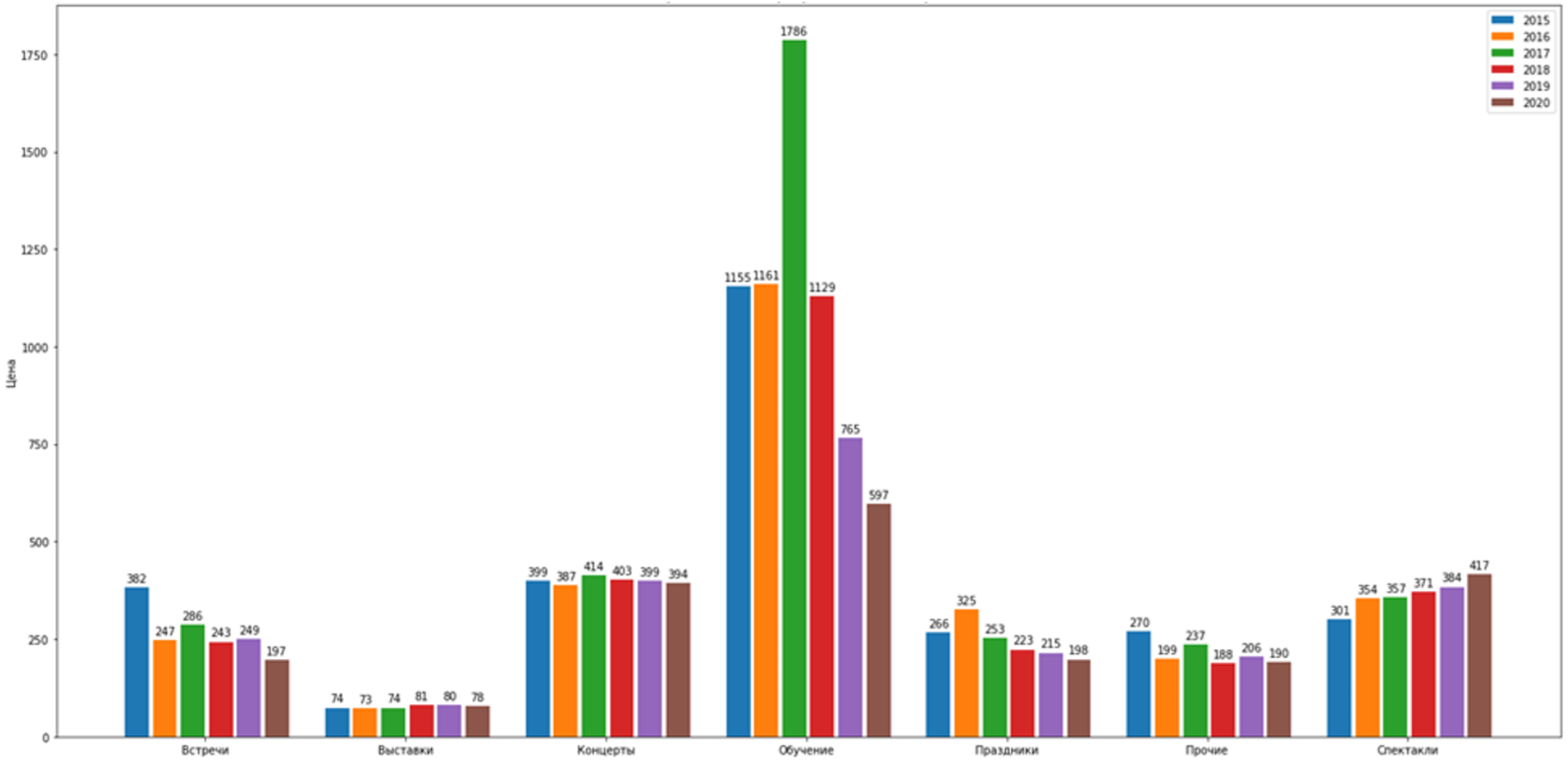
Проверка на данных 2020 года







Средние арифметические цен за 2015-2020 года



Словесное описание итоговой модели

Наиболее загруженные месяца: ноябрь и апрель

Наименее загруженные месяца: январь, июль

Самые загруженные числа: 3 и 4 ноября (День народного единства); 20 числа апреля (День книги 23 апреля) и 1 июня (Всемирный день ребенка)

Наиболее развивающиеся категории: встречи, обучение

Наблюдается снижение развития категорий: спектакли, концерты, выставки, праздники

Постепенное повышение в цене: спектакли

Постепенное понижение в цене: обучение, встречи, праздники, прочие

Рекомендации по дальнейшему развитию

Ситуация с коронавирусной инфекцией показала, что стоит быть готовыми к переносу мероприятий в режим онлайн и стоит развивать данную сферу.

Стоит продолжать развивать «Встречи», так как такие виды мероприятий помогают привлечь больший интерес, при этом охватывая практически все сферы культуры. Более того, данные лекции легко перенести в онлайн пространство.

Следует обратить больше внимания на категории спектаклей, концертов, выставок и мероприятий, связанных с праздниками, так как их количество в процентном соотношении уменьшается по сравнению с другими категориями.

Спасибо за внимание!