Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Орловский государственный университет имени И.С. Тургенева»
Физико-математический факультет
Кафедра информатики

Гулый Михаил Владимирович

Исследование организации работы служб экстренной помощи посредством анализа данных

Курсовая работа по дисциплине
«Проектная деятельность в программировании и научных вычислениях»

Направление подготовки:
01.03.02 Прикладная математика и информатика
Направленность (профиль): Системное программирование
и компьютерные вычисления
Квалификация: бакалавр

Руководитель:

к.ф.-м.н., доц. Дорофеева В.И._____

Оценка _____

Орёл – 2021

Содержание

Введение 2
Глава 1. Постановка задачи. 8
1.1. Постановка задачи для анализа набора данных 8
Глава 2. Анализ данных 10
2.1 Описание используемых инструментов для анализа данных 10
2.2 Подготовительный анализ и исследование набора данных 12
2.2.1 Обзор признаков набора данных и их значение 12
2.2.2 Проверка пропущенных значений 13
2.2.3 Создание дополнительных признаков 14
2.2.4 Анализ вызовов экстренной помощи по типу 14
2.2.5 Определение наиболее характерных причин экстренных вызовов 15
2.2.6 Анализ населённых пунктов, из индексов которых поступало наибольшее количество вызовов 15
2.2.7 Зависимость количества вызовов от дня недели 16
2.2.8 Корреляция между средним количеством вызовов в праздники и в непраздничные дни 18
2.2.9 Зависимость количества вызовов от времени года 20
2.3 Обобщение рекомендаций на основе выполненного анализа 22
Заключение 24
Список использованных источников 25

Введение

Обоснование выбора темы и ее актуальность

Количество вызовов в службу экстренной помощи значительно увеличилось в последнее время. В таких обстоятельствах очень важно быстро ответить на любой звонок службы экстренной помощи. Время имеет существенное значение, и эффективное реагирование зависит от многих факторов, включая, помимо прочего, наличие полицейских сил и машин экстренной помощи. Помня об этом, становится критически важным быстро и правильно обрабатывать любые звонки службы экстренной помощи и быть готовыми к любым ситуациям. Этого можно достичь, выясняя закономерности между временем и днем недели, а также типом звонков. Кроме того, очень эффективным может быть прогнозирование местоположения, в которое может поступать больше всего звонков, с использованием прошлых данных. Поскольку точность алгоритма и скорость анализа являются важными факторами, важно использовать правильную технику обработки, чтобы предоставить нашим силам правильную информацию в ожидании любой ситуации, которая может потребовать немедленного внимания. Для достижения этой цели используется анализ данных.

Рост объемов информации привёл к потребности анализировать эту информацию и выявлять различные закономерности. Так появилась наука, которую сегодня называют наукой о данных (от английского Data Science).

Анализ данных — это необходимая часть любой индустрии сегодня, учитывая огромные массивы данных, которые производятся на сегодняшний день. Популярность науки о данных значительно возросла в последние годы так как компании начали использовать инструменты анализа данных для увеличения доходности своих бизнесов. Использование анализа данных, также должно улучшить эффективность экстренных служб.

Анализ данных является очень важным в этой сфере, поскольку очень важно понимать, с какими проблемами сталкиваются службы и какие решения

стоит предпринимать. Ведь данные сами по себе это всего лишь факты и цифры. Анализ данных группирует, интерпретирует, структурирует и презентует данные в виде полезной информации.

Наука о данных — это область исследования, которая имеет дело с огромными объемами данных с использованием современных инструментов и методов для поиска невидимых на первый взгляд закономерностей, получения значимой информации и принятия деловых решений. Зачастую, в анализе данных используются сложнейшие алгоритмы машинного обучения для построения моделей, используемых для предсказаний.

Данные, используемые в анализе данных, могут быть из разных источников и представлены в разных форматах.

Наука о данных позволяет нам принимать лучшие решения, предсказывать поведение данных в будущем, а также исследовать различные закономерности в данных.

Степень разработанности проблемы

Способность измерять готовность к чрезвычайным ситуациям - прогнозировать вероятную работу систем реагирования на чрезвычайные ситуации в будущих событиях - имеет решающее значение для анализа политики в области внутренней безопасности. Тем не менее, по-прежнему сложно понять, насколько подготовлена система реагирования к крупномасштабным инцидентам, будь то стихийное бедствие, террористическая атака, промышленная или транспортная авария.

Анализ данных уже неоднократно применялся в самых различных сферах. Например, результаты, полученные в ходе анализа данных, способствовали их учету в области создания новых методов в борьбе с лесными пожарами, что значительно уменьшило количество возгораний. Также, анализ данных широко используется в здравоохранении, помогая более точно ставить диагнозы. Ведение статистики является обязательной составляющей любой подобной сферы, ведь именно благодаря статистике и возможен анализ.

Сравнение статистики и анализ данных появились до зарождения письменной истории, но необходимо было пройти несколько важных этапов для превращения аналитики в процесс, каким мы знаем его сегодня.

В 1785 г. Уильям Плейфэр (William Playfair) предложил гистограмму, которая сейчас является одним из основных (и широко используемых) способов визуализации данных. По легенде он изобрел гистограммы, чтобы показывать несколько десятков точек данных.

В 1812 г. картограф Шарль Жозеф Минар (Charles Joseph Minard) изобразил на графике потери армии Наполеона во время похода на Москву. Опираясь на польско-российскую границу, он создал линейную карту из толстых и тонких линий, которая показывала, как потери связаны с суровой зимой и с тем, сколько времени армия была отрезана от путей снабжения.

В 1890 г. инженер Герман Холлерит (Herman Hollerith) изобрел «табулирующую машину», которая записывала данные на перфокартах. Это дало возможность анализировать данные быстрее, что сократило процесс подсчета в Бюро переписи населения США с нескольких лет до 18 месяцев. Тогда и появилась бизнес-потребность постоянно улучшать сбор и анализ данных, что актуально и по сей день.

В 1970-е и 1980-е годы появилось ПО реляционной базы данных и язык программирования SQL, что дало возможность экстраполировать данные для анализа по необходимости.

В конце 1980-х гг. Уильям Г. Инмон (William H. Inmon) предложил концепцию “хранилища данных”, где можно было получать доступ к информации быстро и неоднократно. Кроме того, Говард Дреснер (Howard Dresner), аналитик компании Gartner, ввел термин «бизнес-аналитика» (business intelligence, BI), что подтолкнуло отрасли к анализу данных с целью лучшего понимания бизнес-процессов.

В 1990-х концепция глубинного анализа данных (data mining) дала возможность компаниям анализировать и выявлять закономерности в огромных наборах данных. Аналитики и специалисты по обработке данных

создали множество языков программирования, таких как R и Python, для разработки алгоритмов машинного обучения, работы с большими наборами данных и создания сложных визуализаций данных.

В 2000-х годах инновации в веб-поиске обеспечили разработку MapReduce, Apache Hadoop и Apache Cassandra для помощи в обнаружении, подготовке и представлении информации.

По мере того, как бизнес продвигался от простой доступности данных к их глубокому анализу, развивались средства анализа и их возможности.

Первые наборы аналитических инструментов были основаны на семантических моделях, взятых из ПО для бизнес-аналитики. Они помогали обеспечить эффективное управление, анализ данных и согласованность между инструментами. Одним из недостатков была недоступность своевременных отчетов. Принимающие бизнес-решения руководители не всегда были уверены в том, что результаты соответствовали их исходному запросу. С технической точки зрения эти модели используются в основном локально, что делает их неэффективными по затратам. Кроме того, данные часто оказываются изолированными в разрозненных хранилищах.

В дальнейшем благодаря эволюции в инструментах самообслуживания аналитика данных стала доступной для более широкой аудитории. Эти инструменты способствовали распространению аналитики данных, так как не требовали для работы специальных навыков. Настольная бизнес-аналитика завоевала популярность последние несколько лет, особенно при работе в облаке. Бизнес-пользователи с энтузиазмом исследуют самые разные информационные активы. Легкость использования привлекает, а объединение данных из разных источников и создание «единственной версии достоверных данных» становятся все более сложными. Настольная аналитика данных не всегда может масштабироваться для использования в крупных группах. Существует также риск несогласованных определений.

В последнее время аналитические инструменты обеспечивают более широкое преобразование бизнес-выводов благодаря автоматическому

обновлению и автоматизации процессов обнаружения, очистки и публикации данных. Бизнес-пользователи могут работать на любом устройстве с контекстом, получать информацию в реальном времени и достигать результатов.

Сегодня большая часть работы по-прежнему выполняется людьми, но автоматизация набирает все большее распространение. Данные из существующих источников можно легко объединять. Потребитель выполняет запросы, затем анализирует результаты, взаимодействуя с визуальными представлениями данных, и создает модели для прогнозирования будущих тенденций или выводов. Все это происходит под управлением и контролем людей на глубоком гранулярном уровне. Включение сбора данных, обнаружения данных и машинного обучения обеспечивает конечному пользователю больше вариантов и происходит быстрее, чем раньше [1].

Предмет исследования

Исследование и выявление закономерностей в наборе данных

Объект исследования

Набор данных о вызовах в службу экстренной помощи 911 из округа Монтгомери, штат Пенсильвания. Исследуемый набор данных включает в себя данные в период с 10 декабря 2015 года по 29 июля 2020 года [2].

Цель работы

Проанализировать набор данных по вызовам экстренных служб и дать рекомендации по улучшению ее работы

Основные задачи исследования

1. Исследование задачи на актуальность, анализ источников, соответствующих тематике работы.
2. Обоснование выбранных инструментов для анализа данных.
3. Предобработка данных и описание признаков набора данных.
4. Проведение анализа данных и получение выводов и рекомендаций, исходя из анализа.

Структура работы

Работа состоит из введения, двух глав, заключения, списка источников и приложений.

Во введении рассматривается актуальность работы, ставится цель и обозначаются задачи, необходимые для достижения поставленной цели.

Первая глава представляет собой постановку задачи.

Во второй главе содержится обоснование выбора инструментов для анализа данных, предварительная обработка данных и непосредственно сам анализ данных.

В заключении делаются выводы о проделанной работе.

Приводится список использованных источников.

Глава 1. Постановка задачи.

1.1. Постановка задачи для анализа набора данных

Службы экстренной помощи имеют, несомненно, огромное значение в современном мире. От действий этих служб нередко зависят жизни огромного количества людей. Поэтому очень важно обеспечить соответствующую организацию этих служб. Правильная организация служб экстренной помощи может существенно улучшить уровень жизни людей. К сожалению, достичь этой цели непросто. Нужно вести статистику и учитывать каждый вызов, который был произведён, а также дополнительные данные об этом вызове: где был произведён вызов, по какой причине, в какое время и так далее. Впоследствии, благодаря этим данным, у нас появляется возможность анализа этих данных. Могут быть обнаружены невиданные прежде проблемы, такие как, например наиболее проблемные зоны, которые нуждаются в большей поддержке. Именно благодаря корректному анализу возможно принятие действий, нацеленных на улучшение качества работы служб.

Ежедневно, миллионы людей совершают звонки в службу экстренной помощи 911, в связи с чем ставится задача оптимизации работы соответствующей экстренной службы. Перед нами ставится цель – проанализировать данные и дать рекомендации по организации службы экстренной помощи. Стоит выяснить, наиболее частую причину, по которой звонят люди, места из которых совершается наибольшее количество вызовов, факторы, которые влияют на количество вызовов, и так далее.

Выяснив всё это, мы сможем дать некоторые рекомендации органам, отвечающим за организацию службы экстренной помощи. Это может существенно улучшить эффективность этой службы, улучшив качество жизни людей проживающих в соответствующей области.

Данная работа должна отвечать следующим требованиям:

- наглядность отображения графиков;

- обработка пропущенных значений;
- определение наиболее характерных причин экстренных вызовов;
- выявление проблем и предложения путей по их решению.

Глава 2. Анализ данных

2.1 Описание используемых инструментов для анализа данных

При попытке ответить на вопросы об оптимизации работы служб мы столкнёмся с необходимостью правильно анализировать наш огромный набор данных. Для этого нам нужно выбрать подходящие инструменты, в противном случае процесс анализа данных будет затрачивать много времени и усилий.

Большую часть анализа данных обычно производят при помощи таких языков как R и Python. Это два основных языка которые являются незаменимыми инструментами для анализа данных. В данной работе мы ограничимся использованием языка Python и подходящих библиотек для нашей работы.

В чём заключаются главные преимущества языка Python над другими языками программирования? Почему был выбран именно этот язык, а не другой?

- Данный язык является высокоуровневым языком программирования вследствие чего имеет понятный легко читаемый и в тоже время минималистичный синтаксис, понять который куда легче, чем синтаксисы таких языков программирования как: C++, Java, Pascal.
- Язык Python может похвастаться богатым выбором библиотек, например существуют библиотеки для удобного и простого построения графиков.

Теперь упомянем модули (библиотеки) которые будут использованы.

- **Pandas** – это основной и самый главный модуль который будет использован в нашем анализе данных. Этот модуль предоставляет удобные структуры данных для манипулирования и обработки данных. Основной структурой является Dataframe который и

будет использован далее. Dataframe представляет собой объект манипулирования и индексирования массивов двумерных данных.

- Модуль **Matplotlib** используется для визуализации данных. С помощью этого модуля можно удобно строить графики и диаграммы, благодаря которым мы сможем проанализировать наш набор данных.
- Будет использоваться модуль **datetime** для удобной работы с датой и временем.
- Модуль **Seaborn** основанный на вышеупомянутом модуле Matplotlib расширит наши возможности визуализации данных. Этот модуль представляет высокоуровневый интерфейс для создания информативных графиков и различных диаграмм.

2.2 Подготовительный анализ и исследование набора данных

Импортируем нужные нам модули и загрузим наш набор данных

```
import datetime as dt
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
df = pd.read_csv('911.csv')
```

Рис. 1 – импорт модулей

2.2.1 Обзор признаков набора данных и их значение

- **lat** (Latitude): показывает на какой широте был сделан вызов. Пример: 40.2978759
- **lng** (longitude): показывает на какой долготе был сделан вызов. Пример: -75.5812935
- **desc** (description): описание экстренного вызова. Пример: REINDEER CT & DEAD END; NEW HANOVER; Station 332;
- **zip**: почтовый индекс. Каждый индекс содержит в себе пять цифр. Пример: 19525
- **title**: название чрезвычайной ситуации. До символа двоеточия обозначена общая причина вызова, после двоеточия обозначена конкретная причина вызова. Пример: EMS: BACK PAINS/INJURY
- **timeStamp**: дата и время звонка в формате: ГГГГ-ММ-ДД ЧЧ: ММ: СС. Пример: 2015-12-10 17:10:52
- **twp** (Township): область. Например: NEW HANOVER
- **addr** (Address): адрес. Например: REINDEER CT & DEAD END

Как мы можем видеть, признак desc уже включает в себя признаки twp и addr

2.2.2 Проверка пропущенных значений

Для того чтобы корректно анализировать данные, обязательно нужно разобраться с пропущенными значениями

```
print('Всего отсутствующих значений:',df.isnull().values.sum())
print(df.isnull().sum())
```

```
Всего отсутствующих значений: 80492
lat              0
lng              0
desc             0
zip          80199
title            0
timeStamp        0
twp            293
addr             0
dtype: int64
```

Рис. 2 – проверка пропущенных значений

Данный набор данных содержит 80199 пропусков признака zip и 293 пропуска признака twp. Всего в нашем наборе данных содержится 663522 значения. Таким образом, количество строк с пропущенными значениями составляет приблизительно 12% от всего набора данных. Наилучшим решением будет удалить строки с пропущенными данными, поскольку заменить пропущенные значения на что-либо мы не можем.

```
df.dropna(inplace=True)
print('Всего отсутствующих значений:',df.isnull().values.sum())
print(df.isnull().sum())
```

```
Всего отсутствующих значений: 0
lat          0
lng          0
desc         0
zip          0
title        0
timeStamp    0
twp          0
addr         0
dtype: int64
```

Рис. 3 – повторная проверка пропущенных значений

2.2.3 Создание дополнительных признаков

Добавим столбец **reason** обозначающий общую причину вызова.

Всего существует три причины вызова:

1. **EMS** - Вызов скорой помощи
2. **Traffic** - Дорожное происшествие
3. **Fire** - Пожар

Добавим столбец **title_code** обозначающий конкретную причину вызова. Пример значения из этого столбца: BACK PAINS/INJURY

```
df['reason'] = df['title'].apply(lambda title: title.split(':')[0])
df['title_code'] = df['title'].apply(lambda title: title.split(':')[1])
```

Рис. 4 – добавление столбцов

2.2.4 Анализ вызовов экстренной помощи по типу

Посмотрим, чего совершается больше, вызовов скорой помощи, вызовов, связанных с дорожными происшествиями или вызовов, связанных с пожарами.

```
fig, axes = plt.subplots(1,2, figsize=(15, 5))

sns.countplot(x='reason', data=df, order=df['reason'].value_counts().index, ax=axes[0])
axes[0].set_title('Самые частые причины вызовов 911', size=14)
axes[0].set(xlabel='Причина', ylabel='Число вызовов')

df['reason'].value_counts().plot.pie(autopct='%1.2f%%',ax=axes[1],shadow=True)
axes[1].set(xlabel='', ylabel='')

sns.despine(bottom=False, left=True)
```

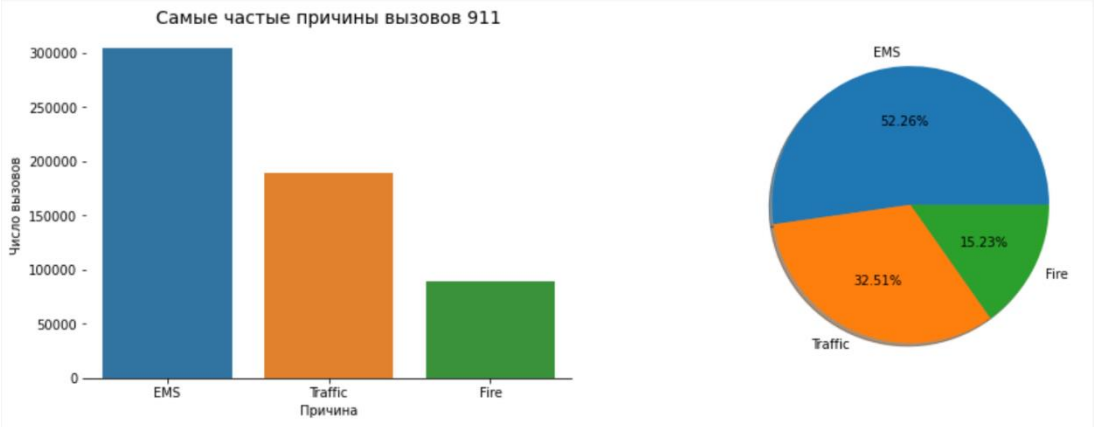

Рис. 5 – Наиболее совершаемые типы вызовов

Как мы можем видеть, количество вызовов скорой помощи преобладает над количеством остальных вызовов.

2.2.5 Определение наиболее характерных причин экстренных вызовов

Узнаем, какие конкретно вызовы совершаются больше всего.

```
fig, axes = plt.subplots(figsize=(10, 5))
sns.countplot(y='title', data=df, order=df['title'].value_counts().index, palette='prism')
sns.despine(bottom=False, left=True)
axes.set_ylim([9, 0])
axes.set_title('Общие причины вызовов 911', size=17)
axes.set_xlabel('Число вызовов 911',fontsize=15)
axes.set_ylabel('Причины',fontsize=15)
plt.tight_layout()
```

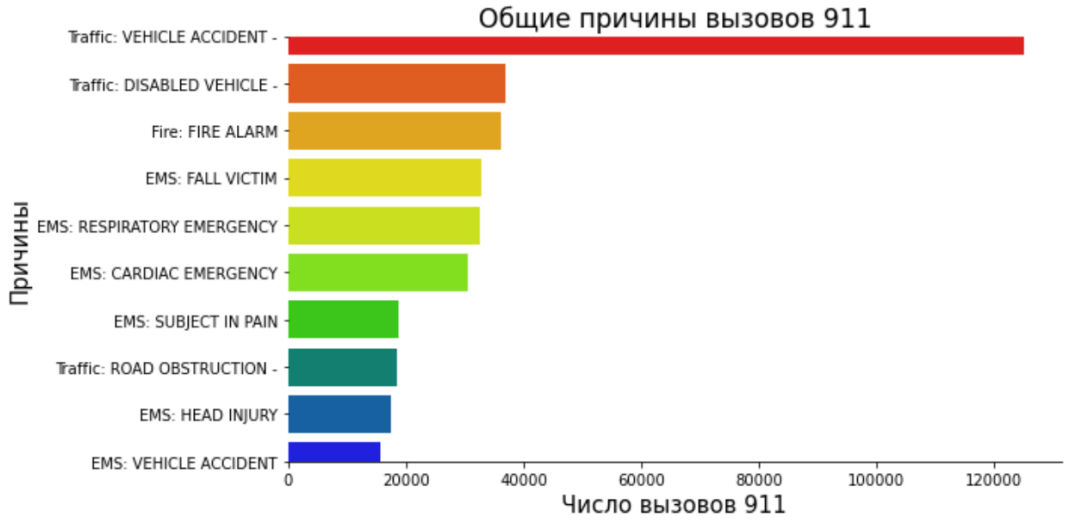

Рис. 6 – Наиболее совершаемые вызовы

Автомобильные аварии являются бесспорно наиболее частой причиной вызовов. Администрации районов стоит всерьёз обдумать эту проблему. Возможно, требуется больше контроля, например увеличение количества дорожного патруля или введение больших ограничений скорости.

2.2.6 Анализ населённых пунктов, из индексов которых поступало наибольшее количество вызовов

Данный код используется для создания таблички из пяти строк, которая содержит индексы и количества вызовов для каждого индекса. Создадим три

таблички, первую для вызовов EMS, вторую для вызовов Traffic и третью для вызовов Fire.

```
df_zip = pd.DataFrame(df.loc[df['reason']=='EMS','zip'].value_counts().head(5))
df_zip.rename(columns = {'zip':'Кол-во вызовов EMS'}, inplace = True)
df_zip.index.name = 'Индекс'
df_zip.index = df_zip.index.map(int)
df_zip.style.background_gradient(cmap='Greens')
```

Рис. 7 – Код, использующийся для создания таблички

Для двух остальных табличек код будет аналогичным.

Индекс	Кол-во вызовов EMS
19401	29594
19464	25986
19403	22408
19446	18320
19406	11773

Индекс	Кол-во вызовов Traffic
19464	10442
19401	9896
19446	9853
19403	8886
19002	7877

Индекс	Кол-во вызовов Fire
19464	7474
19401	6106
19446	4095
19406	3607
19403	3594

Рис. 8 – Индексы, из которых поступает наибольшее количество вызовов

В населённых пунктах с индексами 19464, 19401 и в остальных показанных на табличках совершается огромное количество вызовов. Администрациям населённых пунктов нужно обратить особое внимание на эти районы. Возможно, данная ситуация является нормальной если, например в пунктах с этими индексами живет больше людей чем в остальных и тогда большее количество вызовов закономерно. Но тем не менее, пункты с этими индексами заслуживают особого внимания и если необходимо, то особых мер, как например повышение финансирования.

2.2.7 Зависимость количества вызовов от дня недели

Данный код создаёт тепловую карту, на вертикальной оси расположены дни недели, а на горизонтальной оси расположены часы.

```
df['timeStamp'] = pd.to_datetime(df['timeStamp'])

df['Hour'] = df['timeStamp'].apply(lambda time: time.hour)
df['Month'] = df['timeStamp'].apply(lambda time: time.month)
df['Day of Week'] = df['timeStamp'].apply(lambda time: time.dayofweek)

dayHour = df.groupby(by=['Day of Week', 'Hour']).count()['reason'].unstack()
plt.figure(figsize=(12,6))
y_axis_labels = ['Пн','Вт','Ср','Чт','Пт','Сб', 'Вс']

sns.heatmap(dayHour, cmap='coolwarm', linewidths=0.05, yticklabels=y_axis_labels)
plt.xlabel('Час',fontsize=15)
plt.ylabel('День недели',fontsize=15)
plt.title('Зависимость вызовов 911 от времени и дня недели',fontsize=17)
plt.show()
```

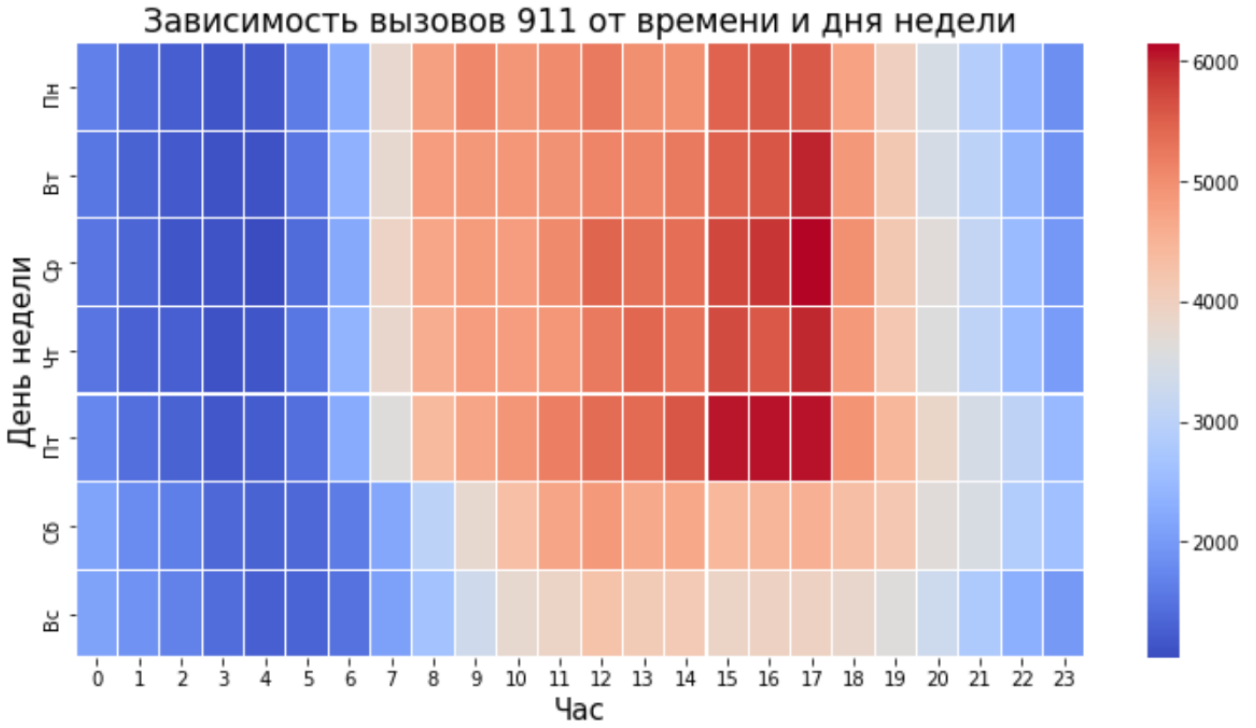

Рис. 9 – Зависимость количества вызовов от времени и дня недели

Как мы можем видеть, наибольшее количество вызовов совершается в будние дни, когда большинство людей работает. Особенно много вызовов совершается по пятницам. В выходные дни большая часть людей предпочитает вести спокойный образ жизни отдыхая дома с семьей. Так же заметно, что в часы, когда большинство людей спит, совершается меньше всего вызовов. Из этого следует, что администрации стоит позаботиться об

этом, например держать наибольшее количество сотрудников службы спасения в часы, когда происходит наибольшее количество вызовов.

2.2.8 Корреляция между средним количеством вызовов в праздники и в непраздничные дни

Для этого создадим дополнительный столбец Holiday, в который будем помещать 1 если в соответствующий день есть праздник и 0 если праздник отсутствует

```
df['timeStamp'].iloc[[0, -1]]
df['Year'] = df['timeStamp'].apply(lambda time: time.year)
df['Day'] = df['timeStamp'].apply(lambda time: time.day)
df['Holiday'] = 0
df['Holiday_name'] = 'NULL'
df.loc[(df['Day'] == 25) & (df['Month'] == 12), 'Holiday'] = 1 # Christmas Day
df.loc[(df['Day'] == 25) & (df['Month'] == 12), 'Holiday_name'] = 'Christmas Day'
```

Рис. 10 – Создание столбцов, обозначающих наличие праздника

Также создадим еще один столбец под названием Holiday_name в котором будем хранить название праздника.

```
df_holiday = df.groupby(df['timeStamp'].apply(lambda x:
    pd.Timestamp(x).strftime('%Y-%m-%d'))).agg({'Holiday':['mean', 'count']})
df_holiday.columns = ['mean', 'count']
df_holiday.columns
ax = sns.barplot(x = df_holiday['mean'], y = df_holiday['count'])
ax.set(xlabel='Наличие праздника', ylabel='Среднее количество вызовов')
plt.show()
```

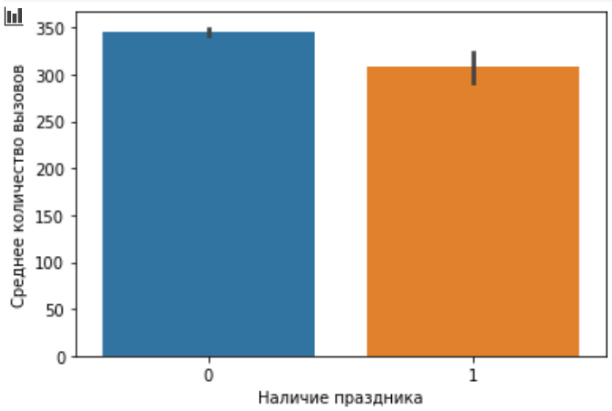

Рис. 11 – Среднее количество совершаемых вызовов в праздники и в непраздничные дни

Как мы видим, в целом, в праздничные дни совершается меньше вызовов. Это опять же таки связанно с тем, что в большинство праздников люди ведут более спокойный образ жизни, например, проводят время дома с семьей.

Посмотрим на данные для каждого праздника.

```
df_holiday_name = df.groupby('Holiday_name')
df_holiday_name = df_holiday_name['Date'].value_counts()
df_holiday_name = df_holiday_name.mean(level=0)
fig, ax = plt.subplots(figsize=(13, 7))
ax = sns.barplot(x = df_holiday_name.index, y = df_holiday_name.values)
ax.set_title('Среднее количество вызовов в каждый праздник',size=17)
ax.set_xlabel('Название праздника', fontsize = 15)
ax.set_ylabel('Количество вызовов', fontsize = 15)
ax.set_xticklabels(labels=df_holiday_name.index, rotation=30)
plt.show()
```

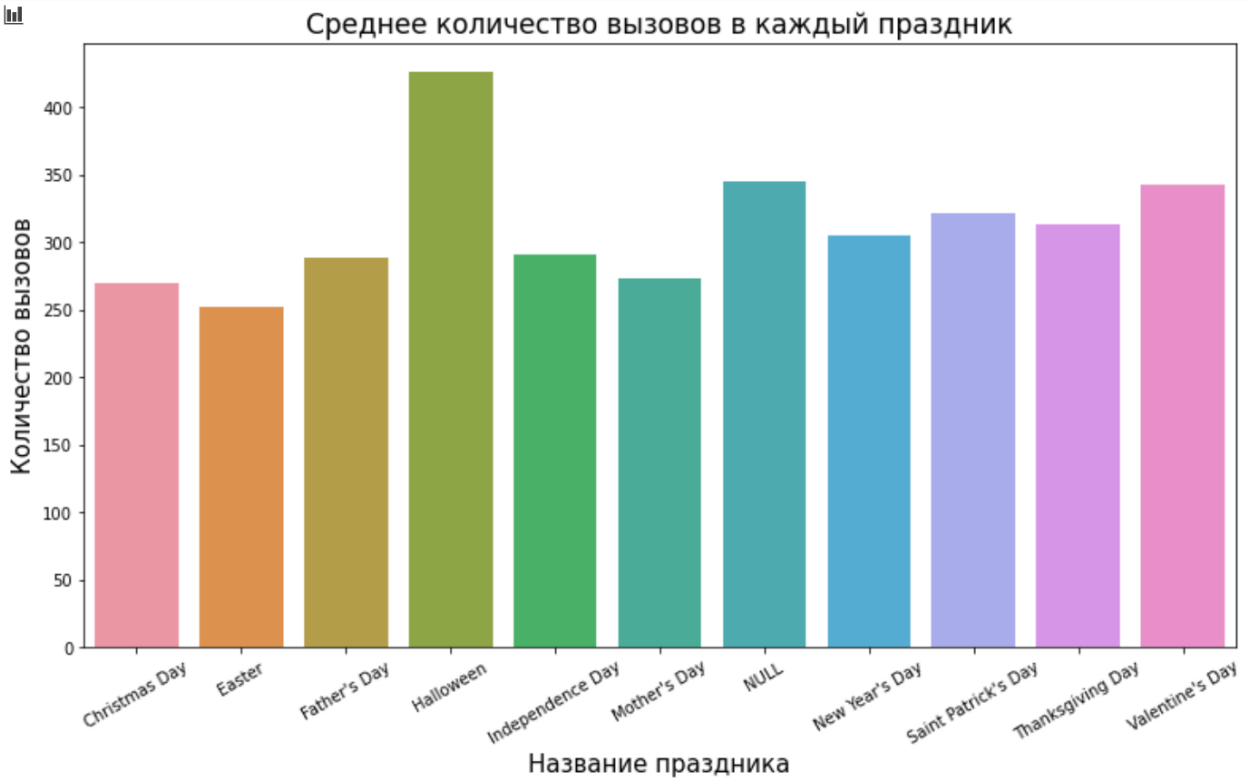

Рис. 12 – Совершаемое количество вызовов в каждый праздник

На этой столбчатой диаграмме видно, что только в Хэллоуин совершается гораздо больше вызовов чем в непраздничные дни. Минимальное количество вызовов совершается на Пасху – день, когда люди ведут себя наиболее спокойно и вероятно никуда не спешат. Службе спасения в Хэллоуин стоит

держать большой штаб сотрудников готовых работать, а вот на Пасху можно наоборот немного расслабиться.

Проанализируем причины вызовов в Хэллоуин.

```
fig, axes = plt.subplots(1,2, figsize=(15, 5))
sns.countplot(x=df[df['Holiday_name'] == 'Halloween']['reason'], data=df,
    order=df['reason'].value_counts().index, ax=axes[0], palette= "rocket_r")
axes[0].set_title('Самые частые причины звонков 911 в Хэллоуин', size=14)
axes[0].set(xlabel='Причина', ylabel='Число вызовов')
colors = ['#e28f71','#b43058','#57274e']
df[df['Holiday_name'] == 'Halloween']['reason'].value_counts().plot.pie(autopct='%1.2f%%',
    ax=axes[1],shadow=True, colors=colors)
axes[1].set(xlabel='', ylabel='')
sns.despine(bottom=False, left=True)
```

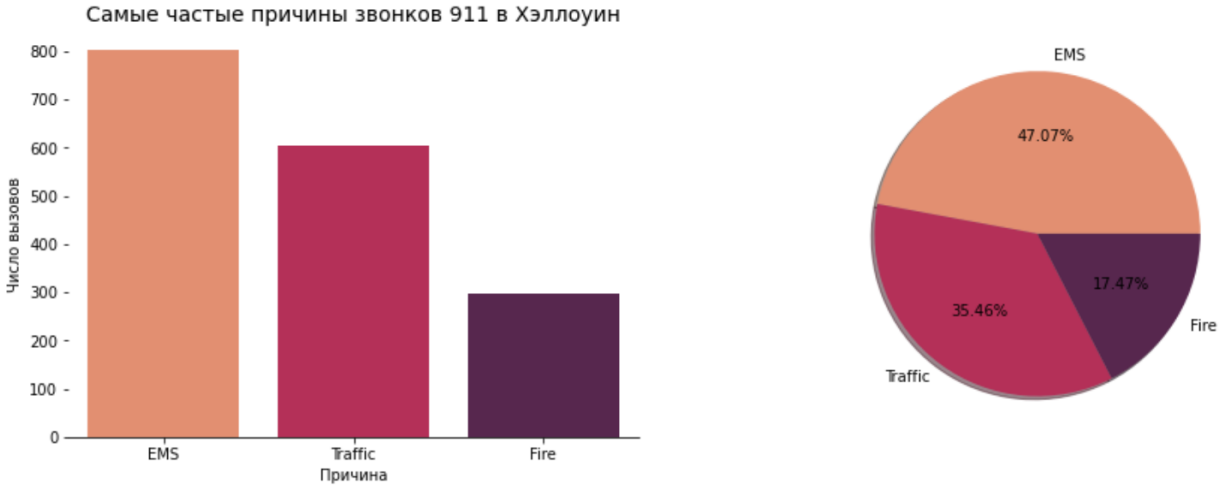

Рис. 13 – Типы вызовов в Хэллоуин

В Хэллоуин совершается больше происшествий связанных с пожарами (17.47% против 15.23% в непраздничные дни) и дорожных происшествий (35.46% против 32.51% в непраздничные дни). Соответственно стоит увеличить количество дежурных пожарных и дорожного патруля в этот день.

2.2.9 Зависимость количества вызовов от времени года

Создадим функцию, возвращающую название времени года и столбец, хранящий соответствующие название времени года

```
def define_season(x):
    if x in (12,1,2):
        return 'Winter'
    elif x in (3,4,5):
        return 'Spring'
    elif x in (6,7,8):
        return 'Summer'
    elif x in (9,10,11):
        return 'Fall'

df['Season'] = df['Month'].apply(define_season)
```

Рис. 14 – Создание функции для определения времени года

Теперь посмотрим на самую первую дату в нашем наборе данных и самую последнюю

```
   df['timeStamp'].iloc[[0, -1]]


0        2015-12-10 17:10:52
663521   2020-07-29 15:52:46
Name: timeStamp, dtype: datetime64[ns]
```

Рис. 15 – Диапазон дат

Можем видеть, что имеется не целое количество лет. Данные нужно ограничить по целым годам, в противном случае мы получим некорректные результаты.

```
   df = df[(df['Year'] > 2015) & (df['Year'] < 2020)]
   df['timeStamp'].iloc[[0, -1]]



7916     2016-01-01 00:10:08
591263   2019-12-31 23:50:12
Name: timeStamp, dtype: datetime64[ns]
```

Рис. 16 – Изменение набора данных с целью получения целых лет

Данные ограничены. Теперь можно приступить к анализу.

```
fig, axes = plt.subplots(1,2, figsize=(15, 5))
sns.countplot(x='Season', data=df, order=df['Season'].value_counts().index, ax=axes[0])
axes[0].set_title('Суммарное количество звонков для каждого времени года', size=14)
axes[0].set(xlabel='Время года', ylabel='Число вызовов')
df['Season'].value_counts().plot.pie(autopct='%1.2f%%',ax=axes[1],shadow=True)
axes[1].set(xlabel='', ylabel='')
sns.despine(bottom=False, left=True)
```

Рис. 17 – Зависимость количества вызовов от времени года

Как мы можем видеть, нет существенной корреляции между количеством вызовов и временем года. В каждое время года совершается примерно одинаковое количество вызовов. Службе спасения всегда стоит быть подготовленными к неожиданным вызовам.

2.3 Обобщение рекомендаций на основе выполненного анализа

Следует начать с того, что автомобильные аварии являются наиболее частой причиной вызовов служб экстренной помощи. Следовательно, необходимо увеличение дорожного патруля и осуществление более тщательного надзора над автомобилистами.

Требуется обратить особое внимание на населённые пункты с индексами 19464, 19401 и принять необходимые меры.

Нельзя оставить без внимания и тот факт, что в период с 15 до 17 часов дня по будням совершается наибольшее количество вызовов. Не помешает увеличение количества дежурных сотрудников в это время.

Также следует отметить то, что в Хэллоуин совершается огромное количество вызовов, гораздо большее чем в среднем. В этот день требуется большее количество дежурных пожарных и дорожного патруля.

Заключение

В работе был представлен анализ данных. Приведено обоснование выбора набора данных и инструментов для анализа этого набора. Описаны все признаки набора данных и произведена предобработка, в ходе которой были удалены строки с пропущенными значениями, а также созданы новые признаки для большего удобства. Проведен анализ данных и построены наглядные графики и диаграммы.

Достигнута цель – проведён тщательный анализ набора данных, а также получены выводы и даны соответствующие рекомендации.

Были выполнены задачи:

1. Исследованы задачи на актуальность, анализ источников, соответствующих тематике работы.

3. Обоснован выбор инструментов для анализа данных.

4. Проведены предобработка данных и описание признаков набора данных.

5. Проведен анализ данных и получены выводы исходя из анализа.

Весь программный код был реализован на высокоуровневом языке Python с применением библиотек для анализа данных и визуализации данных.

Данная работа имеет полезное практическое применение, так как содержит большое количество наглядной информации в виде графиков и диаграмм, а также выводов полученных исходя из анализа набора данных, следование которым могло бы увеличить эффективность службы экстренной помощи.

Список использованных источников

1. Статья: [Электронный ресурс]. — https://www.oracle.com/ru/business-analytics/what-is-analytics/#future — Дата доступа: 11.06.2021.
2. Набор данных: — https://www.kaggle.com/mchirico/montcoalert — Дата доступа: 11.06.2021.
3. Статья: [Электронный ресурс]. — https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85 — Дата доступа: 11.06.2021.
4. Статья: [Электронный ресурс]. — https://habr.com/ru/post/352812/ — Дата доступа: 11.06.2021.