Prediction on London Fire Brigade Data using Machine Learning

Dilruba Khanom Sonia, 4130379. Sukhjinder kaur Bajwa 4126047

1. Introduction

The London Fire Brigade (LFB) is responsible for protecting one of the largest and busiest cities in the world. The organization is tasked with responding to a variety of emergencies, ranging from fires and explosions to natural disasters and terrorist incidents. In order to effectively carry out its mission, the LFB collects and stores vast amounts of data related to its operations.

As data analysts, we will be responsible for following a proper data mining methodology and applying various techniques covered in lectures to analyze the data. Our aim will be to address the raised business concerns and problems related to the London Fire Brigade dataset. We will act as both a business client and a data analyst, and use our skills and knowledge to identify meaningful insights and patterns in the data. The insights and patterns identified through this analysis can then be used to improve the performance and efficiency of the London Fire Brigade.

The objective of this project is to use data mining techniques to identify meaningful patterns and insights that can be used to improve the operations of the LFB. To achieve this objective, we will follow a structured data mining methodology that involves several steps, including problem identification, data understanding, data preparation, modeling, evaluation, and report.

Through this project, we hope to gain valuable insights into the operations of the LFB and provide meaningful recommendations that can help improve the efficiency and effectiveness of the organization.

2. Understanding the Business Environment and Objectives

The London Fire Brigade (LFB) is one of the largest firefighting and rescue organizations in the world, providing services to the Greater London area. With a responsibility to ensure public safety and minimize the loss of life and property from fire and other incidents, LFB faces many challenges, including a high volume of emergency calls, complex rescue operations, and a constantly changing urban environment. To effectively manage its operations and make informed decisions, LFB collects and stores a large amount of data related to incidents, responses, and other relevant factors. However, analyzing and interpreting this data to extract meaningful insights and inform decision-making can be a daunting task. Therefore, in this project, we aim to act as both a business client and a data analyst to identify and address significant concerns and problems related to LFB's data and operations using advanced data mining techniques.

Objective

i. Identify the most common types of incidents attended by LFB and their frequency in each year. This analysis could help LFB to identify the areas where they need to focus more resources to reduce the frequency of these types of incidents.

ii. Analyze the response times of LFB and identify the factors that contribute to the delay in response. This analysis could help LFB to optimize their resources and reduce the response time, which could save lives and minimize damage to properties.

iii. Identify the locations with the highest frequency of incidents and analyze the reasons for this trend. This analysis could help LFB to take proactive measures to prevent such incidents in the future, for example, by increasing fire safety awareness or conducting safety inspections in these areas.

List DM Goals that correspond to Business Objectives
-Make sure to include 1 Descriptive modelling
2 Predictive modelling Tasks

3. Data Exploration

There are 12097 rows and 39 columns in the given dataset. 4 columns were dropped due to redundancy. Those columns are IncGeo_BoroughCode, IncGeo_BoroughName, ProperCase, FRS. All the rows of those 4 columns had the same value.

```
Number of rows:  12097
Number of columns:  35
```

The values of the discarded columns were,

IncGeo_BoroughCode	E09000013
IncGeo_BoroughName	HAMMERSMITH AND FULHAM
ProperCase	Hammersmith And fulham
FRS	London

In the dataset provided, there are several columns with null values. The 'SpecialServiceType' column has the highest number of null values, followed by 'Postcode_full', 'Easting_m', 'Northing_m', 'Latitude', 'Longitude', 'FirstPumpArriving_AttendanceTime', 'FirstPumpArriving_DeployedFromStation', 'SecondPumpArriving_AttendanceTime', 'SecondPumpArriving_DeployedFromStation', 'NumStationsWithPumpsAttending', 'NumPumpsAttending', 'PumpCount', 'PumpHoursRoundUp', 'Notional Cost (£)'. These null values may affect our analysis, especially if we are trying to use these columns for visualization or modeling purposes.

Column	Num of rows with Null values
SpecialServiceType	8210 (67.87%)
Postcode_full	7859 (64.97%)
Easting_m	7859 (64.97%)
Northing_m	7859 (64.97%)
Latitude	7859 (64.97%)
Longitude	7859 (64.97%)
FirstPumpArriving_AttendanceTime	702 (5.80%)

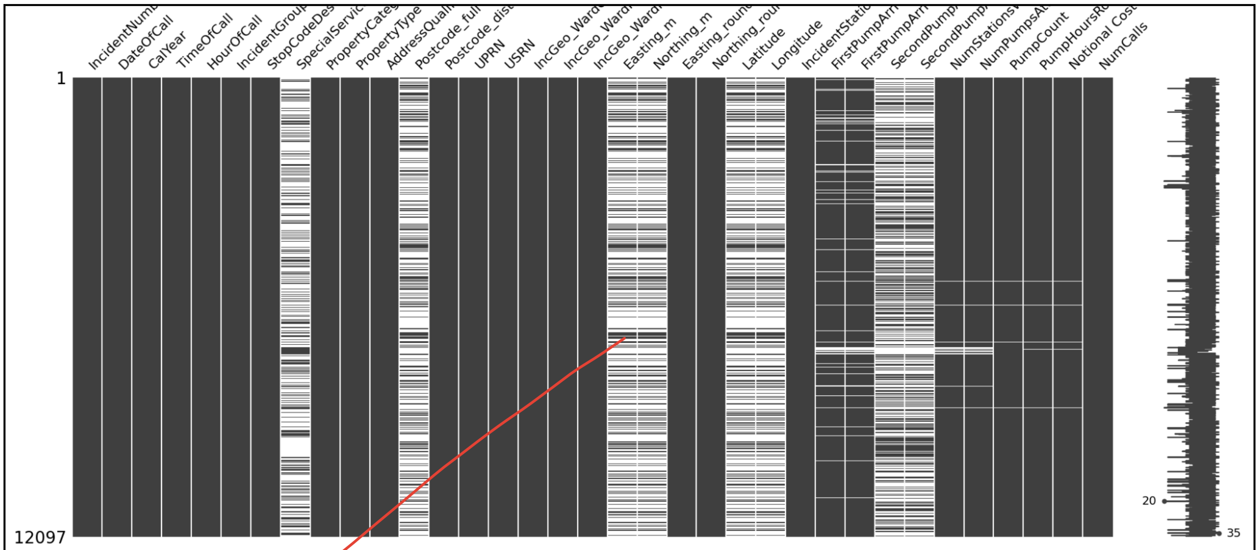
FirstPumpArriving_DeployedFromStation	703 (5.81%)
SecondPumpArriving_AttendanceTime	6626 (54.77%)
SecondPumpArriving_DeployedFromStation	6626 (54.77%)
NumStationsWithPumpsAttending	244 (2.01%)
NumPumpsAttending	244 (2.01%)
PumpCount	77 (0.64%)
PumpHoursRoundUp	83 (0.69%)
Notional Cost (£)	83 (0.69%)

The skewness and kurtosis of each column is as follows,

```
Column: CalYear
Skewness: -0.11071322491430158
Kurtosis: -1.3236670810576634
Column: HourOfCall
Skewness: -0.45501911481300533
Kurtosis: -0.6513907590495518
Column: UPRN
Skewness: 11.393207935730263
Kurtosis: 127.85422665391084
Column: USRN
Skewness: 4.9865456596077475
Kurtosis: 577.3674367279223
Column: Easting_m
Skewness: nan
Kurtosis: nan
Column: Northing_m
Skewness: nan
Kurtosis: nan
Column: Easting_rounded
Skewness: 0.1918570765086197
Kurtosis: -0.8452282247656697
Column: Northing_rounded
Skewness: 0.13841739084422053
Kurtosis: -0.6239522371110389
```

```
Column: Latitude
Skewness: nan
Kurtosis: nan
Column: Longitude
Skewness: nan
Kurtosis: nan
Column: FirstPumpArriving_AttendanceTime
Skewness: nan
Kurtosis: nan
Column: SecondPumpArriving_AttendanceTime
Skewness: nan
Kurtosis: nan
Column: NumStationsWithPumpsAttending
Skewness: nan
Kurtosis: nan
Column: NumPumpsAttending
Skewness: nan
Kurtosis: nan
Column: PumpCount
Skewness: nan
Kurtosis: nan
Column: PumpHoursRoundUp
Skewness: nan
Kurtosis: nan
Column: Notional Cost (£)
Skewness: nan
Kurtosis: nan
Column: NumCalls
Skewness: 18.541439644641812
Kurtosis: 485.8868426628939
```

In Fig 3.1, the missingno matrix plot displays the presence of missing values in the dataset. It visualizes the distribution of missing values across all variables, where white lines indicate the presence of missing values in a specific variable. This plot can help identify variables with high amounts of missing data, as well as patterns in the distribution of missing data that may inform subsequent data cleaning or imputation steps.



nice

Fig: 3.1

In Fig: 3.2, I have mapped out the locations of False Alarm incidents on a map using coordinates from the dataset. By doing this, I have identified potential hotspots where False Alarm incidents are more frequent than other areas. This information can be used by relevant authorities to investigate the reasons behind these false alarms in order to reduce the number of such incidents, and to allocate resources more effectively. The map also helps in visualizing the spatial distribution of False Alarm incidents, and can aid in making better decisions based on geographical patterns. Overall, this mapping exercise has proven to be a useful tool in understanding the distribution of False Alarm incidents in the given area.

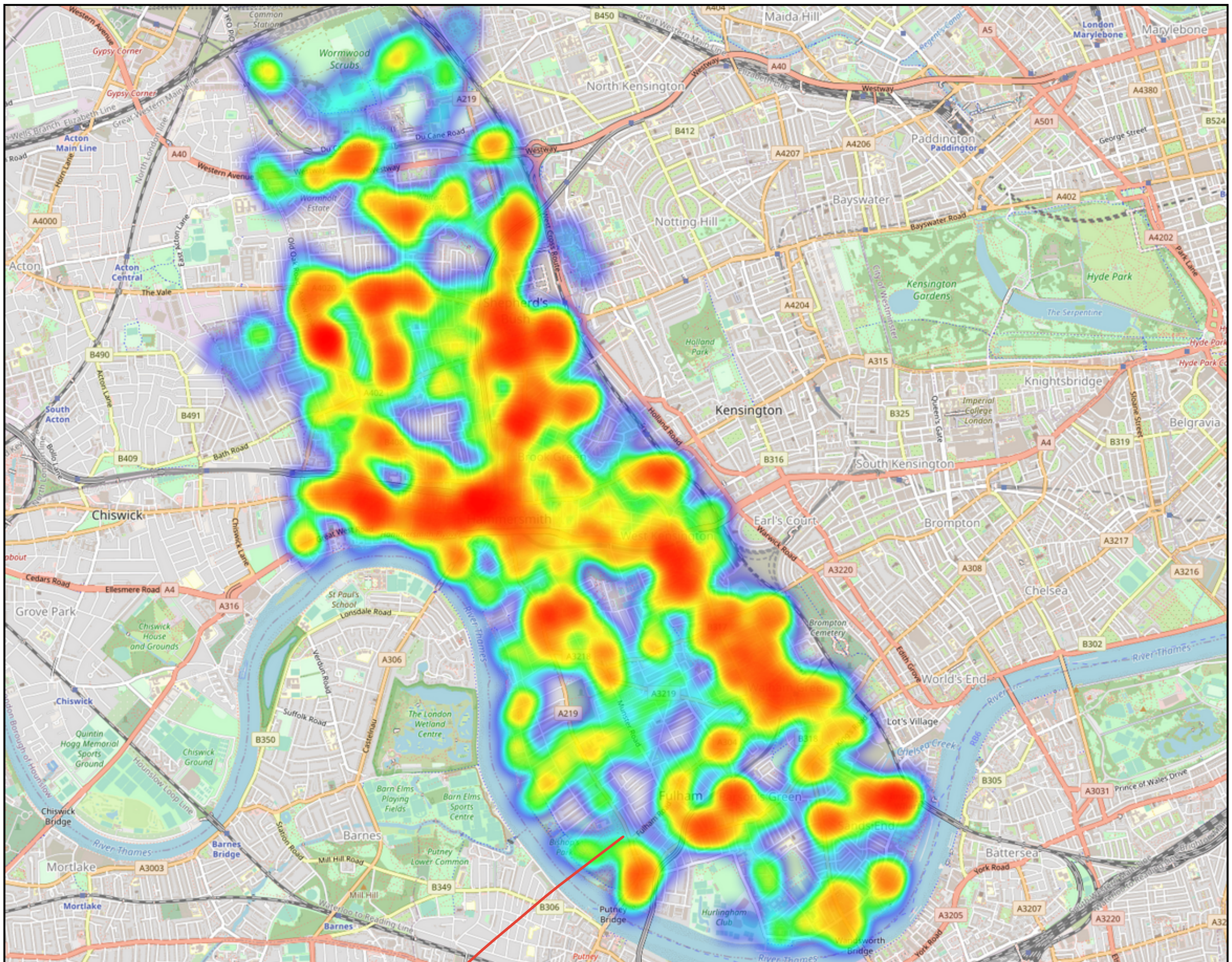


Fig: 3.2

The graph, Fig: 3.3, depicts the frequency of calls made to the Fire Brigade on different hours of the day. The X-axis shows the hour of the day in a 24-hour format, starting from 00:00 to 23:00 hours. The Y-axis represents the frequency of calls received on that particular hour. The graph displays a clear pattern of the peak hours during which the Fire Brigade receives the maximum number of calls. This data can be used to allocate resources more effectively and manage staffing levels accordingly. Additionally, the graph can provide insight into the underlying causes of the increase in calls during certain hours, such as rush hour traffic or popular times for cooking. The information can be used to improve emergency response times and optimize the allocation of resources to improve public safety.

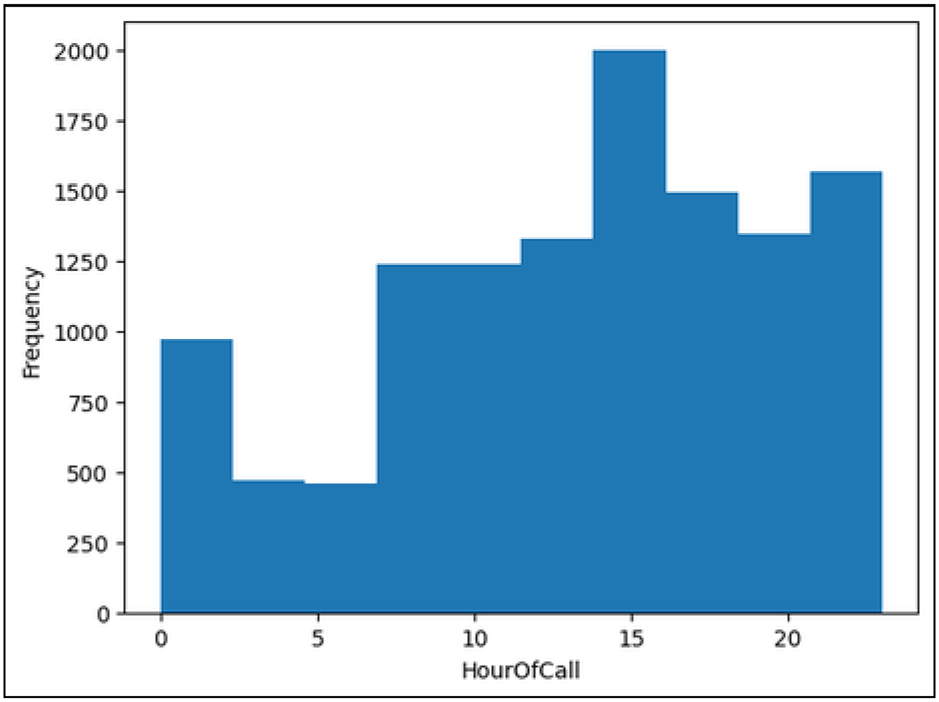


Fig: 3.3

4. Data Cleaning and Transformation

First we convert all column values to their appropriate types.

To handle the null values, we can adopt different strategies depending on the nature of the data and the specific requirements of our analysis. In this case, we will adopt a combination of strategies to ensure the best possible outcome.

i. We used the mean value for the ***Notional Cost (£)*** column to fill in the null values.

5. Modeling

6. Evaluation

7. Conclusion

Please complete your modelling steps and send me a completed draft ASAP. - EDA requires more content

13/4/23