DATA ANALYSIS REPORT

Table of Contents

[1. PROBLEM IDENTIFICATION: 2](#_Toc133513377)

[2. DATA PREPARATION: 8](#_Toc133513378)

[3. MODEL CONSTRUCTION: 9](#_Toc133513379)

[4. MODEL INTERPRETATION AND EVALUATION: 12](#_Toc133513380)

[5. SUMMARY: 12](#_Toc133513381)

[6. REFERENCES: 13](#_Toc133513382)

# PROBLEM IDENTIFICATION:

* 1. Data Description:

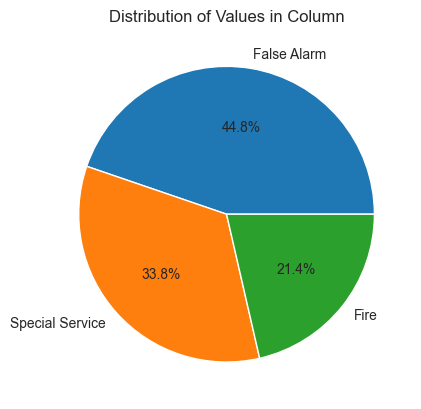
[1] The dataset includes 23 columns with various information about incidents that the London Fire Brigade (LFB) responded between 2019 to 2022. The data provides details on the date and time of call, incident category and description, location information such as postcode and incident station, as well as response information such as pump attendance time and cost. The dataset can be utilized to study the trends in incident types and locations, as well as response times and resources used by the LFB. By conducting further analysis on this dataset, the LFB can gain valuable insights that can help improve their emergency response strategies.

The following Python libraries were utilized: pandas, matplotlib, seaborn, timedelta, and sklearn. Pandas is an effective library that offers tools to manage and analyze large data sets, particularly those in tabular form like databases or spreadsheets. Matplotlib is a library used for creating visualizations in Python that can be static, interactive, or animated. It is a popular choice for data visualization due to its user-friendly features and flexibility. Seaborn is a Python library used to create statistical graphics, which is built on top of Matplotlib and offers a more straightforward way of creating informative and visually appealing graphics with less code. Timedelta is a Python datetime module class that represents the difference between two dates or times. It is widely used in time series analysis for calculating the time between two events or for performing date arithmetic. Scikit-learn is a Python machine learning library that includes tools for classification, regression, clustering, and dimensionality reduction. It is frequently used in the field of data science for constructing predictive models and carrying out statistical analysis.

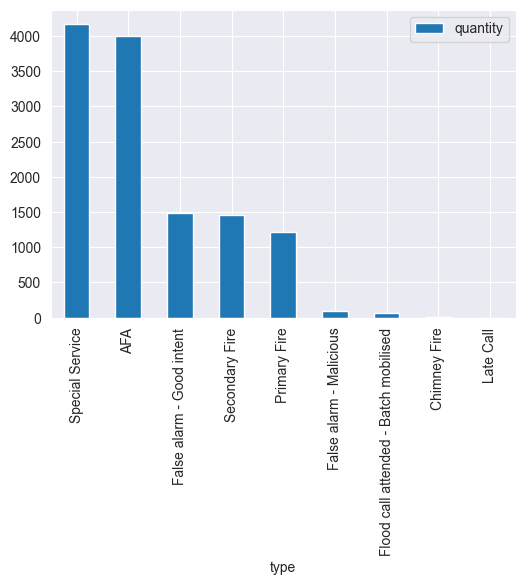
First, WE checked the shape of the dataset which was (12496, 23). In these 23 columns 11 columns contain numerical data while 12 columns contain categorical data. After that WE checked the null values and surprisingly there were many columns which had more than 50% null values. When dealing with categorical data, any null values were substituted with "unknown", whereas null values in numerical data were managed by using the mean value in certain columns and zero in others. To examine the distribution and consistency of the numerical data, skewness, standard deviation, and kurtosis were calculated. Upon assessing the skewness values, only a few columns fell within the normal range, while the majority were highly distorted and asymmetrical. For standard deviation, most of the columns had very low values, indicating that the data was not significantly different from the mean value. However, there was one column that varied significantly from the mean value. As for kurtosis, with a few exceptions, the distribution of the data was close to the mean. The basic statistical values of the categorical data were explored using the ".describe()" method, which provided several significant insights about the data. For instance, the majority of the incidents were found to be false alarms, the majority of the special service type was unknown, and only 50% to 60% of the incident locations were accurately reported. Additionally, most incidents occurred in the BR5 district, and the majority of the service was provided from the Orpington station, whereas the second pump station was found to be scarcely used.

After statistical analysis some data analysis were done to find out the insights and problems of the business.

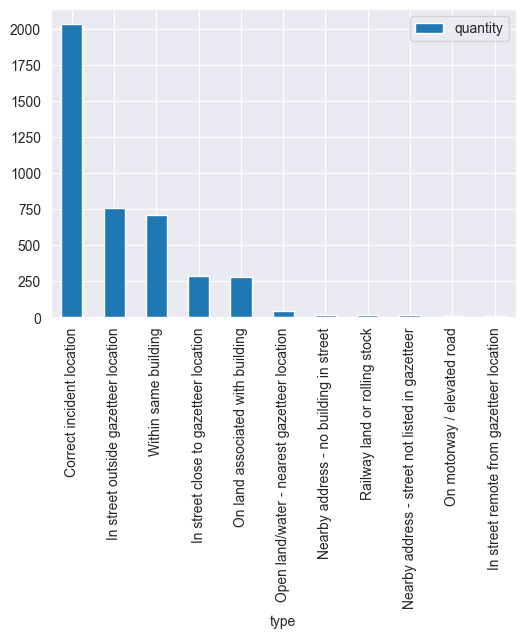
Some of the observations were that most of the incidents are false alarm which cost a lot of money.



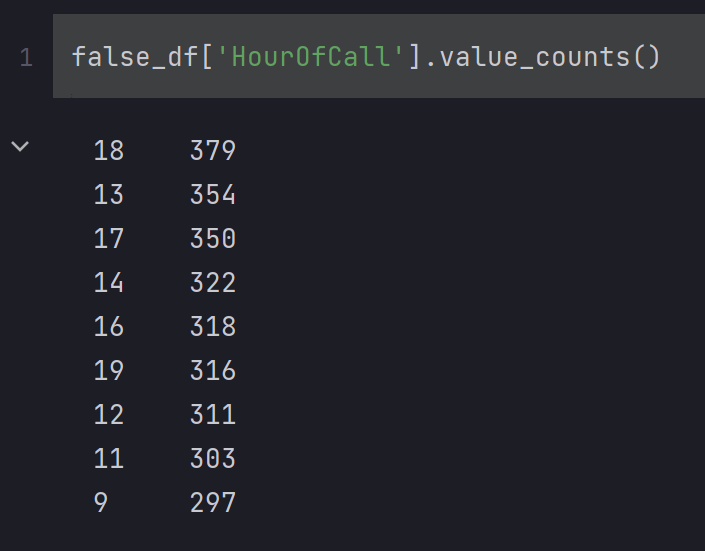
The good thing is most of the false alarm incident were in good intensions.



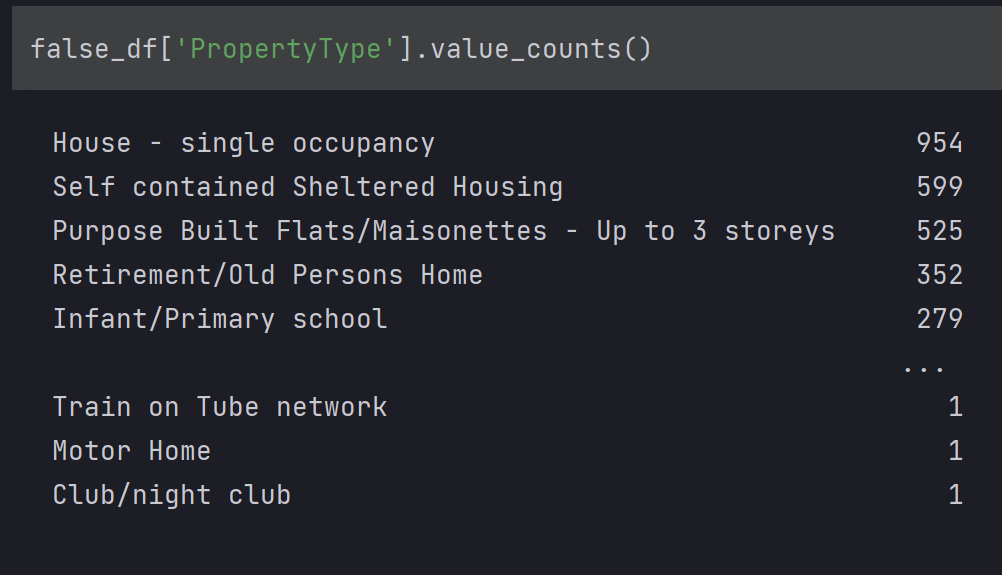
Also the response time from call to incident location was low and most of the time the correct location was given.



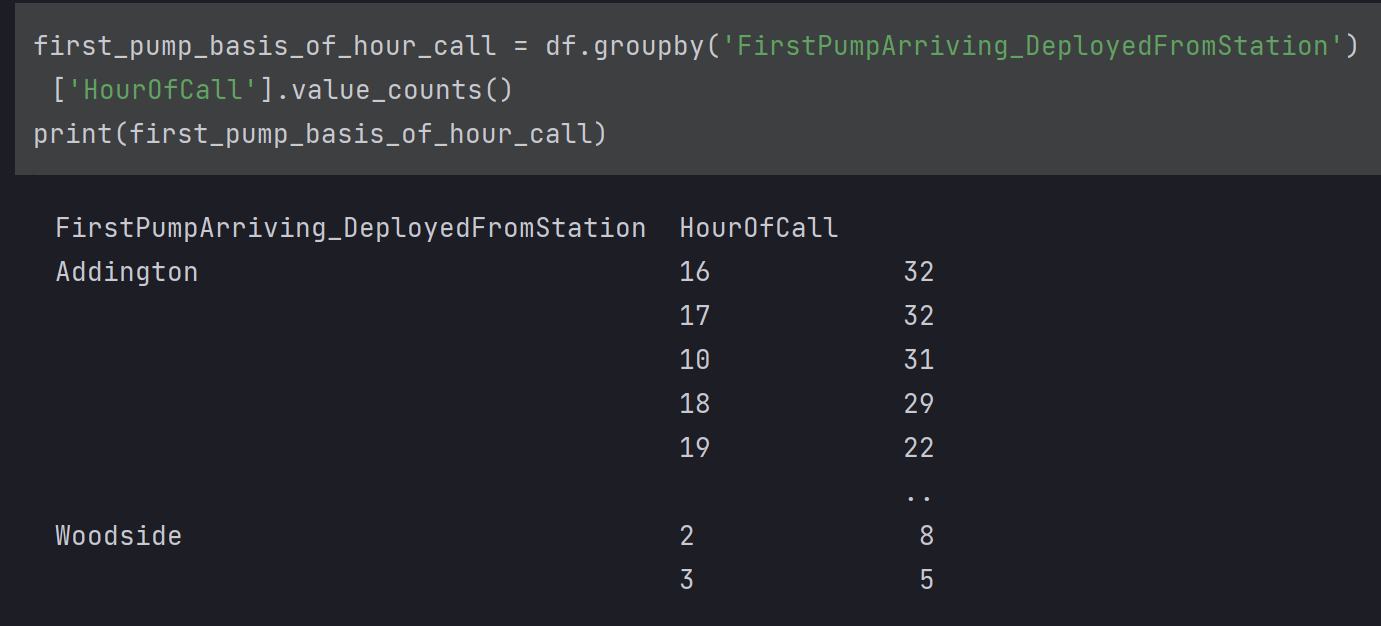
Most of the false alarm calls are done in the 18th hour and in the pm time.



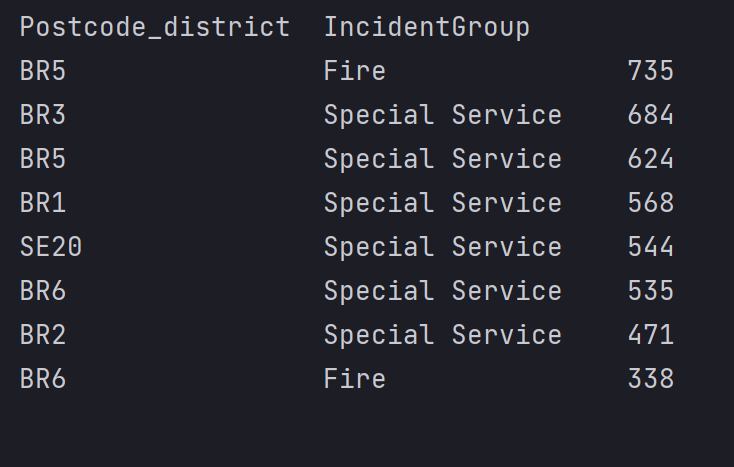
Most of the false alarm property type contains Single house occupancy and self-contained sheltered home. But there are false alarm cases against every property type.



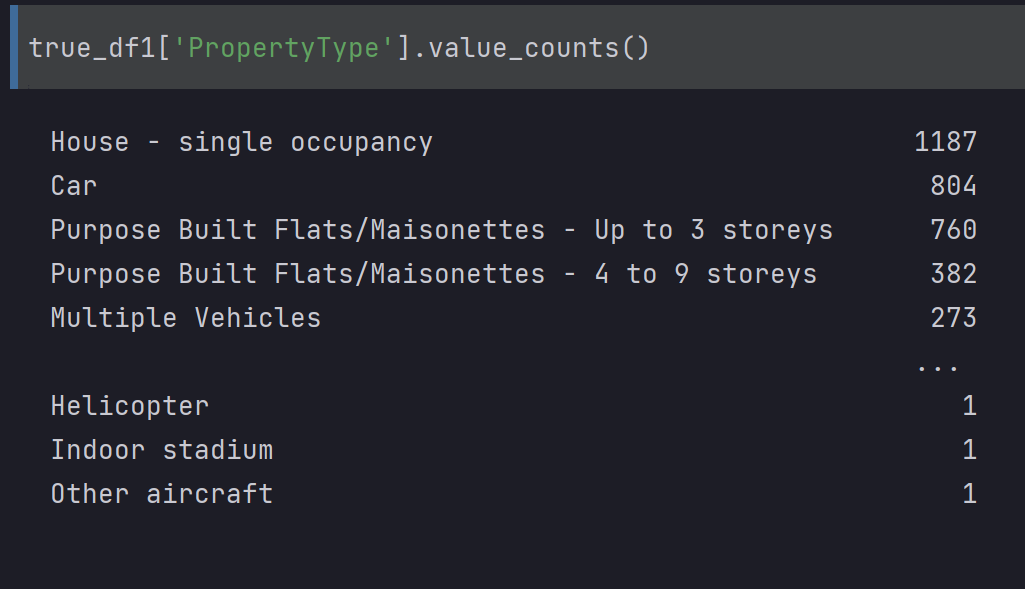
On the 16th and 17th hour call, Addington provided the most services and most of the incidents happen at pm time.



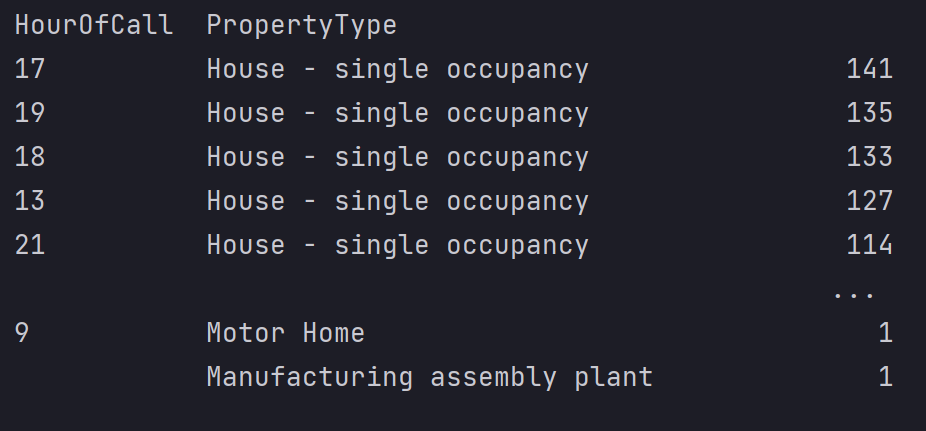
BR5 district is the most affected area by fire and the list shows that there are more districts which have high number of incidents.



Single occupancy house has the most chances of having a fire.



These fires are caused most at afternoon time.



We conclude that there most of the incidents are false alarm which cost a lot of money. The good thing is most of the false alarm incident were in good intensions. Also the response time from call to incident location was low and most of the time the correct location was given. Most of the false alarm calls are done in the 18th hour and in the pm time. Most of the false alarm property type contains Single house occupancy and self-contained sheltered home. But there are false alarm cases against every property type. On the 16th and 17th hour call, Addington provided the most services and most of the incidents happen at pm time. BR5 district is the most affected area by fire and the list shows that there are more districts which have high number of incidents. Single occupancy house has the most chances of having a fire. These fires are caused most at afternoon time. After observing these insights we can take actions according to these analysis to better our business.

* 1. Business Problem:

The most concerning problem for the business is number of incidents of false alarm cases. The amount of resources used and the cost spent is a lot on false alarm cases. Most of the false alarm cases are in the pm time and most real cases also occur in this time which create hindrance in the performance of LFB. Districts which was less affected has more average cost than the more affected districts. Another problem is that the data reported is not well distributed and managed. There were many null values and had large values of skewness and kurtosis. The response time should also decrease even though it is good but it can be better. These are some problems which needs to be handled from the business point of view.

* 1. Possible Solutions:

The biggest problem is the false alarm cases which create hindrance in the performance, a lot of money is wasted, and use a lot of resources. We observed that at what time most false alarm calls are done and from which property type these call are made so from the observations we can take some actions to reduce the false alarm cases. The most incidents occur in BR5 district so the pump stations should be attentive in that area. Also we could increase the number of stations in that area so we could decrease the response time. We can also do this by using variables such as IncidentGroup, StopCodeDescription, PropertyCategory, PropertyType, IncidentStationGround, FirstPumpArriving\_DeployedFromStation, FirstPumpArriving\_AttendanceTime, NumPumpsAttending, Year, Month, and DayOfWeek. The variables can aid in recognizing trends in the occurrence of incidents, the duration taken for the initial pump to arrive, and the count of pumps present. This information can facilitate efficient resource allocation and response time management. Clustering algorithms like k-means, hierarchical clustering, and DBSCAN can be utilized to categorize similar incidents based on these variables. We can collect data after each incident to identify the reasons for the fire and take preventative measures to minimize the occurrence of such incidents in the future. Also instead of leaving columns with null values we can insert some sort of meaningful information.

# DATA PREPARATION:

* 1. Variables For Analysis:

We utilized a majority of the variables in our data analysis by categorizing them into numerical and categorical data. For the numerical data, we conducted statistical analysis involving parameters such as skewness, standard deviation, and kurtosis. For the categorical data, we employed the ".describe()" method to compute fundamental statistical measures. Additionally, we employed various techniques such as plotting of time versus incidents, value counts of different columns, and differentiation between false alarms due to good intentions versus malicious ones to explore relationships between multiple columns.

* 1. Data Preprocessing:

The null values were addressed before the data preprocessing phase. Initially, we detected duplicates by utilizing the ".duplicated()" method. Following that, WE transformed the "DateOfCall" column into datetime format and subsequently created three novel columns to represent the year, month, and day of the week. For the categorical data, WE implemented one-hot encoding and later merged the encoded data with the numerical data. One-hot encoding is a data preprocessing method that transforms categorical information into a numerical format suitable for use in machine learning algorithms. This technique involves representing each distinct category with a binary vector, where the vector's length corresponds to the number of unique categories found in the dataset. For a given data point, the vector has a 1 in the position corresponding to the category of the data point, and 0s everywhere else.

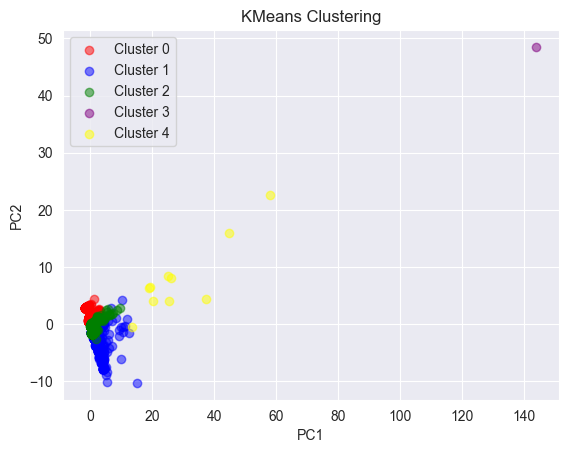
After that standard scalar normalization was used to normalize the data. Standard scalar normalization is a data preprocessing technique used to transform numerical data so that it has zero mean and unit variance. Each feature in this method is separately scaled to achieve an average equals to zero and a standard deviation equals to one. We can achieve this by dividing the result by std (standard deviation) value after subtracting the average of each feature from each data point. The modified data's standard normal distribution has a mean equal to 0 and a variance equal to 1 as a result. After the transformation of data there were 10 null values in the transformed dataframe which were handled by dropping the rows containing null values. We also dropped the last 10 rows of our original dataframe so that the shape of transformed data and original data could be same.

# MODEL CONSTRUCTION:

* 1. Predictive And Descriptive Modelling:

For descriptive modelling we have already done exploratory data analysis where we have found many insights, patterns, observations, and findings. The other model we used is k-means clustering. [2] K-means clustering is a type of unsupervised machine learning algorithm used for partitioning a set of observations or data points into K clusters. In K-means clustering, each data point is assigned to one of the K clusters based on their similarity or proximity to the centroid of the cluster. To begin with, the algorithm selects K initial centroids (known as cluster centers) from the data points randomly. Next, each data point is assigned to the cluster whose centroid is the nearest. Once all the data points have been allocated, the centroids are re-evaluated by computing the average of all the data points in each cluster. This iterative process is repeated until either the centroids stop changing or a designated number of iterations has been completed.

In my case the optimum number of clusters is 5. We manually checked different values of clusters and the optimum result was achieved when the number of clusters were 5.



Before plotting the clusters WE used PCA to reduce the dimensionality of the data to 2 components. [3] PCA stands for Principal Component Analysis. It is a popular dimensionality reduction technique in machine learning and data science. PCA operates by converting the input data into a fresh coordinate system, in which the principal components of the data serve as the axes. The principal components are combinations of the initial characteristics that express the maximum amount of variability in the data. The initial principal component captures the most variance, the second principal component captures the second most variance, and so forth. Following the transformation of the data, We produced a graph by employing a scatter plot with five clusters.

For predictive modelling We used two regression models Random Forest Regressor and SVR (Simple Vector Regressor). [4] Random Forest Regressor is a machine learning algorithm used for regression tasks, which is based on the Random Forest algorithm. It is a type of ensemble learning method, which combines multiple decision trees to make more accurate predictions. In Random Forest Regressor, multiple decision trees are constructed using randomly selected subsets of the training data and features. Each tree is then used to make a prediction, and the final prediction is obtained by averaging the outputs of all the trees. Random Forest Regressor is a popular algorithm for regression tasks because it is relatively simple to implement and provides high accuracy, robustness, and the ability to handle high-dimensional datasets with complex relationships between the features and the target variable. We used 'FirstPumpArriving\_AttendanceTime' as target variable and rest of the data as independent variables. We splat my dataset into 80:20 ratio. 80% for training the model and 20% for testing the model. We instantiated the RandomForestRegressor estimator with 100 decision trees and a random state of 42. Using the fit method, We trained the model on the training data and used the predict method to predict the response time for the test set. To evaluate the accuracy of the model's predictions, We calculated the mean absolute error between the actual and predicted values using the mean\_absolute\_error function.

[5] SVR stands for Support Vector Regression, which is a machine learning algorithm used for regression analysis. It is a variation of the Support Vector Machine (SVM) algorithm that is used for classification tasks. In SVR, the objective is to find the best possible line or hyperplane that can fit the data while maximizing the margin between the hyperplane and the data points. The hyperplane is chosen in such a way that it passes through the maximum number of data points, and the margin is chosen in such a way that it is as wide as possible. SVR is particularly useful when dealing with datasets that have a large number of features or dimensions, as it is able to reduce the dimensionality of the dataset and select the most important features. It is also useful when dealing with non-linear data, as it is able to handle non-linear relationships between the input features and the target variable. We used 'FirstPumpArriving\_AttendanceTime' as target variable and rest of the data as independent variables. WE splat my dataset into 80:20 ratio. 80% for training the model and 20% for testing the model. We instantiated the SupportVectorRegressor estimator with kernel equals to linear. Using the fit method, We trained the model on the training data and used the predict method to predict the response time for the test set. To evaluate the accuracy of the model's predictions, We calculated the mean absolute error between the actual and predicted values using the mean\_absolute\_error function.

* 1. Patterns And Reason Of Model Selection:

In descriptive modelling EDA was already done which gives us so many insights about the business and the patterns helped us find business problems and their respective solutions. The second best choice for descriptive modelling was k-means. The clusters fromed by the k-means model only shows one or two visible outliers and worked really well. For predictive modelling the data was continuous so a regression model was required. The first model we used was RFR with attributes n\_estimators equals to 100 and random\_state equals to 42. The random forest algorithm is a powerful regression model that can handle complex, non-linear relationships between features and target variables while reducing overfitting and being robust to outliers. SVM is one of the best machine learning algorithm so We used support vector regressor as my second predictive model. The performance of SVR was better than the random forest regressor but individually both performed really well.

# MODEL INTERPRETATION AND EVALUATION:

* 1. Interpretation:

Interpreting descriptive models is important for understanding data and drawing insights. In this project, k-means clustering and correlation analysis were applied to the transformed data, and relevant plots were created. The clustering algorithm grouped incidents based on similarity, and a scatter plot showed that the clusters were well-separated. Correlation analysis identified patterns and relationships between features, which were visualized using a heatmap. Relevant plots, such as histograms and box plots, were created to better understand the data distribution and differences among categories.

* 1. Evaluation

We used Random Forest and SVR in predictive models. Since our data is continuous, both the models we used are regressor and we used mean absolute error to check our model. The MAE for RF is 0.795 while for SVR the value is 0.725. We conclude that SVR works better than the RF and also since the error we found was on test data we cannot say that the model is over-fitted. We cannot find accuracy or classification report because our data is continuous and regression models are used.

* 1. Patterns Revealed:

In the descriptive modeling there were many missing values, outliers, large values of skewness and std which shows asymmetric in the dataset. A lot of data preprocessing was done but after that k-means, RF, and SVR models worked fine. K-means clustered the data with very less outliers. The k-means clustering algorithm has effectively grouped similar incidents together based on their features, allowing for a better understanding of incident patterns and resource allocation needs. The correlation analysis has revealed relationships between different variables that can inform decision-making regarding emergency service provision. The predictive model using SVR performed really good in predicting response times based on incident features, which can aid in resource allocation and service planning. These issues in the dataset concern the business to conduct and gather data more carefully.

# SUMMARY:

After a long analysis of this dataset, we found out that most of the incidents are false alarms which is a concern. A lot of cost is wasted on false alarm. Most of the false alarm cases are in the pm time and most real cases also occur in this time which create hindrance in the performance of LFB. Districts which was less affected has more average cost than the more affected districts. Another problem is that the data reported is not well distributed and managed. There were many null values and had large values of skewness and kurtosis. The response time should also decrease even though it is good but it can be better. In order to limit the number of false alarm cases, we examined when the majority of false alarm calls are made and what types of properties are responsible for these calls. The BR5 district is where the majority of incidents take place, so the pump stations there need to be on guard. Additionally, we could add more stations there, which would speed up reaction times. Use of variables like IncidentGroup, StopCodeDescription, PropertyCategory, PropertyType, IncidentStationGround, FirstPumpArriving\_DeployedFromStation, FirstPumpArriving\_AttendanceTime, NumPumpsAttending, Year, Month, and DayOfWeek are other ways we might accomplish this. The variables can help identify patterns in the frequency of events, the amount of time it takes for the first pump to arrive, and the number of pumps that are actually there. The management of reaction times and resource allocation can be made easier with the use of this information. These characteristics can be used to classify related episodes using clustering methods like k-means, hierarchical clustering, etc. After every incident, we may gather data to determine what caused the fire and take preventative action to lessen the likelihood of repeat occurrences. Additionally, we can insert some sort of valuable information rather than leaving columns with null values. Numerous missing values, outliers, and high values of skewness and standard deviation were found in the descriptive modelling, indicating that the dataset was asymmetrical. K-means, RF, and SVR models all performed well after extensive data preprocessing. With extremely few outliers, K-means grouped the data. The k-means clustering method has successfully combined related occurrences based on their characteristics, enabling a better comprehension of incident patterns and resource allocation requirements. The correlation study has identified correlations between many variables that can help in the provision of emergency services. In terms of estimating reaction times based on incident attributes, the predictive model utilizing SVR performed fairly well. This can help with resource allocation and service planning for our business.

# REFERENCES:

[1] London Datastore. (n.d.). Fire incidents attended by London Fire Brigade (LFB) in 2018/19. Greater London Authority. Retrieved April 27, 2023, from https://data.london.gov.uk/dataset/fire-incidents-attended-by-lfb

[2] Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129-137.

[3] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2(11), 559-572.

[4] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[5] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In: Advances in neural information processing systems 9, 155-161. MIT Press.