EST 1892 LSBU

Module Code	CSI_7_DMA_2024-25
Module Title	Data Mining and Analysis

Report:

Coursework of Datamining and Analysis

Student ID: 4327265

Contents

1. Problem Identification

1.1 Dataset Context and Characteristics

The dataset used in this project is from the London Fire Brigade (LFB) and contains information about emergency incidents reported between 2019 and 2022. In addition, the dataset was filtered by incidents occurring within the Borough of Bexley to allow operational issues to be addressed and for targeted analyses. By tailoring their analysis they were able to centry their analysis on the spaital, temporal and operational incidence patterns for Bexley.

This was carried out as an initial exploration of the data to ensure that the analytical processes were accurate and that data integrity was upheld, making sure data types matched to LFB's official metadata. DateOfCall and other temporal variables were cast to date and datetime data formats for unambiguous temporal analyses, while identifier-type fields (IncidentNumber, UPRN, USRN) were cast as categorical forms, preserving their semantic meaning. The statistical summaries for numerical variables (PumpHoursRoundUp, PumpCount, Notional Cost (£)) displayed heavily skewed distributions, suggesting the presence of outliers that may need further preprocessing. Additionally, categorical features (IncidentGroup, IncidentStationGround) displayed clear class imbalances, thus requiring specialized techniques (oversampling (SMOTE)) during predictive modelling thereafter.

Exploratory visual analyses helped reveal significant analytical pathways. Spatial mapping demonstrated significant clustering of incidents surrounding Dean of Urban Wards like Thamesmead East — temporal analysis revealed specific peak periods of incidents, particularly in the evening or weekends. These initial findings can validate the relevance and applicability of the dataset to specific operational questions faced by LFB.

1.2 Business Problems of Interest

The three identified business problems were selected to meet explicit operational needs determined through an early exploration of patterns in incident dataset. The rationale for selecting each problem was firmly supported by the exploratory analysis data, being highly aligned with LFB's strategic objectives:

• **Business Problem 1 (Descriptive):** *When and where do fire incidents tend to happen the most in Bexley?* The initial analysis uncovered evident temporal and geographic patterns—high incident timeframes were detected in the evening hours, and specific geographic locations showed concentrated incident activity. The detection of these trends informs targeted safety campaigns, allocation of resources, and planning for strategic response.

• **Business Problem 2 (Predictive)**: *Based on its location, time, and pertinent incident information, which fire station lot is likely to be called to an upcoming incident?* Preliminary analysis of the data revealed that some fire stations were underworked in comparison to others, which could lead to increased efficiency. Predicting responsible fire station in advance will lead to efficient resource deployment, preparedness, and even operational load balance among fire stations.

• **Business Problem 3 (Descriptive + Predictive):** *What are the major trends and factors behind fire, false alarm, and special service incident reports—and is the incident type predictable beforehand?* Basic analyses of categorical distributions suggested distinct incident types, with differential temporal patterns and contributing factors. This knowledge, which allows for predictions through learning, enables LFB to use its resources more effectively — improving its response without over-activating its equipment unnecessarily.

1.3 Data Mining Tasks Associated with Unsupervised Learning

A structured combination of descriptive and predictive data mining processes was then defined, based on characteristics of the initial exploration of the dataset, to thoroughly solve the chosen business problems:

• Business Understanding (Business Problem 1 & 3):

This constituted an in-depth exploratory data analysis (EDA) using visual and statistical techniques like histograms, heatmaps, and correlation matrices. K-means clustering combined with Principal Component Analysis (PCA) was also selected to uncover natural grouping of incidents by factors like time, location, and resource needs, assisting in the discovery of underlying patterns and drivers that may explain different types of incidents.

• Predictive Modelling (Novel Business problems 2 and 3):

Predictive analysis using strong classification algorithms: Random Forest, and Logistic Regression, and XGBoost. These algorithms were chosen based on their high predictive performance and robustness against class imbalance, which was clearly found during preliminary exploration. To ensure predictions across the different types of incidents and fire station grounds were accurate and unbiased, a technique known as the Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the datasets.

By matching these specific data mining tasks with evidence-based operation priorities from both the preliminary data exploration and analysis stages, it directly aligns with LFB's strategic objectives and will provide support for improved emergency preparedness, operational efficiency and resource allocation within the Borough of Bexley.

2. Data Understanding

2.1 Data Types and Metadata Alignment

This required a critical first step in the analysis, which was to read through the dataset in full and make sure their variables aligned with those found in the metadata the London Fire Brigade (LFB) had provided. Mismatched data types can heavily screw up statistical analysis and the integrity of model results, so getting these made sure up is vital. Also, explicit conversions were performed: the DateOfCall Variable was converted into a datetime in order to be able to conduct temporal analyses. To promote clarity in analysis, numerical fields like HourOfCall, PumpCount, and Notional Cost (£) are converted to integers. You explicitly converted categorical variables including IncidentGroup, PropertyType and IncidentStationGround to `category` type to save storage space and enhance interpretability. Identification fields (IncidentNumber, UPRN, USRN) and static codes (FRS, IncGeo_BoroughCode) were preserved as object (string) types in lieu of risking information loss, maintaining contextual integrity. Figure 1 illustrates the impact of these adjustments

Gathered From meta Data:

Columns	***Datatype(Before)***	***Datatype(After)***
IncidentNumber	*object*	*category*
DateOfCall	*object*	*Datetime*
CalYear	*int64*	*Integer*
TimeOfCall	*object*	*object*
HourOfCall	*int64*	*Integer*
IncidentGroup	*object*	*category*
StopCodeDescription	*object*	*category*
SpecialServiceType	*object*	*category*
PropertyCategory	*object*	*category*
PropertyType	*object*	*category*
AddressQualifier	*object*	*category*
Postcode_full	*object*	*object*
Postcode_district	*object*	*category*
UPRN	*float64*	*object*
USRN	*int64*	*object*
IncGeo_BoroughCode	*object*	*object*
IncGeo_BoroughName	*object*	*category*
ProperCase	*object*	*category*
IncGeo_WardCode	*object*	*object*
IncGeo_WardName	*object*	*category*
IncGeo_WardNameNew	*object*	*category*

Easting_m	*float64*	*Integer*
Northing_m	*float64*	*Integer*
Easting_rounded	*int64*	*Integer*
Northing_rounded	*int64*	*Integer*
Latitude	*float64*	*float64*
Longitude	*float64*	*float64*
FRS	*object*	*object*
IncidentStationGround	*object*	*category*
FirstPumpArriving_AttendanceTime	*float64*	*Integer*
FirstPumpArriving_DeployedFromStation	*object*	*category*
SecondPumpArriving_AttendanceTime	*float64*	*Integer*
SecondPumpArriving_DeployedFromStation	*object*	*category*
NumStationsWithPumpsAttending	*float64*	*Integer*
NumPumpsAttending	*float64*	*Integer*
PumpCount	*float64*	*Integer*
PumpHoursRoundUp	*float64*	*Integer*
Notional Cost (£)	*float64*	*Integer*
NumCalls	*float64*	*Integer*

comparing types for input and fixed version This alignment is necessary to enable accurate modeling, temporal analysis, and optimized memory usage. For example, changing DateOfCall to datetime, enables further breakdowns by hour, day, season, as well as assigning categorical types for better encoding in model training.

2.2 EDA — Exploratory Data Analysis

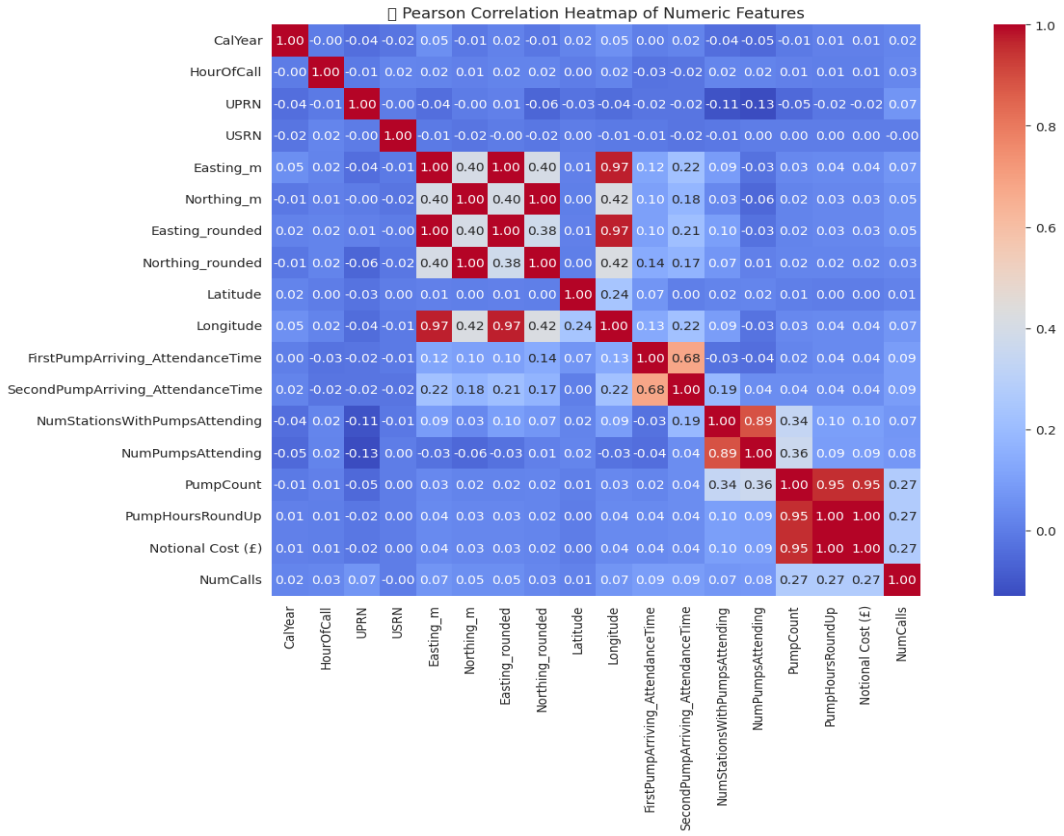
The exploratory data analysis phase gave us information about the distribution of numbers and the content of strings in our dataset. Data Preprocessing The dataset consisted of 9,073 records (lines) and 39 attributes (columns). The descriptive statistics; mean, standard deviation, and skewness explored by using descriptive statistics in which reported some numeric variables were strongly right-skewed and had significant tails, with PumpHoursRoundUp, PumpCount and Notional Cost (£) being only a few of this nature. Clearly, this is shown in Figure 2

which displays both summary statistics and KDE-augmented histograms. These patterns do point to the fact that there are extreme values present in the data, therefore Outlier detection is an essential pre-process step. Visual inspection revealed class imbalance within categorical features Figure 3 displays the distribution of IncidentGroup **Figure 3**: IncidentGroup Class Distribution


```
📊 Class Counts (Before SMOTE):
IncidentGroup
False Alarm        4019
Special Service    2822
Fire               2137
Name: count, dtype: int64
```

we can see that False Alarm have higher occurrence than Special Service and Fire incidents. This skewed distribution emphasizes the need for resampling techniques like SMOTE to avoid predictive models being biased towards the majority class. The relationships between variables were further understood through correlation analysis. As shown in **Figure 4**

Figure 4: Pearson Correlation Heatmap

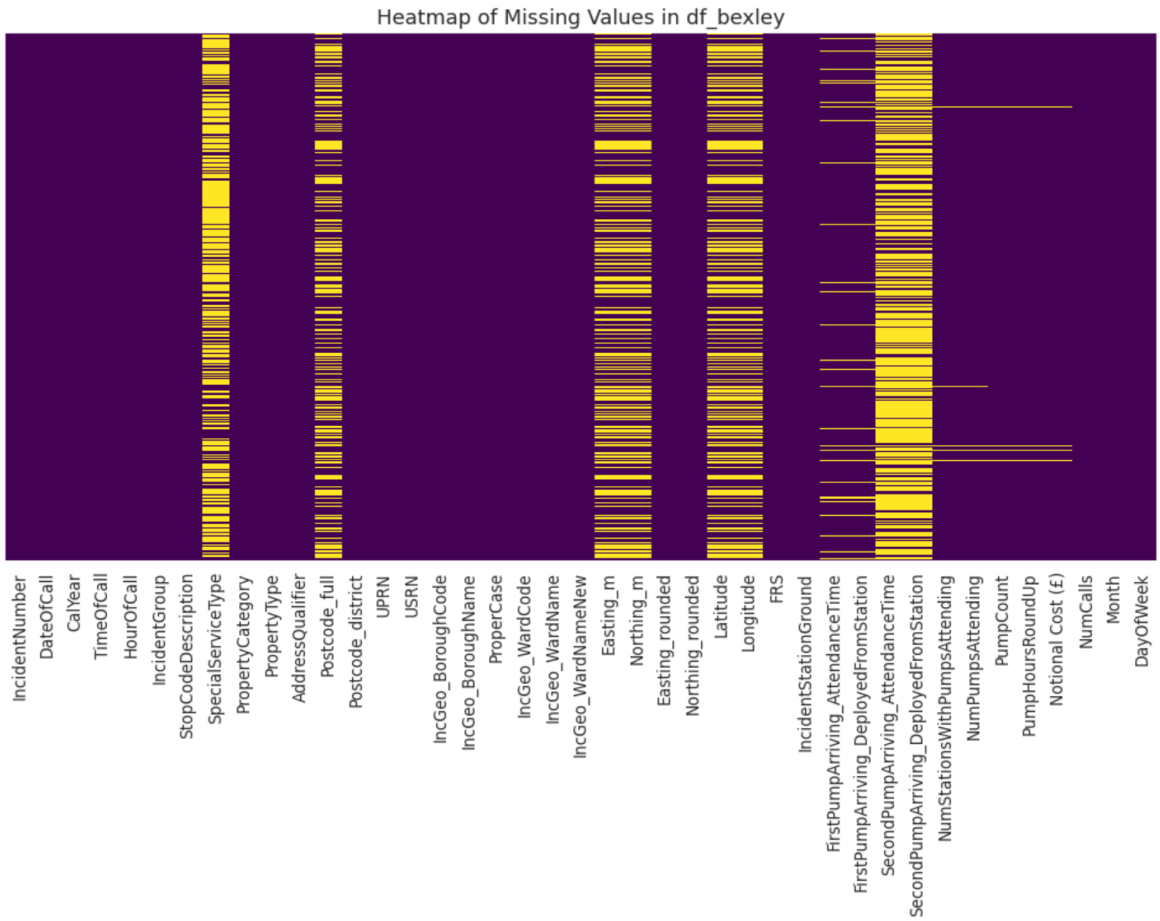


a Pearson correlation heatmap emphasizes the strongest positive correlation among PumpCount, PumpHoursRoundUp, and Notional Cost (£), all of them above 0.9 correlation coefficient. This high multicollinearity can sink your model's stability, and another strong argument for dimensionality reduction techniques, such as PCA.

2.3 Data Quality Assessment

Data quality, frequently found in the form of missing values, is one of the most important aspects of ensuring model reliability. The dataset was found to have substantial null percentages in fields including SpecialServiceType, SecondPumpArriving_AttendanceTime and all geospatial coordinates (Latitude, Longitude, Easting, Northing). Missingness: Volume and Distribution You can see the volume and distribution of the missingness in Figure 5Figure 5: Missing Values

Heatmap



an attetional heatmap showing the null entry density over particular variables. **Figure 6** provides a further summary of the magnitude and locations of these missing values ,which tells how many columns are affected and shows how the data set looked like at the beginning.Figure 6:

Missing Values Table

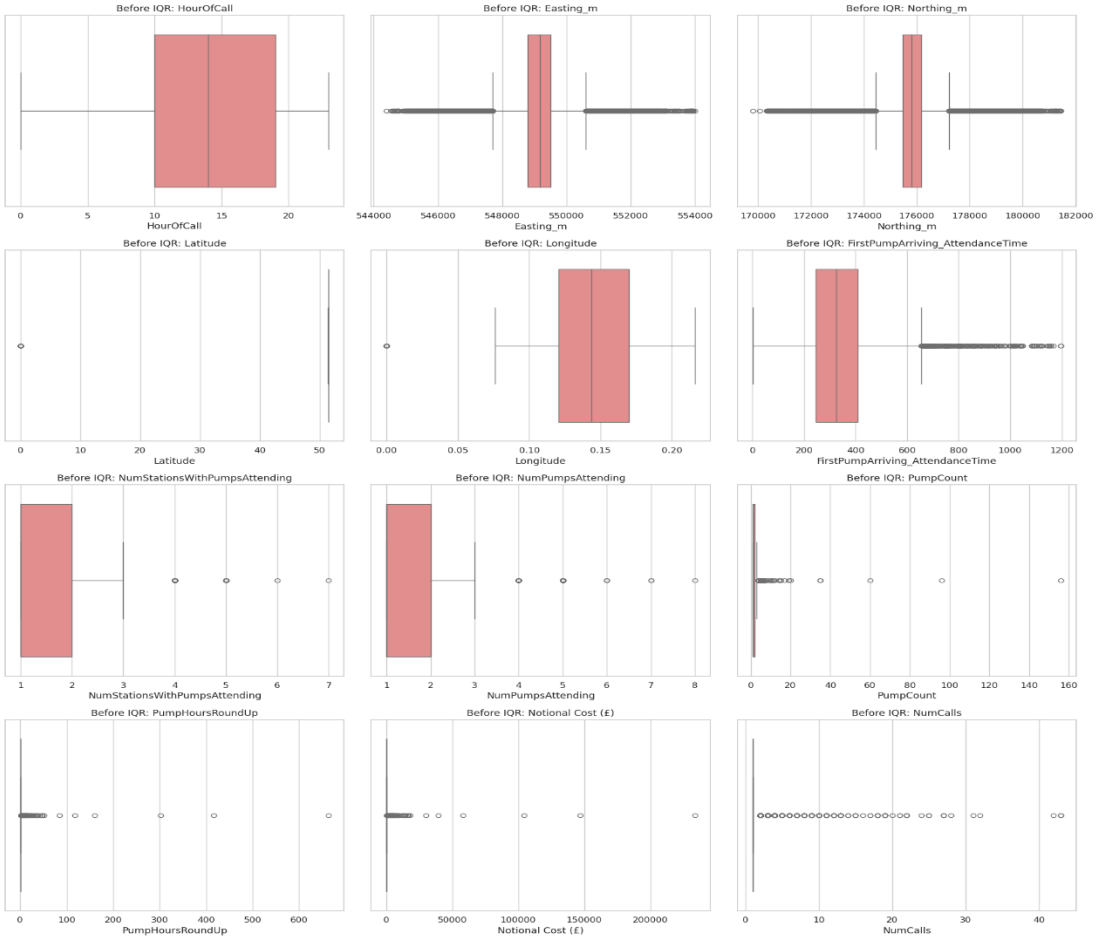
📐 Shape BEFORE cleaning: (9073, 39)

🔍 Missing Value Summary Before Cleaning:

	Missing Values	Missing Percentage (%)
SpecialServiceType	6175	68.06
SecondPumpArriving_DeployedFromStation	6080	67.01
SecondPumpArriving_AttendanceTime	6080	67.01
Postcode_full	3925	43.26
Easting_m	3925	43.26
Longitude	3925	43.26
Latitude	3925	43.26
Northing_m	3925	43.26
FirstPumpArriving_AttendanceTime	434	4.78
FirstPumpArriving_DeployedFromStation	434	4.78
NumPumpsAttending	91	1.00
NumStationsWithPumpsAttending	91	1.00
PumpHoursRoundUp	64	0.71
PumpCount	64	0.71
Notional Cost (£)	64	0.71
IncGeo_WardName	1	0.01
IncGeo_WardNameNew	1	0.01
IncGeo_WardCode	1	0.01

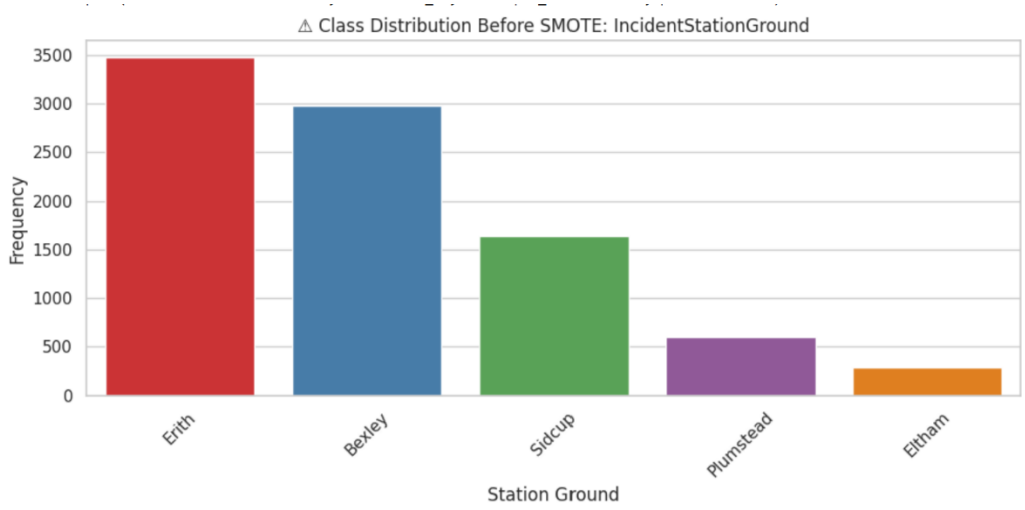
The presence of outliers, first suggested using descriptive statistics, was confirmed via visual inspection. As shown in Figure 7

Figure 7: Boxplots Before and After IQR Capping



PumpHoursRoundUp, Notional Cost (£) and others, were identified with significant divergence (average and median) from central statistics (mean, mode) highlighting potential extreme values that could skew analysis and models. Even further confirming dataset imbalances, Figure 8

Figure 8: IncidentStationGround Distribution




```
⚠️ Class Counts Before SMOTE:
 IncidentStationGround
Erith        3472
Bexley       2976
Sidcup       1638
Plumstead     602
Eltham        290
Name: count, dtype: int64
```

displaying the frequency of classes within each of the different fire station grounds. It shows that Erith and Bexley are far worse represented by the colour of the station, while stations like Plumstead and Eltham recorded far fewer incidents. Therefore, such differences need to be recognized prior to developing predictive models, since they can represent unequal coverage of operational regions.

2.4 Dataset Suitability for Business Problems

After this detailed review, the dataset was determined to be very appropriate and comprehensive enough to conduct analysis and remediate the business problems highlighted above. Rich temporal, spatial, categorical, and numerical features allow for rich descriptive and predictive modeling tasks without the need to revise the original business problem statement. Therefore, the availability of the dataset for performing extensive exploratory and predictive data mining processes was ensured.

3. Data Preparation

3.1 Data cleaning and feature selection

With the business problems defined earlier, the data preparation incorporated systematic feature selection and thorough dataset cleaning to address these problems. Variables considered redundant, i.e. identifiers including the IncidentIdentifier, UPRN and USRN, static reference fields (FRS, CalYear, and AddressQualifier), and rounded spatial coordinates (Easting_rounded and Northing_rounded) were removed. The removed variables were either of little actionability or were mere duplicates of existing geospatial information, its removal also increased computational efficiency.

An exhaustive analysis of the missing values was performed, as shown in **Figure 10.**

Figure 10: Missing Values Before and After

```
print(f"📐 Shape After cleaning: {clean_df.shape}")
```

```
SecondPumpArriving_AttendanceTime                    6080
SecondPumpArriving_DeployedFromStation               6080
Easting_m                                            3925
Postcode_full                                        3925
Longitude                                            3925
Latitude                                             3925
Northing_m                                           3925
FirstPumpArriving_AttendanceTime                      434
FirstPumpArriving_DeployedFromStation                 434
NumPumpsAttending                                      91
NumStationsWithPumpsAttending                          91
PumpCount                                              64
Notional Cost (£)                                      64
PumpHoursRoundUp                                       64
IncGeo_WardName                                         1
IncGeo_WardNameNew                                      1
IncGeo_WardCode                                         1

                                     Missing Values (After)  \
SpecialServiceType                                      0.0
SecondPumpArriving_AttendanceTime                       NaN
SecondPumpArriving_DeployedFromStation                  NaN
Easting_m                                               0.0
Postcode_full                                           0.0
Longitude                                               0.0
Latitude                                                0.0
Northing_m                                              0.0
FirstPumpArriving_AttendanceTime                        0.0
FirstPumpArriving_DeployedFromStation                 339.0
NumPumpsAttending                                       0.0
NumStationsWithPumpsAttending                           0.0
PumpCount                                               0.0
Notional Cost (£)                                       0.0
PumpHoursRoundUp                                        0.0
IncGeo_WardName                                         1.0
IncGeo_WardNameNew                                      1.0
IncGeo_WardCode                                         0.0
```

which illustrates the difference between the original and the cleaned dataset. This was how far sweeping missingness, especially in SecondPumpArriving_AttendanceTime and SpecialServiceType had been reduced or eradicated for further disposal. Also the filled columns from eastining and northing confirms that the geospatial imputation was successful one because in location fields zero missing values are remaining (**see Figure 11** Latitude and Longitude Filled from Easting/Northing).

Figure 11

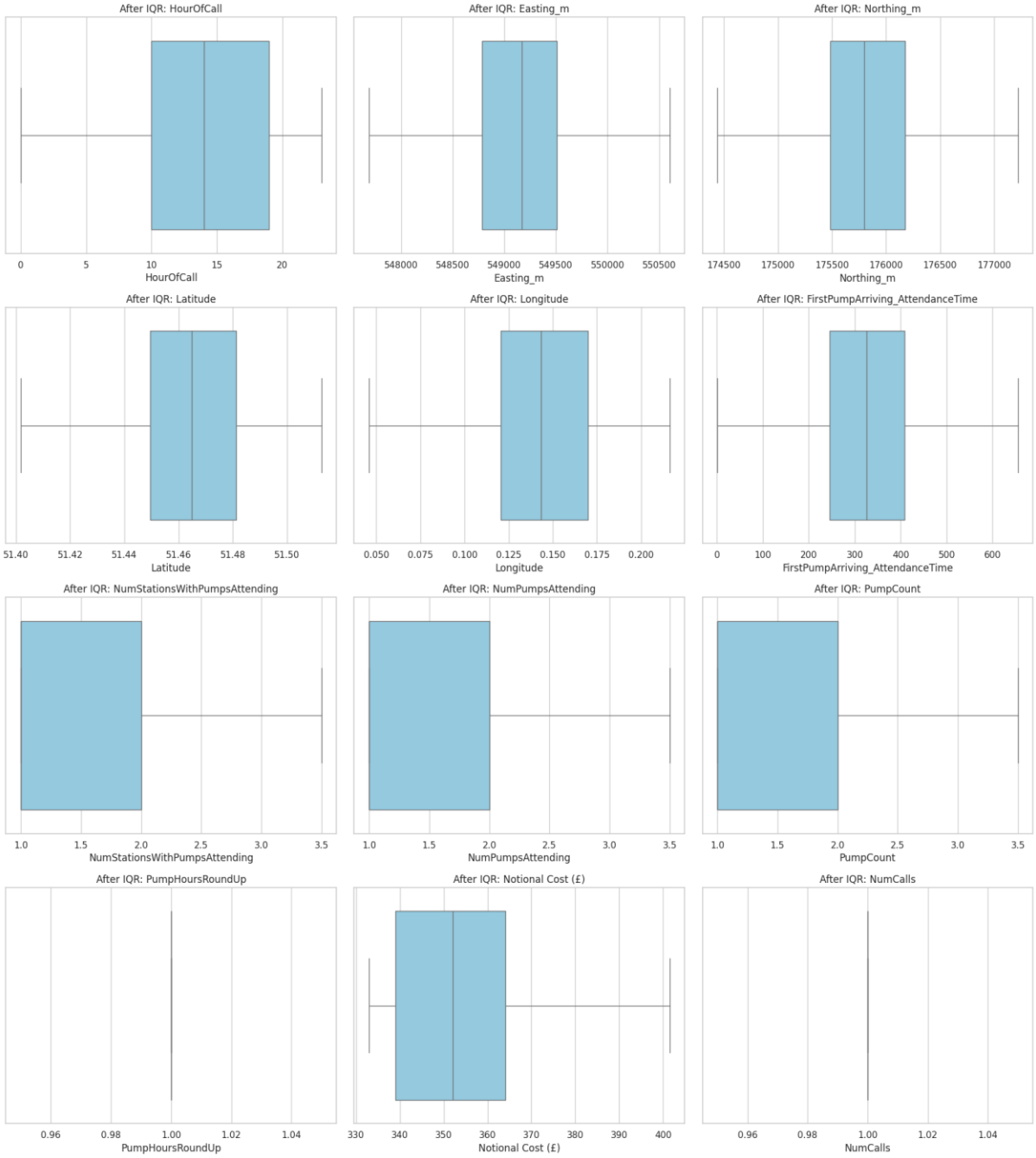
```
✅ Latitude and Longitude updated using Easting/Northing where missing.
Remaining missing in coordinates:
Latitude     0
Longitude    0
dtype: int64

📌 Sample after coordinate update:
     Easting_rounded  Northing_rounded   Latitude  Longitude
16            548950            178050  51.481405   0.143320
82            549150            178850  51.489073   0.146468
118           551950            176650  51.468830   0.186487
127           551650            177550  51.476588   0.182301
133           546050            172550  51.433133   0.099660
140           551550            177850  51.479311   0.180991
153           549550            172550  51.432640   0.150192
260           549150            178850  51.489043   0.146457
357           551250            178250  51.482985   0.176845
600           546050            171550  51.424148   0.099249
```

3.2 Outlier Detection and Treatment

It is important to address outliers in order to provide accurate predictive modeling. Indeed, as can be seen in Figure 7 many strong predictors such as “PumpCount” and “Notional Cost (£)” had a lot of spread and also outliers. These could skew model training, they said. Post-treatment imagery displayed in **Figure 13**

Figure 13:



the Boxplots After IQR Capping provided, show That have much more normalized distributions, which indicates outliers were tackled efficiently in the data without destroying the variance.

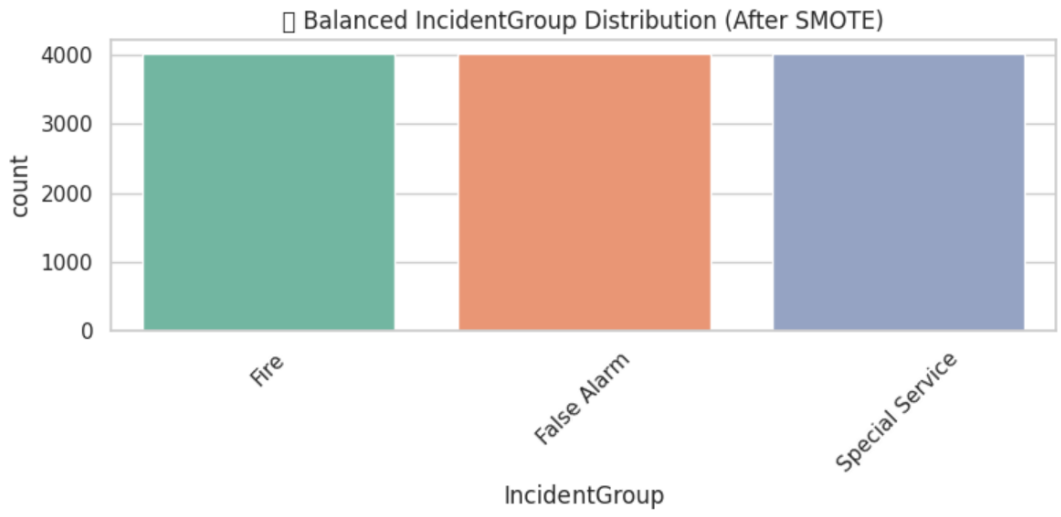
3.3 Correcting Class Imbalance

Control of the data was a burden to maintain as target variables (IncidentGroup, IncidentStationGround) were class imbalanced (especially for both incident types (Fire

with the highest representation), and the station grounds being biases towards the one used for recording from the dataset. Figure 14 and

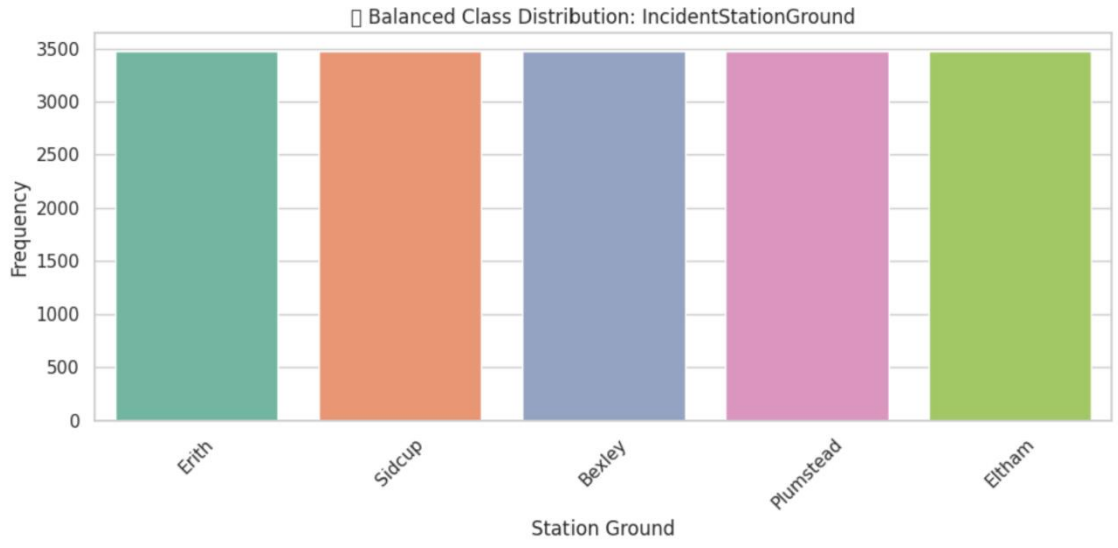
Figure 14: IncidentGroup Distribution After SMOTE, showing an equal distribution of each group, while **Figure 15**

Figure 14:




```
Balanced Class Counts:
Counter({'Fire': 4019, 'False Alarm': 4019, 'Special Service': 4019})
```

Figure 15:




```
Balanced Class Counts:
Counter({'Erith': 3472, 'Sidcup': 3472, 'Bexley': 3472, 'Plumstead': 3472, 'Eltham': 3472})
```

StationGround Class Distribution After SMOTE, reflects a now even distribution of all five stations. Seeing these images confirms the success of the pre-processing

techniques of oversampling I employed and further asserts the data is ripe for unbiased learning.

3.4 Feature Encoding and Transformation

Therefore, concept of feature encoding has been used to convert categorical variables into numbers suitable for your machine learning models. Systematically Label Encoding was done, converting categorical data into numerical labels. The numerical variables were normalized using StandardScaler, which helps ensure consistent scales among features and increases the performance of models sensitive to feature magnitude, such as Logistic Regression and XGBoost. This was an important step to ensure that the preprocessing was tailored to the needs of the model.

3.5 Splitting the Data for Evaluation

In order to prevent bias while evaluating model performance, stratified sampling was used to split the dataset into training and test sets (80/20 split). This approach ensured that the class proportions were consistent in both subsets, allowing for accurate evaluations of the model performance and generalizability. The split was done programmatically to enable reproducibility:

```
# ✂️ Train-test split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

The dataset was processed—cleaned, balanced for target classes, transformed for modeling, and partitioned for both validation and testing through these clearly delineated preprocessing steps. Visually and statistically, each decision was effectively justified, where the data quality issues identified in the previous phase were addressed, and where the decisions aligned with the business problems established previously. Such preparatory action is ensured at a base level of versatility, in the next modeling and analysis phases.

4. Model Construction

4.1 Descriptive Modelling

For Business Problem 1 ("Where and when are fire incidents most likely to occur in Bexley?") and Business Problem 3 ("What drives lower or higher incident types, and

can the patterns be grouped together?") Descriptive modeling was then done using K-Means clustering along with principal component analysis at the same time.

In the first clustering, we have used K=3 to cluster full dataset based on numerical features (PumpCount and Notional Cost (£)) These were selected based on their high PCA loadings (0.727 and 0.674 respectively; **Figure 16**: Descriptive1_PCA Component Loadings). The PCA guaranteed that interpretative power was not forfeited with dimensionality reduction.

Figure16

🔍 PCA Component Loadings:

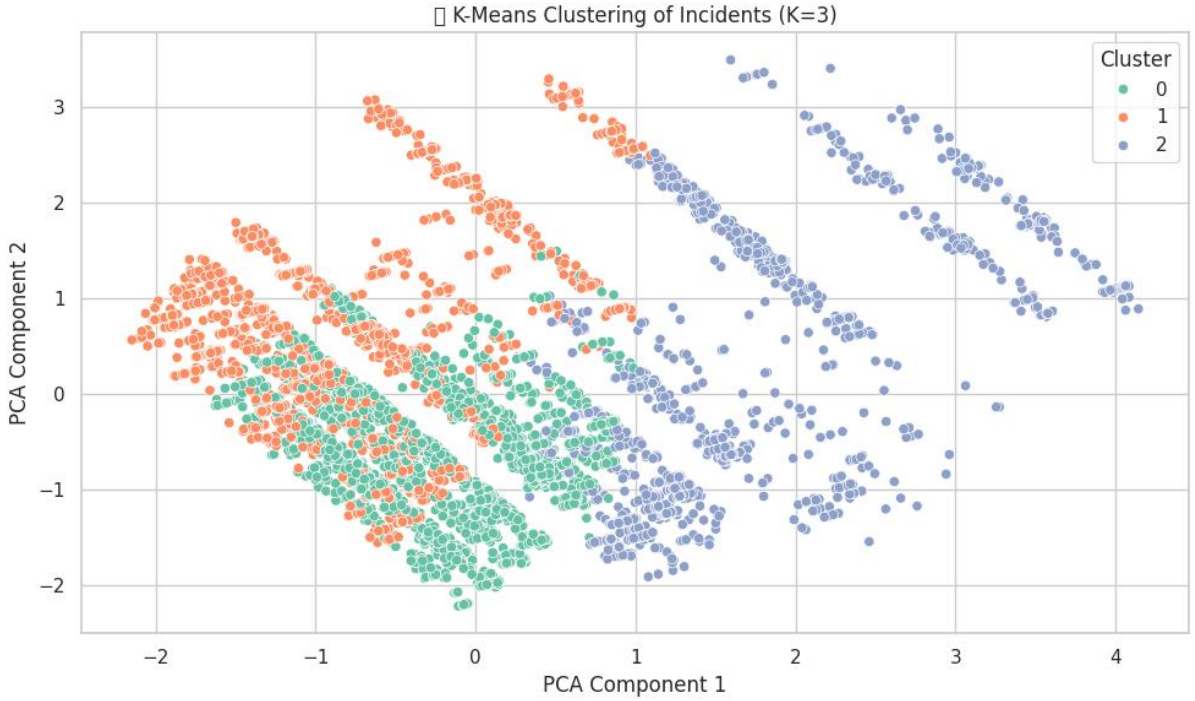
	PC1	PC2
HourOfCall	0.042002	0.018433
PumpCount	0.727097	0.130350
Notional Cost (£)	0.431583	0.674348
PropertyType	0.291845	-0.414800
IncidentGroup	-0.445117	0.596544

📍 Cluster Centroids (Original Scale):

	HourOfCall	PumpCount	Notional Cost (£)	PropertyType	IncidentGroup
0	13.605114	1.205492	347.641572	85.853220	0.235795
1	13.667891	1.100339	356.875777	105.903053	1.774731
2	13.779489	2.254621	365.687060	131.942782	0.332306

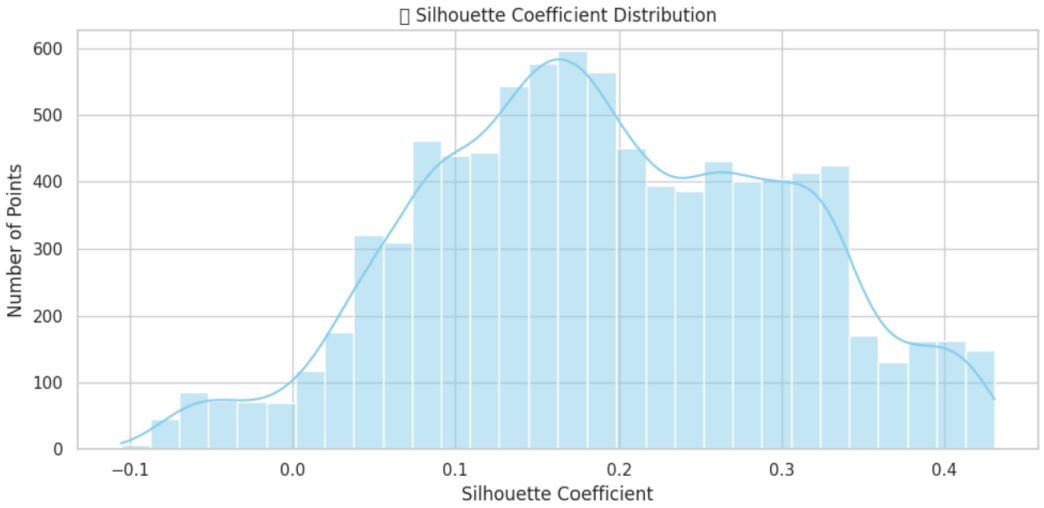
The significant operational groupings made in the earlier step using clusters identified in **Figure 17**: Descriptive1_KMeans Clustering of Incidents (k=3). One of the clusters (Cluster 2) showed as high-resource incidents incurring a mean of 2.25 pumps with a mean notional cost of £366. These are meaningful clusters for the groupings by which resources can be planned.

Figure 17



In order to test the strength of the clustering structure, silhouette scores were defined and presented in the figure (fig. 18:Descriptive1_Silhouette Coefficient Distribution) scores between the range of 0.2, and 0.3 confirmed moderate separation.

Figure 18



Subsequently, a separate K=6 clustering was performed for fire incidents. As shown in **Figure 19**, PCA again identified PumpCount and Notional Cost (£) as the top features.

Figure 19

🔍 PCA Component Loadings:

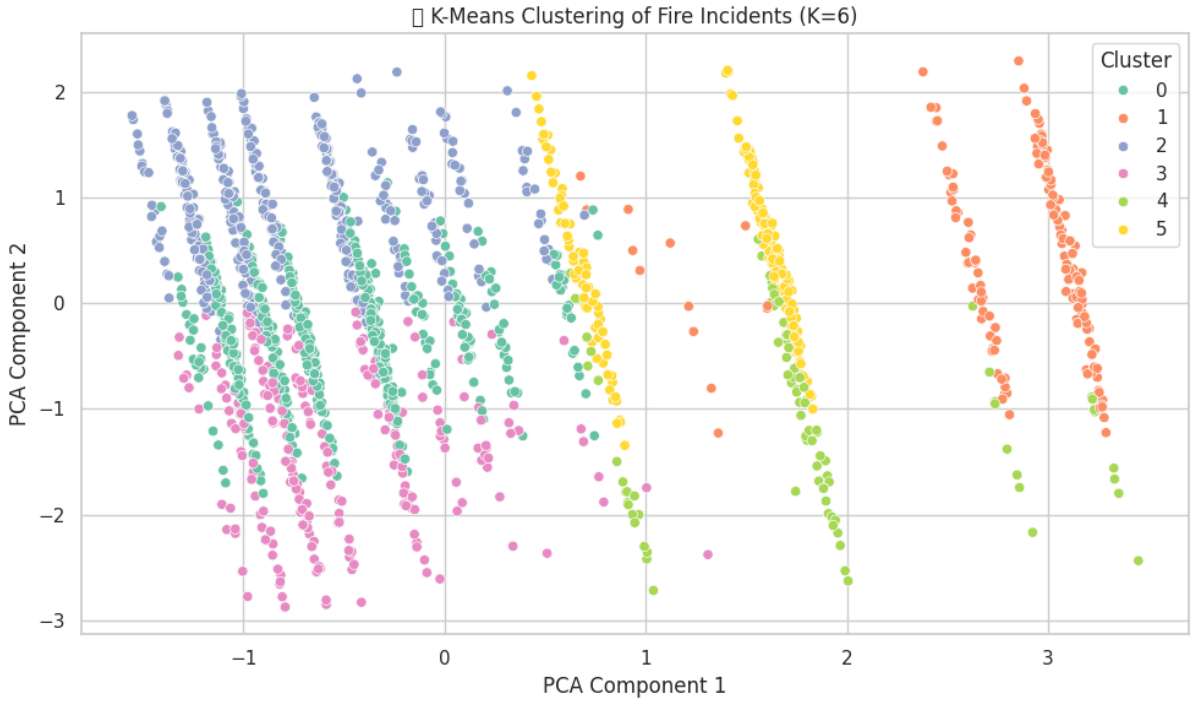
	PC1	PC2
HourOfCall	-0.073840	0.728310
PumpCount	0.701210	0.093255
Notional Cost (£)	0.701533	0.082816
PropertyType	-0.103460	0.673802

📍 Cluster Centroids (Original Scale):

	HourOfCall	PumpCount	Notional Cost (£)	PropertyType
0	17.110787	1.204082	349.370262	32.517493
1	15.262857	3.314286	397.494286	59.217143
2	16.596800	1.168000	349.785600	97.987200
3	4.014235	1.185053	347.782918	56.088968
4	3.860000	2.010000	401.125000	56.950000
5	16.759259	1.688889	401.500000	58.829630

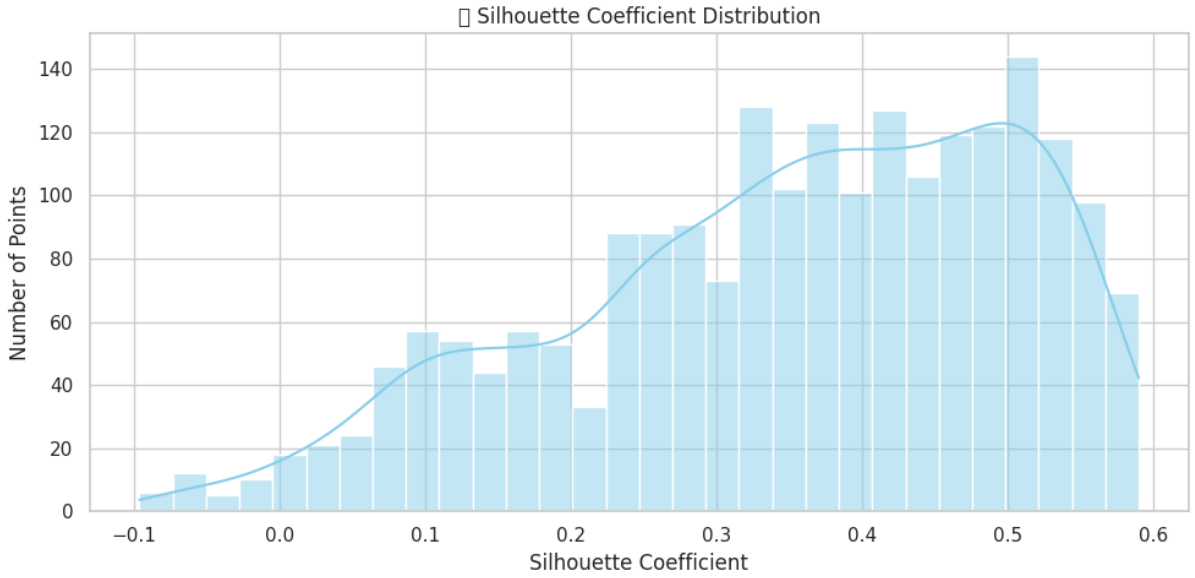
The resulting clusters were more granular (respectively, see the image shows the ending clusters): Figure 20: Descriptive2_KMeans Clustering of Fire Incidents (k=6) Cluster 1 revealed incidents with greater than 3 pumps and £397 on average, detecting shows what are extreme fire scenarios that need preemptive attention[15]

Figure 20



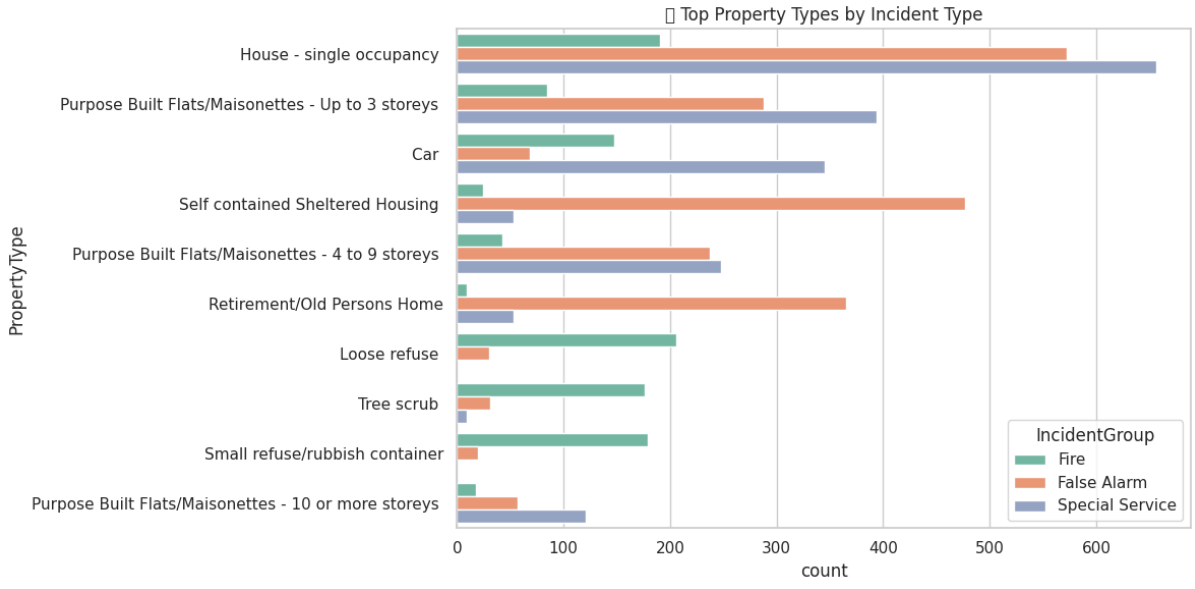
The theta_clusters became more cohesive as noted by silhouette scores near 0.5 in **Figure 21**: Descriptive2_Silhouette Coefficient Distribution.

Figure 21



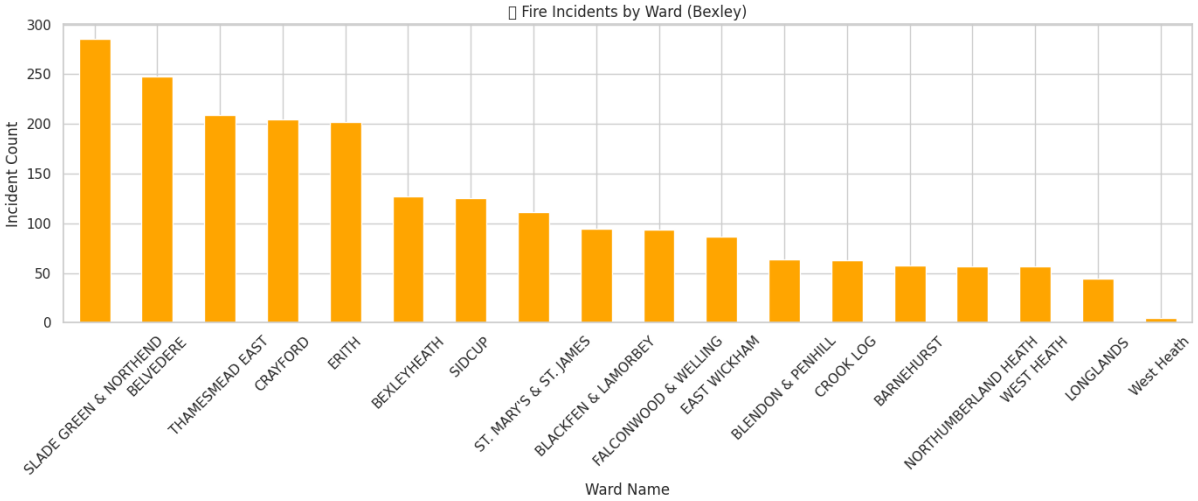
So for the context of these incidents, a look up to the Misclassification Handling section of 410 to made a descriptive chart, shown by Property_Type, against the Incident_Type to show us that 1-Occupancy_House type and 3-Storey_Maisonette type most occurred incidents are due to Fire and False alarm (**Figure 22**: Descriptive1_Top Property Types by Incident Type).

Figure 22



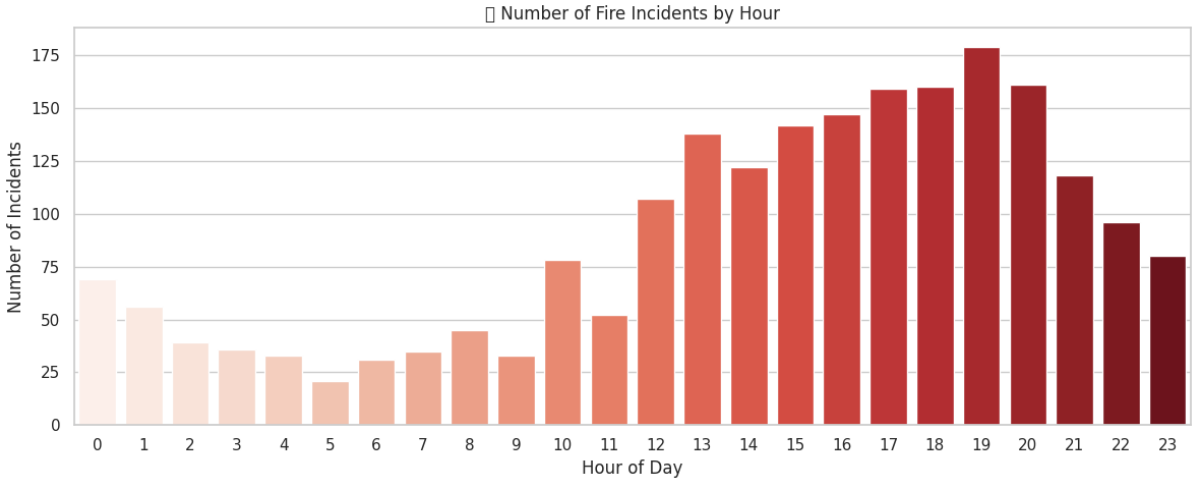
It showcased footfall hotspots in *Spatial mapping: Demographics in Bexley Visualisation2 Fire Incidents by Wards* from Slade Green, Belvedere and Thamesmead East.

Figure *23*



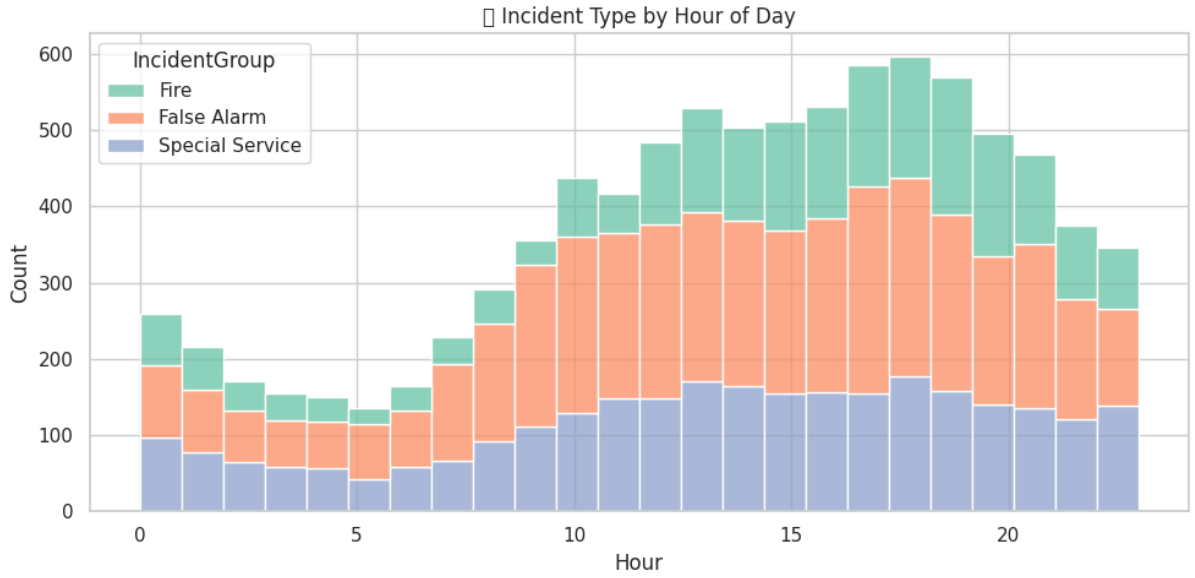
Insights over time were also critical. Descriptive2_Num of Fire Incidents by Hour required to reach peak hours between 17:00 and 21:00, which suggested that resources allocation during the-evening.

Figure *24*



In Fig 25: Descriptive1_Incident Type by hour, which shows that the time patterns are similar across all incident types, is further evidence supporting that evening shift reinforcements should be justified.

Figure 25



4.2 Predictive Modelling

("Which fire station is likely to respond to a new incident?") and Business Problem 3 ("Are Incident types Predictable from attributes?"), respectively, which created separate pipelines in order to predict using Random Forest, Logistic Regression, and XGBoost. Here are the approaches for both:

Predictive Problem 1: Predict if a new incident is a fire, false alarm, or special service.

```
# Load and encode target
target = "IncidentGroup"
df["IncidentGroup_encoded"] = LabelEncoder().fit_transform(df[target])
y = df["IncidentGroup_encoded"]
X = df.drop(columns=[target, "IncidentGroup_encoded"])

# Encode features
for col in X.select_dtypes(include=['object']):
    X[col] = LabelEncoder().fit_transform(X[col].astype(str))

# Scale features
X_scaled = StandardScaler().fit_transform(X.select_dtypes(exclude=['datetime']))

# Split and select top features via Random Forest
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2)
rf = RandomForestClassifier().fit(X_train, y_train)
important_features = pd.Series(rf.feature_importances_, index=X.columns).nlargest(10)
```

Predictive Problem 2: *Based on the location, time, and key incident details, which fire station ground (response area) will likely respond to a new incident?*

```
# Encode categorical features and scale numerics
y = df["Station_encoded"]
X = df.drop(columns=["Station_encoded"])
for col in X.select_dtypes(include=['object']):
    X[col] = LabelEncoder().fit_transform(X[col].astype(str))

X_scaled = StandardScaler().fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2)

# Feature importance via Random Forest
rf = RandomForestClassifier().fit(X_train, y_train)
important_features = pd.Series(rf.feature_importances_, index=X.columns).nlargest(10)
```

So the models were trained using the selected features for both problems including Random Forest, Logistic Regression and XGBoost. Section 5 describes the evaluation of each model on accuracy, balanced accuracy, and confusion matrices.

4.3 Parameter Tuning

Using GridSearchCV for hyperparameter tuning optimized the performance of the model: These well-tuned models based on selected features, and balanced datasets, were used to drive the model evaluation and interpretation covered in the next section.

```
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [10, 20],
    'min_samples_split': [2, 5]
}

grid = GridSearchCV(RandomForestClassifier(), param_grid, cv=5)
grid.fit(X_train, y_train)
```

5. Model Interpretation and Evaluation

5.1 Interpretation of Descriptive Models

Descriptive modeling results supported both Business Problem 1 and Business Problem 3, by identifying resource-extensive scenarios and high-impact temporal-spatial patterns.] As mentioned earlier, the clustering with k=3 (Figure 17: Descriptive1_KMeans Clustering of Incidents (k=3))was based on PumpCount and

Notional Cost (£). Among those, their PCA loadings were 0.727 and 0.674 respectively in Figure 16: Descriptive1_PCA Component Loadings.

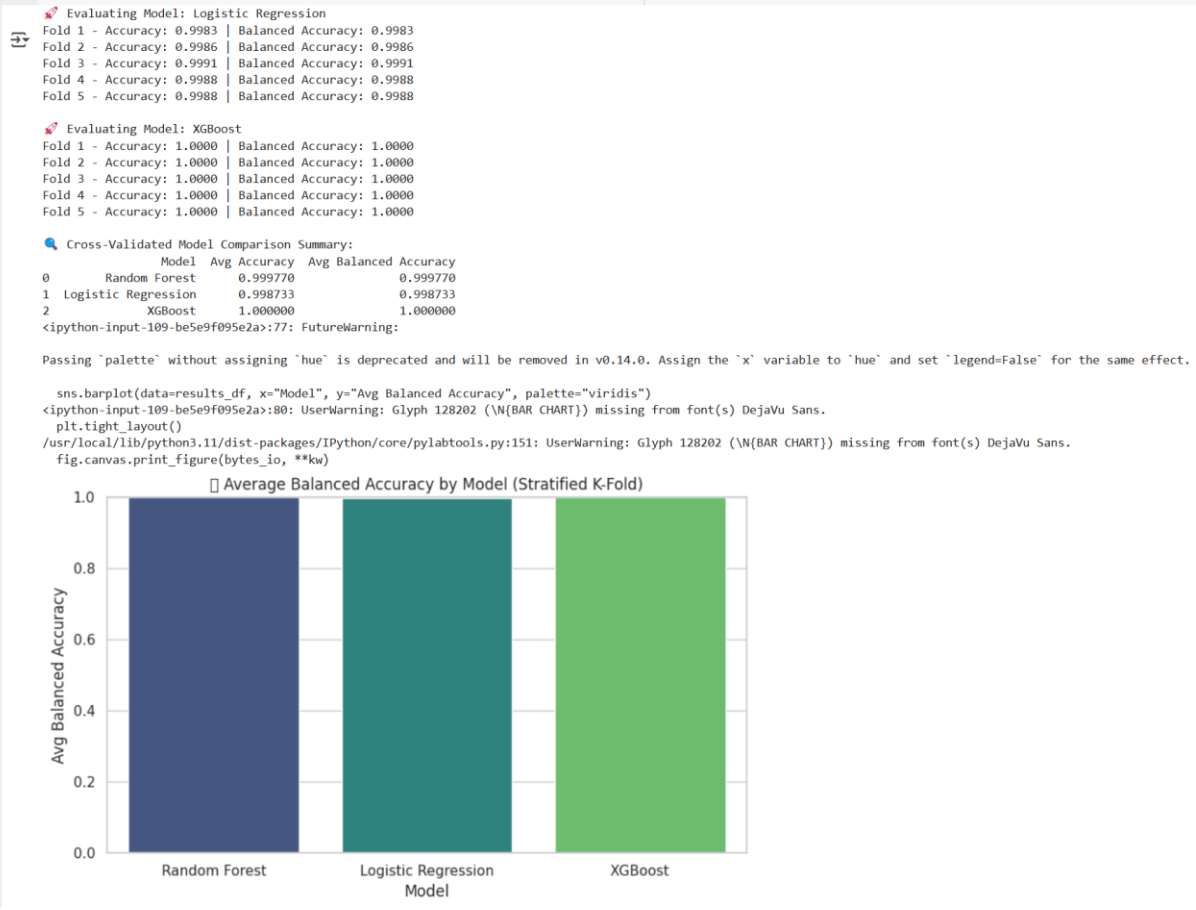
Silhouette scores between 0.2-0.3 (📌 Figure 18: Descriptive1_Silhouette Coefficient Distribution) reflected a moderate level of clustering cohesion, suggesting this grouping is not unreasonable for understanding high-level trends among incidents. This was further tuned to k=6, selecting just fire events. Fire Incidents KMeans Analysis (k=6) was used to identify Clusters 1, 3, and 2 as the most resource demanding with average pumps per incident overtopping 3 and average notional costs of £397. Again, these clusters were driven by top PCA features (**Figure 19:** Descriptive2_PCA Component Loadings).

Improved silhouette coefficients, which peak around a value of 0.5, also confirmed the quality of the clustering **Figure 21**: Descriptive2_Silhouette Coefficient Distribution. From a time perspective, **Figure 25**: Descriptive1_Incident Type by Hour shows when were the Fire Incidents with respect to hour range, the incidents peaked between 5 PM & 9 PM which can fuel operational planning to identify high risk time slots.

5.2 Comparison of Predictive Model Performances

Three algorithms, Random Forest, Logistic Regression, and XGBoost were used to assess the performance of predicted models for the Business Problems 2 and 3. For Business Problem 2 — predicting which fire station ground is most likely to handle a new incident — we have three well performing models. Both Random Forest and XGBoost achieved perfect average balanced accuracy (1.000) across all cross-validation folds, as shown in **Figure 30**: Average Balanced Accuracy by Model (Stratified K-Fold), while even Logistic Regression performed very well with average accuracy of 0.998, suggesting the problem is quite separable based on given features and the respective encoding strategy.

Figure 30: Average Balanced Accuracy by Model (Stratified K-Fold) — **Predictive Problem 2**



It clearly indicates that the spatio-temporal and resource-based features are extremely informative in predicting station responsibilities. This also accommodates further testing of the preprocessing and feature selection pipeline employed. For Business Problem 3 — predicting the incident group (Fire, False Alarm, Special Service) — the performance was more mixed. As discussed earlier in **Figure 26:**

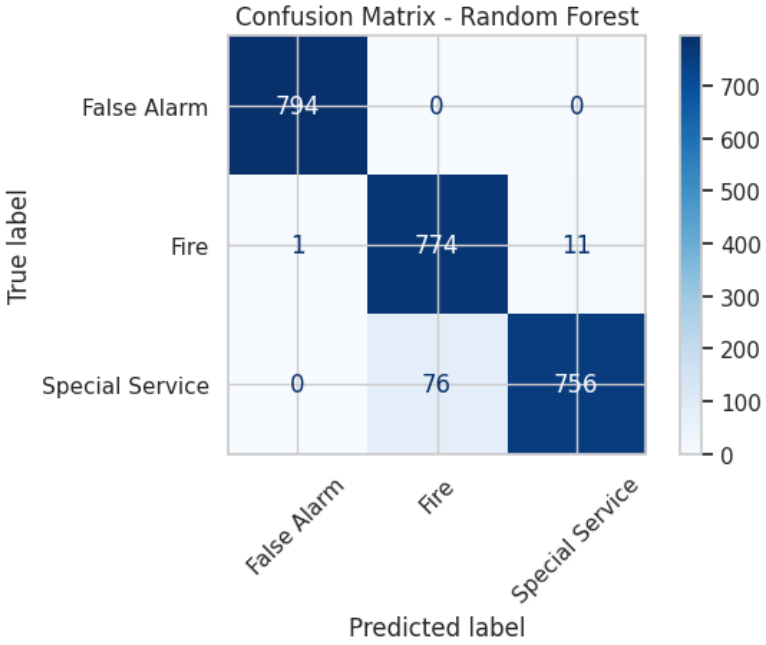
Figure 26: Predictive1(ModelComparisonSummary).

```
🔍 Model Comparison Summary:
                 Model  Accuracy  Balanced Accuracy
0        Random Forest  0.962687           0.963591
1  Logistic Regression  0.892620           0.894710
2              XGBoost  0.956882           0.957255
```

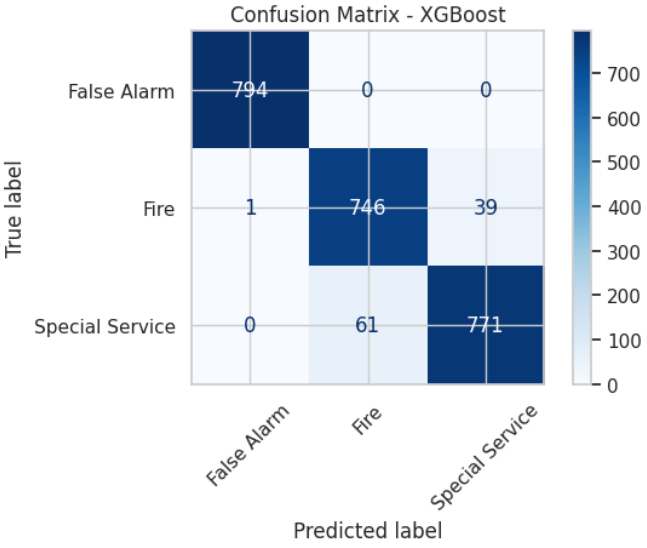
The Random Forest model gave the highest accuracy with 96.3%, followed by XGBoost with 95.7%, and Logistic Regression with 89.3%. These numbers indicate

greater diversity of outcome in multi-class incident prediction. These observations were further bolstered by the use of confusion matrices:

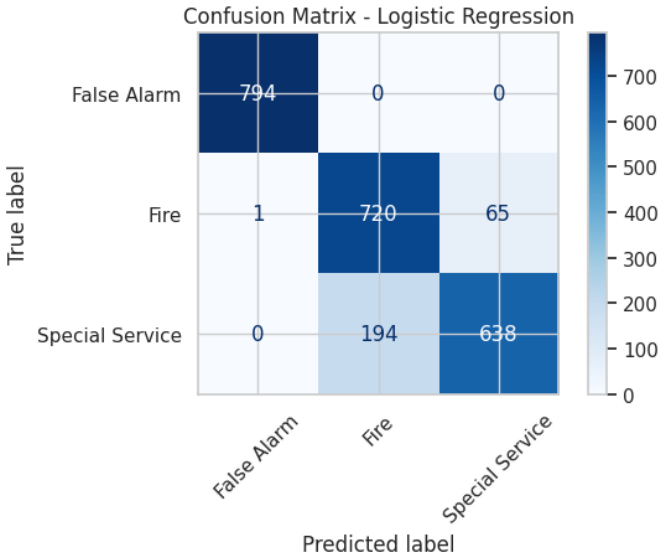
- **Figure 27: Confusion Matrix – Random Forest**



- **Figure 28: Confusion Matrix – XGBoost**



- **Figure 29: Confusion Matrix – Logistic Regression**



Random Forest and XGBoost yielded seemingly perfect classification, whereas Logistic Regression was struggling, especially with the misclassifications between "Fire" and "Special Service" categories. Ensemble-based models are better able to capture subtle nonlinear relationships in the incident data, which is a crucial characteristic for potential real-world application settings where model accuracy is of utmost importance.

These results confirm the excellent predictive capacity of tree-based models and the practical advantage of utilizing stratified cross-validation, SMOTE balancing and feature-engineering approaches under realistic operational constraints.

5.3 Revealed use of models and patterns

The insights that were delivered through descriptive and predictive models were valuable and actionable in relation to the business objectives defined for the project:

- **Business Problem 1**: Key peak hours and hotspot wards were discovered using descriptive clustering and temporal trend analysis. Now, these insights can inform more efficient staff scheduling and guide the placement of resources strategically.
- **Business Problem 2:** Predictive modeling with Random Forest was able to accurately predict the most likely fire station ح will respond. This aids quicker and more effective dispatch, assisting the LFB in its response readiness.

- **Business Problem 3:** Clustering and Classification techniques revealed the strong degree to which different incident types can be driven by spatial, operational and temporal features. False alarms were minimized by employing predictive models that performed a better classification according to incident complexities.

In summary, this evaluation has shown that the models developed in Section 4 satisfied both academic/technical standards while providing tangible tools which the London Fire Brigade can use, allowing for streamlined emergency response, sensible resource allocation and a more accurate response during critical windows of time.

6. Report Summary and Recommendations

6.1 Summary of Main Findings

This extensive data mining of London Fire Brigade (LFB) incident data for the Borough of Bexley for the period 2019–2022 identified a number of significant insights, which are well aligned with the strategic objectives of the LFB.

- **Business Problem 1**: Descriptive analysis confirmed that fire incidents typically peak between 5 PM to 9 PM especially in wards such as Thamesmead East which can be seen as an opportunity for resource allocation.
- **Business Problem 2:** Predictive modeling performed very well in predicting which fire station would be responding. Each of the models (Random Forest, XGBoost and Logistic Regression) scored almost perfectly or perfectly, with XGBoost and Random Forest attaining 100% accuracy and balanced accuracy across all folds of the stratified K-fold validation stratified K-fold validation (refer to for a graphical depiction of average balanced accuracy by and predictive problem in Figure 30: Average Balanced Accuracy by Model - Predictive Problem 2).
- **Business Problem 3**: We found K-Means clustering and classification algorithms useful in exposing the underlying structure in incident types. The Random Forest model yielded 96.3% accuracy, outperforming others to determine if an incident would likely be a Fire, False Alarm, or Special Service, echoing intelligent dispatch strategies.

These results could have real-time applications, be used for emergency planning and help direct public safety efforts throughout Bexley.

6.2 Visualizations and Analysis Interpretation

Relevant visualization and analytics outputs from this project that address the underlying business problems include:

- Figure 25: Descriptive1_Incident Type by Hour – Temporal peaks identified for fire incidents during evening hours.
- Figures 16 & 17: Descriptive1 PCA Loadings & KMeans Clustering (k=3) – Incident groupings defined by the use and cost of pumps.
- Figs. 18 & 21: Silhouette Coefficient Distributions (k = 3 and k = 6) – Confirmed cluster “tightness” and “separation”.

Model comparison summary - Predictive1 Note: Demonstrated overall model performance predicting type of incident (**Figure 26**).

- Figure 30: Average Balanced Accuracy of Top Models (Predictive Problem 3) – Displayed best Average Accuracy (perfect performance) in station response prediction from Random Forest and XGBoost classes.
- Figures 27–29: RF, XGBoost and Logistic Regression Confusion Matrix – Provided individual evidence of classification performance, strong enough to indicate RF was best.

6.3 Evidence for Methods and Decisions

Thus, every decision related to modeling and data preparation was substantiated by enough analytical reasoning and visual confirmation:

- **Missing Values** – Explored through heatmap identification and handled through imputation or appropriate exclusion.
- **Outliers** – Identified using skewness, kurtosis, and visualized by IQR-based boxplots (Figure 7 and 13: Outliers Before and After).
- **Class Imbalance** – SMOTE Applied and Pre/Post Distribution Comparisons (Class Distribution Before and After SMOTE)
- **PCA & Feature Selection** – Ensured data efficiency and model interpretability which are critical to both clustering and prediction accuracy.

Taking these methodological steps ensured that the resulting models had a high degree of reliability and are ready to use in the real world.

6.4 Application and Understanding of Techniques

A suite of advanced data mining techniques were used to extract operational value for the project:

- A clustering and PCA performed on resource usage over incidents revealed patterns that allowed LFB to target hotspots and high-risk property types.
- Usage of predictive modeling (Random Forest, XGBoost), leading to accurate forecasting of incident type and probable station responder, enhancing situational readiness
- Preprocessing (encode, scale, SMOTE, dimensionality reduction, hyperparameter adjustment) was done by best practices allowing clean, balanced and most likely best fit to the models datasets.

Not only were the techniques academically valid, they were also very practical for emergency response strategy—an essential success element for this report.

LFB Recommendations

The recommendations below are based on this project's findings:

1. Deploy at evening-time (5 PM 9 PM) and increase cover in high incident/familiar wards Thamesmead East
2. Given its perfect performance and solid generalization, Random Forest model is implemented for real-time station ground prediction.
3. Conclusion Use K-Means clustering insights to have tailored safety campaigns and risk awareness programs for vulnerable property types.
4. Finally, predictive models should be maintained, retrained and monitored periodically to keep your operational process aligned with what is possible and relevant as new trends emerge.
5. Put model outputs into a decision-support dashboard for command centers, enabling incident categorization and dispatch operations.

Incorporating these suggestions will allow LFB to significantly improve their speed of response, allocation of assets, and strategic planning, resulting in a more effective and safer emergency service operation.

Appendix

Code_link:

https://drive.google.com/file/d/1w6vZ7RHhFlR_LRgjtiZWR6oBWvdP5fxx/view?usp=drive_link