Advanced Data Management – Using Pig/Hadoop to Analyse Datasets

In 2017, London experienced one of its worst fires in the Grenfell Tower tragedy. With the fire at Grenfell came a renewed focus on how London Fire Brigade operates. The motivation behind selecting this dataset was to analyse data in PIG in order to provide recommendations to London Fire Brigade to help them improve their overall service and to help make the London fire services more efficient, in order so London does not see another tragedy. From the analysis of the two datasets, It is hoped that key recommendations can be made to the London Fire Brigade in order to help them improve their operations such as where to allocate staff at any given time, how key Borough statistics (such as population and housing densities) affect the number of callouts they receive and how to focus efforts in order to reduce the amount of False Alarms the London Fire Brigade receives each year.

The datasets chosen are as follows:

Dataset 1 – London Fire Brigade Calls - https://www.kaggle.com/jboysen/london-fire

This dataset includes data pertaining to every single call made to the London Fire Brigade for the first 4 months of 2017. For the analysis conducted, data including which Borough the call was made to (e.g. Camden), what type of incident (e.g. False Alarm) and the Hour of the call was obtained from this dataset and utilised as necessary. The data contained a large amount of unnecessary columns from the get-go so only key columns were imported to use in PIG for analysis.

After pre-processing the columns (pre-processed data CSV uploaded as londonfiredata.csv), the data obtained from the dataset for use in Apache Pig was structured as follows:

Attribute	**Type**	**Meaning**
borough_code_lfd	chararray	London Borough Code
borough_name_lfd	chararray	London Borough Name
cal_year	chararray	Calendar Year of Callout
date_lfd	chararray	Date of Callout
Hour	chararray	Hour of Callout
incident	chararray	Incident Type (False Alarm, Special Service or Fire)

This dataset interests me as it provides detailed by-the-hour data pertaining to the callouts that London fire brigade receive per Borough. This dataset can be used to investigate the distribution of callouts for the whole of London/for each Borough in order to decide where to properly allocate resources at any given time, as well as to link the callouts per Borough to other datasets.

Dataset 2 – Housing in London - https://www.kaggle.com/justinas/housing-in-london?select=housing_in_london_yearly_variables.csv

This dataset includes Borough information for each of the Boroughs in London. Originally the dataset was created for analysing housing data in each Borough. It includes key metrics such as population per Borough, number of houses per Borough etc. For this analysis, only data pertaining to 2017 was utilised.

After pre-processing (pre-processed data CSV uploaded as boroughVariables.csv), the data obtained from the dataset for use in Apache pig was structured as follows:

Attribute	**Type**	**Meaning**
borough_code_lbd	chararray	London Borough Code
borough_name_lbd	chararray	London Borough Name
date_lbd	chararray	Date (1st Jan 2017)
median_salary	chararray	Median Salary of Borough
life_satisfaction	chararray	Life Satisfaction of Borough
mean_salary	chararray	Mean Salary of Borough
recycling_pct	chararray	Recycling Percent of Borough
population_size	int	Population Size of Borough
number_of_jobs	chararray	Number of Jobs in Borough
area_size	int	Size of Borough
no_of_houses	int	Number of Houses in Borough

This dataset interests me as it provides a way of linking key borough statistics to the London Fire Brigade Calls dataset in order to analyse how factors such as population density and housing density affect the callouts of the fire service to each Borough. Factors such as population density and housing density can be obtained from this dataset by using the Borough area size, population size and number of houses for the year 2017.

The two datasets both include the Borough code, ensuring that they can be joined together by code for further analysis and insights.

The three analysis tasks conducted on the datasets are as follows:

Analysis 1 – Join the two data sets together and investigate the relationship between the number of callouts London Fire Brigade receive per Borough and the Population/Housing Densities of each Borough. Perform a calculation using area size, total population and overall number of houses per Borough to obtain these two densities for analysis.

The motivation of analysis 1 is to investigate how the population density and the housing density of each London Borough affect the number of callouts the London Fire Brigade receives for each Borough. Based on the information obtained, recommendations can be made to London Fire Brigade on where to allocate resources.

The raw results of analysis 1 were as follows:

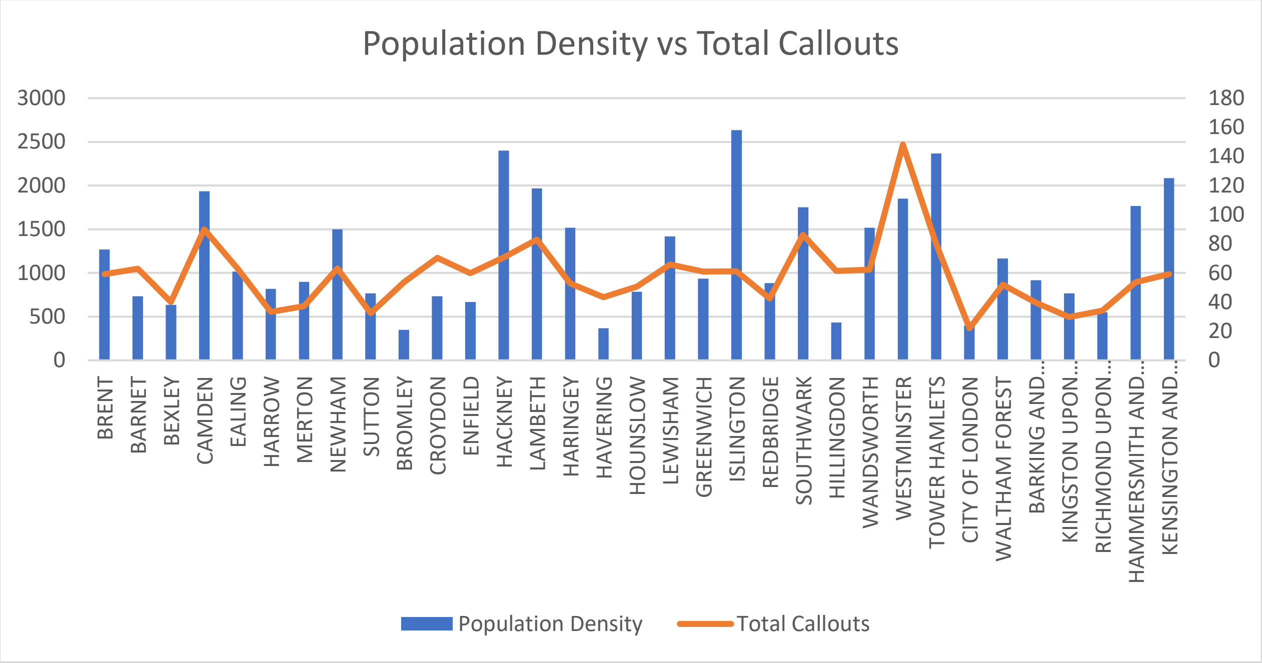
The format for the following is (Borough Name, Population Density, Housing Density, Total Callouts)

```
(BRENT,76,27,988)

(BARNET,44,17,1049)

(BEXLEY,38,15,666)

(CAMDEN,116,47,1499)

(EALING,61,23,1054)

(HARROW,49,17,555)
```

```
(MERTON,54,22,621)

(NEWHAM,90,29,1053)

(SUTTON,46,18,538)

(BROMLEY,21,9,893)

(CROYDON,44,18,1174)

(ENFIELD,40,15,996)

(HACKNEY,144,57,1177)

(LAMBETH,118,50,1383)

(HARINGEY,91,36,879)

(HAVERING,22,8,723)

(HOUNSLOW,47,17,845)

(LEWISHAM,85,35,1097)

(GREENWICH,56,22,1017)

(ISLINGTON,158,68,1019)

(REDBRIDGE,53,18,709)

(SOUTHWARK,105,44,1435)

(HILLINGDON,26,9,1024)

(WANDSWORTH,91,40,1036)

(WESTMINSTER,111,55,2469)

(TOWER HAMLETS,142,54,1330)

(CITY OF LONDON,24,20,367)

(WALTHAM FOREST,70,26,867)

(BARKING AND DAGENHAM,55,19,659)

(KINGSTON UPON THAMES,46,17,496)

(RICHMOND UPON THAMES,33,14,569)

(HAMMERSMITH AND FULHAM,106,50,893)

(KENSINGTON AND CHELSEA,125,70,988)
```

In absence of a Pig Data Visualisation Tool (Such as Apache Zeppelin), excel can be used to visualise the results/output of the PIG script for interpretation and recommendation purposes.

On the following page is a graph of the Population Density vs The total amount of Callouts for each Borough obtained from the analysis using PIG.

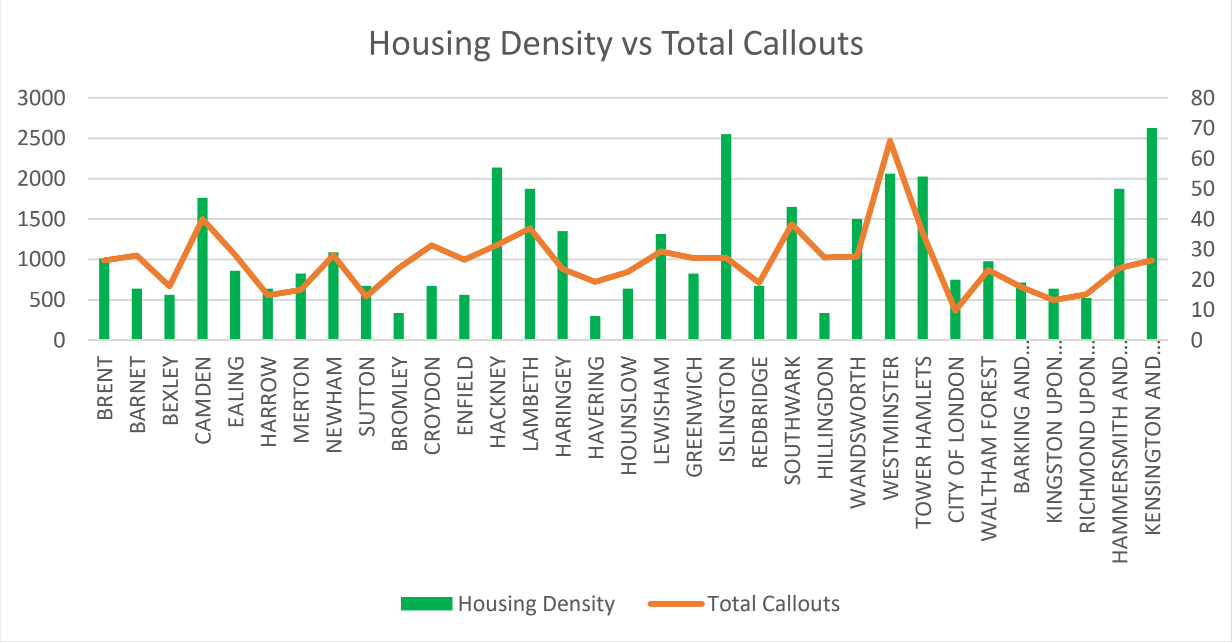


The graph above provides a visual way to interpret the results obtained from the PIG script. It can be seen that the total number of callouts trends with the overall population density of the Boroughs – however there are a few key outliers which are as follows:

Westminster – It can be seen that the Borough of Westminster experienced a large amount of callouts relative to its Population density.

Islington – It can be seen that the Borough of Islington experienced a low level of callouts relative to its high Population Density.

Below is a graph of the Housing Density vs The total amount of Callouts for each Borough.



Similar results as the previous graph are obtained with Westminster again displaying a large amount of total callouts relative to its overall housing density as well as Islington displaying a low amount of callouts relative to its density. Interestingly, Kensington and Chelsea (the Borough that Grenfell tower is in) highlights a similar trend to Islington – namely that there is a fairly high Housing density compared to the total number of callouts.

From this analysis a few conclusions can be drawn

1. Westminster experiences a large amount of callouts relative to its population and housing density. This is likely due to Westminster's popularity with tourists and the amount of commercial premises in Westminster. It's important that London Fire Brigade recognises that due to Westminster's non-permanent footfall and large amount of commercial premises, a large amount of calls relative to the overall population/housing density is likely to occur here.
2. Islington and Kensington/Chelsea have high population and housing densities relative to their low amount of total callouts. As we have seen since with the Grenfell tower incident, this may not necessarily suggest that everything within these Boroughs is going well – London Fire Brigade will need to monitor for changes in this trend since Grenfell and ensure they do not ignore these areas based on good statistics.

Analysis 2 – Investigate the Distribution of callouts over the course of 24 hours for the whole of London and for the key tourist Boroughs of Westminster and Camden, making a recommendation for where London Fire Brigade should allocate resources.

The motivation of analysis 2 is to investigate how callouts differ for each hour for the whole of London and for key tourist Boroughs over the course of a day. Combined with the analysis conducted in analysis 1, this will further help London Fire Brigade allocate resources efficiently.

The raw results of analysis 2 were as follows:

The format for the following is (Hour, London Total Callouts, Westminster Callouts, Camden Callouts)

(0,976,83,52)

(1,879,85,47)

(2,645,56,41)

(3,549,45,27)

(4,485,27,25)

(5,510,32,33)

(6,653,55,34)

(7,876,71,53)

(8,1207,102,62)

(9,1377,136,69)

(10,1525,152,82)

(11,1641,127,62)

(12,1671,126,61)

(13,1780,141,76)

(14,1805,126,85)

(15,1871,151,83)

(16,1880,110,85)

(17,1942,146,71)

(18,2187,158,92)

(19,1950,143,81)

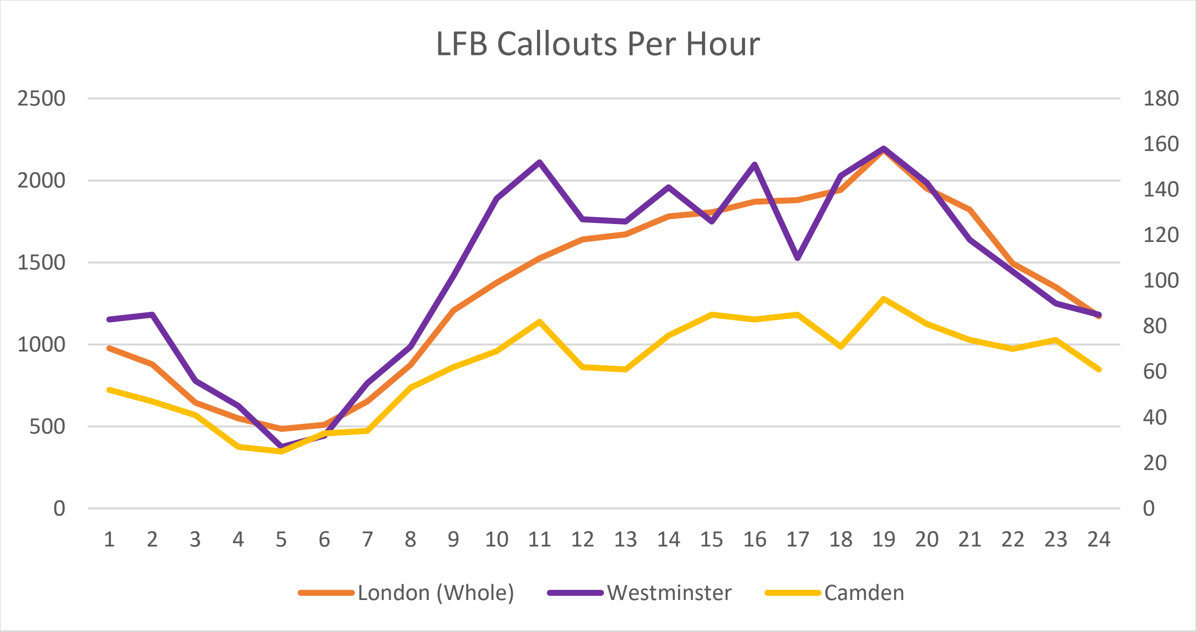
(20,1822,118,74)

(21,1494,104,70)

(22,1350,90,74)

(23,1172,85,61)

Below is a graph of the callouts per hour for the whole of London, Westminster Borough and Camden Borough over the course of an average 24 hour day.



It can be seen from the graph above that relative to the whole of London, Westminster experiences a large amount of callouts between 9am and 11am, peaking at around 11am. Camden experiences a similar trend to Westminster. All areas of London follow a trend of dipping to their lowest callouts around 5am in the morning and peaking around 8pm in the evening.

From this analysis a key conclusion can be drawn

1. London Fire Brigade should allocate more resources steadily throughout the day for the whole of London, from a low point at 5am, to peaking at around 8pm. The exceptions to this rule are

in tourist Boroughs like Westminster where higher resources should be allocated earlier in order to cover the 9am to 11am callouts.

Analysis 3 – Investigate the rate of False Alarm callouts per Borough and make recommendations on where efforts should be focused on tackling False Alarm callouts.

London Fire Brigade has a problem with False Alarm callouts – the motivation of Analysis 3 is to investigate the Boroughs with the most False Alarm callouts in order to provide recommendations to London Fire Service on which Boroughs to target in order to reduce False Alarms and to ensure London Fire Brigade makes the best use of its resources.

The raw results of analysis 3 were as follows:

The format for the following is (Borough Name, False Alarm %, Special Service %, Fire %)

```
(CITY OF LONDON,79,14,8)

(WESTMINSTER,67,19,14)

(KENSINGTON AND CHELSEA,62,28,10)

(HAMMERSMITH AND FULHAM,60,28,13)

(CAMDEN,59,28,14)

(KINGSTON UPON THAMES,55,26,19)

(RICHMOND UPON THAMES,54,27,19)

(LEWISHAM,52,32,17)

(BARNET,51,29,20)

(HARROW,51,28,21)

(SUTTON,51,25,25)

(ISLINGTON,50,33,18)

(EALING,50,28,22)

(SOUTHWARK,47,36,17)

(HILLINGDON,47,27,26)

(MERTON,47,33,20)

(WALTHAM FOREST,47,27,26)

(HOUNSLOW,46,31,23)

(HACKNEY,46,39,15)

(WANDSWORTH,46,33,20)

(HAVERING,45,28,28)
```

(BRENT,45,32,23)

(HARINGEY,45,35,20)

(REDBRIDGE,44,32,24)

(TOWER HAMLETS,43,36,21)

(CROYDON,42,34,23)

(GREENWICH,42,34,24)

(LAMBETH,42,44,14)

(BEXLEY,42,28,30)

(BROMLEY,40,30,30)

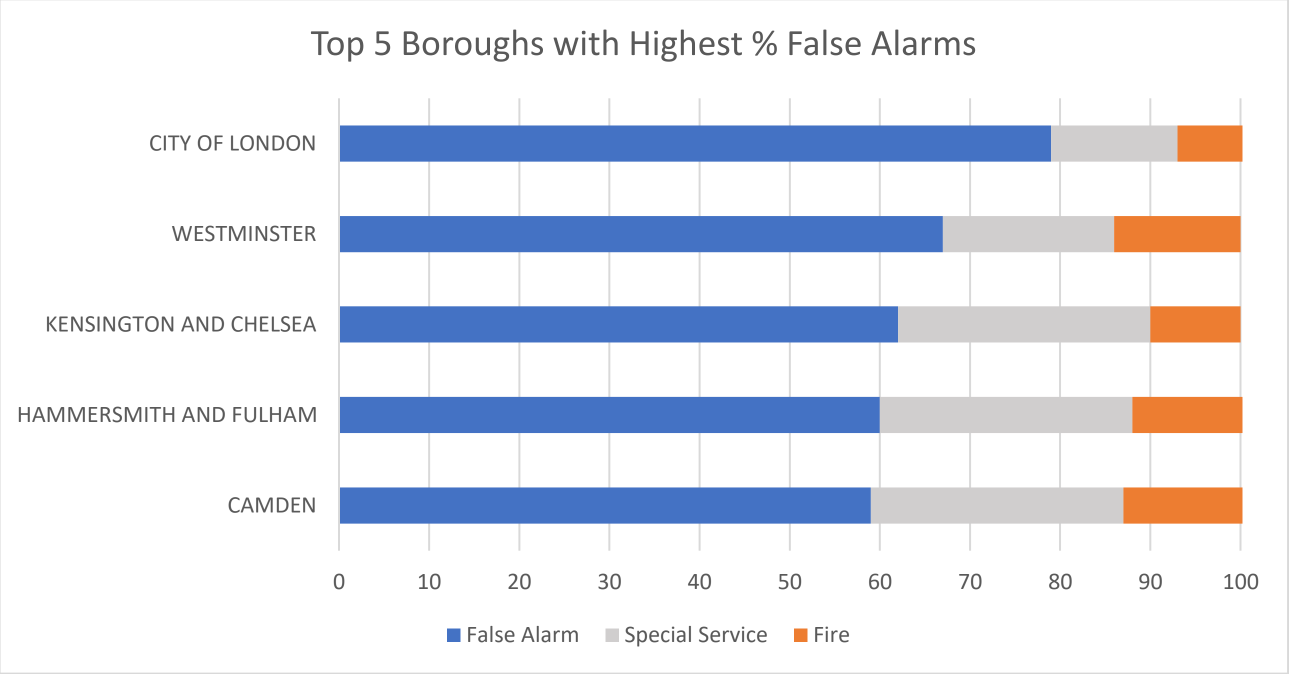
(ENFIELD,38,39,23)

(NEWHAM,36,42,22)

(BARKING AND DAGENHAM,35,34,30)

(NOT GEO-CODED,35,30,35)

Below is a chart highlighting the Top 5 Boroughs that have the Highest Percentage of False Alarm calls.



It can be seen from the chart that the City of London, followed by Westminster and then Kensington/Chelsea have the highest proportion of False alarm calls. The City of London has an exceptionally high proportion of False alarm calls. The City of London is of course famed for being the financial capital of the UK so it is likely that the high False Alarm rate is due to the amount of office premises in this area.

Interestingly again, the Borough of Kensington and Chelsea, which houses Grenfell Tower, is the 3rd highest Borough with the most False alarms. This indicates that Boroughs that have high false alarm rates can still experiences fires that require huge rapid response.

It is crucial that London Fire Brigade use these results obtained from the PIG script in order to target their reduction in False Alarm callouts in order to ensure that key firefighter resources are not wasted. By targeting these areas and reducing the overall number of False Alarm callouts (maybe by expanding the commercial fine scheme that London Fire Brigade has started to implement – where commercial premises that have too many false alarms are fined), more firefighters can be ready for if there is an actual fire.

Key Recommendations/Conclusions to the London Fire Brigade

With the 3 analysis conducted via PIG scripting and the interpretation of the results visualised and discussed, key recommendations can be made to the London Fire Brigade in order to improve their service. They are as follows:

1. Ensure there is adequate knowledge of the situation in the Borough of Westminster where a large amount of callouts are expected relative to its population/housing density – mainly due to there being a large amount of tourism and commercial property in the area. It must be ensured that there is adequate staff available to deal with the level of calls coming from the Westminster Borough, in order not to draw key services away from other Boroughs.
2. London Fire Brigade should ensure that their staffing levels ramp up through the day, starting from a low point at around 5am in the morning, peaking at around 8pm in the evening. London Fire Brigade should ensure that key tourist and business Boroughs are adequately staffed in the morning around 9am to mid-day, relative to other quieter areas.
3. London Fire Brigade should target the Boroughs of “City of London”, “Westminster”, “Kensington and Chelsea”, “Hammersmith and Fulham” and “Camden” for programs that will reduce the overall amount of False Alarm callouts to the service. This will ensure that key resources aren’t being used up unnecessarily and are able to respond to real threats.