Completed for the Google Data Analytics Certification

September, 2021- Hunter Faulkner

Introduction

Precursory information: This is not a research paper. This is a challenge to work with a single dataset and produce basic investigative/exploratory results.

This report is structured specifically using Google's method for data analytics as taught on Coursera. The resulting report is a capstone project to demonstrate my skills attained from completing Google's course and becoming a certified data analyst. I chose the topic, the metric's and the dataset using Kaggle. The report is done over a short period and is specifically timed to test quickness and efficiency. The main idea was to work with a single chosen dataset and come up with the best exploratory analysis possible with the available resources and timeframe. On reflection, I would have included more data and research into the topic if the time frame were longer than several days. As such, the Act phase of this report would realistically include a cyclical repeat of this process to conduct further analysis and data collection. This practice would produce more accurate and meaningful results.

The tech stack I will use includes BigQuery (SQL), the BigQuery API for accessing the datasets, and R as directed by the Google course. Python is also used because I have more experience with it and would like to make a heatmap for the geo-spatial data.

Topic

I chose to use a dataset on Crime in London Boroughs and a supplementary dataset on the population of each borough for the time period specified. The makeup of each dataset will be detailed in a future section of the report. The set was easily accessed and is public on BigQuery's API. It is also listed on kaggle, where I originally located it.

Analyzing crime in London will allow for the demonstration of many important techniques that I have learned in this course as well as a bit of further practice in crime economics (in which I have previous experience as a research assistant). The remainder of the report will outline the process and discuss each topic as if reporting to stakeholders. In reality, such insights would be useful to the city of London and law enforcement officials including any governing bodies concerned with crime. The key metrics and sample presentation demonstrate how I would approach a real world project on the job for these specific tasks and datasets. For the purposes of using this data, the most recent year in the set is 2016. Therefore, these particular results will be outdated, but the analysis could be easily applied to more recent data to provide useful insights for the present.

Table of Contents

London Crime Data (BigQuery Dataset)

Metadata

Usage

License: CC0: Public Domaininfo
Visibility: Public

Source

https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

Version

1.0 - 2018-04-20 (No Updates Expected)

Size

1.1 GB

Shape

7 Cols x 13.5 M Rows

Variables (Columns)

1. **ISOA_CODE** - Lower Layer Super Output Area code according to the Office of National Statistics
2. **borough** - Neighborhood of London
3. **major_category** - General Crime Category
4. **minor_category** - Specific Crime Category
5. **value** - Count of Crimes in each per month/year of each major_category/minor_category
6. **year**
7. **month**

Observations (Rows)

13.5 Million

Location

Google BigQuery API - London Data Authority
Reported Crime in London by borough and LSOA

Data Context

Variables of Context

1. borough
2. major_category
3. minor_category
4. year

Years - 9 Total

2008 - 2016

Boroughs - 33 Total

Barking and Dagenham
Barnet
Bexley
Brent
Bromley
Camden
City of London
Croydon
Ealing
Enfield
Greenwich
Hackney
Hammersmith and Fulham
Haringey
Harrow
Havering
Hounslow
Islington
Kensington and Chelsea
Kingston upon Thames
Lambeth
Lewisham
Merton
Newham
Redbridge
Richmond upon Thames
Southwark
Sutton
Tower Hamlets
Waltham
Wandsworth Forest
Westminster

Data Context Continued

Major & Minor Categories of Crimes

1. Burglary
    a. Burglary in a Dwelling
    b. Burglary in Other Buildings
2. Criminal Damage
    a. Criminal Damage To Dwelling
    b. Criminal Damage To Motor Vehicle
    c. Criminal Damage To Other Building
    d. Other Criminal Damage
3. Drugs
    a. Possession of Drugs
    b. Drug Trafficking
    c. Other Drugs
4. Fraud or Forgery
    a. Counted per Victim
    b. Other Fraud & Forgery
5. Robbery
    a. Personal Property
    b. Business Property
6. Sexural Offences
    a. Rape
    b. Other Sexual
7. Theft and Handling
    a. Theft from Shops
    b. Theft/Taking of Pedal Cycle
    c. Theft/Taking of Motor Vehicle
    d. Theft From Motor Vehicle
    e. Other Theft Person
    f. Other Theft
    g. Motor Vehicle Interference & Tampering
    h. Handling Stolen Goods
8. Violence Against the Person
    a. Common Assault
    b. Assault with Injury
    c. Harassment
    d. Murder
    e. Offensive Weapon
    f. Other Violence
    g. Wounding/GBH
9. Other Notifiable Offenses
    a. Going Equipped
    b. Other Notifiable

Ask

Organization Goals (Stakeholders)

Target and reduce crime

Insights

1. Identify Crime Trends
    a. Regional Crime Growth
    b. Type of Crime Growth
2. Define Useful Metrics for Exploring Crime Data
    a. Extract Insights of Data
    b. Determine Monthly Effects (if any)

Key Stakeholders / Audience

(Theoretically) London Police and London's Governing Bodies

Topics of Investigation

Are certain months more prevalent for crime or certain types of crime?

Which areas are experiencing crime growth/decline?

In areas with rising crime, are there any correlations between types of crime? For example, correlated rise in Theft/Burglary/Robbery

Key Metrics

1. Crimes per month or year
2. Crimes per month per borough
3. Change in Crimes per month per borough
4. Correlation of changes per month per borough
5. Correlation of changes per month per borough per crime
    a. By Major Category of Crime
    b. By Minor Category of Crime

Prepare

Timeline

One Week

1. Plan & Process
2. Clean and Wrangle
3. Analyze
4. Visualize
5. Write Report
6. Present to Stakeholders

Access

Public Data and Analysis

Key Questions

1. What are the ongoing trends (if any) in criminal activity across the London Metropolitan area?

2. Which borough has the highest crime rate?
   What trends have taken place there over the time period?

3. What is the overall likelihood of being victimized by a crime?

4. Are there any regional, rising-crime trends that could be identified as areas of interest for targeted policing strategies?

Mock Deliverables to Stakeholders

1. Report on findings and documentation on cleaning process

2. Visualizations illustrating key trends

3. Slide presentation on findings for stakeholder meeting

4. Recommendation for action to stakeholders based on findings

Process

Cleaning

Cleaning the data is mostly unnecessary as it is fairly clean. One thing to keep in mind is that the data for the “City of London” borough is sparse and seemingly incomplete. It will be mostly excluded in the analysis. Otherwise, the data is well organized. This probably has to do with the nature that it is a public dataset provided by a government organization. There are few missing data points in the set so it is nearly complete. Note, the size of the dataset (13.5 M) is somewhat misleading due to the fact that the value variable detailing the number of occurrences of a given crime during a given month is **NOT** non-zero. A row in the dataset is used for 0 values in which no such crime occurred during that given month. Such rows will not contribute to any sums.

Wrangling, Purpose-Oriented Approach

Given that calculating the chosen metrics for the analysis is fairly simple, the bulk of the work lies in processing the data.

This includes writing the appropriate queries via BigQuery, but is relatively simple as well. Keeping the key metrics in mind, I extract counts across each subset of concern for the comparative analysis. This includes summing the occurrences of a crime across boroughs for both months and years to check for growth trends. Next, each month can be analyzed for time effects (if any). Correlation will also be helpful to determine if related categories of crime are moving together. Finally, supplementary data on each borough’s population will be added to calculate citizens’ likelihood of being victimized by a given criminal act. This is done be comparing the number of crimes to the relative population of a borough. However, a figure that might be more accurate is the likelihood of being victimized across the entire metropolitan given citizens’ mobility or local commuting.

All queries and code will be included in the Github Repository for this project.

Process

Ethical Management of Sensitive Data

For private data, the processing step would also first entail the securitization of the raw data for the duration of the analysis. This includes restricting access and implementing a controlled environment for any personal information. This dataset does not include any such sensitive or personally identifiable data.

Queries and Wrangling Examples - Implemented in BigQuery

(Query for the Number of each Minor Category of Crimes Reported per Month by Borough)

```
SELECT
    borough,
    year,
    month,
    major_category,
    minor_category,
    SUM(value) AS no_crimes
FROM `bigquery-public-data.london_crime.crime_by_lsoa`
GROUP BY
    minor_category,
    major_category,
    month,
    year,
    borough
ORDER BY
    borough,
    year,
    month,
    major_category,
    minor_category
```

In terms of Queries, all others were a slight variation of this. For Example, total crimes per month by borough would exclude the major and minor category variables and group by the borough, then year, then month; which is displayed in figure 1. Other than this, other wrangling is done in the calculation phase during the Analyze step.

Analyze & Share

Procedure

After wrangling the data appropriately to form several subsets of particular interest, I exported each as a CSV and imported these into R and Python. There, I completed each visual and analysis of the data. The Figure 1 code was used to produce the metrics and graphs in R. R and Python were both used for practice, but my experience with Python exceeds R. The heatmaps in particular were made with Python for ease-of-use relative to my experience.

Following the export of the query results from BigQuery, I used Pandas to further wrangle the data as well. This was done using both join and masking to filter for what was necessary. I primarily analyzed the outlier Westminster and its crimes of concern. Namely, these were thefts and violent crimes which were examples that are would be of particular interest to policing efforts given the data.

Figure 1 Code (R)

Total Crimes per Month by Borough

```
ggplot(data = total_permonth,
       aes(x = date,
           y = no_crimes,
           group = borough)) +
  geom_line(aes(color=borough)) +
  geom_text(x = 2015,
            y = 4500,
            label = "Westminster",
            size = 8) +
  geom_text(x = 2015,
            y = 250,
            label = "City of London",
            size = 8) +
  labs(
    title="Total Crimes per Borough per Month",
    subtitle="2008-2016",
    x = "Date",
    y = "Number of Reported Crimes",
    color = "Boroughs of London"
  ) +
  theme_minimal()
```

Analyze & Share

Image 2 Code (Python)

Total Crimes per Month by Borough with Geo Data

```
#%% Join Map with Crimes per year by Borough

merge_mapcrimes = londonmap.set_index('NAME').join(
total_peryear[total_peryear['year']==2016].set_index('borough'))

#%% Create Heatmap of London Crimes
fig, ax = plt.subplots(1, figsize=(32, 28))
plt.title("Total Crime in London Boroughs\n2016",
          size=40)
merge_mapcrimes.plot(column='no_crimes',
                     cmap="Reds",
                     ax=ax,
                     linewidth=1,
                     edgecolor='0')
ax.axis('off')
```

Summary Statistics (EDA)

Crime in Westminster was by far the highest out of any borough throughout all periods observed. The relative change over the period was flat with high seasonal variation. Upon investigation of Westminster's crime data the category of "other theft" shows the highest value by a large margin. In fact, most of this might be attributed to tourist-related incidents given that this borough contains many of London's attractions. This specifically refers to pickpocketing, but that is mostly speculation as the dataset does not break down the category "other theft" more specifically. A more complete analysis would investigate more data on Westminster to determine the culprit behind this high number. However, it is worth mentioning that the dataset used indicates that the number of crimes of this category peaked and has since declined for the last several years of the observation window. The trend seems prevalent and would be worth further investigation to determine the cause of this decline.

Analyze & Share

Aggregating across Major and Minor Categories of Crime in Westminster

```
#%% Westminster Analysis

minor_westminster = minor_permonth[minor_permonth['borough'
                            ]=='Westminster'].groupby(['year',
                               'minor_category'])['no_crimes'].sum()

major_westminster = minor_permonth[minor_permonth['borough'
                            ]=='Westminster'].groupby(['year',
                               'major_category'])['no_crimes'].sum()
```

An example of the table produced by this code is given as follows:

Westminster's Crime in 2012 & 2016 by Major Category (The peak year and the latest year)

36	2012	Burglary	4083
37	2012	Criminal Damage	2254
38	2012	Drugs	4654
39	2012	Fraud or Forgery	0
40	2012	Other Notifiable Offences	641
41	2012	Robbery	2312
42	2012	Sexual Offences	0
43	2012	Theft and Handling	38152
44	2012	Violence Against the Person	7130

72	2016	Burglary	3218
73	2016	Criminal Damage	2179
74	2016	Drugs	2049
75	2016	Fraud or Forgery	0
76	2016	Other Notifiable Offences	708
77	2016	Robbery	1822
78	2016	Sexual Offences	0
79	2016	Theft and Handling	27520
80	2016	Violence Against the Person	10834

While the occurrence of violent crimes is still high, it is clear that the majority of crimes in this area are theft related. Notice that the peak number of thefts in 2012 is quite high and corresponds with the year of the London olympics. I think further investigation should be done to determine if this outlying value is due to that event or due to actual trends in crime. Tourists are often targeted in pick-pocketing so it is entirely possible that 2012's numbers are inflated in relation to London's olympic games. Regardless of this, the down-trend mentioned previously occurs even with the inclusion of this outlier.

Analyze & Share

48	2012-01-01 00:00:00	Theft and Handling	3339
49	2012-02-01 00:00:00	Theft and Handling	2757
50	2012-03-01 00:00:00	Theft and Handling	3484
51	2012-04-01 00:00:00	Theft and Handling	2885
52	2012-05-01 00:00:00	Theft and Handling	3253
53	2012-06-01 00:00:00	Theft and Handling	3163
54	2012-07-01 00:00:00	Theft and Handling	3501
55	2012-08-01 00:00:00	Theft and Handling	3109
56	2012-09-01 00:00:00	Theft and Handling	2708
57	2012-10-01 00:00:00	Theft and Handling	3188
58	2012-11-01 00:00:00	Theft and Handling	3305
59	2012-12-01 00:00:00	Theft and Handling	3460

Westminster's "Theft and Handling"

The table on the left examines the evolution of thefts over 2012 to determine if they peaked near the time of the olympic games. In fact, it appears that they did, but not by a large margin. Several other months outside of the olympic window come close to the peak. Therefore, it does not seem as though the olympics directly affected thefts, however such results are speculation and inconclusive. Further research in the effects of the olympic games could be completed with other data that details tourism counts from 2008-2016. It could be that tourism in general might result in increased pickpocketing thefts.

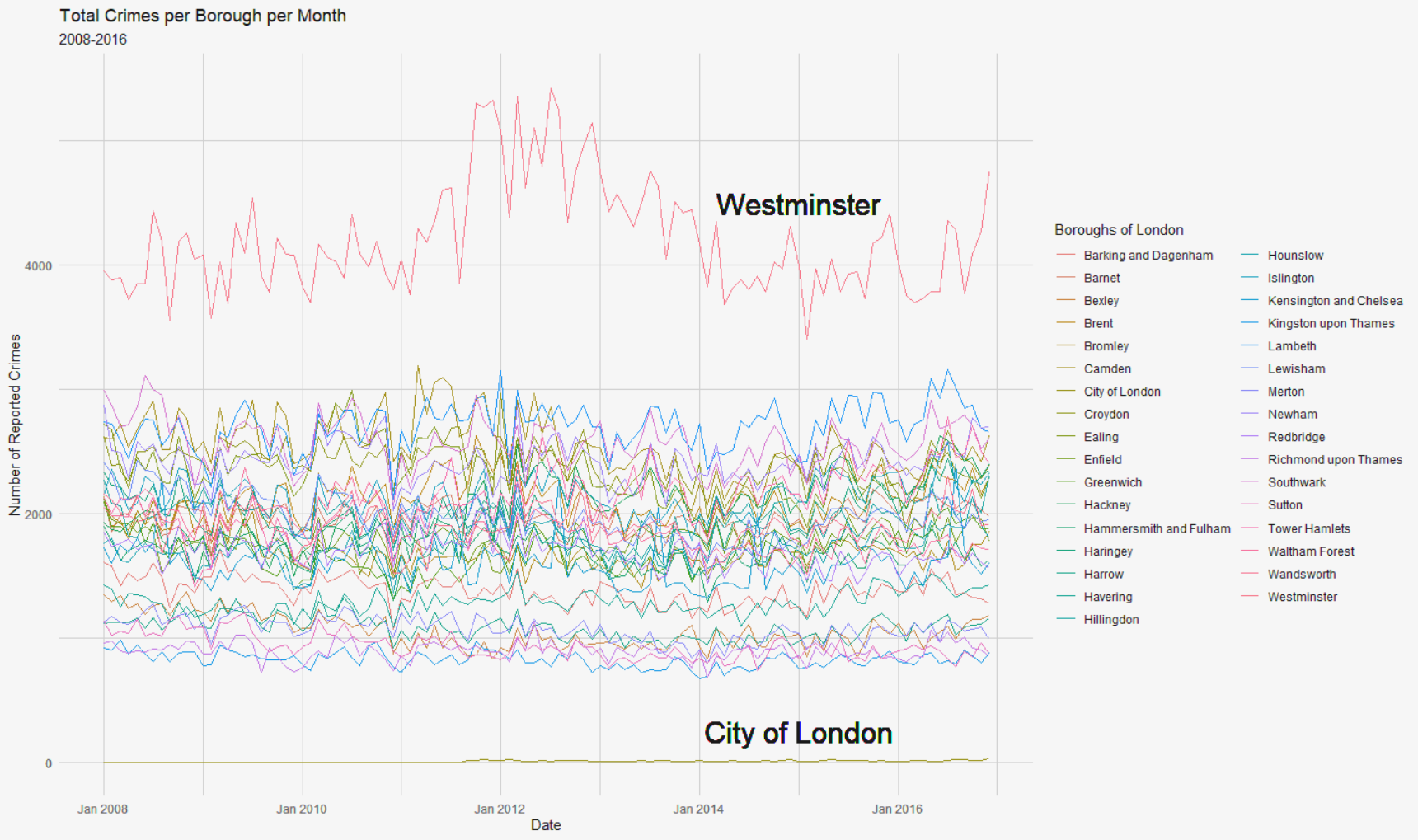
Violent Crime in Westminster

Another area of concern is that the rate of violent crimes has risen considerably. This table illustrates that there has been a steady rise in violent crime year-on-year. This is more concerning whereas thefts showed a slight decrease over the period, violent crimes have steadily increased. Figure 2 below illustrates this trend for the duration of the time window.

	date	major_category	no_crimes
0	2008-12-01 00:00:00	Violence Against the Person	590
1	2009-12-01 00:00:00	Violence Against the Person	572
2	2010-12-01 00:00:00	Violence Against the Person	507
3	2011-12-01 00:00:00	Violence Against the Person	681
4	2012-12-01 00:00:00	Violence Against the Person	576
5	2013-12-01 00:00:00	Violence Against the Person	689
6	2014-12-01 00:00:00	Violence Against the Person	737
7	2015-12-01 00:00:00	Violence Against the Person	911
8	2016-12-01 00:00:00	Violence Against the Person	963

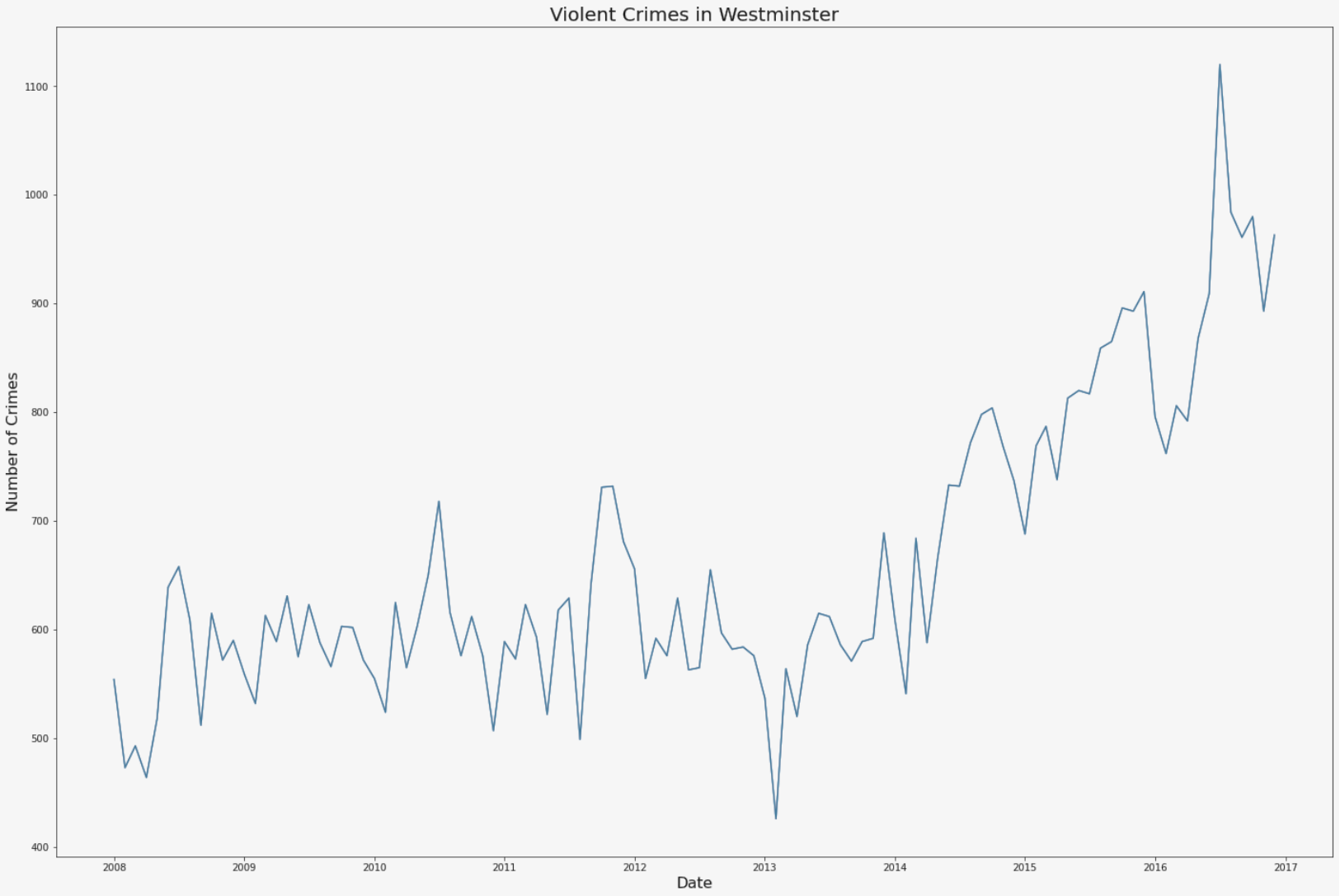
Analyze & Share

Figure 1 - Total Crimes per Borough per Month



Analyze & Share

Figure 2 - Total Crimes per Borough per Month



Analyze & Share

Results

Figure 1 provides an overview of monthly trends where each line represents the evolution of total crimes reported per month in each borough of London from January, 2008 until December, 2016. The most notable insights of figure 1 include the minimum and maximum observations of crime for the City of London and Westminster boroughs, respectively. In fact, Westminster and the City of London display as clear outliers form the bulk of the data.

The situation regarding the City of London could be that Central Business Districts typically experience lower rates of crime due to the fact that they are not trafficked as much by anyone besides working commuters. Furthermore, the City of London itself is quite small compared to the other boroughs thus reducing the probability of encountering illicit activity within the borough. Refer to the Image 1 below for a visual reference on the boroughs' size:

Image 1

Source: https://th.bing.com/th/id/OIP.Q-Un-udMBJHvYzYHYNewNwHaFt?pid=ImgDet&rs=1

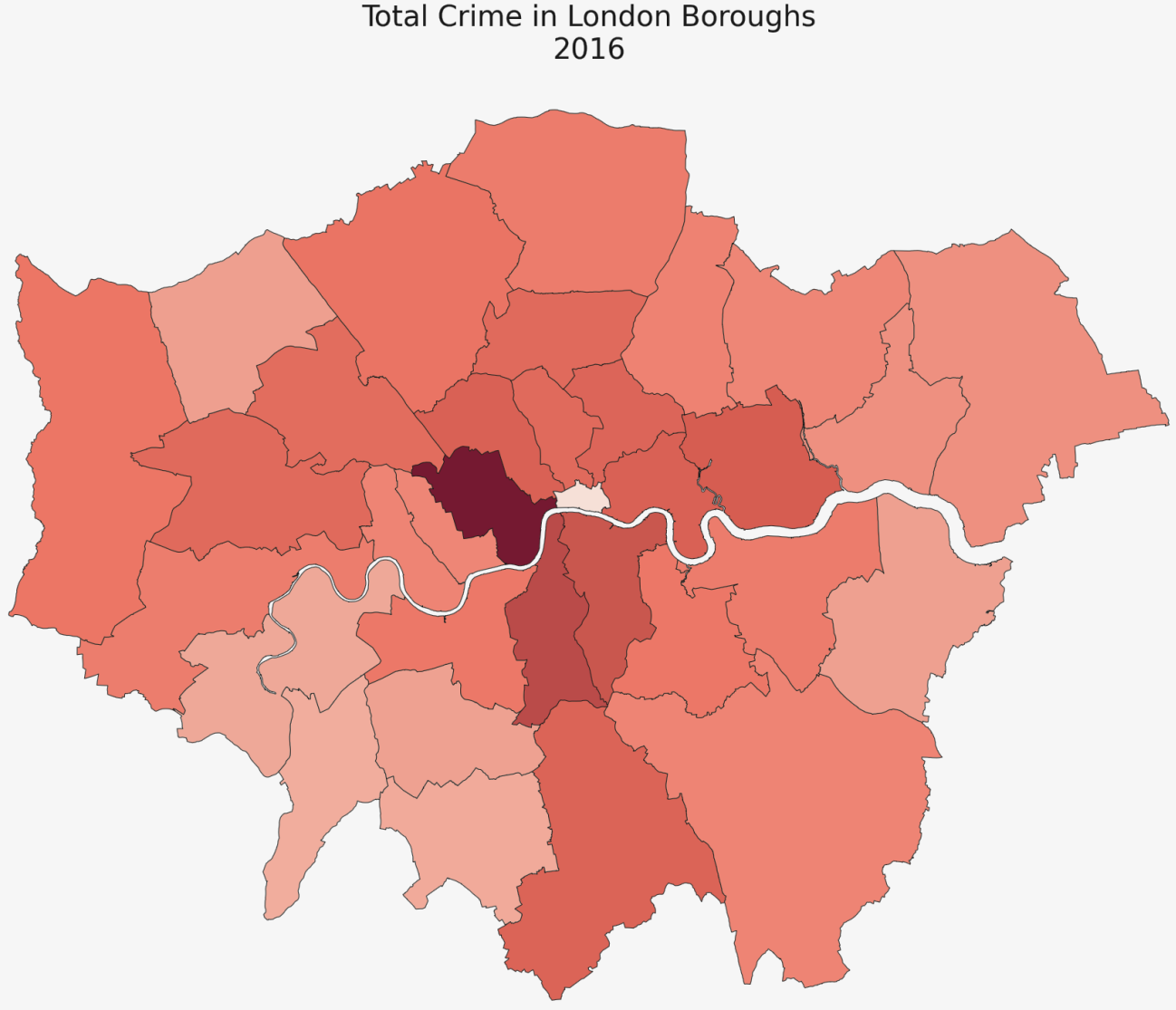
Analyze & Share

Results

Figure 3 below illustrates a heatmap of total crime in London in 2016. This provides a clear visual that crime seems somewhat concentrated around (but not in) the city center. Westminster is shown as the darkest shade of red signifying its high crime count as the outlier. This makes it the perfect candidate for any targeted efforts by police to reduce crime.

This heatmap was built in Python using GeoPandas and Pandas for working with the crime dataset concatenated with a geographical dataset. This tool was then used to produce the color mapping based on the number of crimes in each borough. That is, the darker the shade of red, the more total crime.

Figure 3

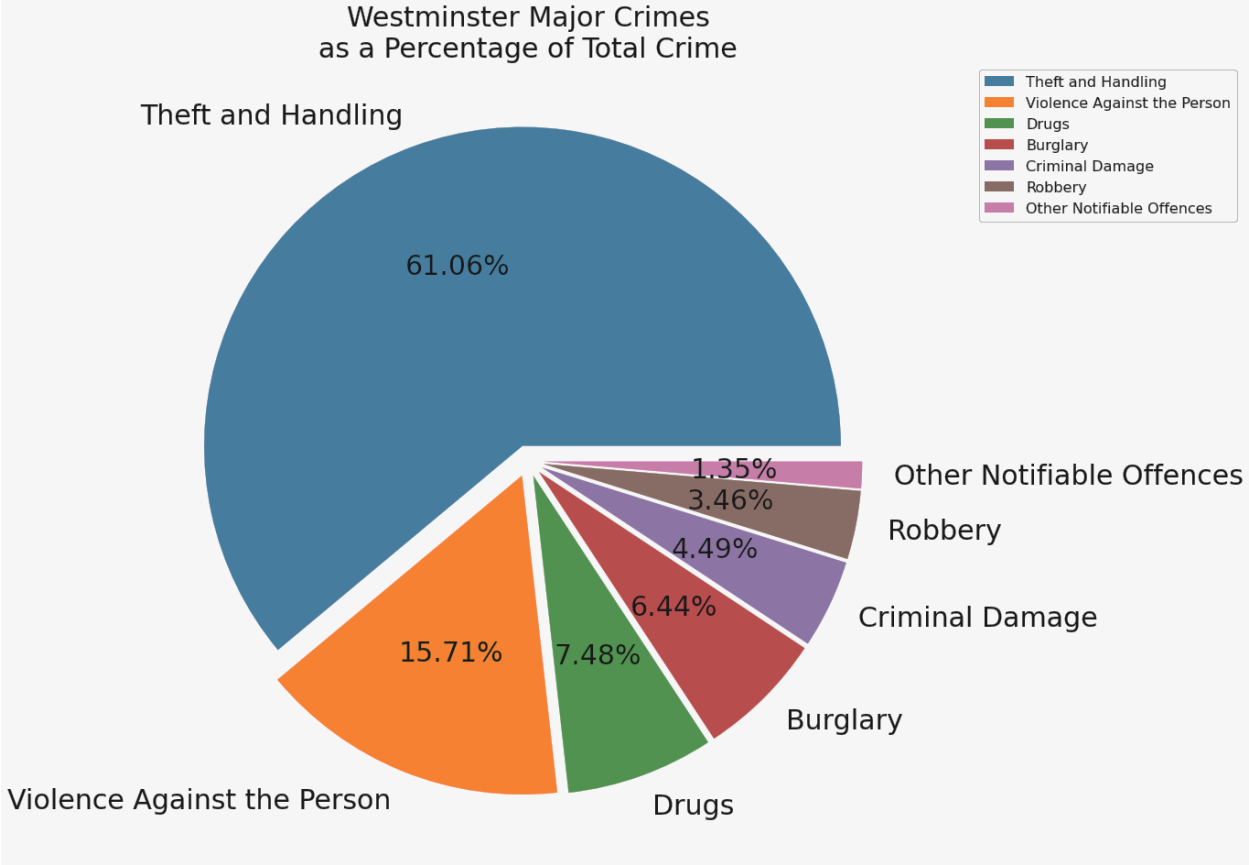


Analyze & Share

Results

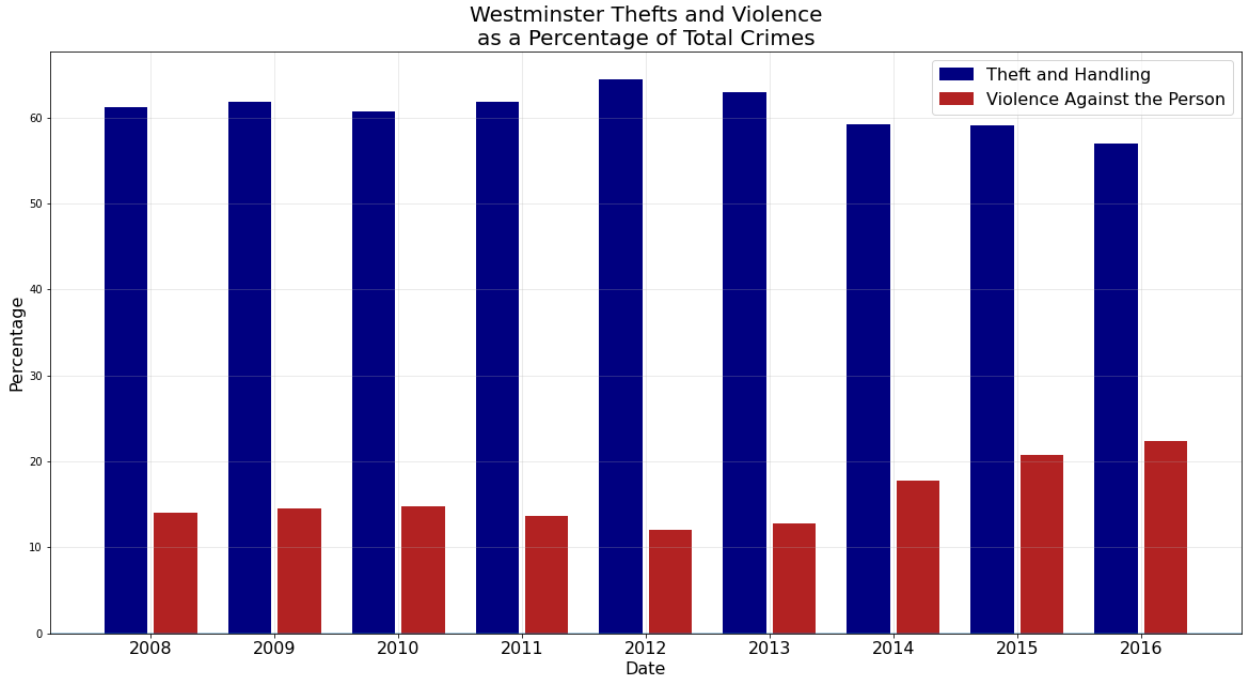
The final metric for consideration in this report will be in regards to the proportion of each crime category as a percentage of Westminster's total observed crimes. This aggregation was done in BigQuery and the visualization was completed in Python.

Figure 4



Percentages for each Major Category are calculated as a percentage of the total number of crimes over the entire observation period. Two of the major categories, Sexual Offences and Fraud and Forgery, were excluded for clarity due to the percentages being less than 1. This chart offers a useful view of just how much Theft and Handling eclipsed the other categories. With that being said, Violence Against the Person should also be an area of concern given the nature of this category's offences. It might come as a surprise that this category for violent crime ranks higher than several other non-violent categories. In the public's view, this could be the most concerning finding as it represents a more direct risk-to-self than theft.

Figure 5



Finally, I want to illustrate how the percentages in the figure 4 change over time for the two most prevalent categories. With more time, a dashboard with animated and interactive charting would have been my preference, but I decided to use the static display in figure 5 above.

The histogram above details Theft and Handling in navy and Violence Against the Person in red. This contrasts each category's evolution over time as the percentage of the total number of crimes committed per year. Again, violent crimes look to be rising through the latter year of the observation window. Thefts appear to be declining slightly during the same years that violent crime was increasing. This could be due to the change in violent crime, more-so than the reduction in thefts given that these are percentages of the total count. Given this dependency, the main takeaway should not be the actual trend for each individual category, but rather, the focus should be on how best to allocate re-sources to better match these changes in proportionality. That is, the proportion of policing resources allocated at fighting each crime can change with the change in proportion of each category of crime coupled with any category-specific goals set for targeted crime reduction.

Act

For recommendations, the investigation of crime trends indicates that Westminster is a solid target for increased surveillance and policing to reduce thefts. Violent crime should be policed, but I would like to see more descriptive data regarding violent crime specifically to have a more justifiable solution.

Conclusion

Overall this project could have benefitted from more data to make accurate conclusions. In terms of exploring the set, all of the goal metrics were successfully analyzed and investigated. Though the analysis was not in-depth, I think as a timed project it was successful in that I learned just how many insights can be found through a simple investigation of the data. If given more time, I would have investigated seasonal trends to determine if the time of year displays any correlation to the levels of crime in the boroughs. I also would have improved visuals, created a dashboard for interactive views of the time-series, and, upon identifying these key issues in criminal activity, would have performed an assessment of supplementary data regarding policing strategies to offer a more qualifiable recommendation.