PH125.9x Data Science - Capstone - Project (for IDV learners) : Kaggle - London Crime Data, 2008-2016

Nigel Chan

02/11/2020

1. Introduction

This is the final course of HarvardX Data Science Professional Certificate. This capstone project is for IDV learners only, and it applies machine learning techniques that go beyond standard linear regression and have opportunity to use a publicly available dataset of your choice and explore new data.

It is strongly discouraged from using datasets that have been used as examples in previous courses or are similar to them (such as the iris, titanic, mnist, or movielens datasets, among others).

"The ***UCI Machine Learning Repository*** and ***Kaggle*** are good place to seek out a dataset. ***Kaggle*** also maintains a ***curated list of dataset*** that are cleaned and ready for machine learning analyzes."[1]

This project uses the below Kaggle available dataset:

London Crime Data, 2008-2016, 13 million rows of Crime Counts, by Borough, Category, and Month.

London Crime Data Description:

Crime in major metropolitan areas, such as London, occurs in distinct patterns. This data covers the number of criminal reports by month, LSOA borough, and major/minor category from Jan-2008 to Dec-2016.

reference:

1. PH125.9x: Capstone Project: IDV Learners, Project Overview: Choose Your Own!

2. Method / Analysis

2.1 Create London Crime dataset

Note:
1. R version 4.0.2 (2020-06-22) is using in this project
2. Platform using in this project: Windows 10 pro with 32GB RAM, system type: x86_64, mingw32
3. It takes some time to load London Crime dataset with 13.5M rows of data.
4. This project uses the saved dataset called "london_crime_dataset.R".

2.2 Data Analysis

London Crime dataset

The London Crime dataset has 13490604 rows and 7 columns.

Table 1: Summary of London Crime dataset

	count
total_rows	13490604
total_columns	7
total_lsoa_code	4835
total_borough	33
total_major_category	9
total_minor_category	32
total_crimes	6447758

Data Structure of London Crime dataset

```
## tibble [13,490,604 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ lsoa_code     : chr [1:13490604] "E01001116" "E01001646" "E01000677" "E01003774" ...
##  $ borough       : chr [1:13490604] "Croydon" "Greenwich" "Bromley" "Redbridge" ...
##  $ major_category: chr [1:13490604] "Burglary" "Violence Against the Person" "Violence Against the Pe
##  $ minor_category: chr [1:13490604] "Burglary in Other Buildings" "Other violence" "Other violence"
##  $ value         : num [1:13490604] 0 0 0 0 0 0 0 0 0 1 ...
##  $ year          : num [1:13490604] 2016 2016 2015 2016 2008 ...
##  $ month         : num [1:13490604] 11 11 5 3 6 5 7 4 9 8 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   lsoa_code = col_character(),
##   ..   borough = col_character(),
##   ..   major_category = col_character(),
##   ..   minor_category = col_character(),
##   ..   value = col_double(),
##   ..   year = col_double(),
##   ..   month = col_double()
##   .. )
```

Table 2: Data Class

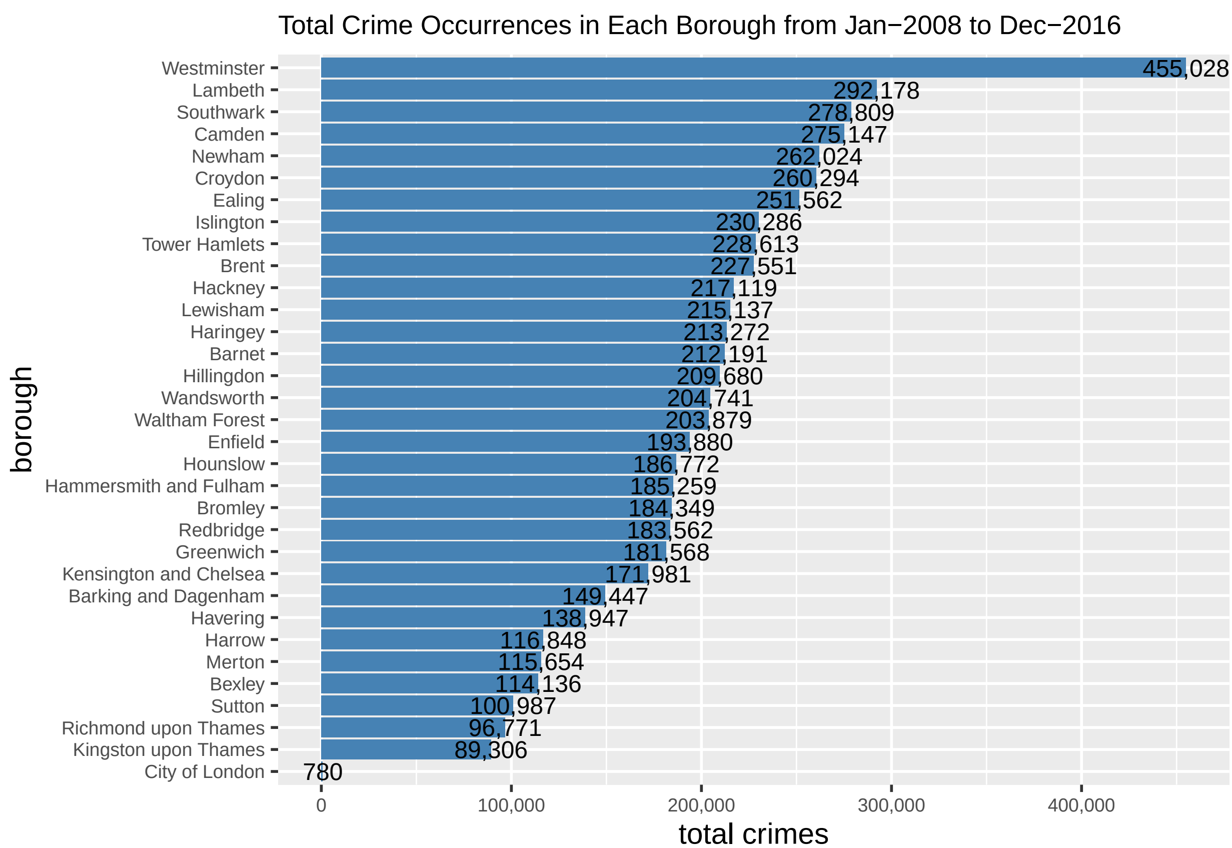
	x
lsoa_code	character
borough	character
major_category	character
minor_category	character
value	numeric
year	numeric
month	numeric

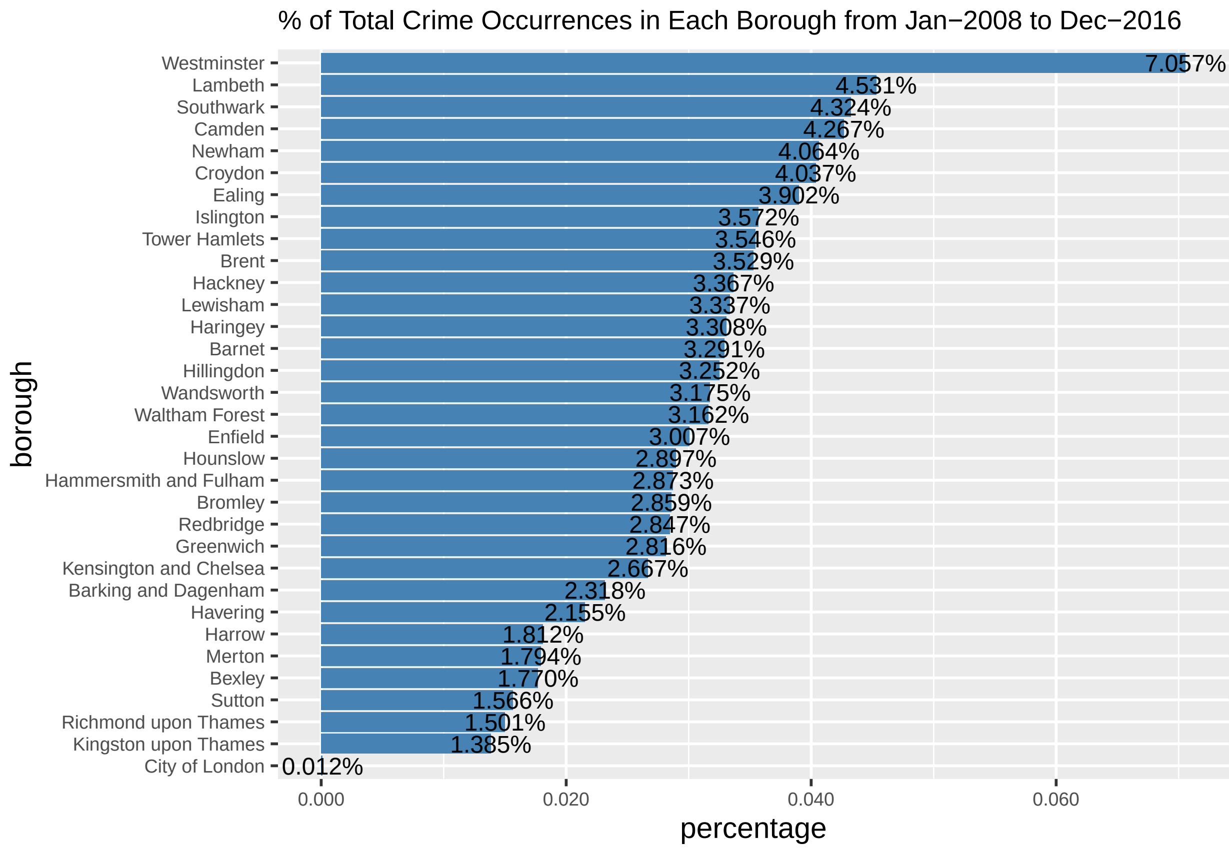
Table 3: Column Description:

Column	Description
lsoa_code	code for Lower Super Output Area in Greater London
borough	Common name for London borough
major_category	High level categorization of crime
minor_category	Low level categorization of crime within major category
value	monthly reported count of categorical crime in given borough
year	Year of reported counts, 2008-2016
month	Month of reported counts, 1-12

Summary of Borough

There are total 33 boroughs in this dataset. The below 2 bar graphs show ***Westminster*** has the highest crime occurrences and highest percentage, and ***City of London*** has the lowest crime occurrences and lowest percentage. *(Surprisingly, City of London has the lowest crime occurrence which is not very far from Westminster!)*





Let's explore top *major category* and *minor category* of crime occurrence in each borough.

1. ***Theft and Handling*** is the top crime occurrences in *major category* in all boroughs.

Table 4: Top Major Category of Crime Occurrences In Each Borough

borough	major_category	count
Westminster	Theft and Handling	277617
Camden	Theft and Handling	140596
Lambeth	Theft and Handling	114899
Southwark	Theft and Handling	109432
Islington	Theft and Handling	107661
Newham	Theft and Handling	106146
Kensington and Chelsea	Theft and Handling	95963
Ealing	Theft and Handling	93834
Wandsworth	Theft and Handling	92523
Croydon	Theft and Handling	91437
Hackney	Theft and Handling	91118
Tower Hamlets	Theft and Handling	87620
Barnet	Theft and Handling	87285
Hammersmith and Fulham	Theft and Handling	86381
Haringey	Theft and Handling	83979
Hillingdon	Theft and Handling	80028
Waltham Forest	Theft and Handling	77940

borough	major_category	count
Brent	Theft and Handling	72523
Redbridge	Theft and Handling	71496
Lewisham	Theft and Handling	70382
Enfield	Theft and Handling	70371
Hounslow	Theft and Handling	70180
Bromley	Theft and Handling	69742
Greenwich	Theft and Handling	64923
Havering	Theft and Handling	52609
Barking and Dagenham	Theft and Handling	50999
Merton	Theft and Handling	44128
Richmond upon Thames	Theft and Handling	40858
Harrow	Theft and Handling	40800
Bexley	Theft and Handling	40071
Sutton	Theft and Handling	39533
Kingston upon Thames	Theft and Handling	38226
City of London	Theft and Handling	561

Table 5: Distinct Major Category of Table 4

major_category	total_count
Theft and Handling	33

2. ***Other Theft*** is the top crime occurrences in *minor category* in most boroughs, except ***Enfield*** which has ***Theft From Motor Vehicle*** and ***Harrow*** which has ***Burglary in a Dwelling*** are the top crime occurrences in two different boroughs.

Table 6: Top Minor Category of Crime Occurrences In Each Borough

borough	minor_category	count
Westminster	Other Theft	142032
Camden	Other Theft	64265
Lambeth	Other Theft	44006
Southwark	Other Theft	42879
Kensington and Chelsea	Other Theft	42217
Islington	Other Theft	37330
Newham	Other Theft	33289
Croydon	Other Theft	33021
Tower Hamlets	Other Theft	32995
Hillingdon	Other Theft	30488
Hackney	Other Theft	30267
Barnet	Other Theft	29966
Wandsworth	Other Theft	29956
Ealing	Other Theft	29165
Hammersmith and Fulham	Other Theft	28082
Haringey	Other Theft	27263
Waltham Forest	Other Theft	25462
Lewisham	Other Theft	24807
Brent	Other Theft	24779

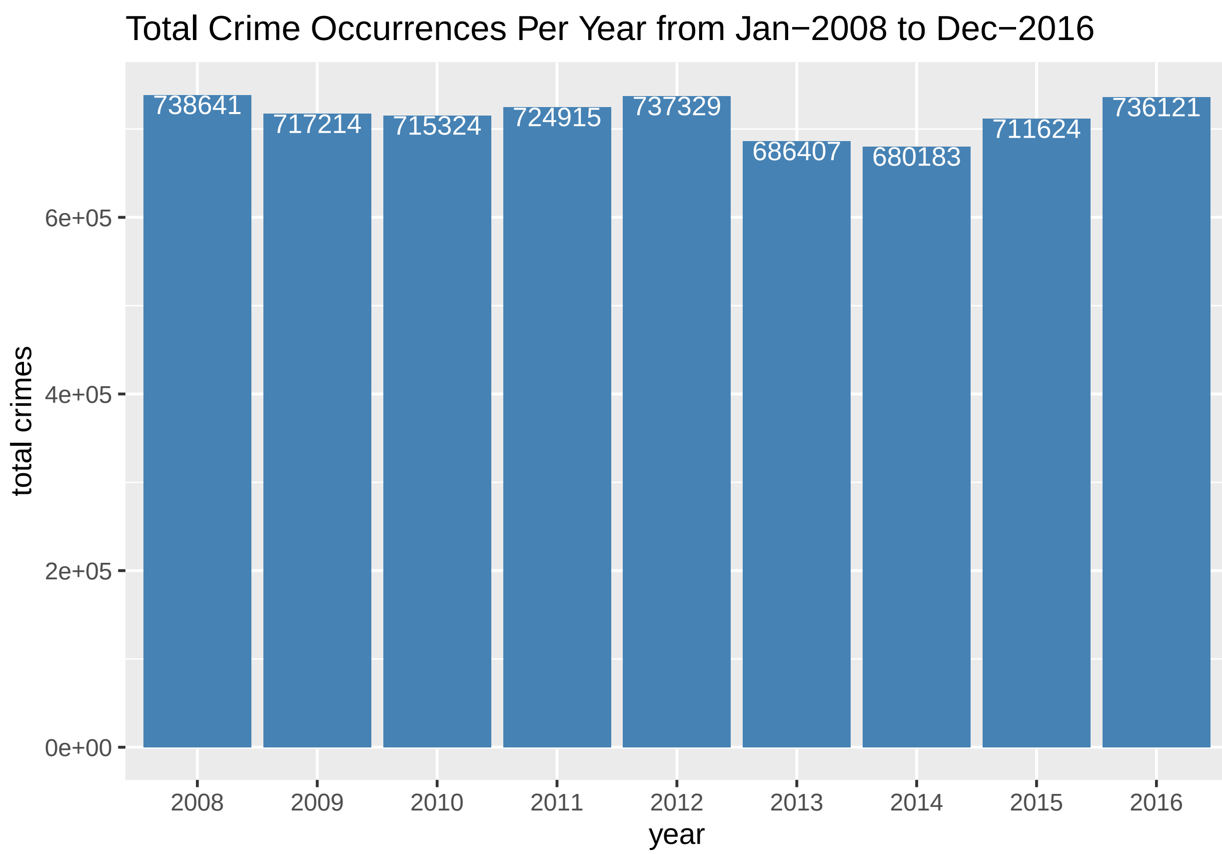
borough	minor_category	count
Bromley	Other Theft	23935
Hounslow	Other Theft	23157
Enfield	Theft From Motor Vehicle	23042
Greenwich	Other Theft	22425
Redbridge	Other Theft	21760
Havering	Other Theft	17716
Barking and Dagenham	Other Theft	16740
Harrow	Burglary in a Dwelling	14918
Merton	Other Theft	14700
Kingston upon Thames	Other Theft	13346
Richmond upon Thames	Other Theft	13108
Bexley	Other Theft	12909
Sutton	Other Theft	12348
City of London	Other Theft	270

Table 7: Distinct Minor Category of Table 6

minor_category	total_count
Other Theft	31
Burglary in a Dwelling	1
Theft From Motor Vehicle	1

Summary of Total Crime Occurrences Per Year

The below bar graph shows crime occurrences from Jan-2008 to Dec-2016, and there is no significant change year by year and all figures remain almost the same. The lowest crime occurrence is in 2014 but the figure is still remain high.

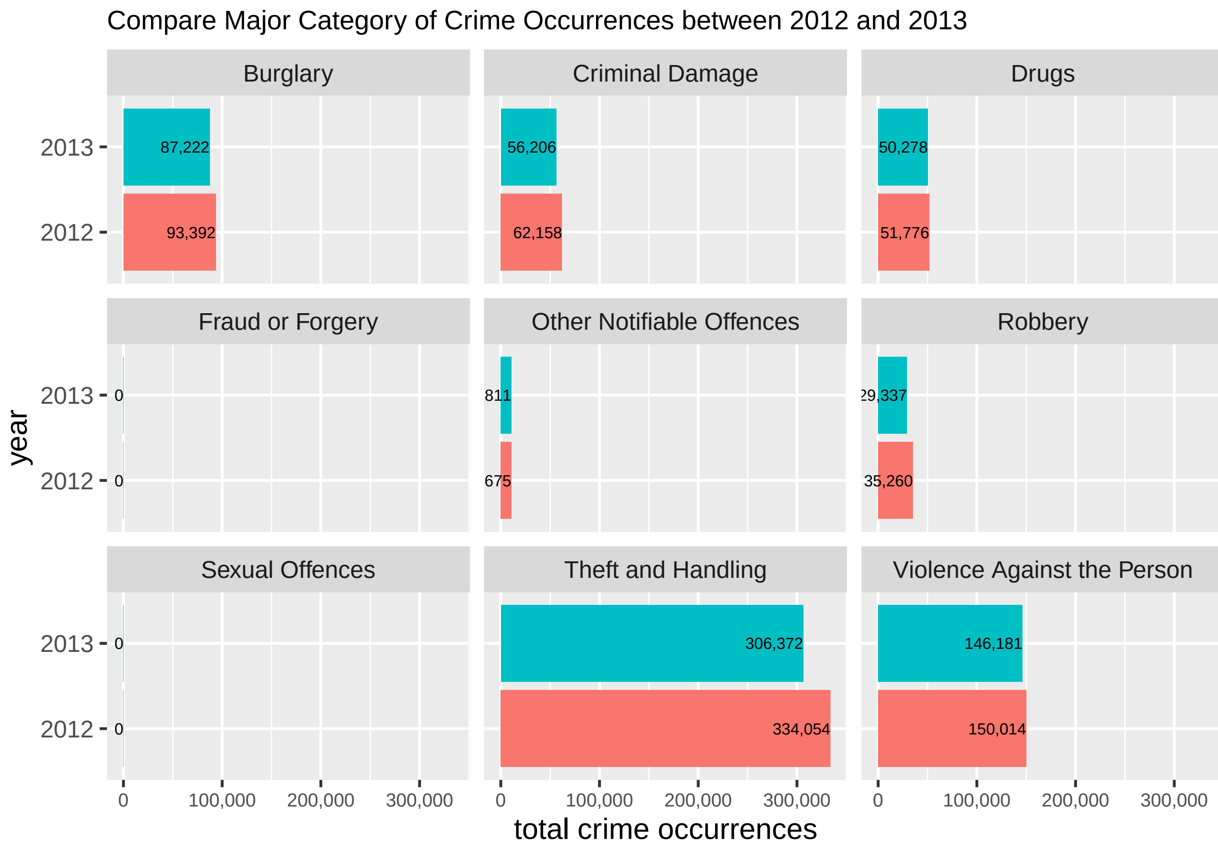


The below table shows the percentage of crime occurrences increase/decrease from previous year.

Table 8: % Change Year-By-Year

year	crimes per year	% change year by year
2008	738641	0.00
2009	717214	-2.90
2010	715324	-0.26
2011	724915	1.34
2012	737329	1.71
2013	686407	-6.91
2014	680183	-0.91
2015	711624	4.62
2016	736121	3.44

There is a big decrease in percentage from 1.71 to -6.91 between 2012 and 2013, let's explore *major category* of crime occurrence on these two years.



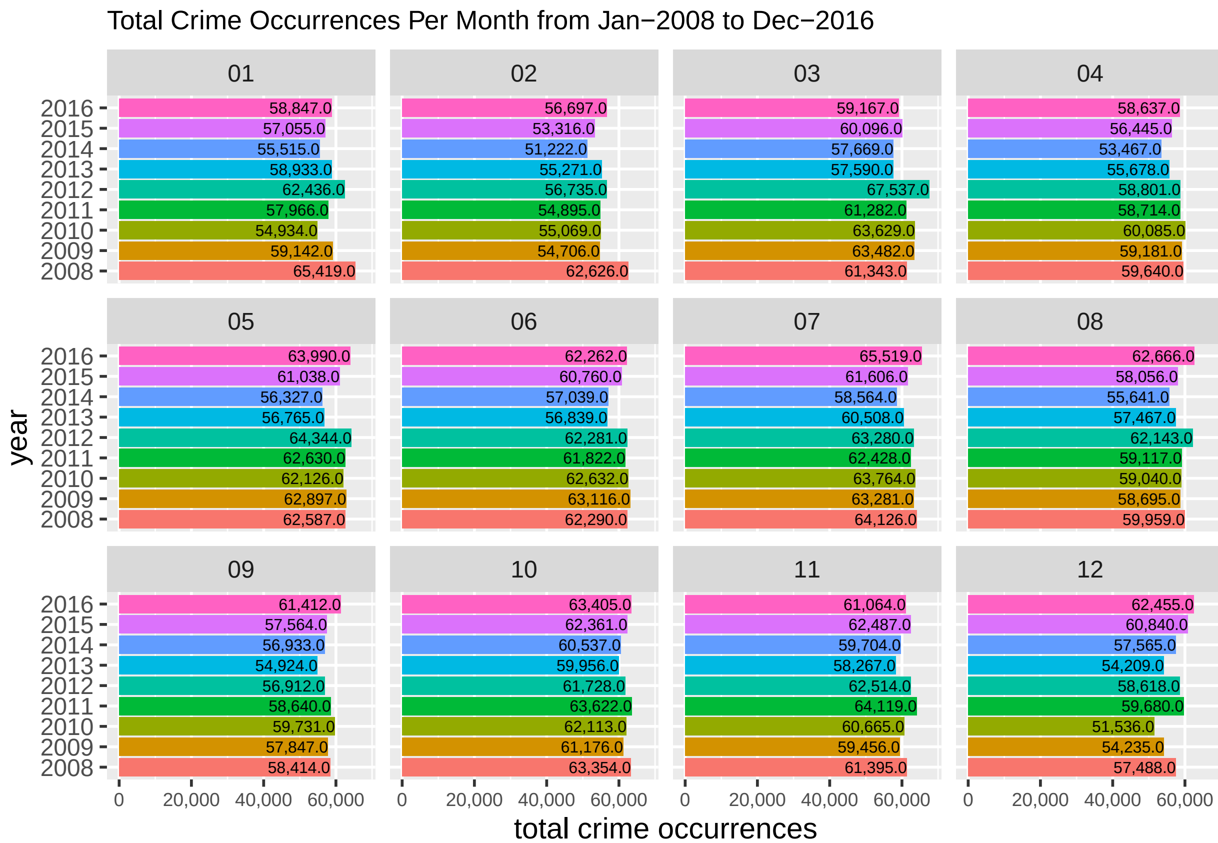
The below table shows the difference in *major category* between 2012 and 2013. The biggest difference in *major category* is ***Theft and Handling*** which reduced -2.7682×10^4 in total. However, there is an increase in ***Other Notifiable Offences*** with 136 crime occurrences.

Table 9: Difference of Total Crime Occurrences between 2012 and 2013

major_category	count
Other Notifiable Offences	136
Fraud or Forgery	0
Sexual Offences	0
Drugs	-1498
Violence Against the Person	-3833
Robbery	-5923
Criminal Damage	-5952
Burglary	-6170
Theft and Handling	-27682

Summary of Total Crime Occurrences Per Month

Let's explore the maximum year-month total crime occurrences. The below table shows monthly total per year.



The below table shows the maximum year-month in each borough with the top *major category* occurrences.

Theft and Handling is the top *major category* in most boroughs, except ***Greenwich*** which has ***Violence Against the Person*** is the top *major category* in 2016-07.

Table 10: Maximum Year_Month of Total Crime occurrences In Each Borough

borough	year_month	major_category	total_count
Westminster	2011-12	Theft and Handling	3634
Camden	2011-03	Theft and Handling	1817
Lambeth	2012-01	Theft and Handling	1394
Newham	2012-03	Theft and Handling	1385
Southwark	2011-10	Theft and Handling	1347
Hackney	2012-03	Theft and Handling	1336
Croydon	2012-01	Theft and Handling	1314
Wandsworth	2012-03	Theft and Handling	1292
Islington	2012-07	Theft and Handling	1277
Kensington and Chelsea	2012-08	Theft and Handling	1262
Ealing	2011-05	Theft and Handling	1170
Tower Hamlets	2016-07	Theft and Handling	1089
Barnet	2012-03	Theft and Handling	1034
Hammersmith and Fulham	2011-11	Theft and Handling	1031
Haringey	2013-06	Theft and Handling	991
Waltham Forest	2011-05	Theft and Handling	984
Hounslow	2012-01	Theft and Handling	923
Hillingdon	2016-12	Theft and Handling	906

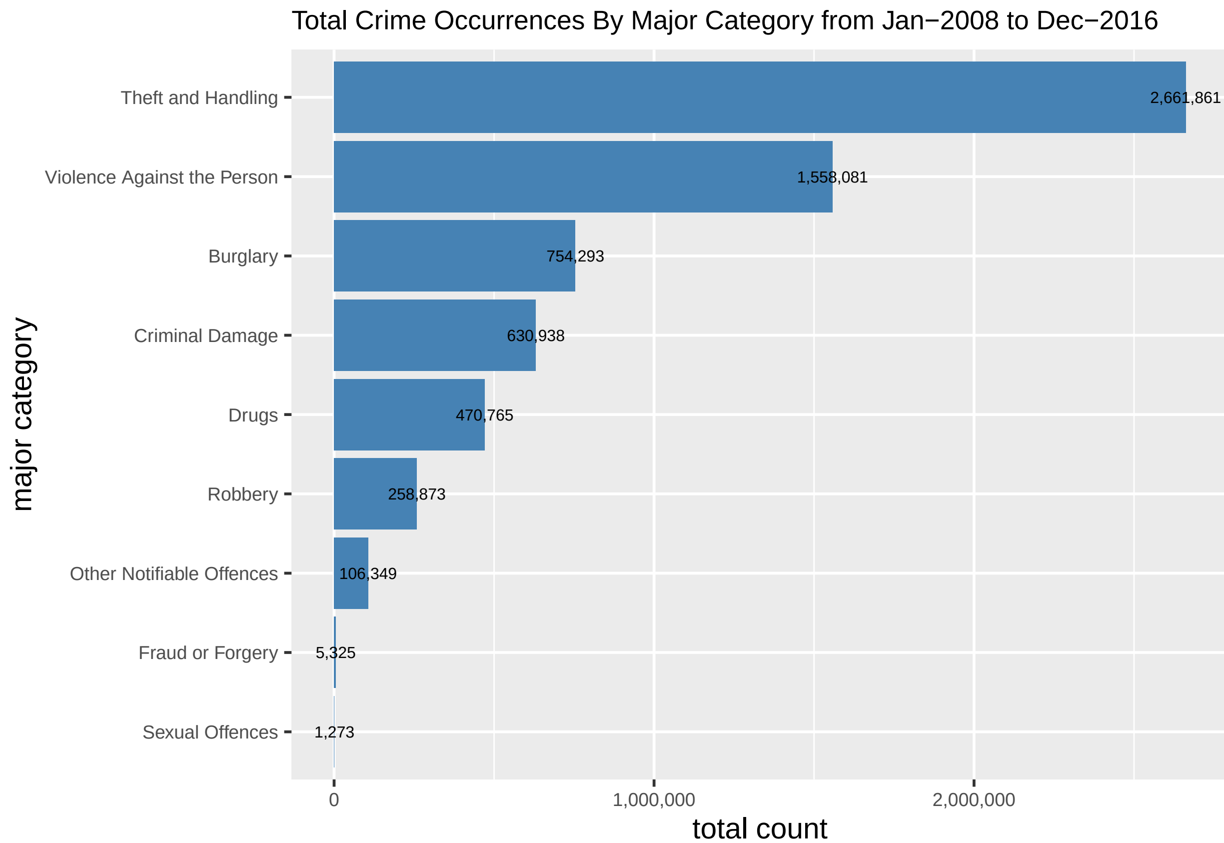
borough	year_month	major_category	total_count
Lewisham	2012-03	Theft and Handling	892
Brent	2015-10	Theft and Handling	890
Bromley	2008-12	Theft and Handling	857
Redbridge	2011-05	Theft and Handling	835
Enfield	2013-06	Theft and Handling	830
Greenwich	2016-07	Violence Against the Person	790
Havering	2012-03	Theft and Handling	671
Barking and Dagenham	2012-03	Theft and Handling	622
Sutton	2008-10	Theft and Handling	536
Harrow	2010-05	Theft and Handling	517
Merton	2012-05	Theft and Handling	516
Bexley	2008-11	Theft and Handling	512
Richmond upon Thames	2012-05	Theft and Handling	500
Kingston upon Thames	2011-10	Theft and Handling	469
City of London	2016-12	Theft and Handling	31

Table 11: Distinct Major Category of Table 10

major_category	total_count
Theft and Handling	32
Violence Against the Person	1

Summary of Major Category

Let's explore all crime occurrences in *major category*.



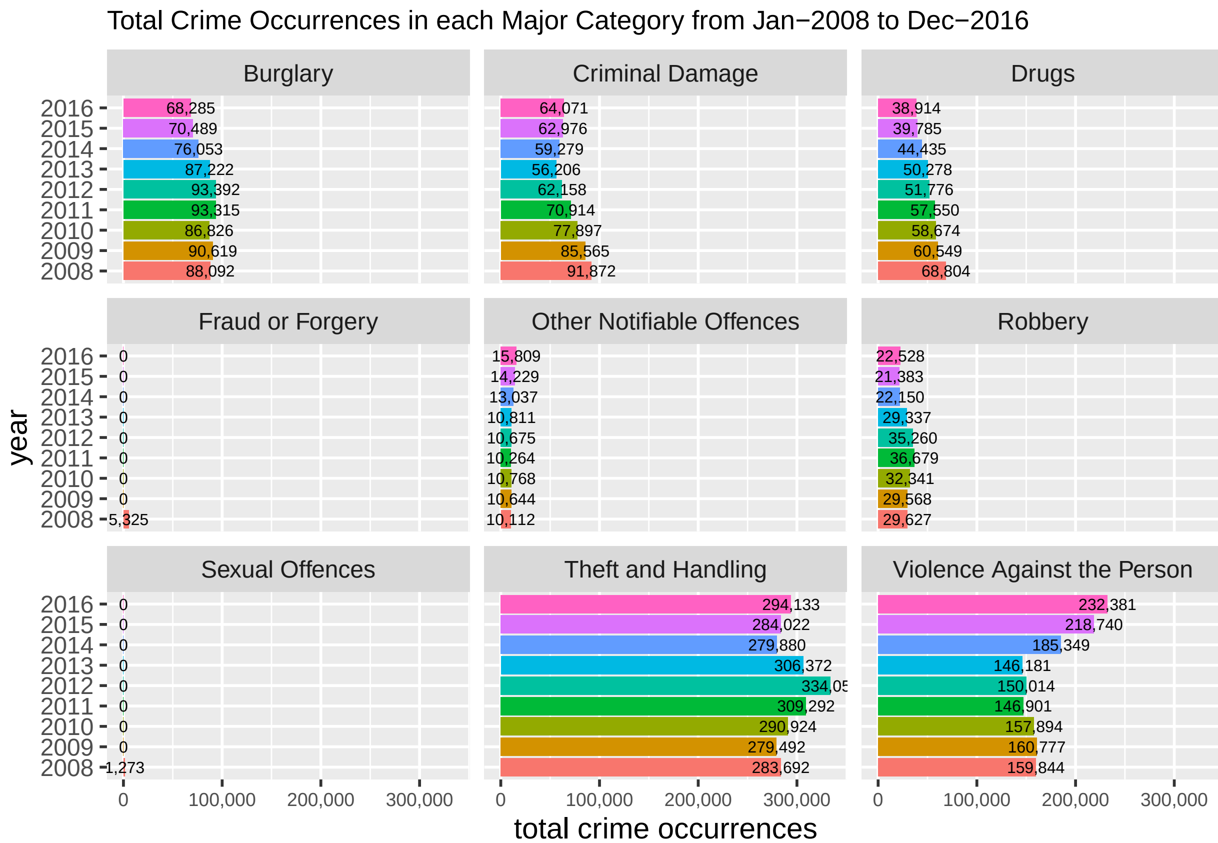
The below bar graph shows total crime occurrences in each *major category* per year.

1. ***Theft and Handling*** and ***Violence Against the Person*** are dominating in total crime occurrences and they are increasing from 2014 to 2016.

2. ***Burglary*** and ***Drugs*** are slightly decreasing.

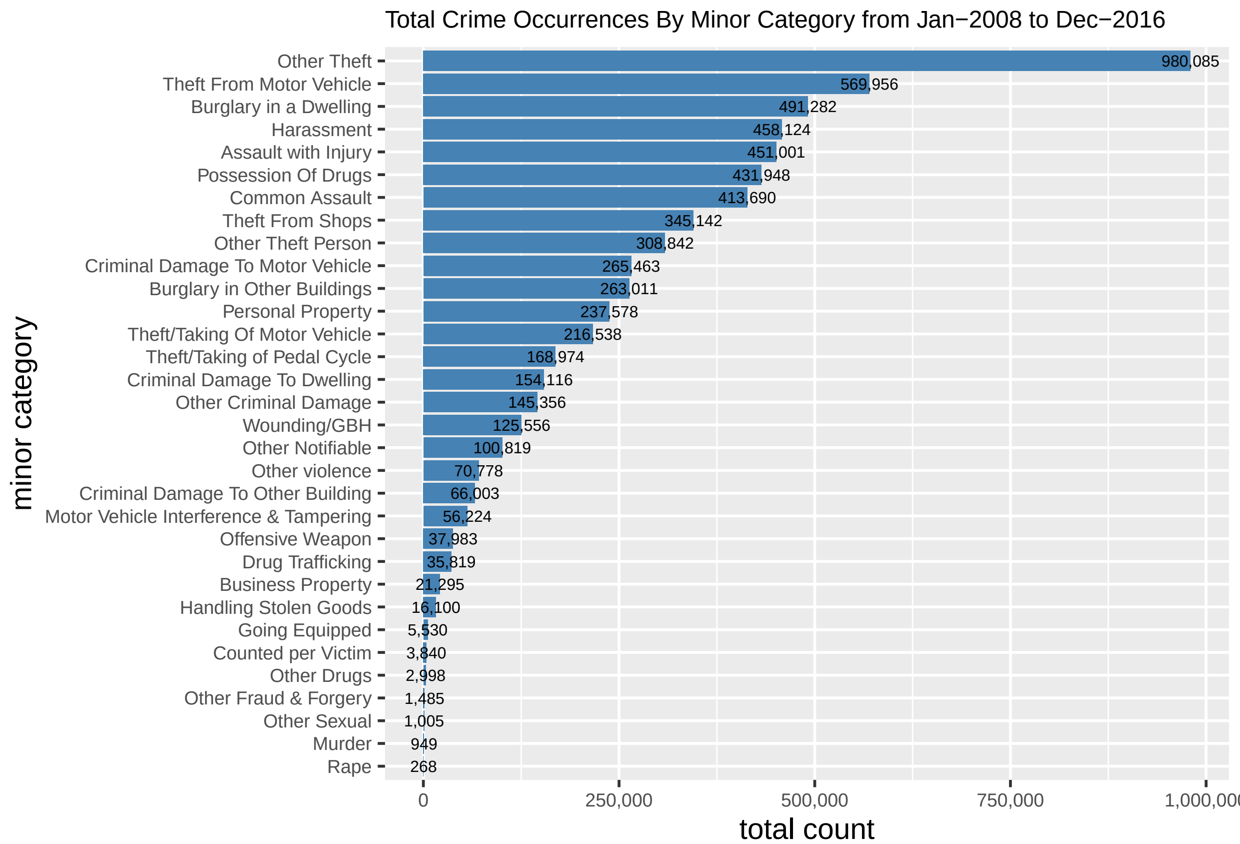
3. ***Criminal Damage*** , ***Other Notifiable Offences*** and ***Robbery*** are tended to increase from 2013 onward.

4. **Surprisingly**, there are no records on these two crimes: ***Fraud or Forgery*** and ***Sexual Offences*** from 2009 onward. (*Is data correct? Is it under different categories? Have these 2 crimes been under control?*)



Summary of Minor Category

Let's explore all crime occurrences in *minor category*.

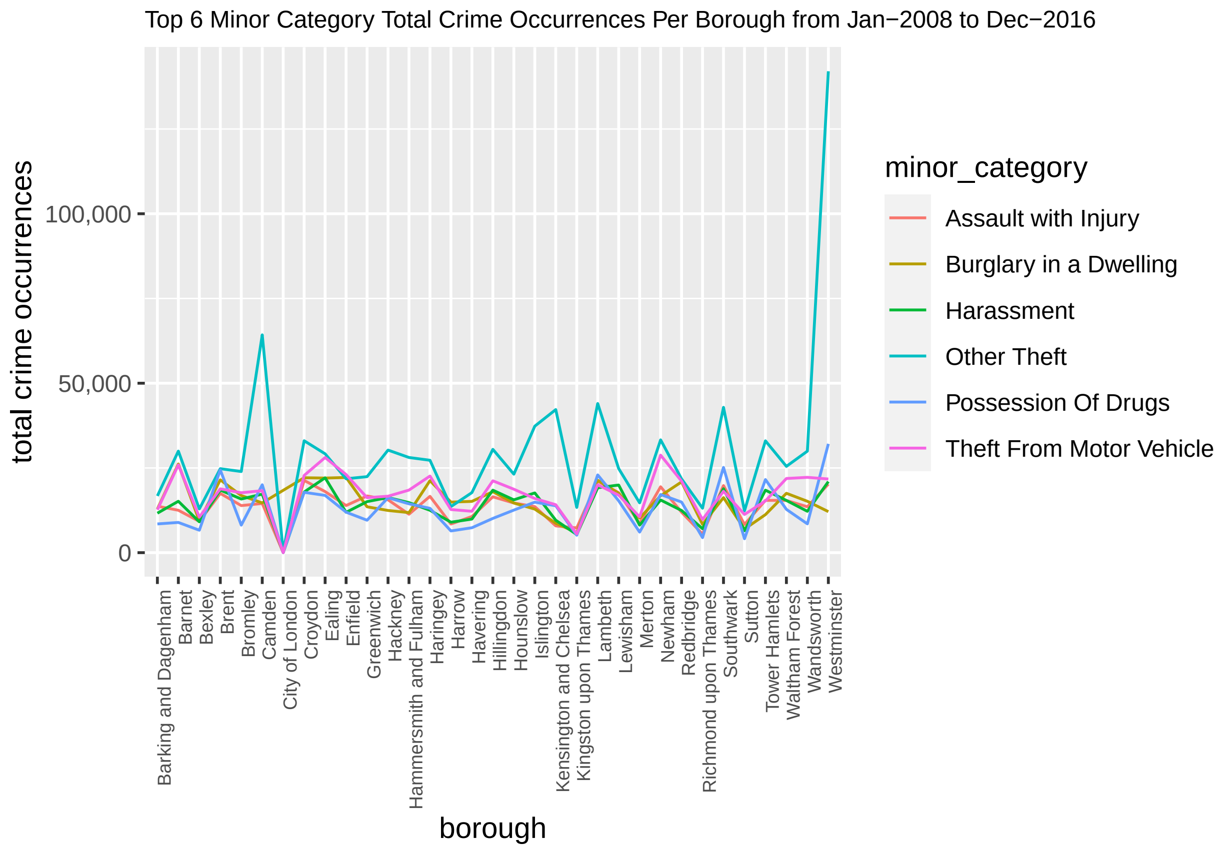


The below line graph shows total crime occurrences in *top 6 minor category*.

1. ***Westminster*** has the top crime occurrences in all *top 6 minor category*. ***(This is a very popular tourist area.)***

2. ***City of London*** has the lowest crime occurrences in all *top 6 minor category*.

3. ***Other Theft*** is the dominating crime in all *top 6 minor category*.

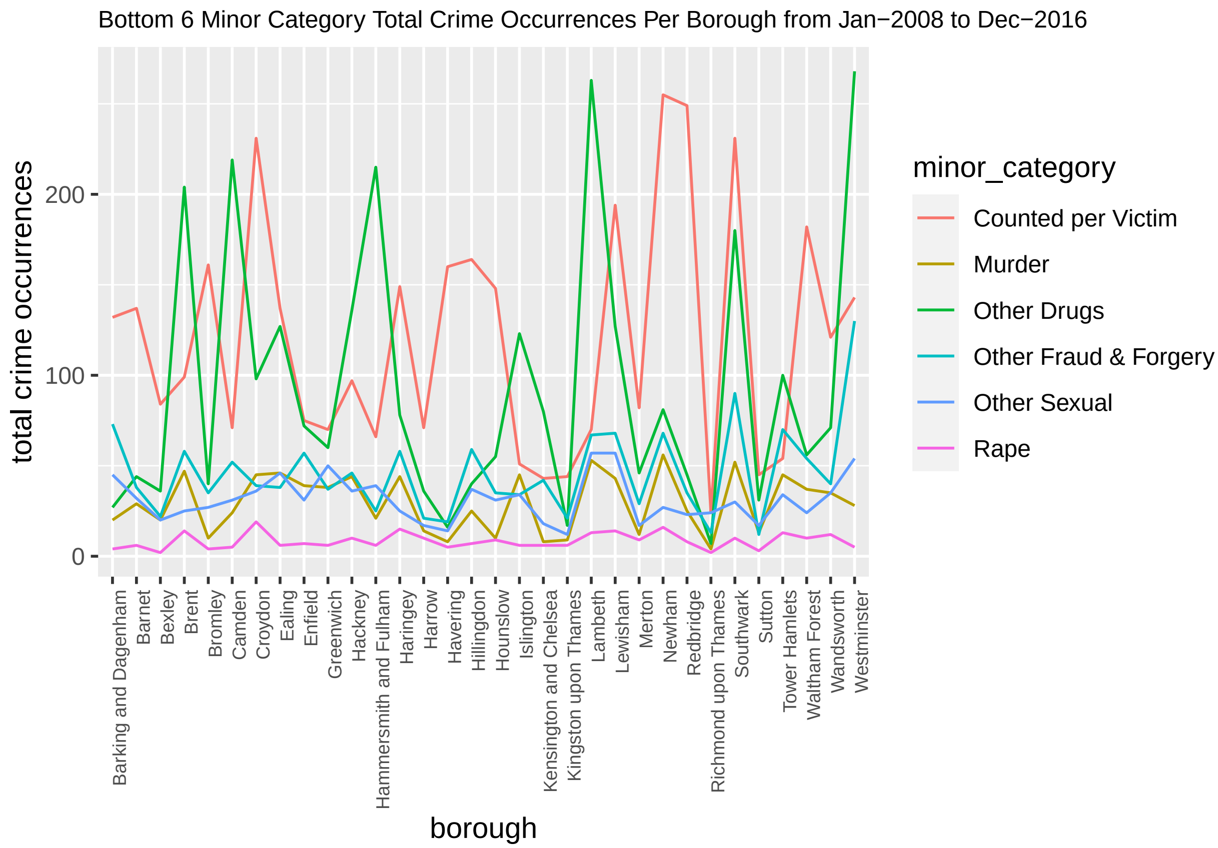


The below line graph shows total crime occurrences in *bottom 6 minor category*.

1. ***Other Drugs*** is the top crime occurrences in *bottom 6 minor category* in these 2 boroughs ***Westminster*** and ***Lambeth***.

2. ***City of London*** disappeared in the *bottom 6 minor category*.

3. ***Counted per Victim*** and ***Other Drugs*** are dominating crimes in *bottom 6 minor category* and ***Rape*** has the lowest crime occurrences across all boroughs.



3. Result

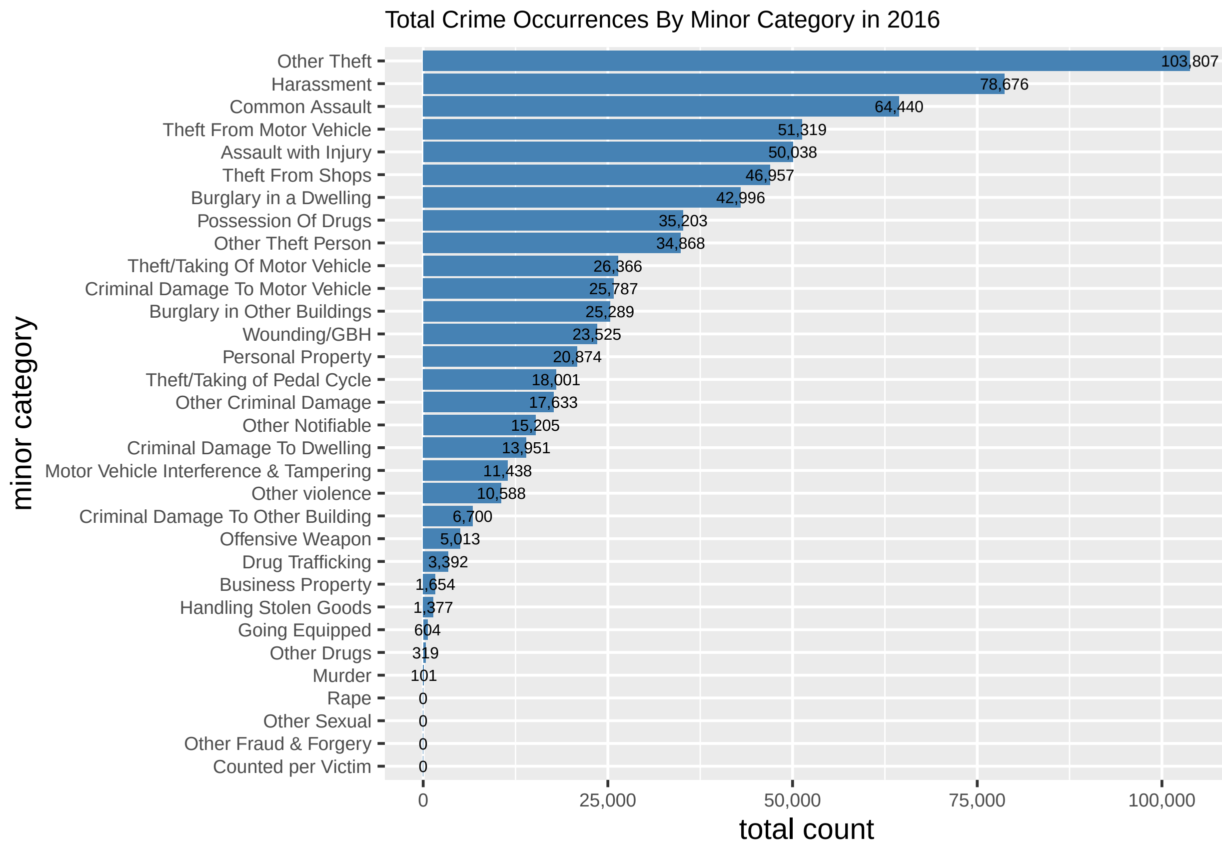
The below models use 2016 data (a subset from London Crime dataset) called "data_2016" for prediction, which has 1498956 rows and 4 columns.

```
## tibble [1,498,956 x 4] (S3: tbl_df/tbl/data.frame)
##  $ borough       : Factor w/ 33 levels "1","2","3","4",..: 8 11 26 29 17 21 23 5 14 28 ...
##  $ major_category: Factor w/ 9 levels "1","2","3","4",..: 1 9 1 8 8 8 9 2 2 8 ...
##  $ minor_category: Factor w/ 32 levels "1","2","3","4",..: 3 24 3 31 30 29 5 8 9 29 ...
##  $ value         : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 2 1 2 ...
```

Table 12: Data Class

	x
borough	factor
major_category	factor
minor_category	factor
value	factor

The below bar graph shows all *minor category* crime occurrences in 2016. ***Other Theft*** is still top of the list and there are zero crime occurrence in ***Rape***, ***Other Sexual***, ***Other Fraud & Forgery*** and ***Counted per victim*** in 2016.



3.1 Naive Bayes Model

"Naive Bayes is a Supervised Machine Learning Algorithm based on the Bayes Theorem that is used to solve classification problems by following a probabilistic approach. It is based on the idea that the predictor variables in a Machine Learning model are independent of each other. Meaning that the outcome of a model depends on a set of independent variables that have nothing to do with each other."[1]

Table 13: Summary of Data_2016

borough	major_category	minor_category	value
8 : 66900	8 :440700	5 : 58020	0:1106914
2 : 63648	9 :352416	22 : 58020	1: 392042
9 : 61044	2 :229908	28 : 58020	NA
5 : 58212	3 :131052	3 : 58008	NA
22 : 57672	1 :115956	13 : 58008	NA
10 : 56796	6 :104376	30 : 58008	NA
(Other):1134684	(Other):124548	(Other):1150872	NA

Naive Bayes Model **uses 2016 dataset on** ***minor category*** **to predict crime occurrences.**

The ***Naive Bayes Model*** achieved **75.65%** accuracy with a p-value of less than 1.

The below table shows the Confusion Matrix output and overall model statistics.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 200951  52582
##          1  20431  25826
##
##                Accuracy : 0.7565
##                  95% CI : (0.7549, 0.758)
##     No Information Rate : 0.7385
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2733
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9077
##             Specificity : 0.3294
##          Pos Pred Value : 0.7926
##          Neg Pred Value : 0.5583
##              Prevalence : 0.7385
##          Detection Rate : 0.6703
##    Detection Prevalence : 0.8457
##       Balanced Accuracy : 0.6185
##
##        'Positive' Class : 0
##
```

3.2 Decision Tree Model

"A Decision Tree is a supervised learning predictive model that uses a set of binary rule to calculate a target value. It is used for either **classification (categorical target variable)** or **regression (continuous target variable)**. Hence, it is also known as **CART (Classification & Regression Trees)**.

Decision trees have three main parts:

1. Root Node: the node that performs the first split.

2. Terminal Nodes/Leaves: nodes that predict the outcome.

3. Branches: arrows connecting nodes, showing the flow from question to answer.

The root node is the starting point of the tree, and both root and terminal nodes contain questions or criteria to be answered. Each node typically has two or more nodes extending from it. For example, if the question in the first node requires a *yes* or *no* answer, there will be one leaf node for the *yes* response, and another node for *no*."[2]

Table 14: Summary of Data_2016

borough	major_category	minor_category	value
8 : 66900	8 :440700	5 : 58020	0:1106914
2 : 63648	9 :352416	22 : 58020	1: 392042
9 : 61044	2 :229908	28 : 58020	NA

borough	major_category	minor_category	value
5 : 58212	3 :131052	3 : 58008	NA
22 : 57672	1 :115956	13 : 58008	NA
10 : 56796	6 :104376	30 : 58008	NA
(Other):1134684	(Other):124548	(Other):1150872	NA

Decision Tree Model **uses 2016 dataset on** ***minor category*** **to predict crime occurrences.**

The ***Decision Tree model*** achieved **75.84%** accuracy with a p-value of less than 1.

The below table shows the Confusion Matrix output and overall model statistics.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 201488  52095
##          1  20309  25820
##
##                Accuracy : 0.7584
##                  95% CI : (0.7569, 0.76)
##     No Information Rate : 0.74
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2764
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9084
##             Specificity : 0.3314
##          Pos Pred Value : 0.7946
##          Neg Pred Value : 0.5597
##              Prevalence : 0.7400
##          Detection Rate : 0.6723
##    Detection Prevalence : 0.8461
##       Balanced Accuracy : 0.6199
##
##        'Positive' Class : 0
##
```

3.3 Random Forest Model

"Random Forest is a learning method for classification. It is based on generating a large number of decision trees, each constructed using a different subset of your training set. These subsets are usually selected by sampling at random and with replacement from the original data set. The decision trees are then used to identify a classification consensus by selecting the most common output(mode). While random forest can be used for other applications (i.e. regression)."[3]

Table 15: Summary of Data_2016

borough	major_category	minor_category	value
8 : 66900	8 :440700	5 : 58020	0:1106914
2 : 63648	9 :352416	22 : 58020	1: 392042
9 : 61044	2 :229908	28 : 58020	NA
5 : 58212	3 :131052	3 : 58008	NA
22 : 57672	1 :115956	13 : 58008	NA
10 : 56796	6 :104376	30 : 58008	NA
(Other):1134684	(Other):124548	(Other):1150872	NA

Random Forest Model **uses 2016 dataset on *minor category* to predict crime occurrences and use ntree value = 100.**

The ***Random Forest model*** achieved **75.68%** accuracy with a p-value of less than 1.

The below table shows the Confusion Matrix output and overall model statistics.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 200852  52557
##          1  20370  26108
##
##                Accuracy : 0.7568
##                  95% CI : (0.7553, 0.7584)
##     No Information Rate : 0.7377
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2762
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9079
##             Specificity : 0.3319
##          Pos Pred Value : 0.7926
##          Neg Pred Value : 0.5617
##              Prevalence : 0.7377
##          Detection Rate : 0.6698
##    Detection Prevalence : 0.8450
##       Balanced Accuracy : 0.6199
##
##        'Positive' Class : 0
##
```

3.4 KNN Model - prediction

"K-NN is an example of a supervised learning method, which means we need to first feed it data so it is able to make a classification based on that data (this is called the training phase). Upon training the algorithm on the data we provided, we can test our model on an unseen dataset (where we know that what class each observation belongs to), and can then see how successful our model is at predicting the existing classes.

K-NN is a non-parametric technique that stores all available cases and classifies new cases based on a similarity measure (distance function). Therefore when classifying an unseen dataset using a trained K-NN algorithm, it looks through the training data and finds the **k** training examples that are closest to the new example. It then assigns a class label to the new example based on a majority vote between those **k** training examples. This means if **k** is equal to 1, the class label will be assigned based on the nearest neighbour. However if K is equals to 3, the algorithm will select the three closest data points to each case and classify it based on a majority vote based on the classes that classes that those three adjacent points hold."[4]

Table 16: Summary of Data_2016

borough	major_category	minor_category	value
8 : 66900	8 :440700	5 : 58020	0:1106914
2 : 63648	9 :352416	22 : 58020	1: 392042
9 : 61044	2 :229908	28 : 58020	NA
5 : 58212	3 :131052	3 : 58008	NA
22 : 57672	1 :115956	13 : 58008	NA
10 : 56796	6 :104376	30 : 58008	NA
(Other):1134684	(Other):124548	(Other):1150872	NA

KNN Model uses 2016 dataset on ***minor category*** to predict crime occurrence and use the optimal k vaule = 94.

Note: There is a limitation when loading a large of number records into knn model. The maximum size is 11000 records. (I tested 80% of training set from "data_2016" dataset with the optimal k value = 1 and it encountered an error, then I reduced the size of "data_2016" dataset multiple times and it came up with maximum size = 11000.)

The ***KNN Model*** achieved **75.59%** accuracy with a p-value of less than 1.

The below table shows the Confusion Matrix output and overall model statistics.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1481  395
##          1  142  182
##
##                Accuracy : 0.7559
##                  95% CI : (0.7374, 0.7737)
##     No Information Rate : 0.7377
##     P-Value [Acc > NIR] : 0.0271
##
##                   Kappa : 0.2654
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9125
##             Specificity : 0.3154
##          Pos Pred Value : 0.7894
##          Neg Pred Value : 0.5617
##              Prevalence : 0.7377
##          Detection Rate : 0.6732
```

```
##    Detection Prevalence : 0.8527
##       Balanced Accuracy : 0.6140
##
##        'Positive' Class : 0
##
```

reference:

1. edureka! : A Comprehensive Guide to Naive Bayes in R
2. Analytics Vidhya
3. OPIG
4. Intro to Machine Learning in R (***K Nearest Neighbours algorithm***)

4. Conclusion

According to the above different model predictions for 2016 London Crime dataset on *minor category*:

1. ***KNN model*** achieved **75.59%** accuracy which is the lowest of 4 models prediction. Note, there is a downside when testing on the ***KNN model***, because it has a limitation on "data_2016" dataset (maximum 11000 records) when processing knn model, perhaps, it works better with a smaller dataset.
2. ***Naive Bayes model*** achieved **75.65%** accuracy and it is better than ***KNN model***
3. ***Random Forest model*** is a slightly better than ***Naive Bayes model***, and it achieved **75.68%** accuracy. (Note: also tested with ntree = 200 but it achieved accuracy rate as in ntree = 100.)
4. ***Decision Tree model*** achieved **75.84%** accuracy. This model has the best prediction among other models also has better performance.

Table 17: Accuracy Results

Method	Accuracy	p-Value	Sensitivity	Specificity	Balanced Accuracy
Naive Bayes Model	0.7564529	0	0.9077116	0.3293797	0.6185456
Decision Tree Model	0.7584214	0	0.9084343	0.3313868	0.6199105
Random Forest Model	0.7568184	0	0.9079206	0.3318884	0.6199045
KNN Model	0.7559091	0	0.9125077	0.3154246	0.6139662