Crime Analysis through Machine Learning

Presented by

Mohammad Ayyaz Azeem

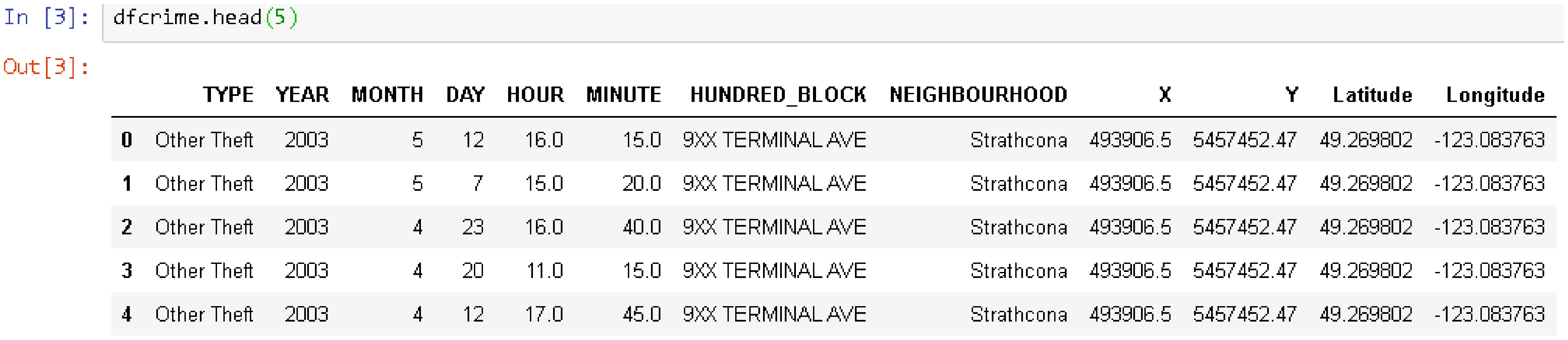
CMS 32175

INTRODUCTION

- Dataset
- Steps taken
- Procedure/Steps taken

DATASET EXPLANATION

- 560,000 records
- Vancouver Police Department crimes dataset [2003-2017]
  - https://vancouver.ca/police/

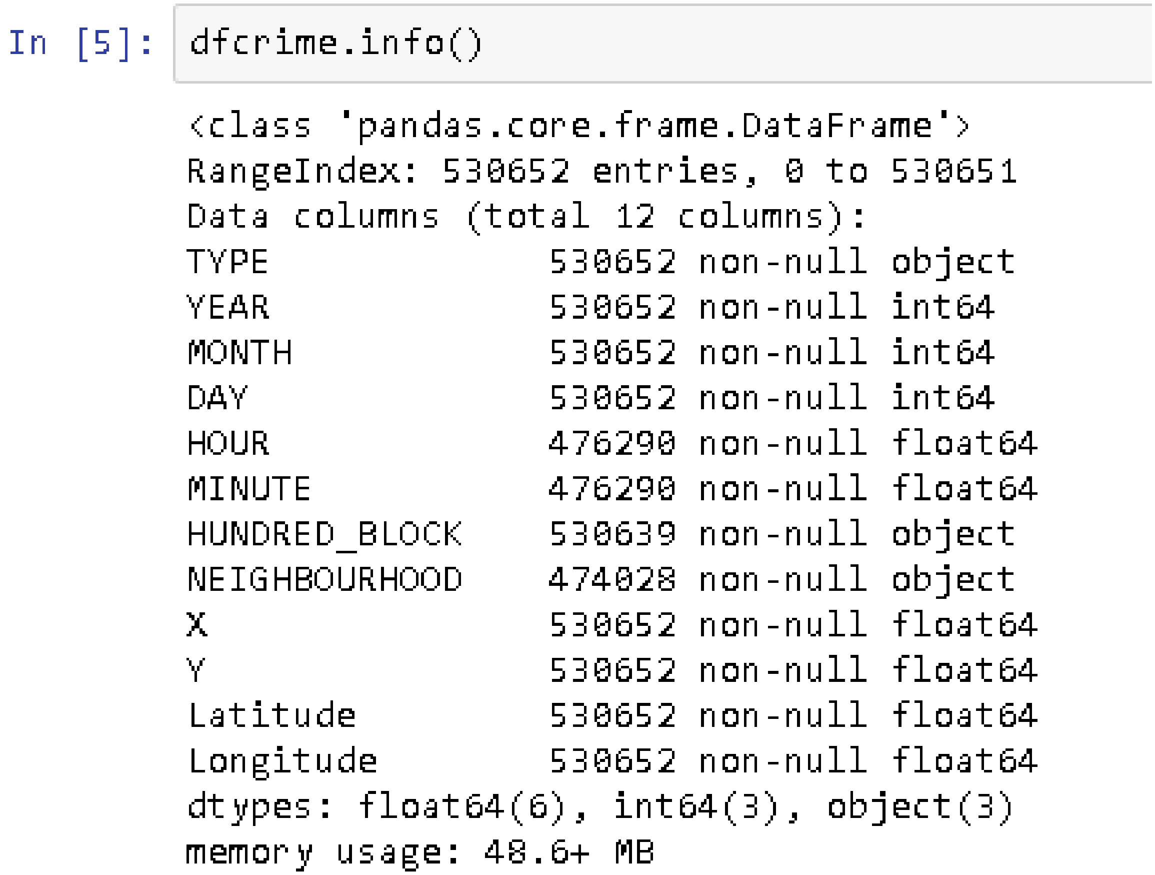


```
In [3]: dfcrime.head(5)
```

Out[3]:

	TYPE	YEAR	MONTH	DAY	HOUR	MINUTE	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y	Latitude	Longitude
0	Other Theft	2003	5	12	16.0	15.0	9XX TERMINAL AVE	Strathcona	493906.5	5457452.47	49.269802	-123.083763
1	Other Theft	2003	5	7	15.0	20.0	9XX TERMINAL AVE	Strathcona	493906.5	5457452.47	49.269802	-123.083763
2	Other Theft	2003	4	23	16.0	40.0	9XX TERMINAL AVE	Strathcona	493906.5	5457452.47	49.269802	-123.083763
3	Other Theft	2003	4	20	11.0	15.0	9XX TERMINAL AVE	Strathcona	493906.5	5457452.47	49.269802	-123.083763
4	Other Theft	2003	4	12	17.0	45.0	9XX TERMINAL AVE	Strathcona	493906.5	5457452.47	49.269802	-123.083763

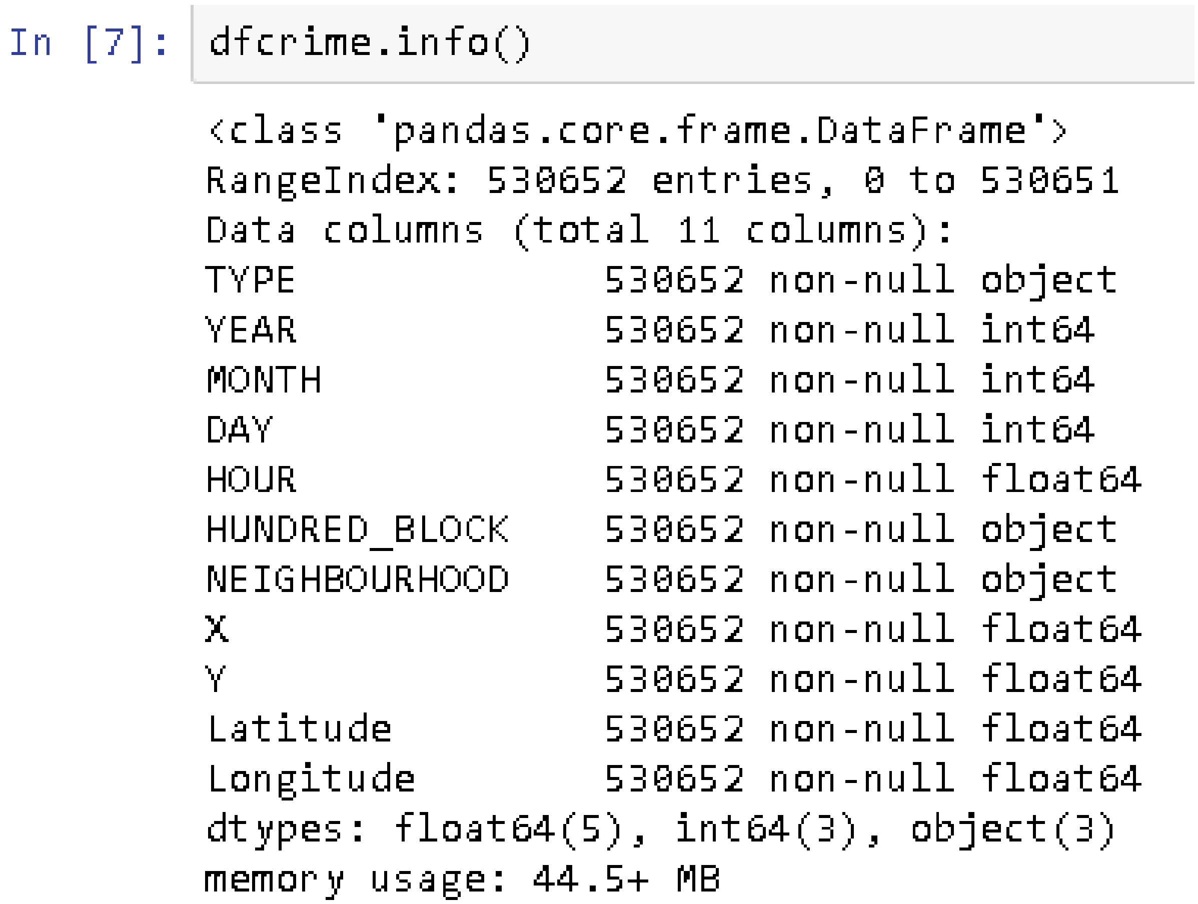
Data overview



```
In [5]: dfcrime.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 530652 entries, 0 to 530651
Data columns (total 12 columns):
TYPE                 530652 non-null object
YEAR                 530652 non-null int64
MONTH                530652 non-null int64
DAY                  530652 non-null int64
HOUR                 476290 non-null float64
MINUTE               476290 non-null float64
HUNDRED_BLOCK        530639 non-null object
NEIGHBOURHOOD        474028 non-null object
X                    530652 non-null float64
Y                    530652 non-null float64
Latitude             530652 non-null float64
Longitude            530652 non-null float64
dtypes: float64(6), int64(3), object(3)
memory usage: 48.6+ MB
```

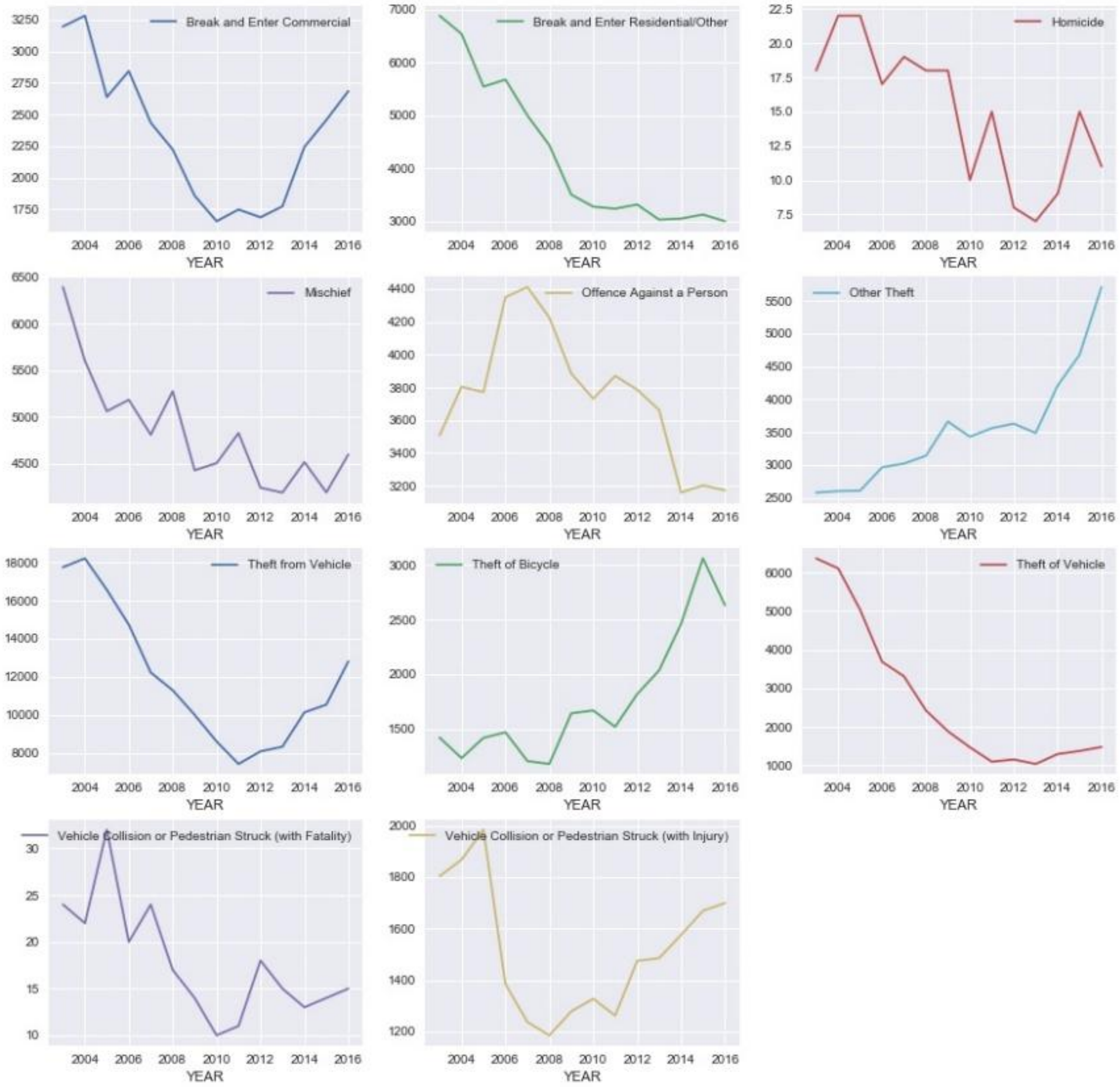
Handling Missing Values



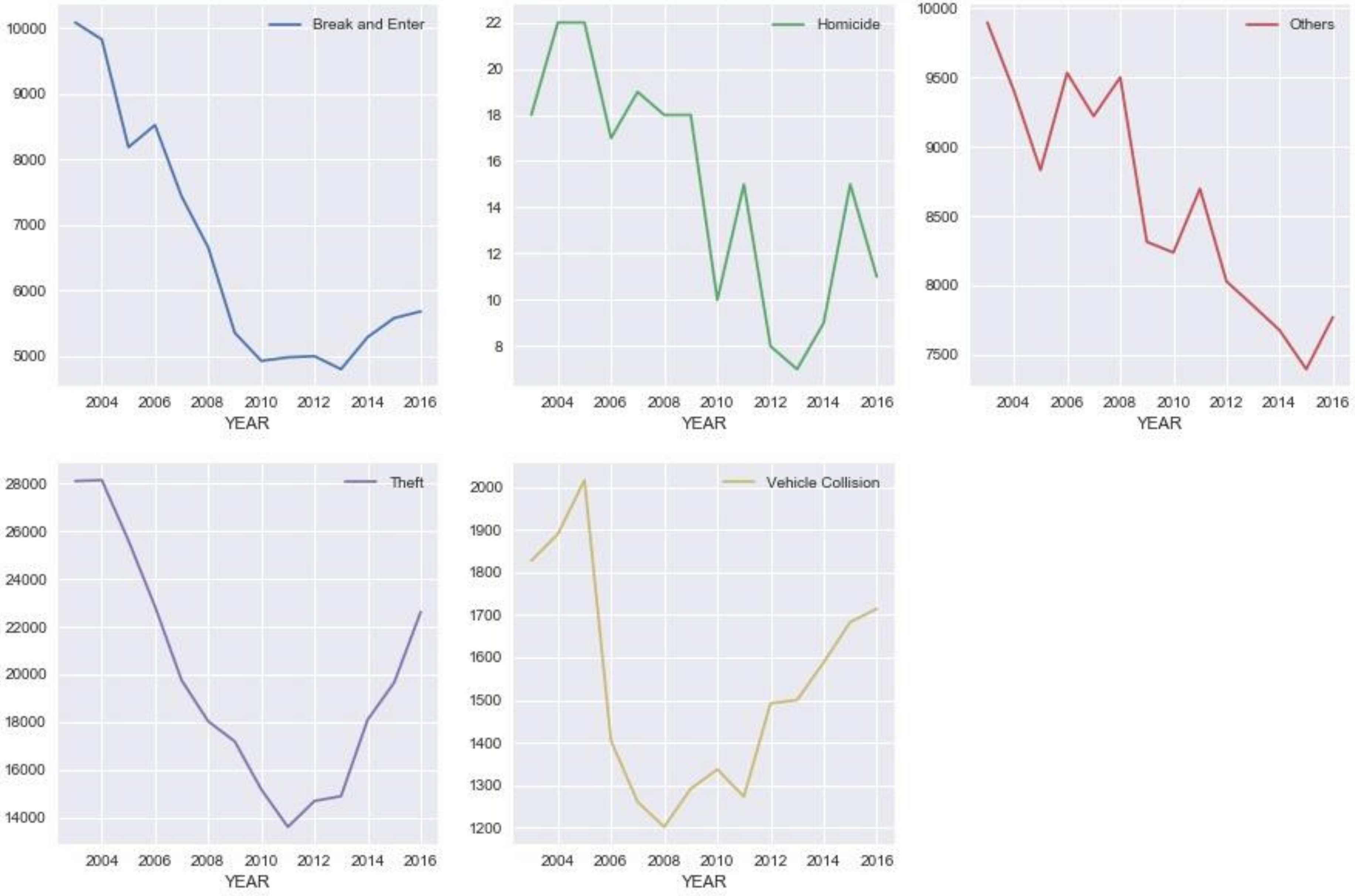
```
In [7]: dfcrime.info()
```

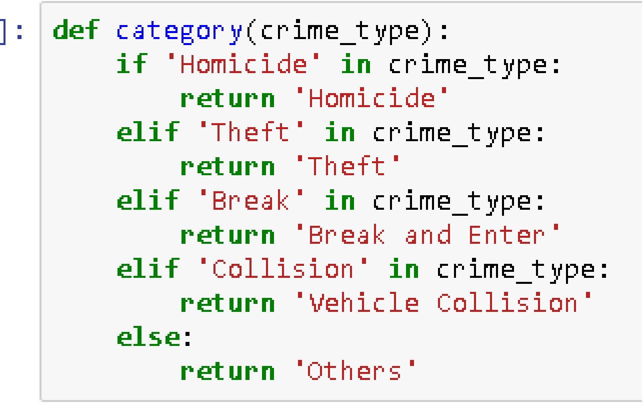
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 530652 entries, 0 to 530651
Data columns (total 11 columns):
TYPE                 530652 non-null object
YEAR                 530652 non-null int64
MONTH                530652 non-null int64
DAY                  530652 non-null int64
HOUR                 530652 non-null float64
HUNDRED_BLOCK        530652 non-null object
NEIGHBOURHOOD        530652 non-null object
X                    530652 non-null float64
Y                    530652 non-null float64
Latitude             530652 non-null float64
Longitude            530652 non-null float64
dtypes: float64(5), int64(3), object(3)
memory usage: 44.5+ MB
```

Crime Types yearly analysis



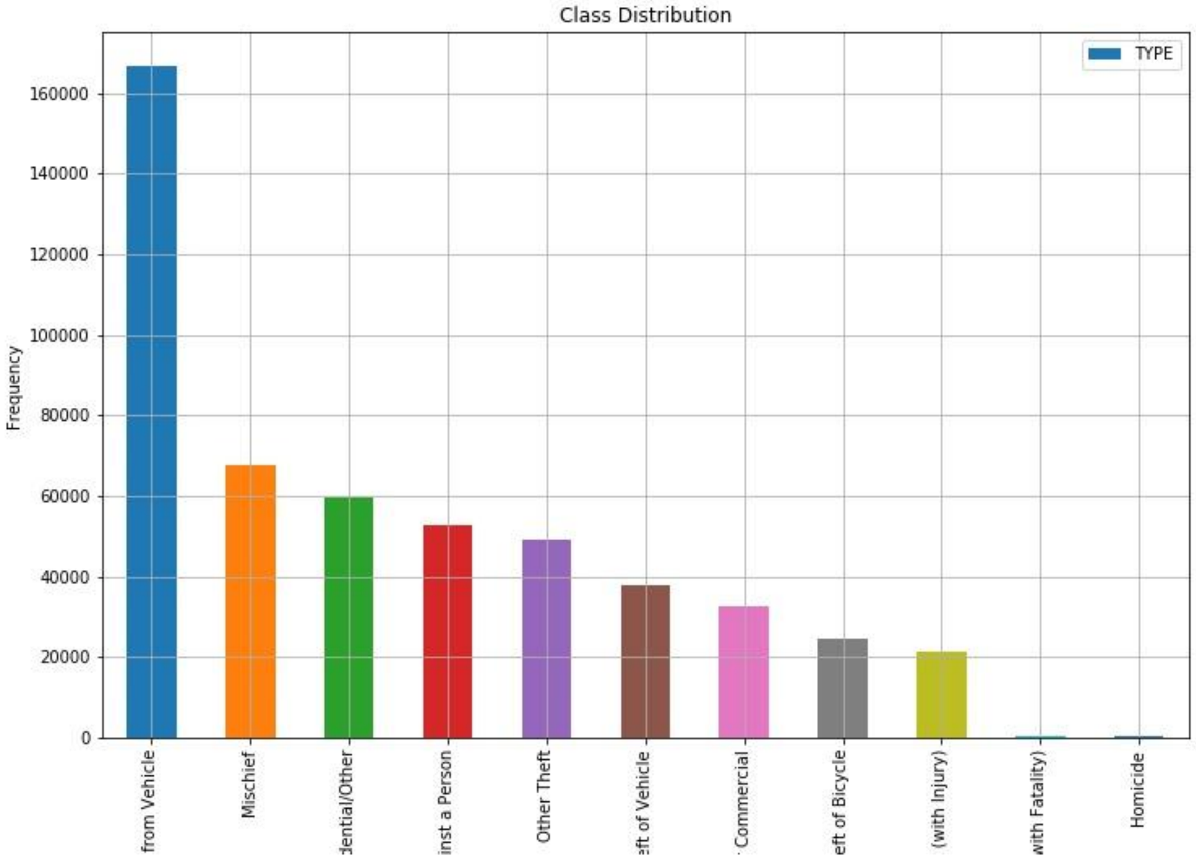
Category wise yearly analysis



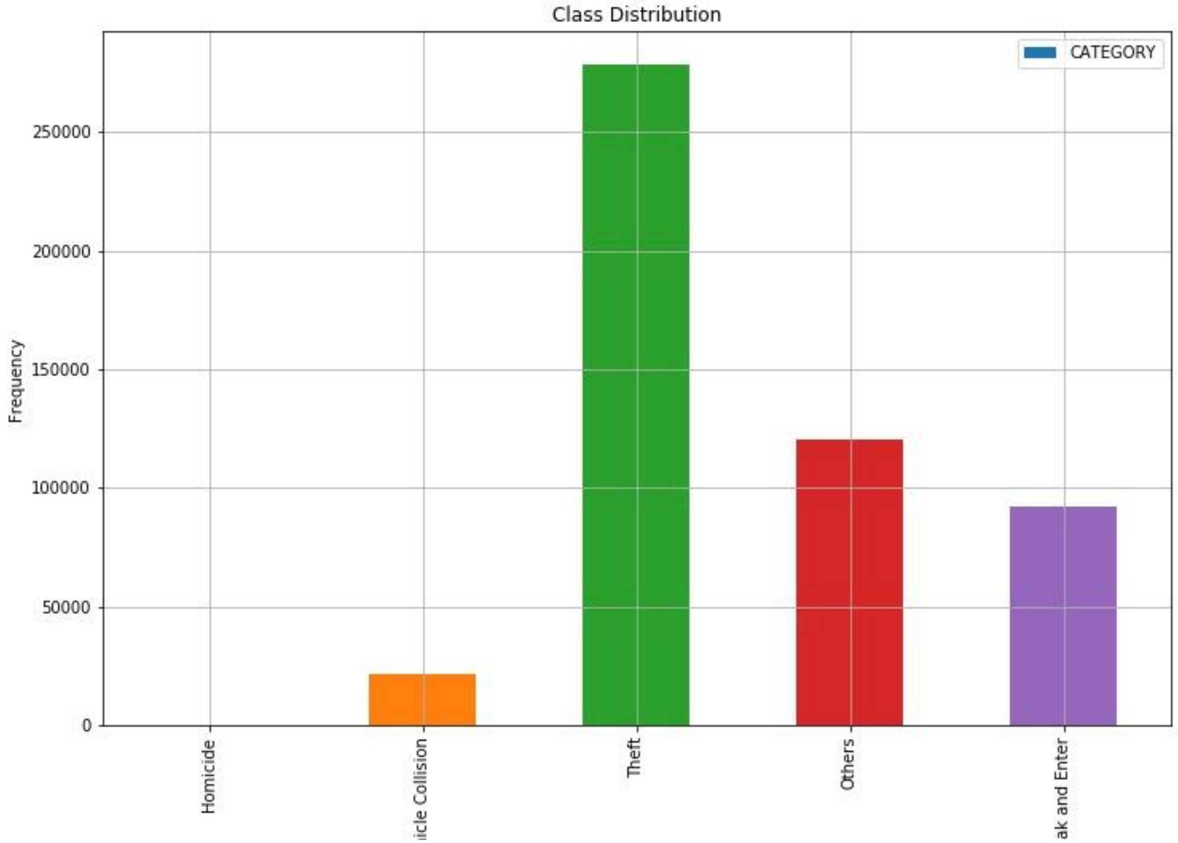


```
def category(crime_type):
    if 'Homicide' in crime_type:
        return 'Homicide'
    elif 'Theft' in crime_type:
        return 'Theft'
    elif 'Break' in crime_type:
        return 'Break and Enter'
    elif 'Collision' in crime_type:
        return 'Vehicle Collision'
    else:
        return 'Others'
```

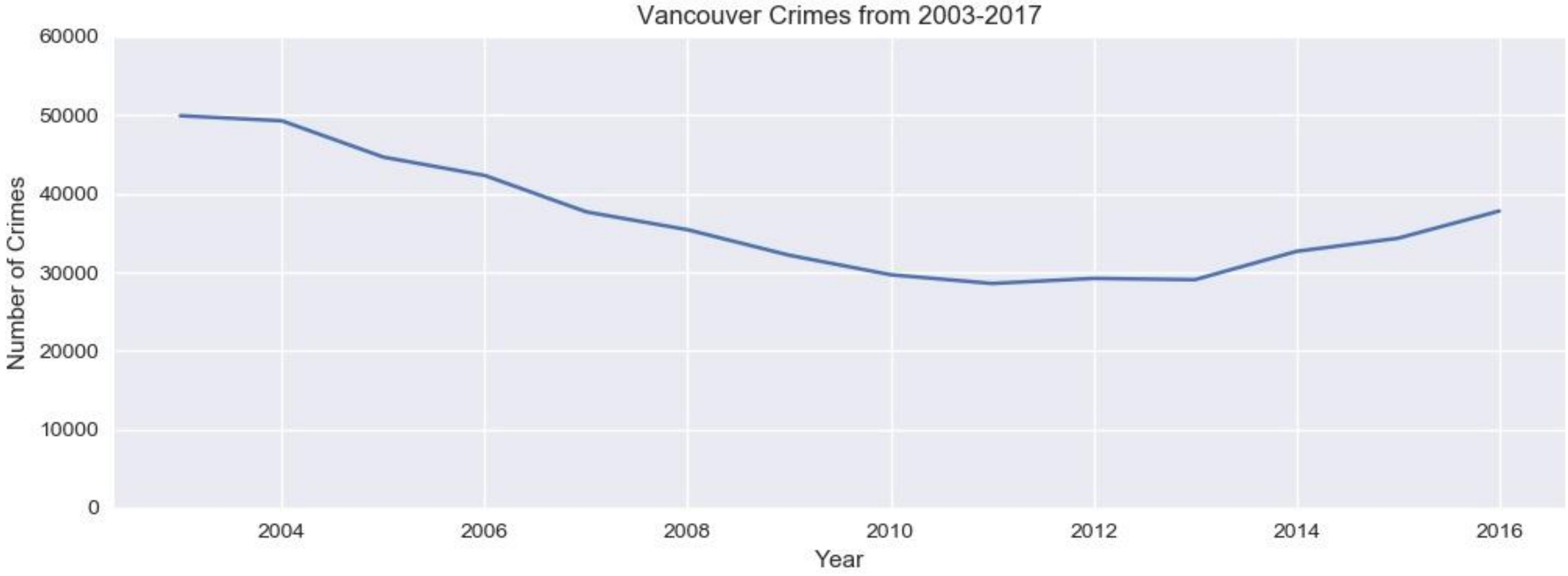
Crimes by Types



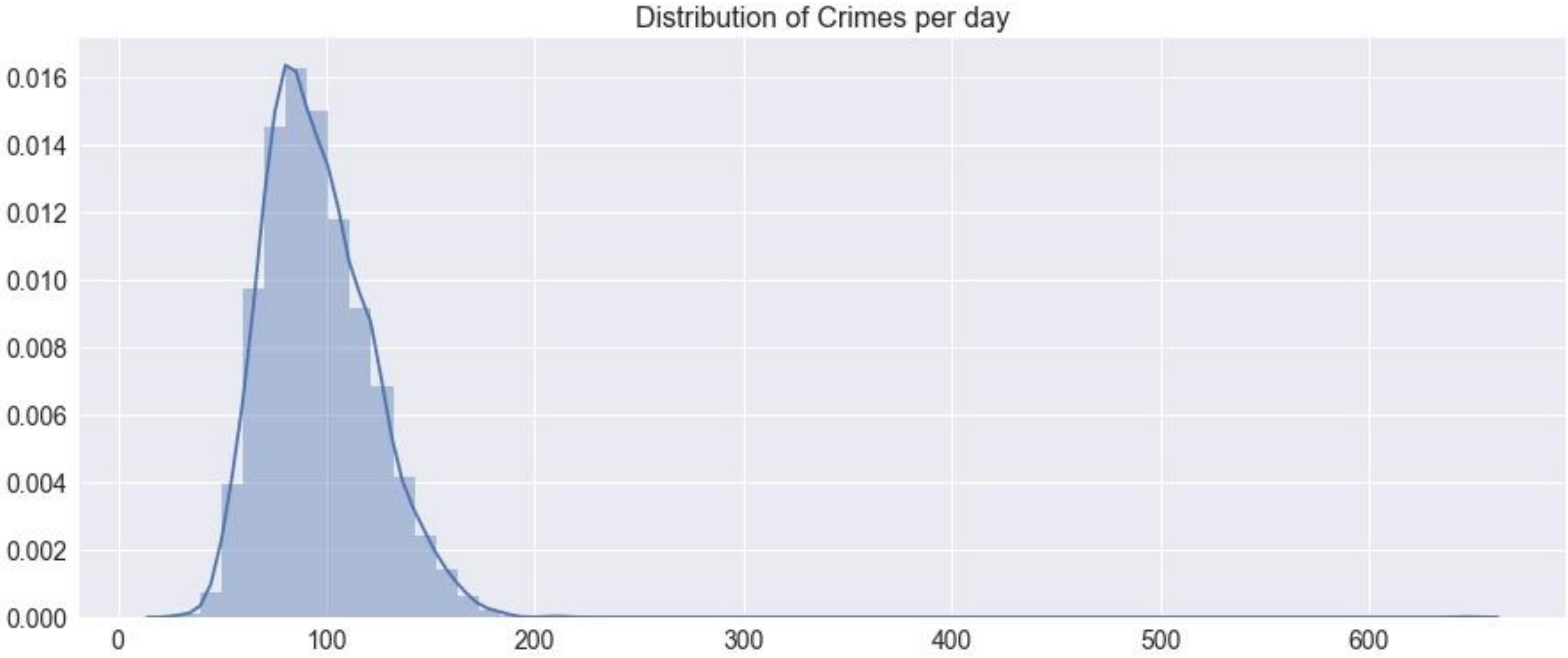
Crimes by Category



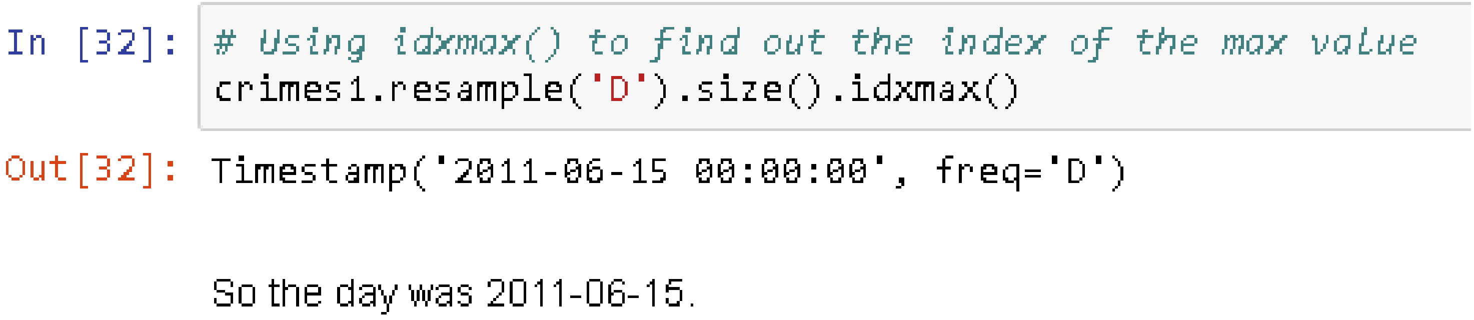
Number of Crimes Pattern/Trend



Crimes Distribution Per Day

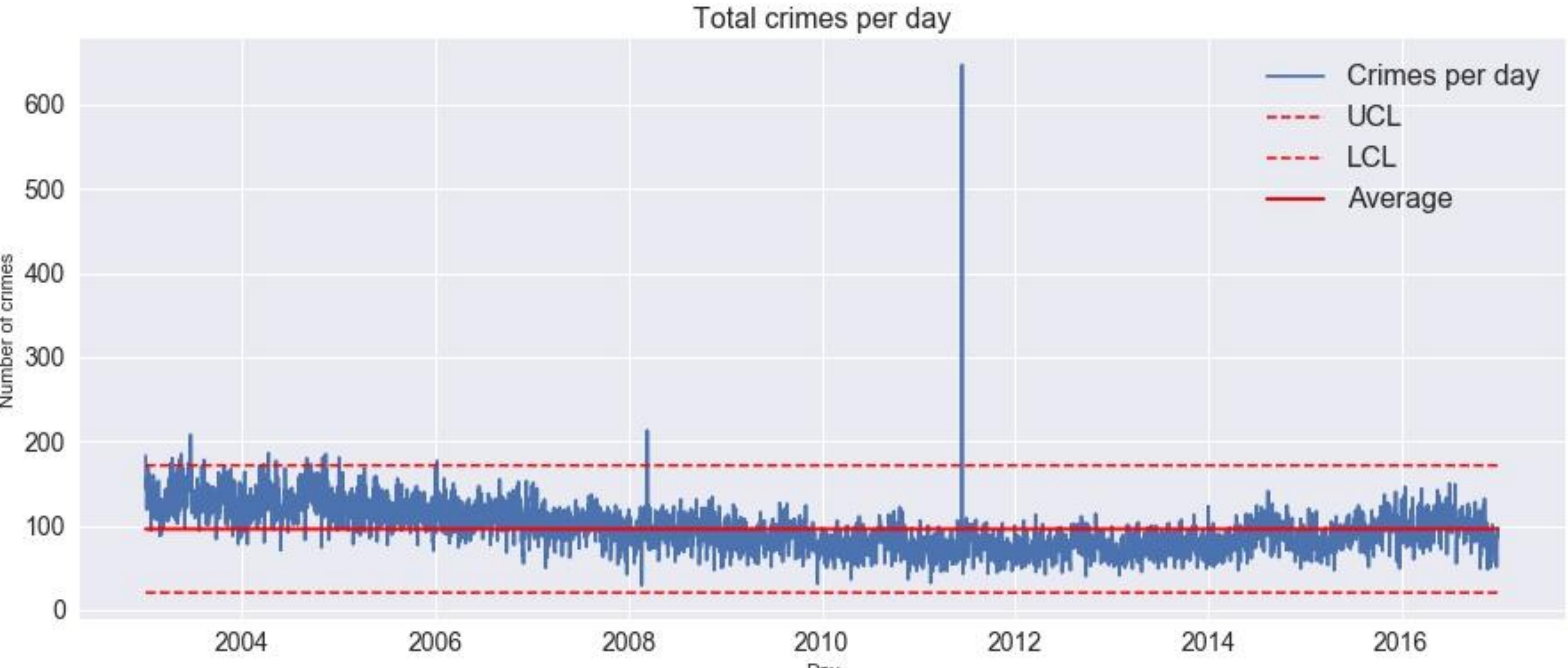


Crimes Per day

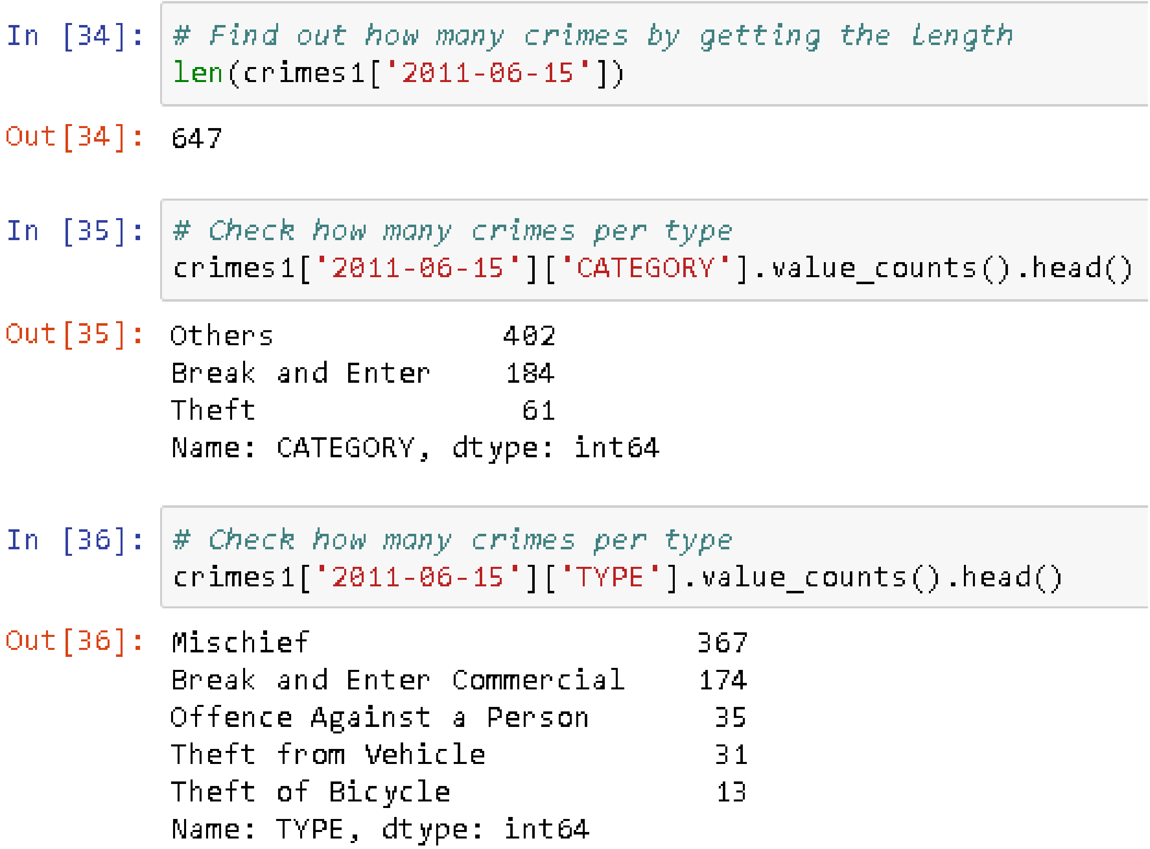


```
In [32]: # Using idxmax() to find out the index of the max value
         crimes1.resample('D').size().idxmax()

Out[32]: Timestamp('2011-06-15 00:00:00', freq='D')
```

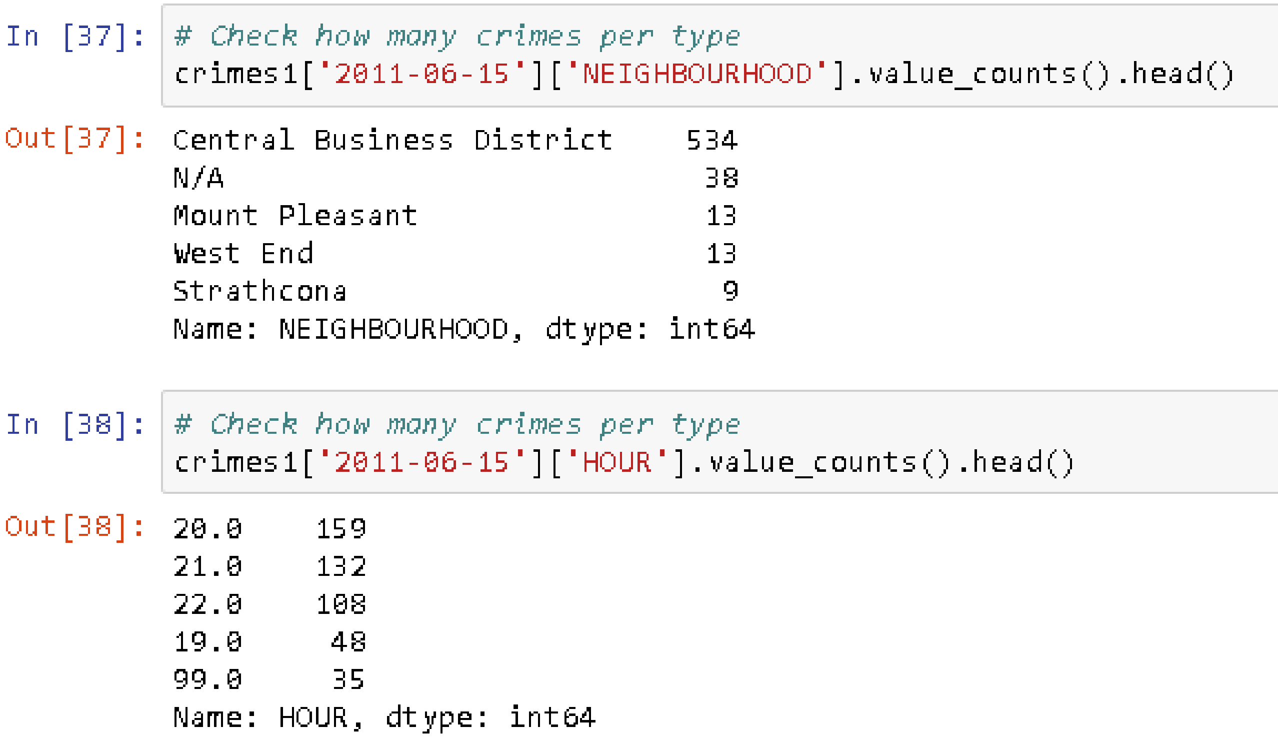
So the day was 2011-06-15.


Total crimes per day

Outlier



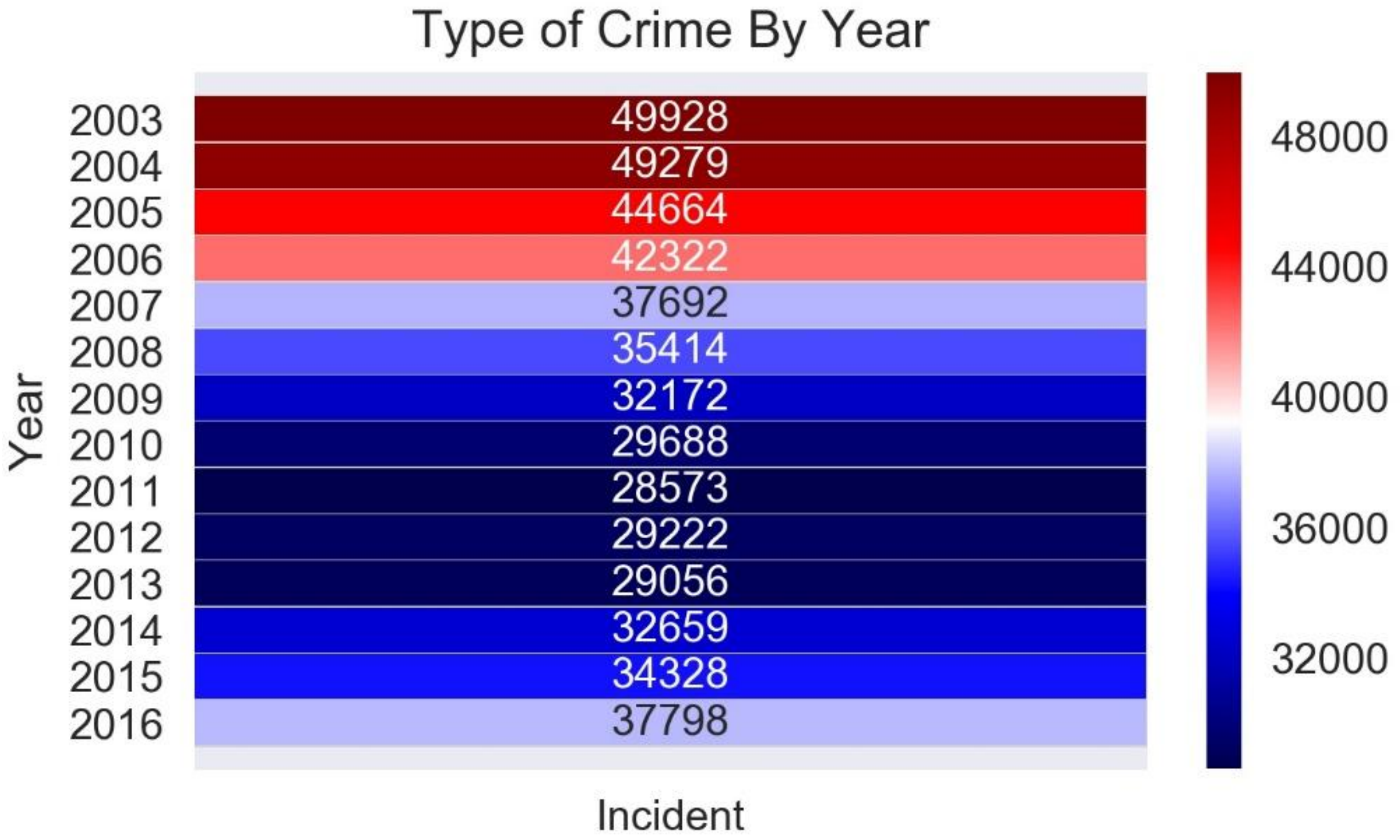
```
In [34]: # Find out how many crimes by getting the length
         len(crimes1['2011-06-15'])

Out[34]: 647

In [35]: # Check how many crimes per type
         crimes1['2011-06-15']['CATEGORY'].value_counts().head()

Out[35]: Others              402
         Break and Enter     184
         Theft                61
         Name: CATEGORY, dtype: int64

In [36]: # Check how many crimes per type
         crimes1['2011-06-15']['TYPE'].value_counts().head()

Out[36]: Mischief                        367
         Break and Enter Commercial      174
         Offence Against a Person         35
         Theft from Vehicle               31
         Theft of Bicycle                 13
         Name: TYPE, dtype: int64
```



```
In [37]: # Check how many crimes per type
         crimes1['2011-06-15']['NEIGHBOURHOOD'].value_counts().head()

Out[37]: Central Business District    534
         N/A                           38
         Mount Pleasant                13
         West End                      13
         Strathcona                     9
         Name: NEIGHBOURHOOD, dtype: int64

In [38]: # Check how many crimes per type
         crimes1['2011-06-15']['HOUR'].value_counts().head()

Out[38]: 20.0    159
         21.0    132
         22.0    108
         19.0     48
         99.0     35
         Name: HOUR, dtype: int64
```

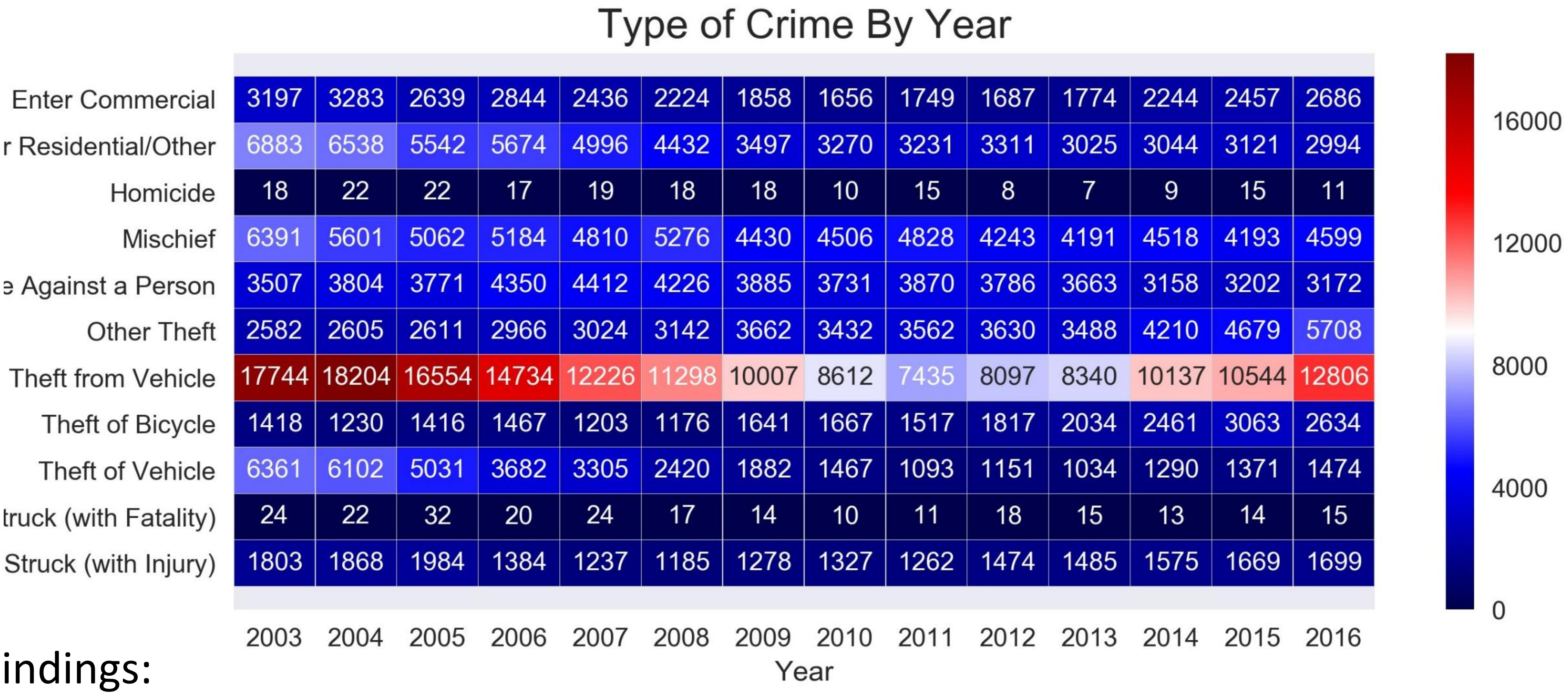
- There are 647 occurrences, mostly mischief type, in Central Business District, around 20:00-22:00.

Crimes By Year



Findings: Initial Year from 2003-2006 were the worst in case of crimes

Crimes By Year By Type



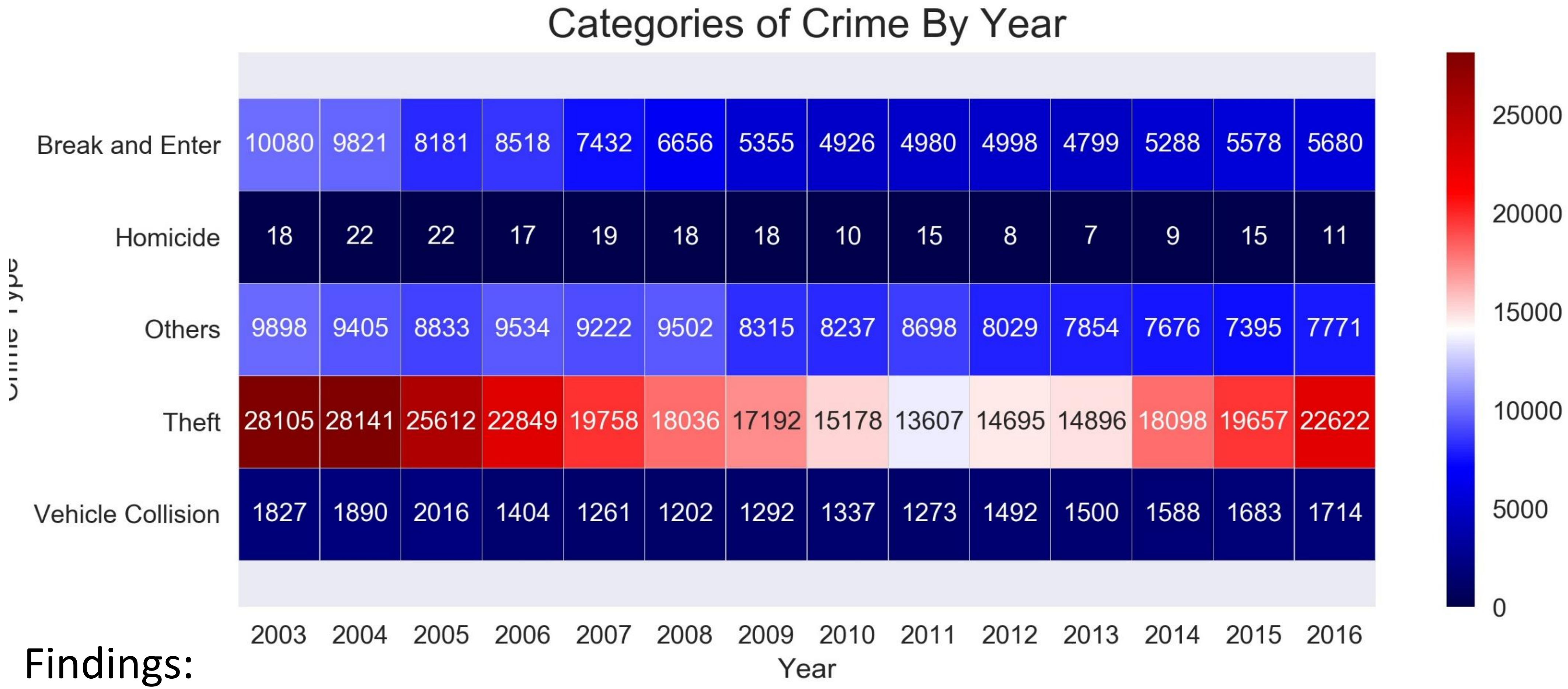
Type of Crime By Year

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Enter Commercial	3197	3283	2639	2844	2436	2224	1858	1656	1749	1687	1774	2244	2457	2686
r Residential/Other	6883	6538	5542	5674	4996	4432	3497	3270	3231	3311	3025	3044	3121	2994
Homicide	18	22	22	17	19	18	18	10	15	8	7	9	15	11
Mischief	6391	5601	5062	5184	4810	5276	4430	4506	4828	4243	4191	4518	4193	4599
e Against a Person	3507	3804	3771	4350	4412	4226	3885	3731	3870	3786	3663	3158	3202	3172
Other Theft	2582	2605	2611	2966	3024	3142	3662	3432	3562	3630	3488	4210	4679	5708
Theft from Vehicle	17744	18204	16554	14734	12226	11298	10007	8612	7435	8097	8340	10137	10544	12806
Theft of Bicycle	1418	1230	1416	1467	1203	1176	1641	1667	1517	1817	2034	2461	3063	2634
Theft of Vehicle	6361	6102	5031	3682	3305	2420	1882	1467	1093	1151	1034	1290	1371	1474
truck (with Fatality)	24	22	32	20	24	17	14	10	11	18	15	13	14	15
Struck (with Injury)	1803	1868	1984	1384	1237	1185	1278	1327	1262	1474	1485	1575	1669	1699

Findings:

1. Theft from Vehicle are the worst kind of crime.
2. Least occurring crimes are 'Homicide' and 'Vehicle Collision(with Fatality)'

Crimes By Year By Categories

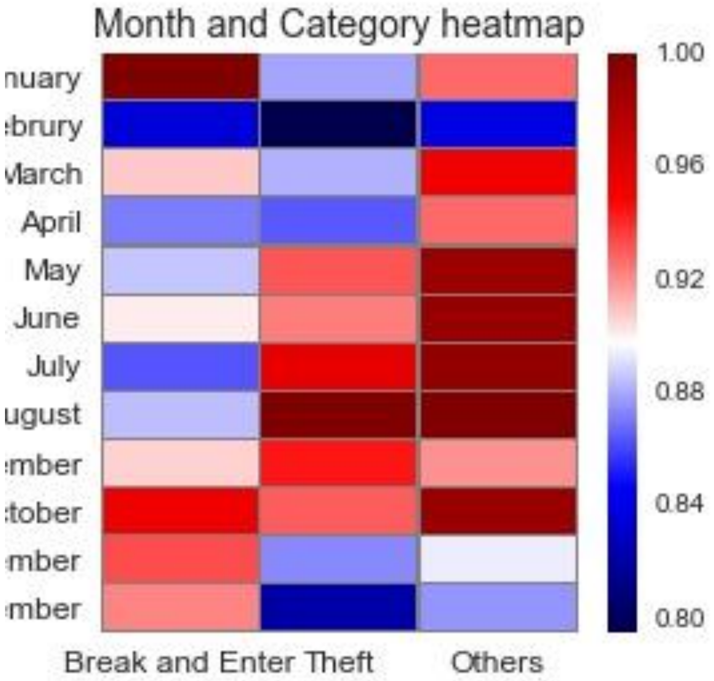


Categories of Crime By Year

Crime Type	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Break and Enter	10080	9821	8181	8518	7432	6656	5355	4926	4980	4998	4799	5288	5578	5680
Homicide	18	22	22	17	19	18	18	10	15	8	7	9	15	11
Others	9898	9405	8833	9534	9222	9502	8315	8237	8698	8029	7854	7676	7395	7771
Theft	28105	28141	25612	22849	19758	18036	17192	15178	13607	14695	14896	18098	19657	22622
Vehicle Collision	1827	1890	2016	1404	1261	1202	1292	1337	1273	1492	1500	1588	1683	1714

Findings:

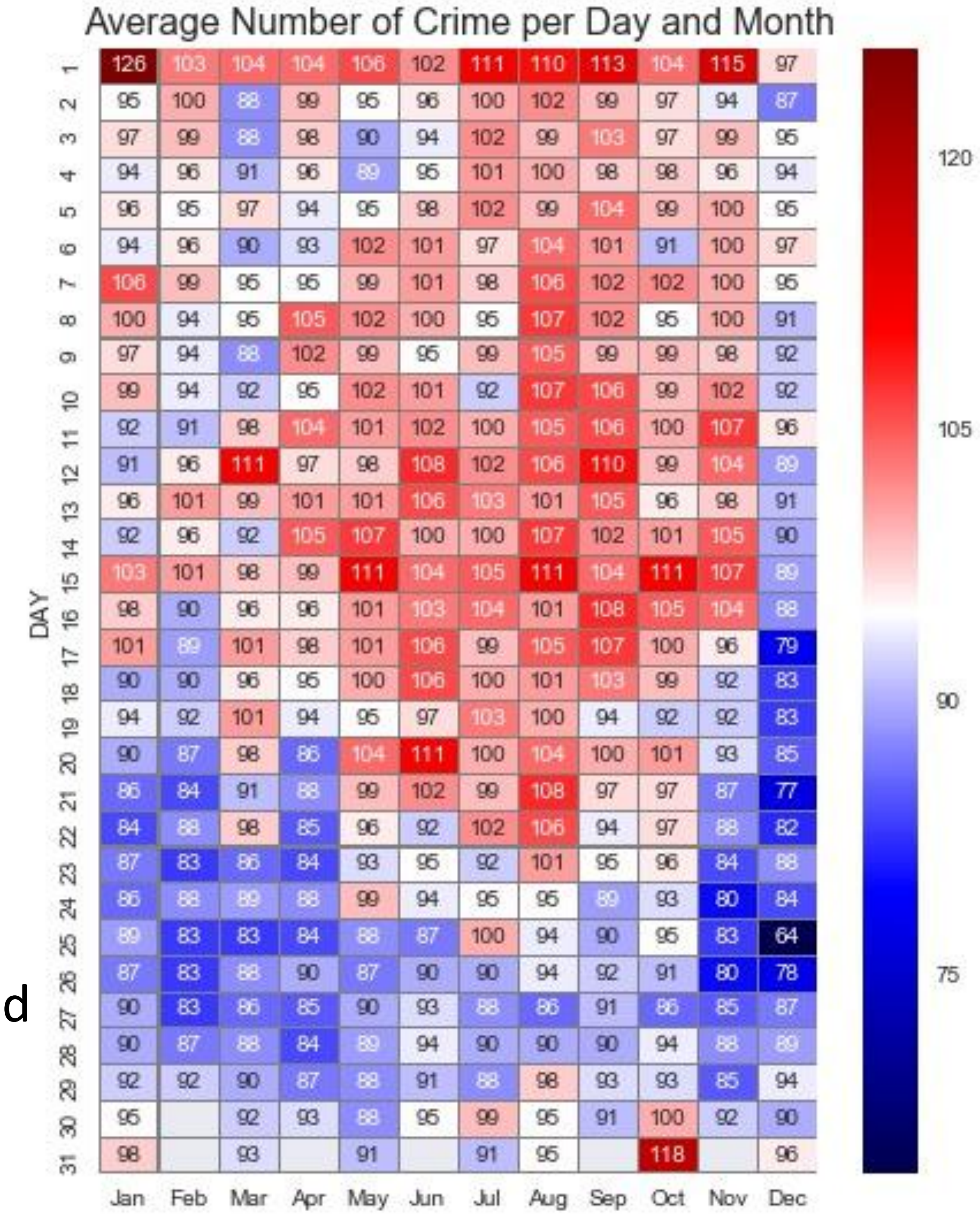
1. Theft is the most occurring Crime
2. Homicide is the least occurring Crime



Findings:

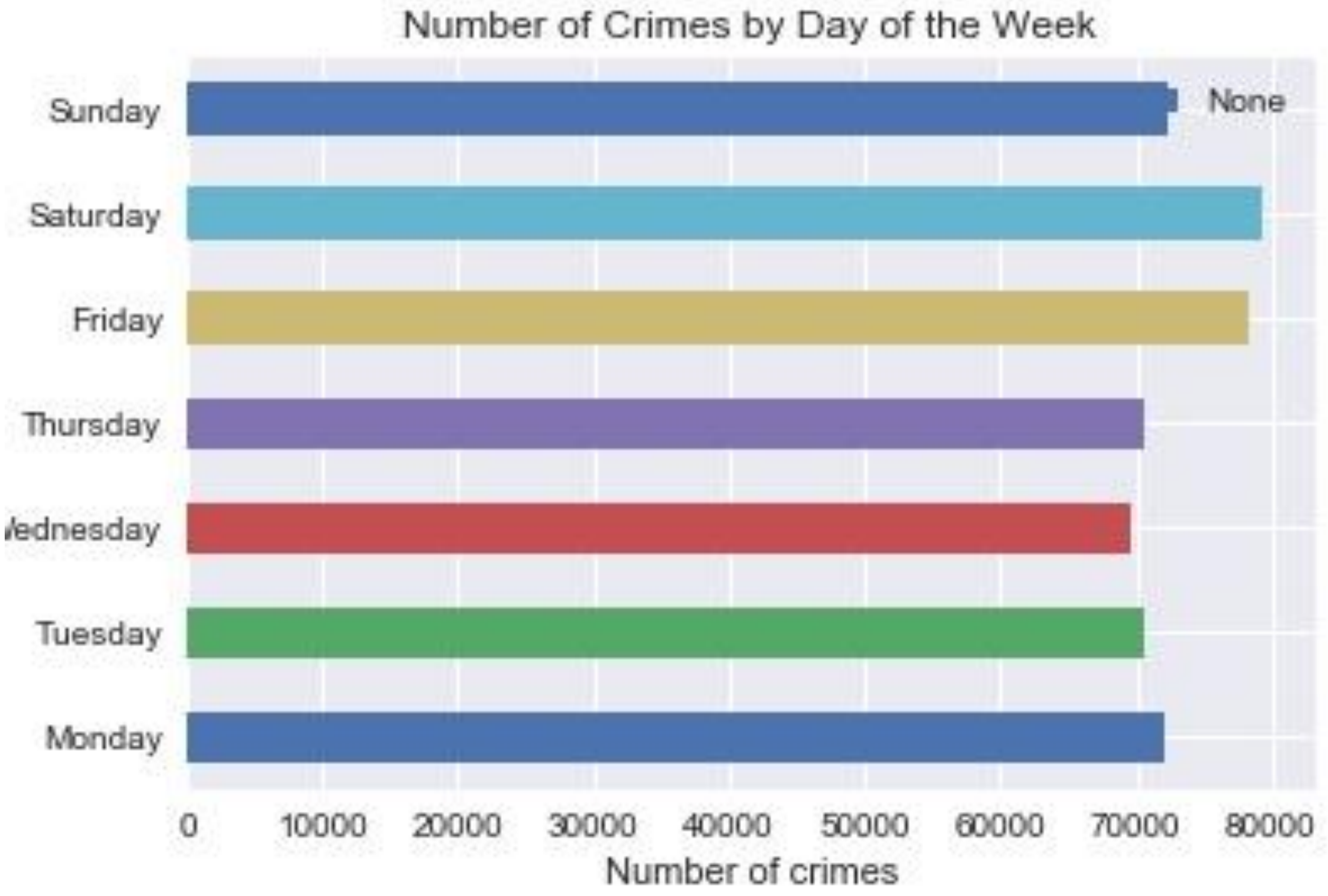
Blue means good days. Red means bad days.
White average days.

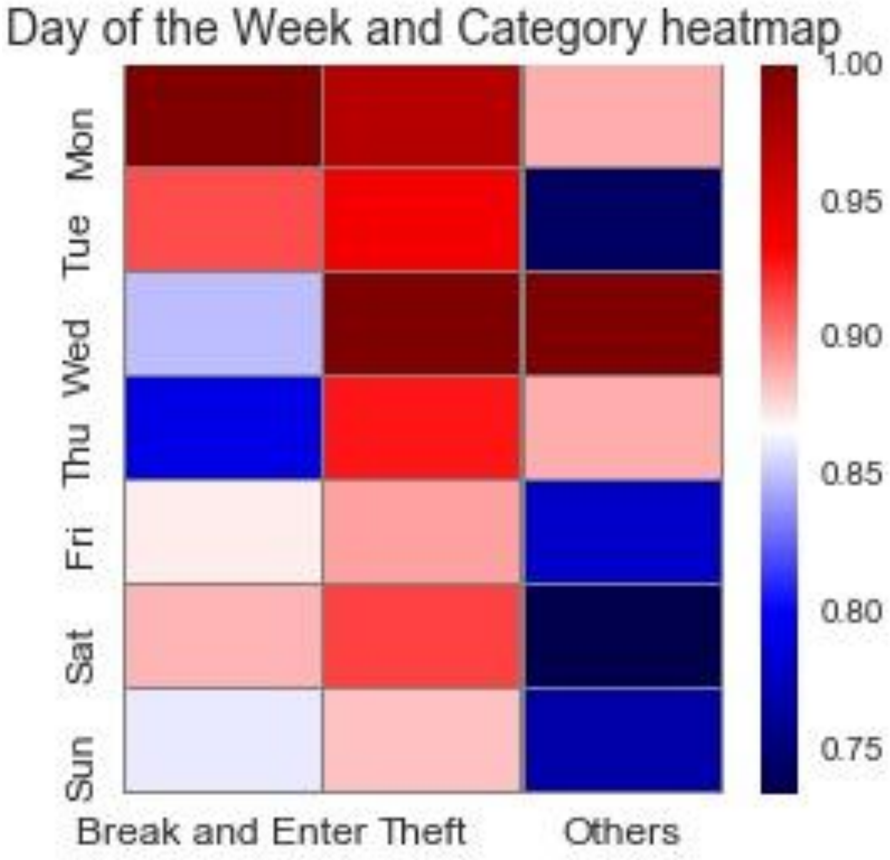
- The Calmest day of crime is Christmas Day. December 25(30% below average).
- The worst day is New Year's Day, January 1 and October 30-November 1 (Halloween).
- The first day of the month is a busy day for all month.



DAY	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	126	103	104	104	106	102	111	110	113	104	115	97
2	95	100	88	99	95	96	100	102	99	97	94	87
3	97	99	88	98	90	94	102	99	103	97	99	95
4	94	96	91	96	89	95	101	100	98	98	96	94
5	96	95	97	94	95	98	102	99	104	99	100	95
6	94	96	90	93	102	101	97	104	101	91	100	97
7	106	99	95	95	99	101	98	106	102	102	100	95
8	100	94	95	105	102	100	95	107	102	95	100	91
9	97	94	88	102	99	95	99	105	99	99	98	92
10	99	94	92	95	102	101	92	107	106	99	102	92
11	92	91	98	104	101	102	100	105	106	100	107	96
12	91	96	111	97	98	108	102	106	110	99	104	89
13	96	101	99	101	101	106	103	101	105	96	98	91
14	92	96	92	105	107	100	100	107	102	101	105	90
15	103	101	98	99	111	104	105	111	104	111	107	89
16	98	90	96	96	101	103	104	101	108	105	104	88
17	101	89	101	98	101	106	99	105	107	100	96	79
18	90	90	96	95	100	106	100	101	103	99	92	83
19	94	92	101	94	95	97	103	100	94	92	92	83
20	90	87	98	86	104	111	100	104	100	101	93	85
21	86	84	91	88	99	102	99	108	97	97	87	77
22	84	88	98	85	96	92	102	106	94	97	88	82
23	87	83	86	84	93	95	92	101	95	96	84	88
24	86	88	89	88	99	94	95	95	89	93	80	84
25	89	83	83	84	88	87	100	94	90	95	83	64
26	87	83	88	90	87	90	90	94	92	91	80	78
27	90	83	86	85	90	93	88	86	91	86	85	87
28	90	87	88	84	89	94	90	90	90	94	88	89
29	92	92	90	87	88	91	88	98	93	93	85	94
30	95		92	93	88	95	99	95	91	100	92	90
31	98		93		91		91	95		118		96

Number of Crimes by Day of the Week



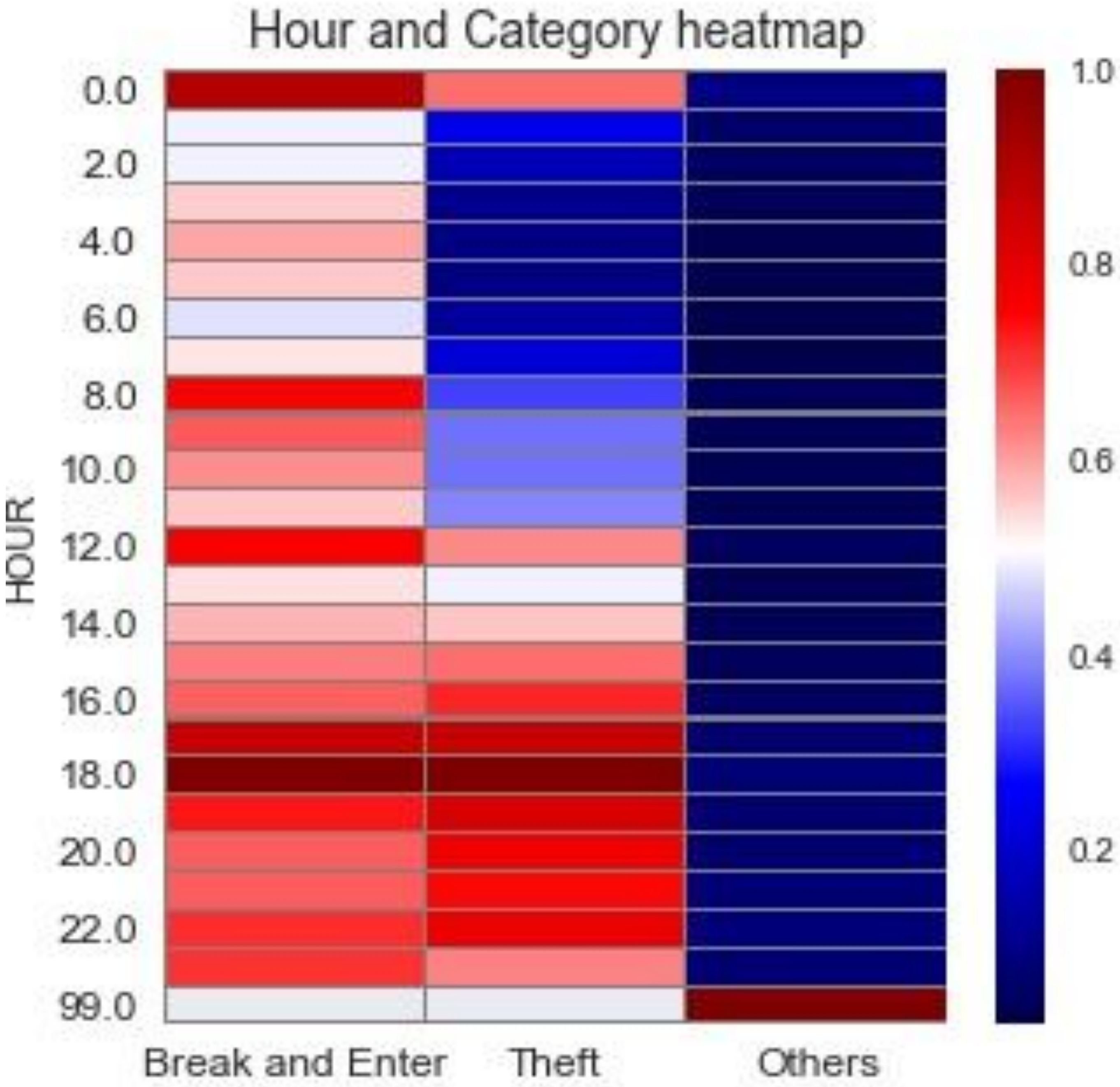


Findings: Weekend +Friday are most prone to Crimes

What hours do crime happen?

Findings

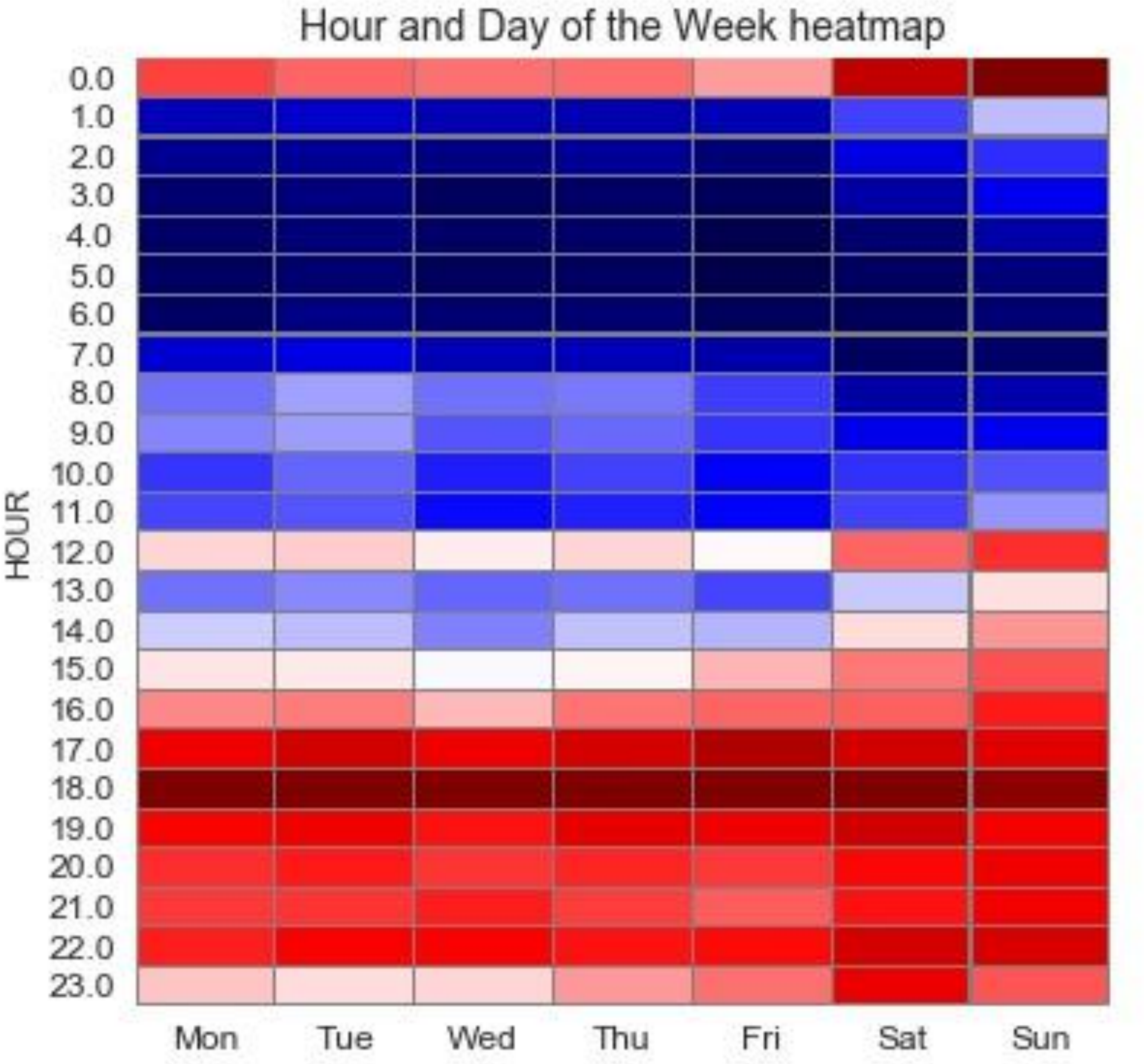
1. Most crimes happen between 17:00-01:00
2. Category Others doesn't have Hours mentioned in the dataset in most cases



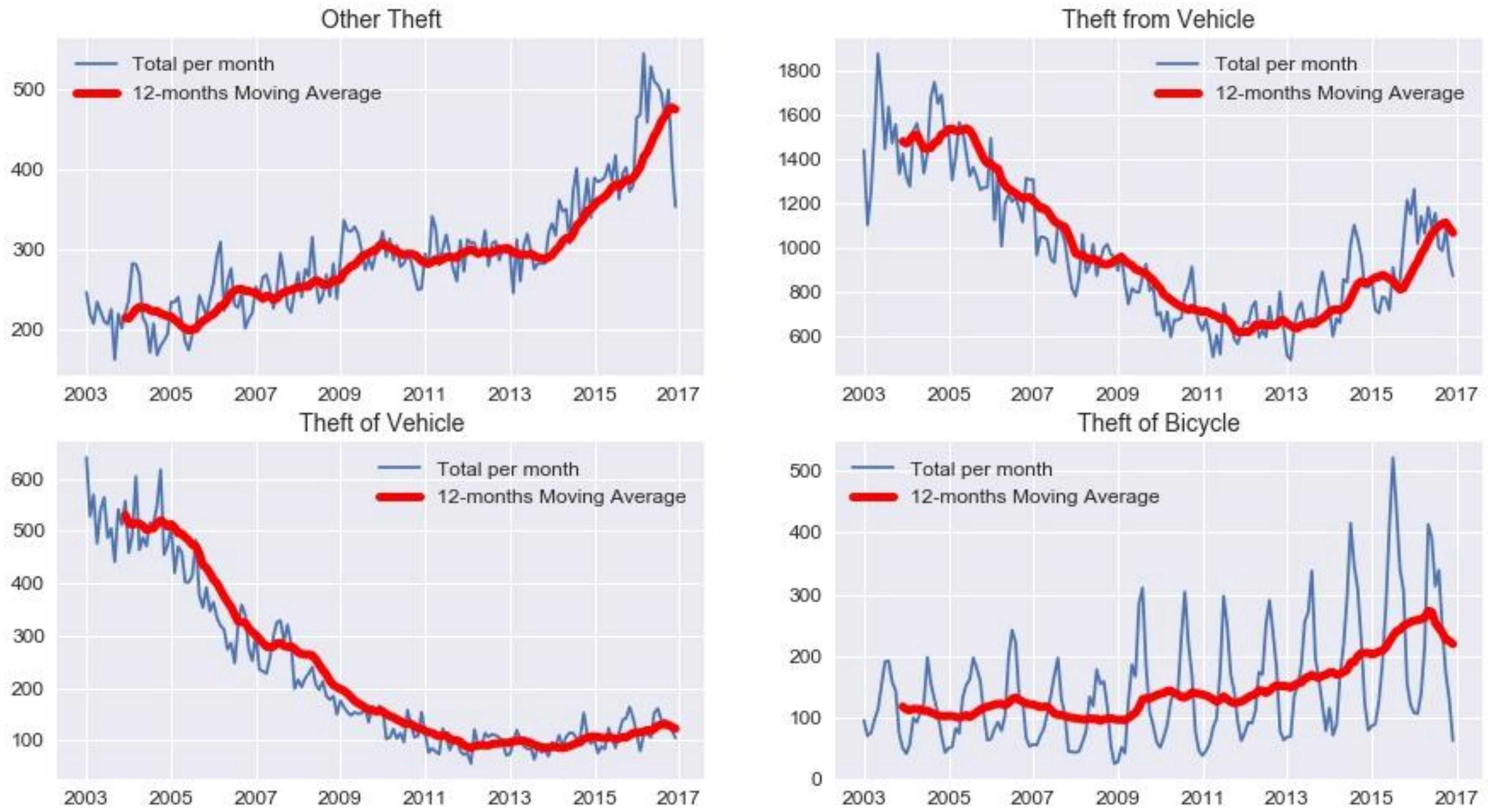
Do Crimes happen in the same hour for each day of the week?

Findings

1. on Weekends: the crimes activity starts at 15:00-00 peaks at 17:00-23:00
2. on weekdays: the crimes activity starts at 16-00:00 peaks at 17:00-22:00



Each type of crimes general trend



Findings

Other Theft

- 1. This trend has been increasing. from around 200 to almost 500 crimes per month.

Theft from Vehicle

- 1. it is the most frequent type of crime.
- 2. This trend has been decreasing till 2012 from 1600 to 600 and then it has increased to 1200 in 2017.

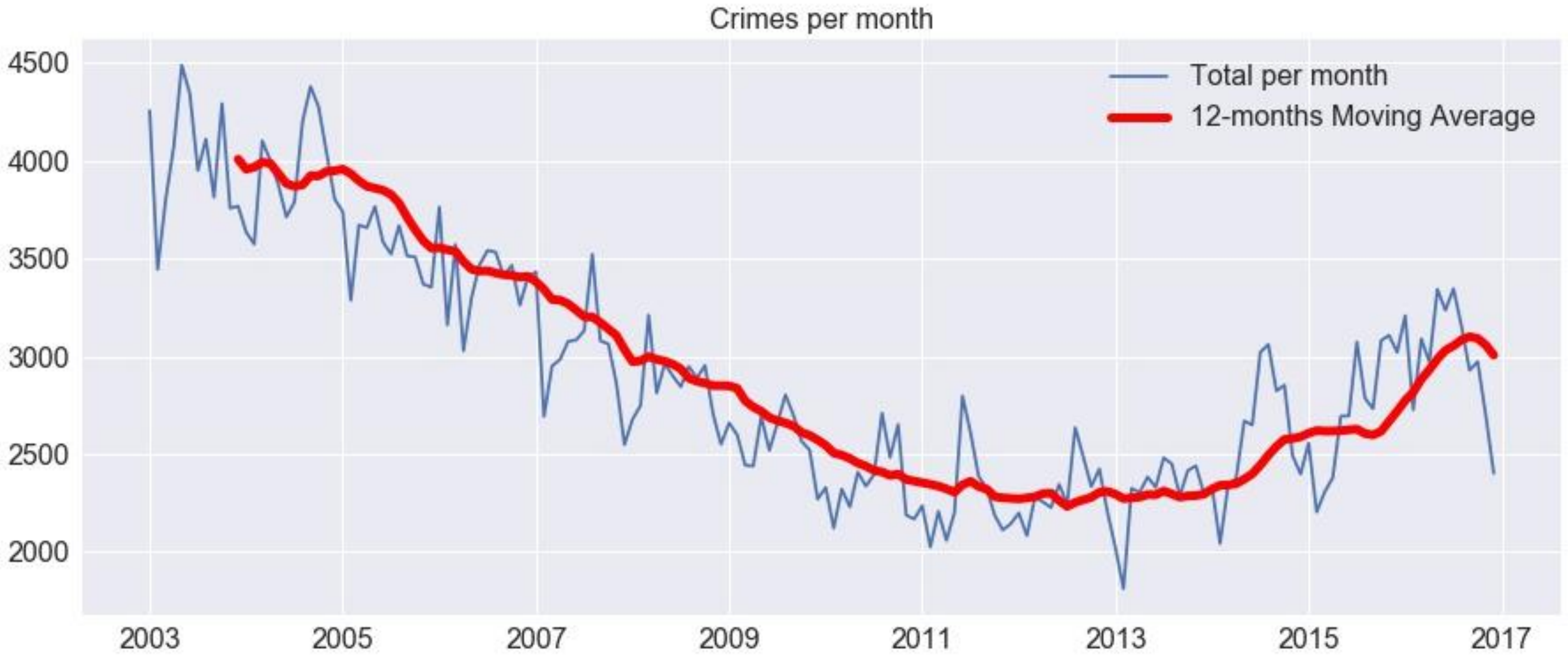
Theft of Vehicle

- 1. this crime has decreased from 600ish to almost 100

Theft of Bicycle

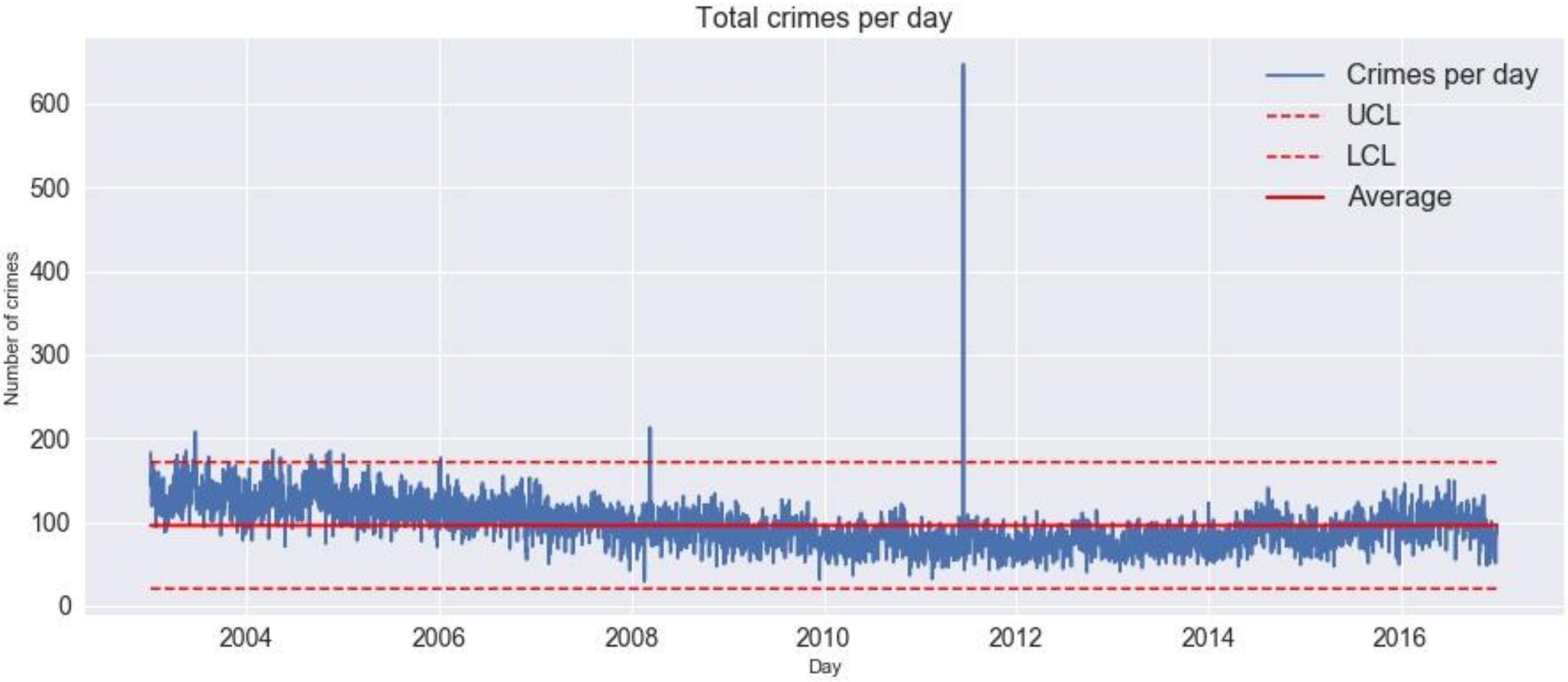
- 1. we can see the trend from graph that this crimes peak during the mid of year: summer.
- 2. The average has also been increasing.

Is the Crime decreasing or increasing-Year/Month

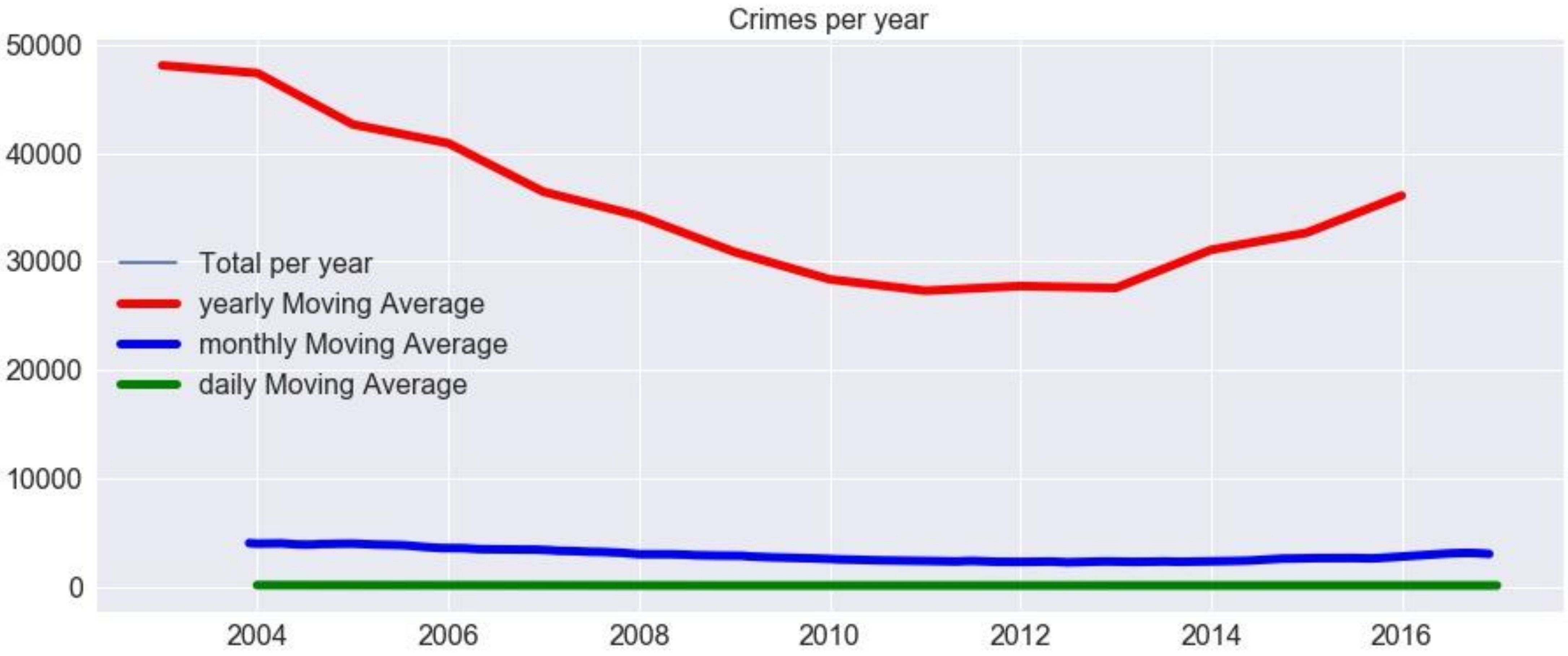


From 2003 to 2011 the average number of crimes per month decreased from 4000 crimes per month to arround 2400.
From 2011 to 2014, the moving average was around the same.
From 2014 to 2015 the average has increased.
From 2016 reached similar levels of 2008

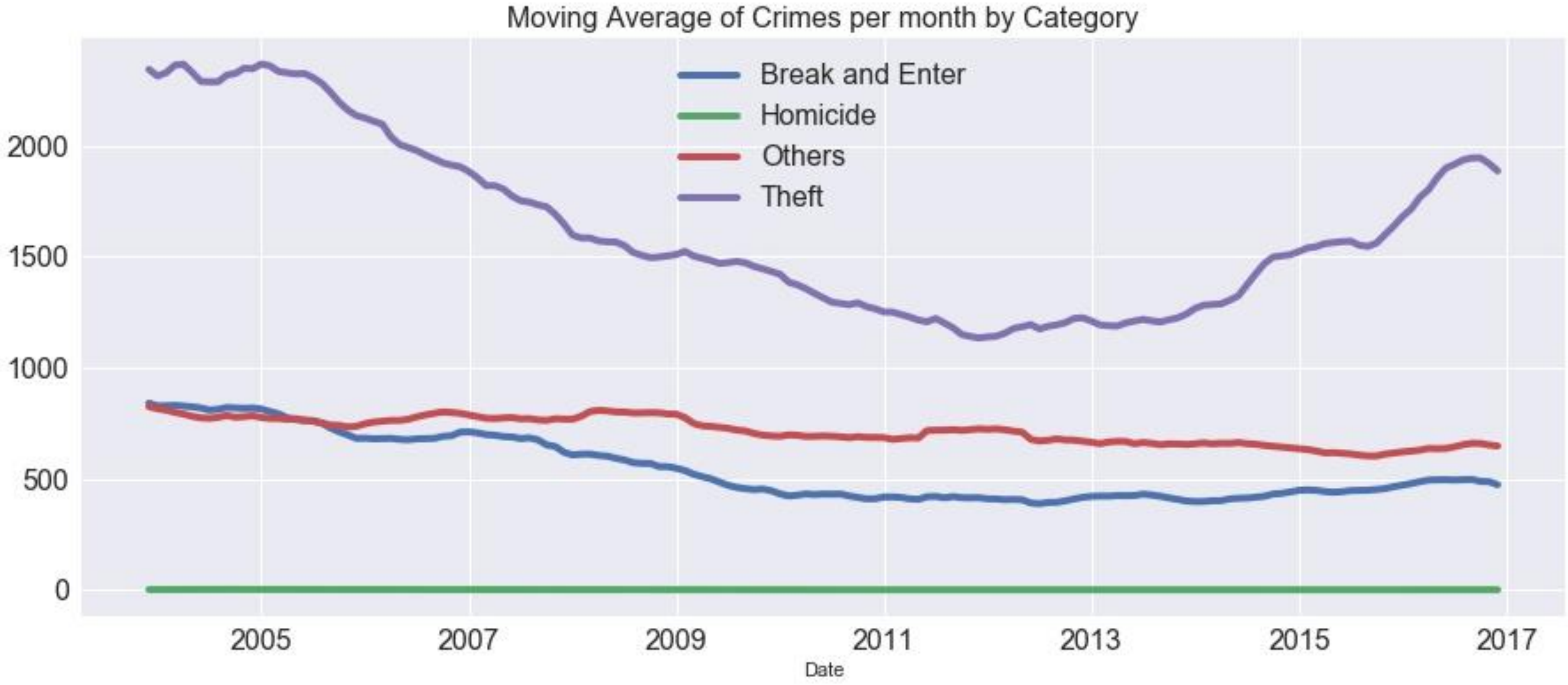
Moving average Crimes Per Day Data



Moving average Crimes Data (Year/Month/Day)



Is the trend the same for all categories?

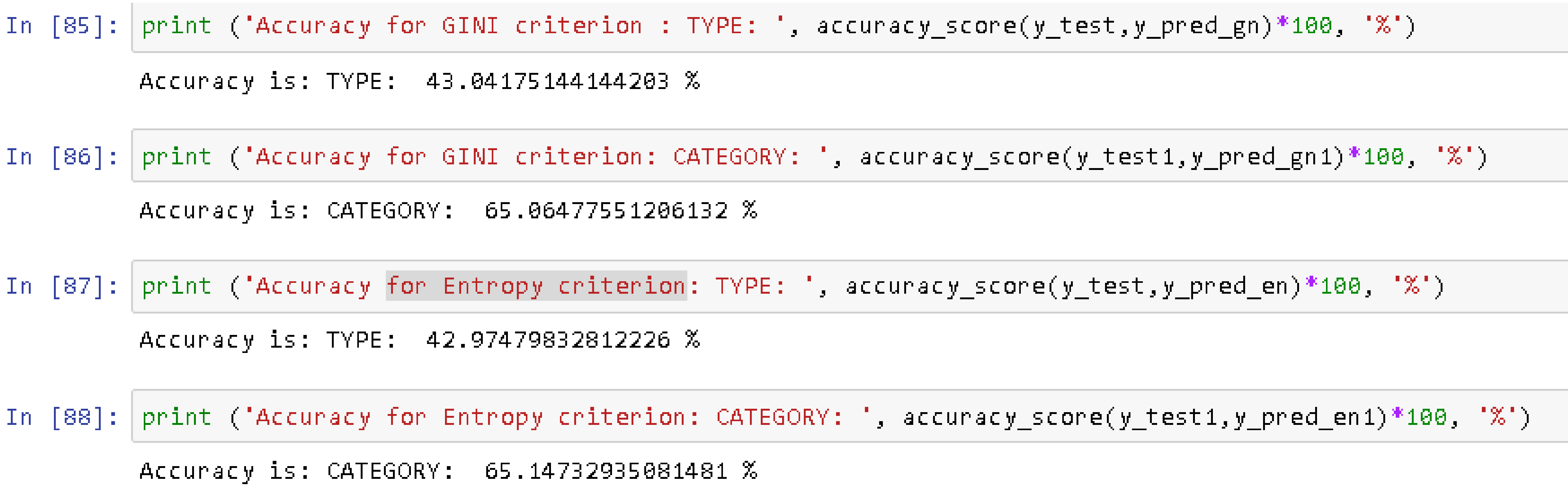


Findings: Theft is the major category.

The decrease and increase that we saw in the average number of crimes per month was mainly because of the variations in this category.

PROCEDURE/TECHNIQUES USED

- Decision Tree Classifier Applied



```
In [85]: print ('Accuracy for GINI criterion : TYPE: ', accuracy_score(y_test,y_pred_gn)*100, '%')

         Accuracy is: TYPE:  43.04175144144203 %

In [86]: print ('Accuracy for GINI criterion: CATEGORY: ', accuracy_score(y_test1,y_pred_gn1)*100, '%')

         Accuracy is: CATEGORY:  65.06477551206132 %

In [87]: print ('Accuracy for Entropy criterion: TYPE: ', accuracy_score(y_test,y_pred_en)*100, '%')

         Accuracy is: TYPE:  42.97479832812226 %

In [88]: print ('Accuracy for Entropy criterion: CATEGORY: ', accuracy_score(y_test1,y_pred_en1)*100, '%')

         Accuracy is: CATEGORY:  65.14732935081481 %
```