Is Vancouver, CA Safe?

Disha Saha

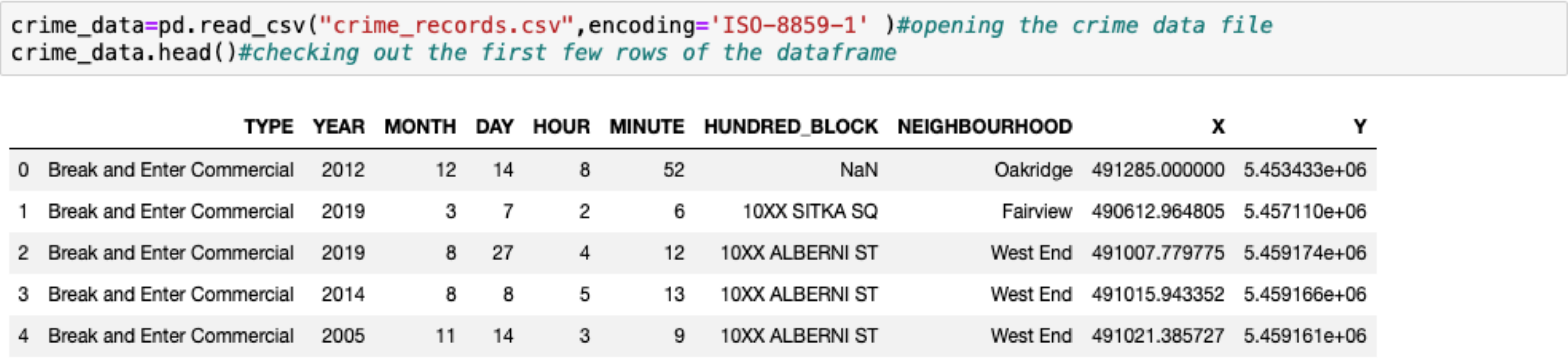
DSC530-T301 Data Exploration and Analysis

Goals of the project

For my final project I want to figure out how safe is Vancouver City, therefore I want to investigate the following questions:

- Is there a particular crime in Vancouver, CA that has become more prevalent?
- Is crime increasing throughout the years in Vancouver, CA?
- Do colder months have less crime prevalence?
- Which neighborhoods have the most crime?
- Can we predict when crimes are most likely to occur?

Looking at the data set



```
crime_data=pd.read_csv("crime_records.csv",encoding='ISO-8859-1' )#opening the crime data file
crime_data.head()#checking out the first few rows of the dataframe
```

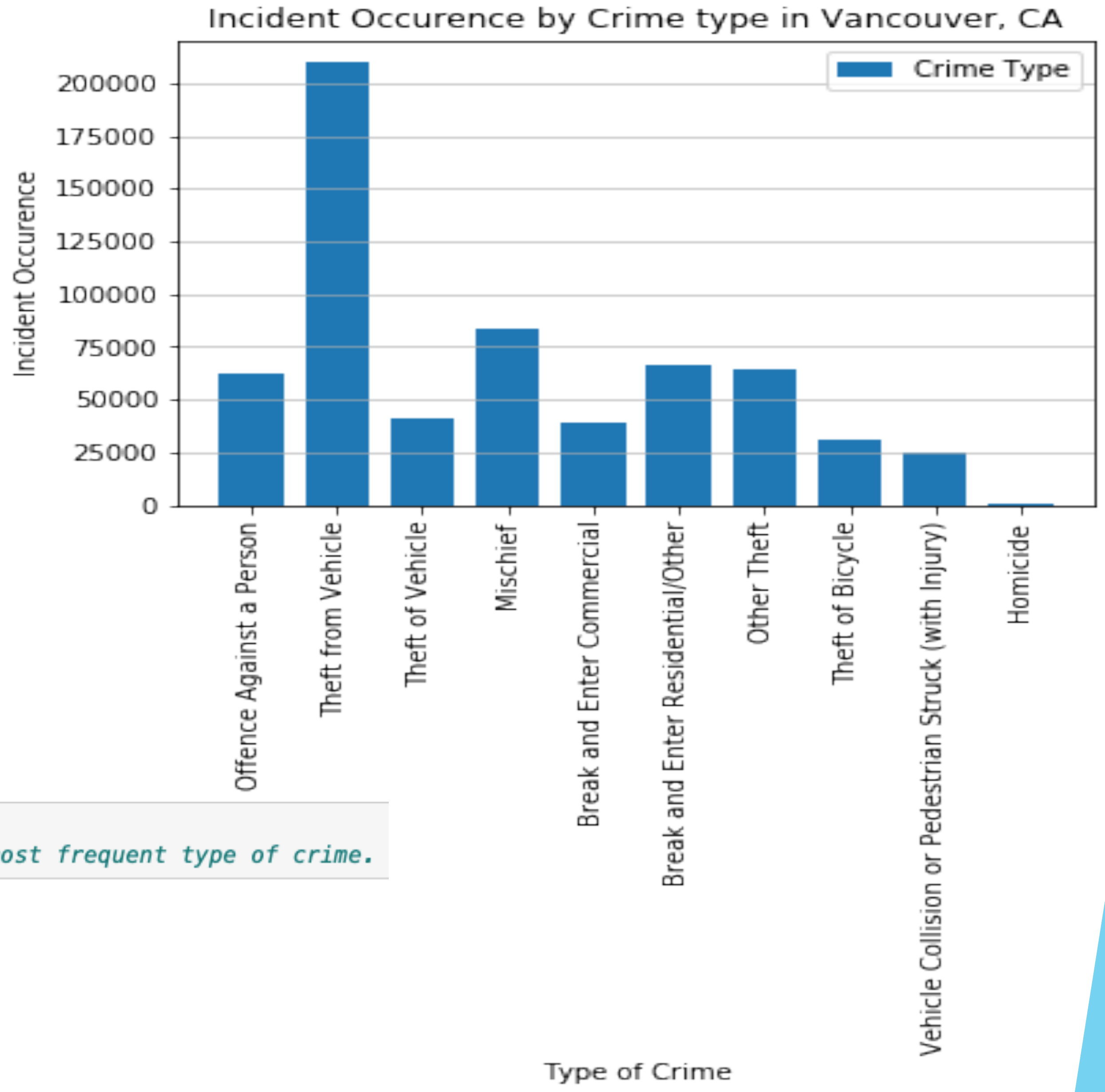
	TYPE	YEAR	MONTH	DAY	HOUR	MINUTE	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y
0	Break and Enter Commercial	2012	12	14	8	52	NaN	Oakridge	491285.000000	5.453433e+06
1	Break and Enter Commercial	2019	3	7	2	6	10XX SITKA SQ	Fairview	490612.964805	5.457110e+06
2	Break and Enter Commercial	2019	8	27	4	12	10XX ALBERNI ST	West End	491007.779775	5.459174e+06
3	Break and Enter Commercial	2014	8	8	5	13	10XX ALBERNI ST	West End	491015.943352	5.459166e+06
4	Break and Enter Commercial	2005	11	14	3	9	10XX ALBERNI ST	West End	491021.385727	5.459161e+06

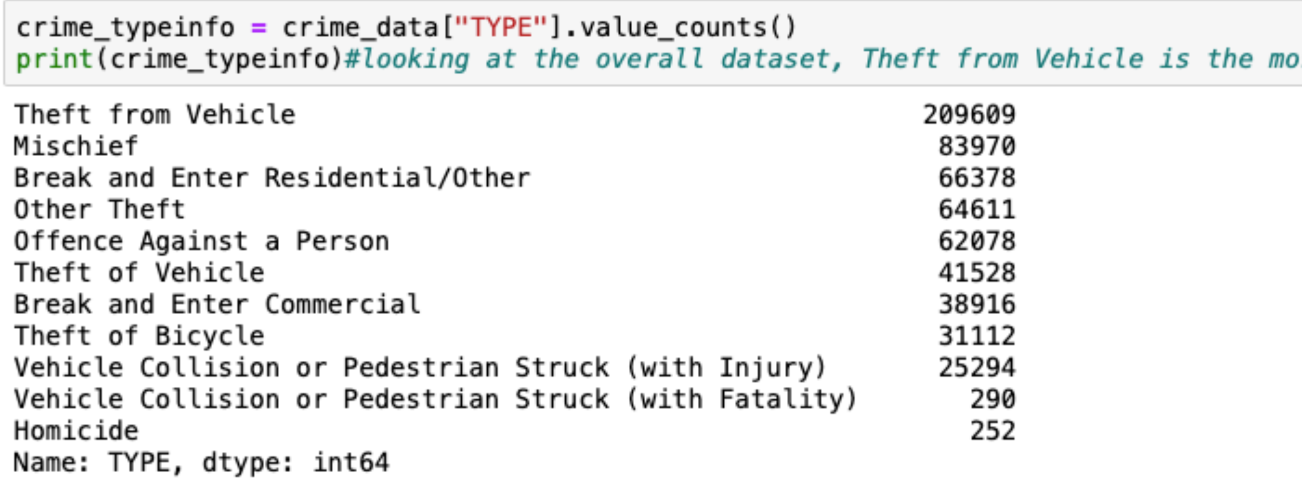
Variables used in analysis

- Type- Type column contains the types of crime committed
- Year- Numerical value of the year for the crime
- Month- Numerical value of the month for the crime
- Day- Numerical value of the date of the crime
- Hour- Numerical value of the Hour for the crime
- Neighborhood- Neighborhood name of the crime

Type Variable

For the overall dataset, it looks likes that Theft from Vehicle has the most incident occurrence. No outliers present from this graph. Descriptive statistics shown below.



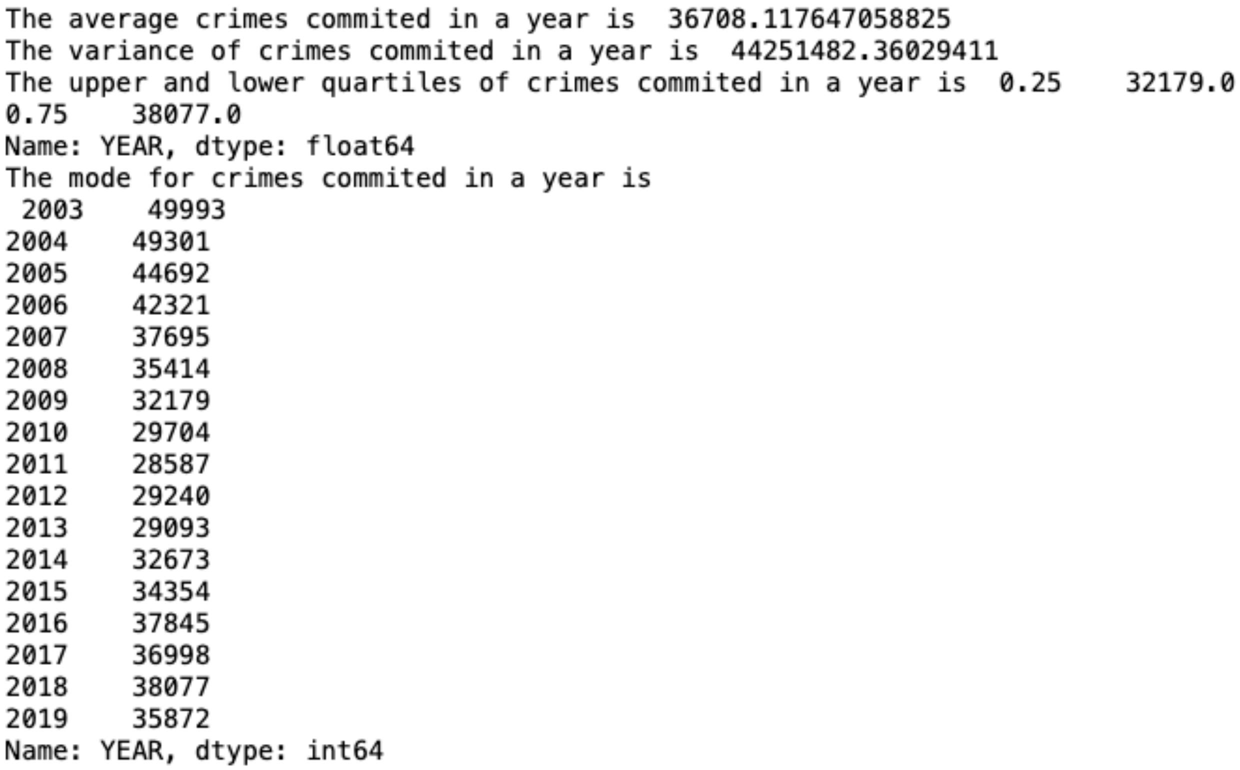


```
crime_typeinfo = crime_data["TYPE"].value_counts()
print(crime_typeinfo)#looking at the overall dataset, Theft from Vehicle is the most frequent type of crime.
```

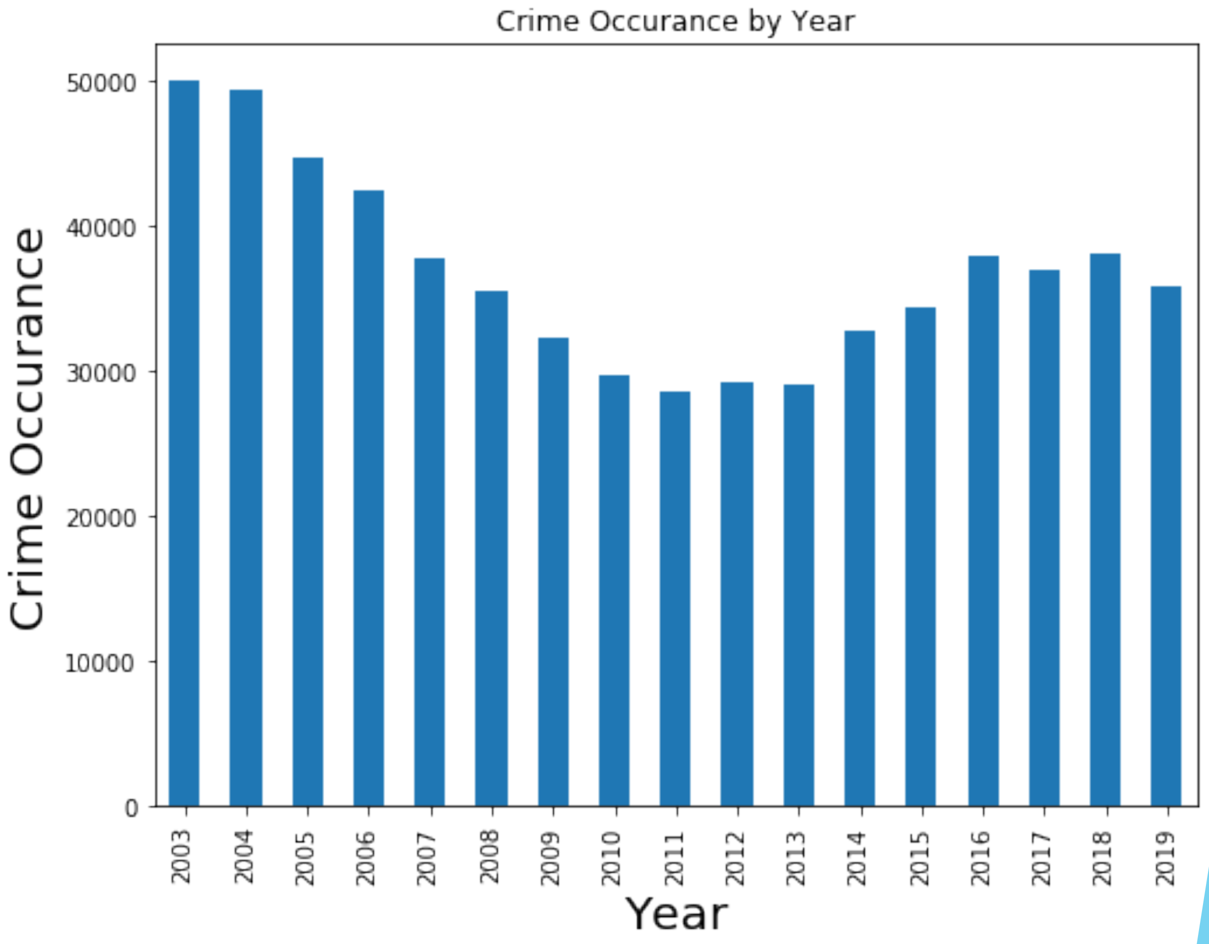
```
Theft from Vehicle                                      209609
Mischief                                                 83970
Break and Enter Residential/Other                        66378
Other Theft                                              64611
Offence Against a Person                                 62078
Theft of Vehicle                                         41528
Break and Enter Commercial                               38916
Theft of Bicycle                                         31112
Vehicle Collision or Pedestrian Struck (with Injury)     25294
Vehicle Collision or Pedestrian Struck (with Fatality)     290
Homicide                                                   252
Name: TYPE, dtype: int64
```

Year Variable

For the overall dataset, it looks likes that 2003 had the most incident occurrence. No outliers present from this graph. Descriptive statistics shown below.

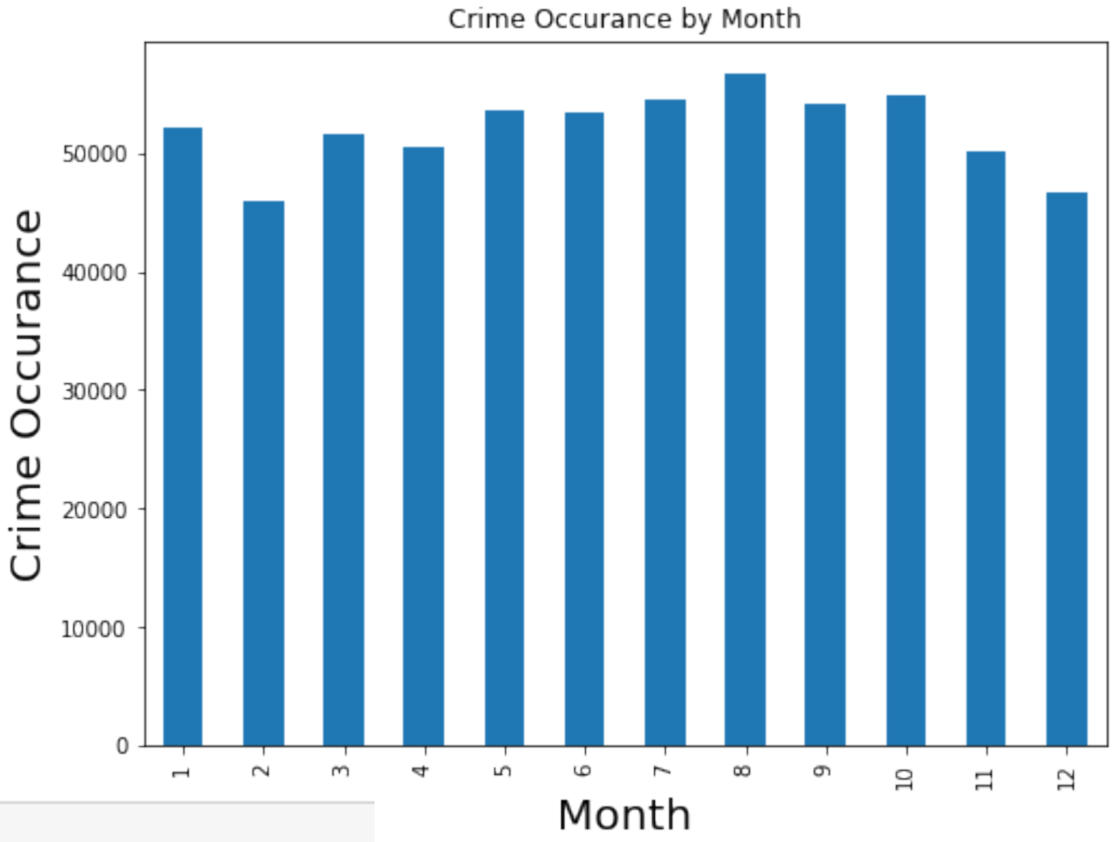


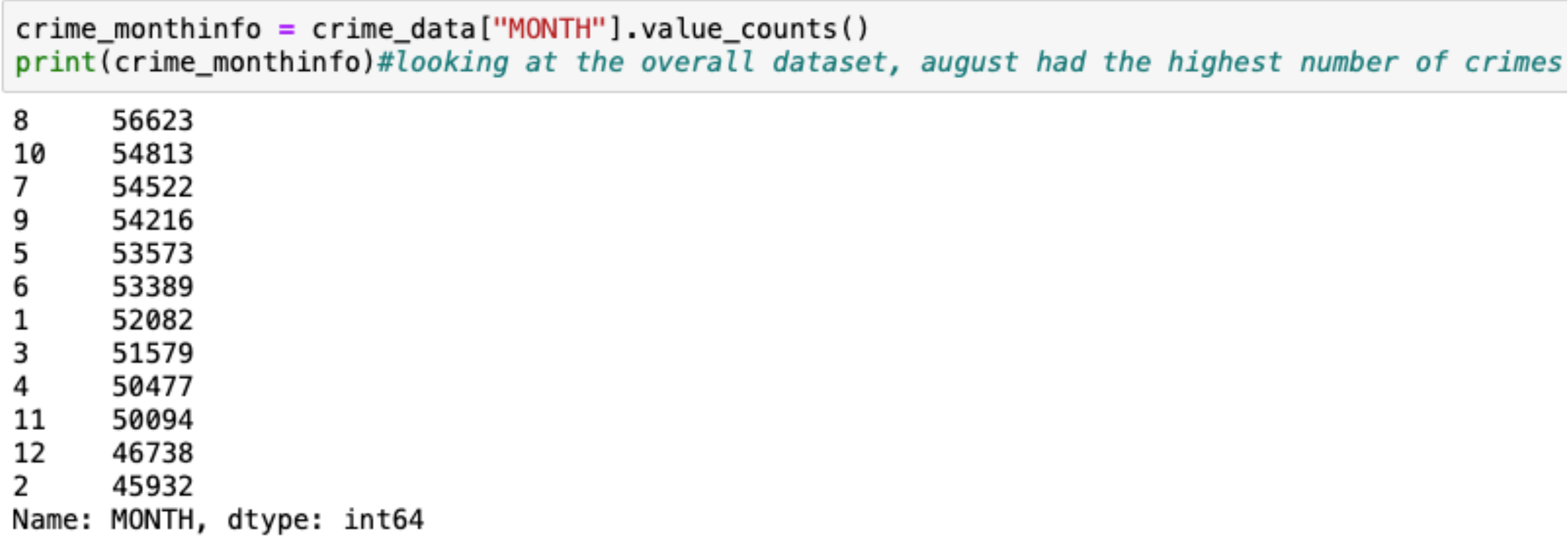
```
The average crimes commited in a year is  36708.117647058825
The variance of crimes commited in a year is  44251482.36029411
The upper and lower quartiles of crimes commited in a year is  0.25    32179.0
0.75    38077.0
Name: YEAR, dtype: float64
The mode for crimes commited in a year is
 2003    49993
2004    49301
2005    44692
2006    42321
2007    37695
2008    35414
2009    32179
2010    29704
2011    28587
2012    29240
2013    29093
2014    32673
2015    34354
2016    37845
2017    36998
2018    38077
2019    35872
Name: YEAR, dtype: int64
```



Month Variable

For the overall dataset, it looks likes that August had the most incident occurrence. No outliers present from this graph. Descriptive statistics shown below.



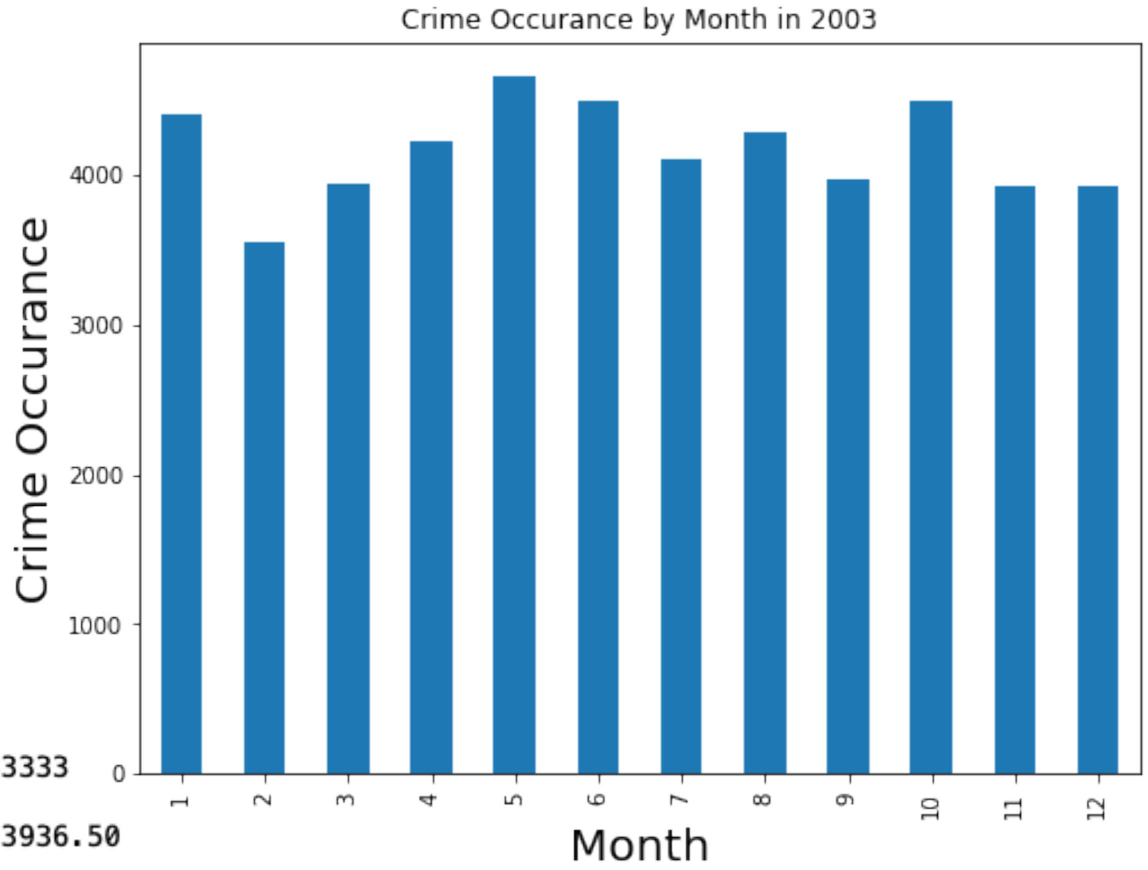


```
crime_monthinfo = crime_data["MONTH"].value_counts()
print(crime_monthinfo)#looking at the overall dataset, august had the highest number of crimes
```

```
8     56623
10    54813
7     54522
9     54216
5     53573
6     53389
1     52082
3     51579
4     50477
11    50094
12    46738
2     45932
Name: MONTH, dtype: int64
```

Month in 2003

Since 2003 had the most crimes throughout the years, I was curious to see which month from 2003 had the highest crime incidents. It looks likes that May had the most incident occurrence. No outliers present from this graph. Descriptive statistics shown below.

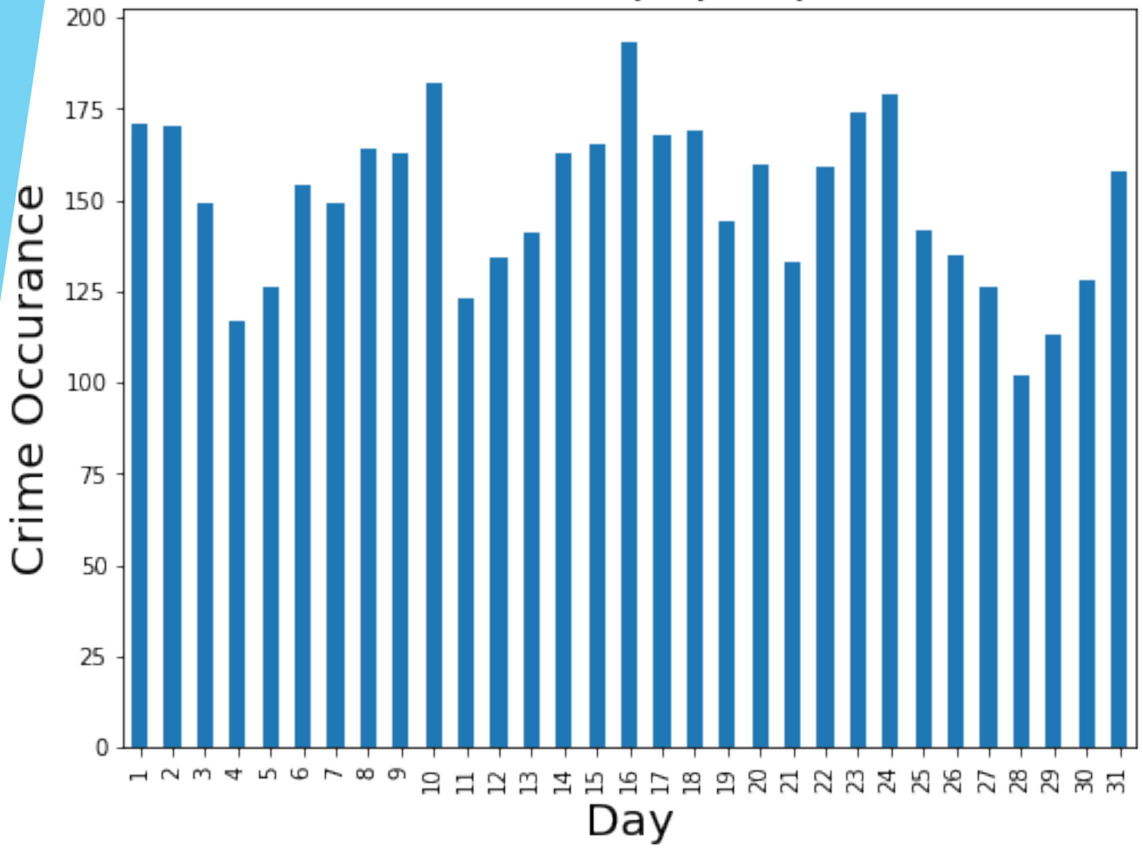


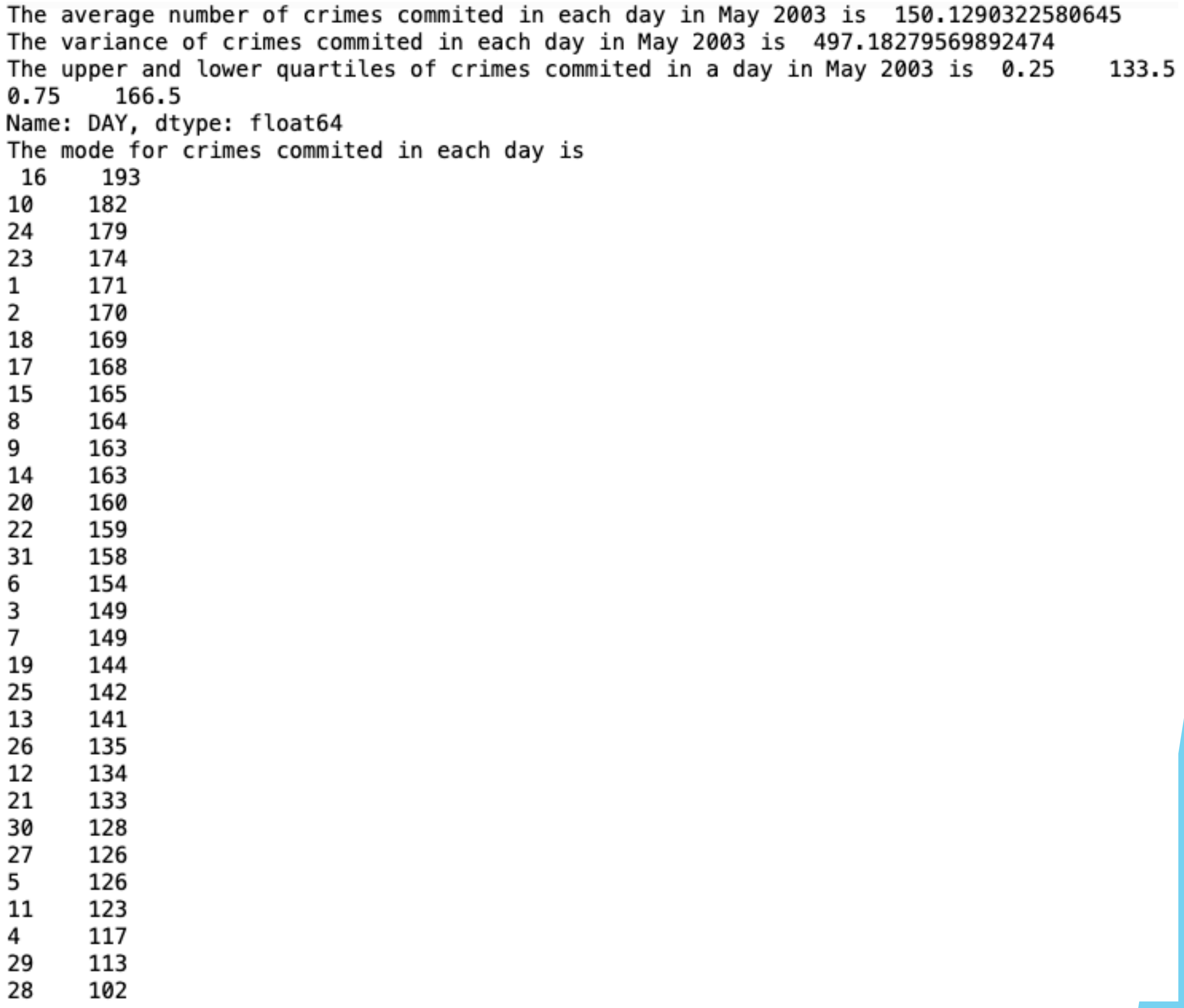
```
The average number of crimes commited in each month in 2003 is  4166.083333333333
The variance of crimes commited in each month in 2003 is  100697.71969696971
The upper and lower quartiles of crimes commited in each in 2003 is 0.25    3936.50
0.75    4429.75
Name: MONTH, dtype: float64
The mode for crimes commited in each month in 2003 is
 1      4410
2      3556
3      3939
4      4220
5      4654
6      4489
7      4107
8      4291
9      3977
10     4497
11     3924
12     3929
Name: MONTH, dtype: int64
```

Day in May of 2003

Since May of 2003 had the most crimes throughout the years, I was curious to see which day from May of 2003 had the highest crime incidents. It looks likes that the 16th had the most incident occurrence. No outliers present from this graph. Descriptive statistics shown.

Crime Occurance by Day in May 2003

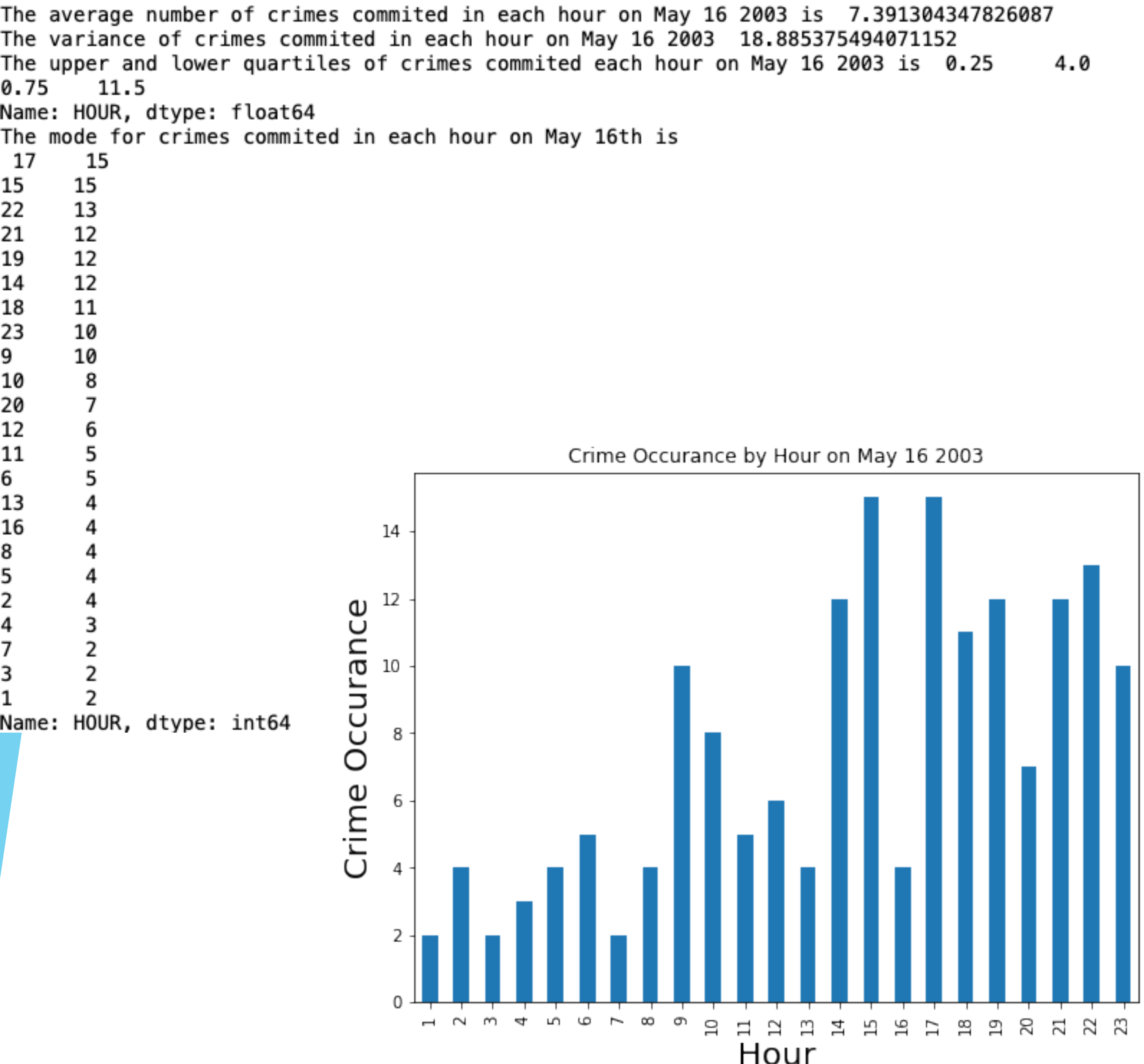




```
The average number of crimes commited in each day in May 2003 is  150.1290322580645
The variance of crimes commited in each day in May 2003 is  497.18279569892474
The upper and lower quartiles of crimes commited in a day in May 2003 is  0.25    133.5
0.75    166.5
Name: DAY, dtype: float64
The mode for crimes commited in each day is
 16    193
10    182
24    179
23    174
1     171
2     170
18    169
17    168
15    165
8     164
9     163
14    163
20    160
22    159
31    158
6     154
3     149
7     149
19    144
25    142
13    141
26    135
12    134
21    133
30    128
27    126
5     126
11    123
4     117
29    113
28    102
```

Hour of May 16, 2003

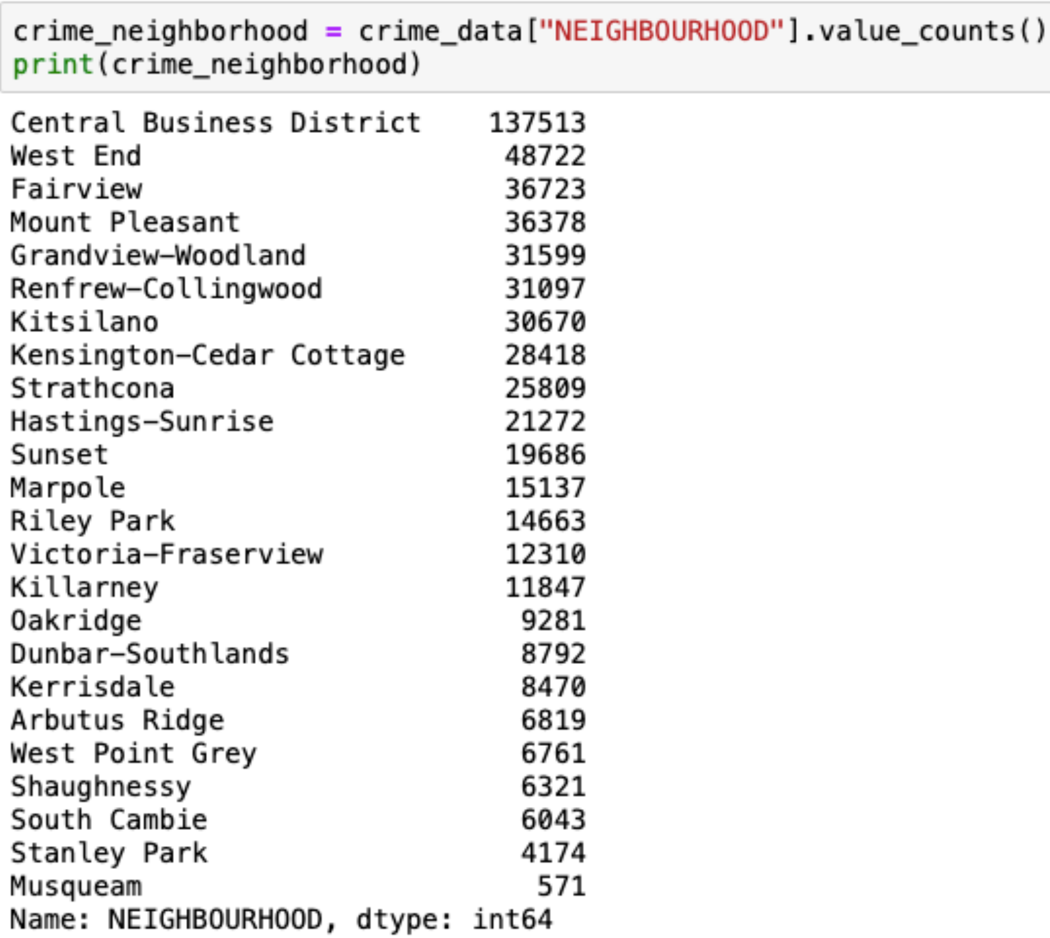
Since May 16th of 2003 had the most crimes throughout the years, I was curious to see which Hour from May 16th of 2003 had the highest crime incidents. It looks likes that the 17th and 15th hour had the most incident occurrence. We found that 0 was an outlier that was present in the initial graph for the hour, but we removed to get a better understanding. Descriptive statistics shown below.



```
The average number of crimes commited in each hour on May 16 2003 is  7.391304347826087
The variance of crimes commited in each hour on May 16 2003  18.885375494071152
The upper and lower quartiles of crimes commited each hour on May 16 2003 is  0.25     4.0
0.75    11.5
Name: HOUR, dtype: float64
The mode for crimes commited in each hour on May 16th is
 17    15
15    15
22    13
21    12
19    12
14    12
18    11
23    10
9     10
10     8
20     7
12     6
11     5
6      5
13     4
16     4
8      4
5      4
2      4
4      3
7      2
3      2
1      2
Name: HOUR, dtype: int64
```

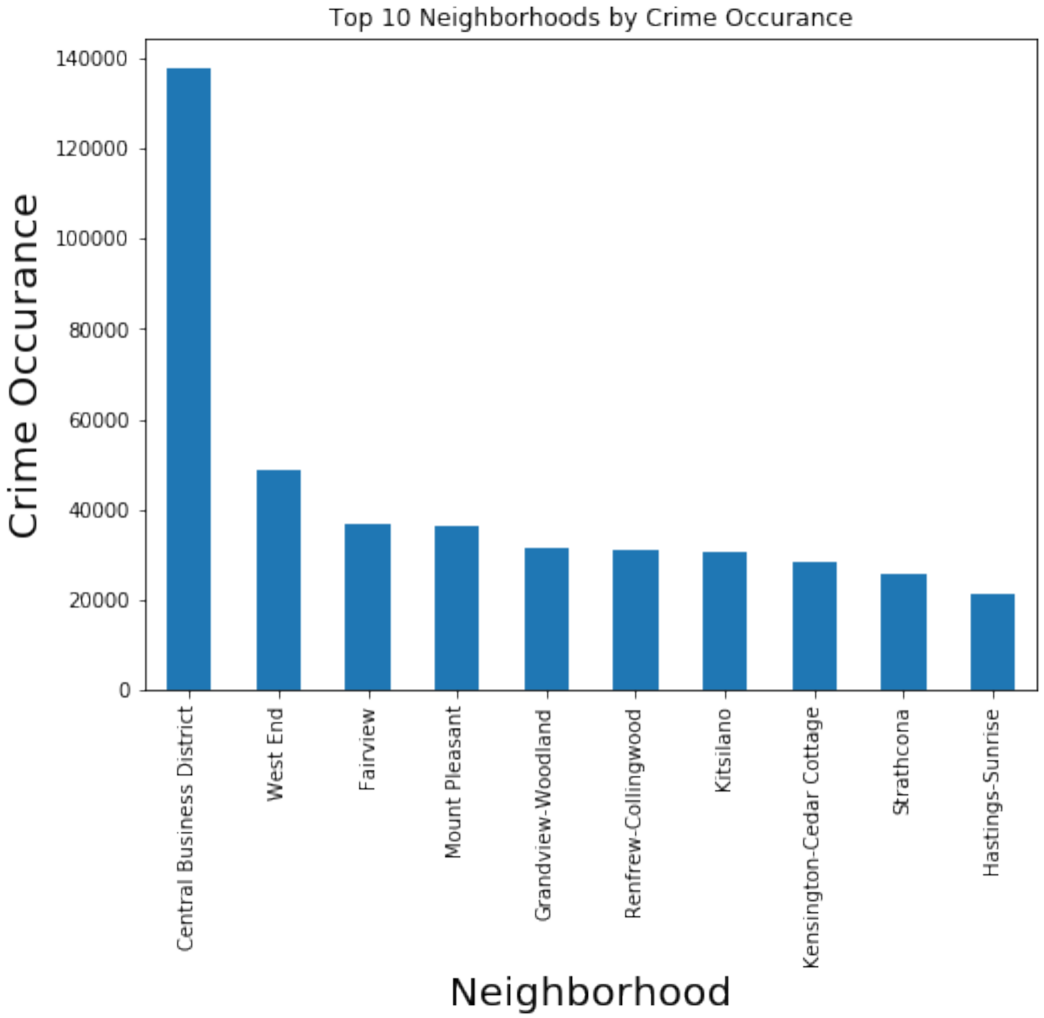
Neighborhood Variable

For the overall dataset, it looks likes that the neighborhood of Central Business District had the most incident occurrence. No outliers present from this graph. Descriptive statistics shown below.



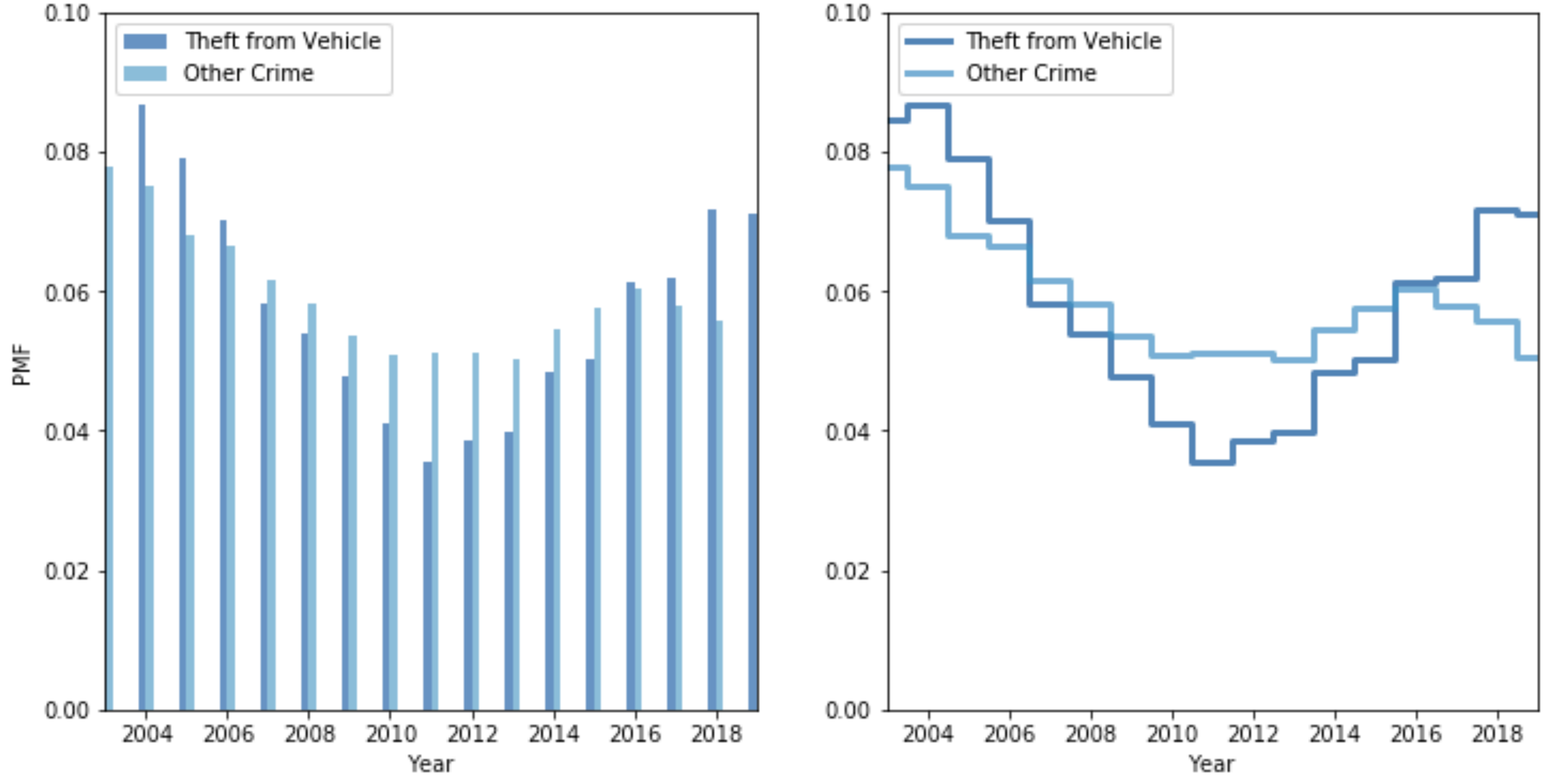
```
crime_neighborhood = crime_data["NEIGHBOURHOOD"].value_counts()
print(crime_neighborhood)
```

```
Central Business District    137513
West End                      48722
Fairview                      36723
Mount Pleasant                36378
Grandview-Woodland            31599
Renfrew-Collingwood           31097
Kitsilano                     30670
Kensington-Cedar Cottage      28418
Strathcona                    25809
Hastings-Sunrise              21272
Sunset                        19686
Marpole                       15137
Riley Park                    14663
Victoria-Fraserview           12310
Killarney                     11847
Oakridge                       9281
Dunbar-Southlands              8792
Kerrisdale                     8470
Arbutus Ridge                  6819
West Point Grey                6761
Shaughnessy                    6321
South Cambie                   6043
Stanley Park                   4174
Musqueam                        571
Name: NEIGHBOURHOOD, dtype: int64
```



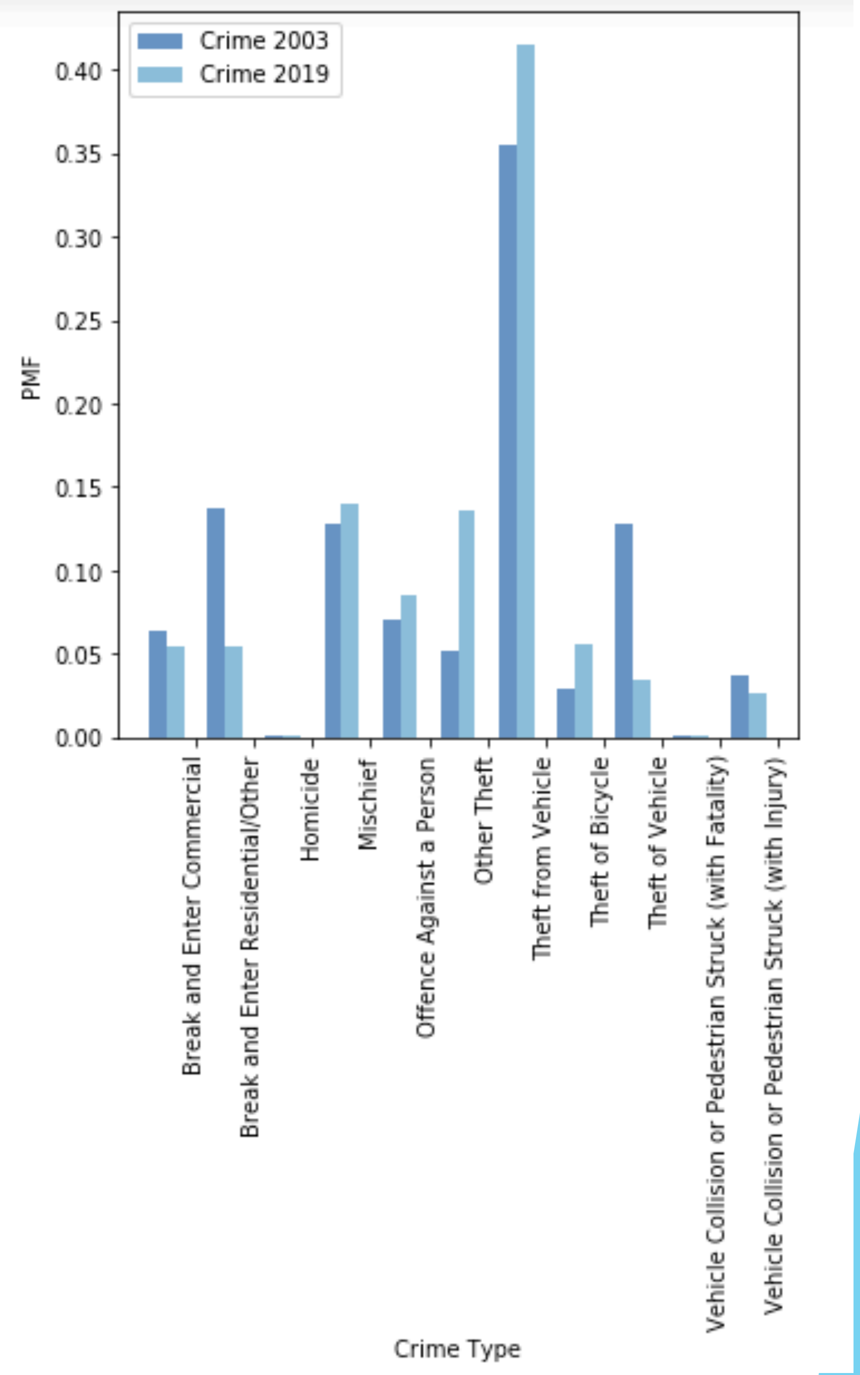
PMF-First Graph

Comparing Theft from Vehicles to Other Types of Crimes



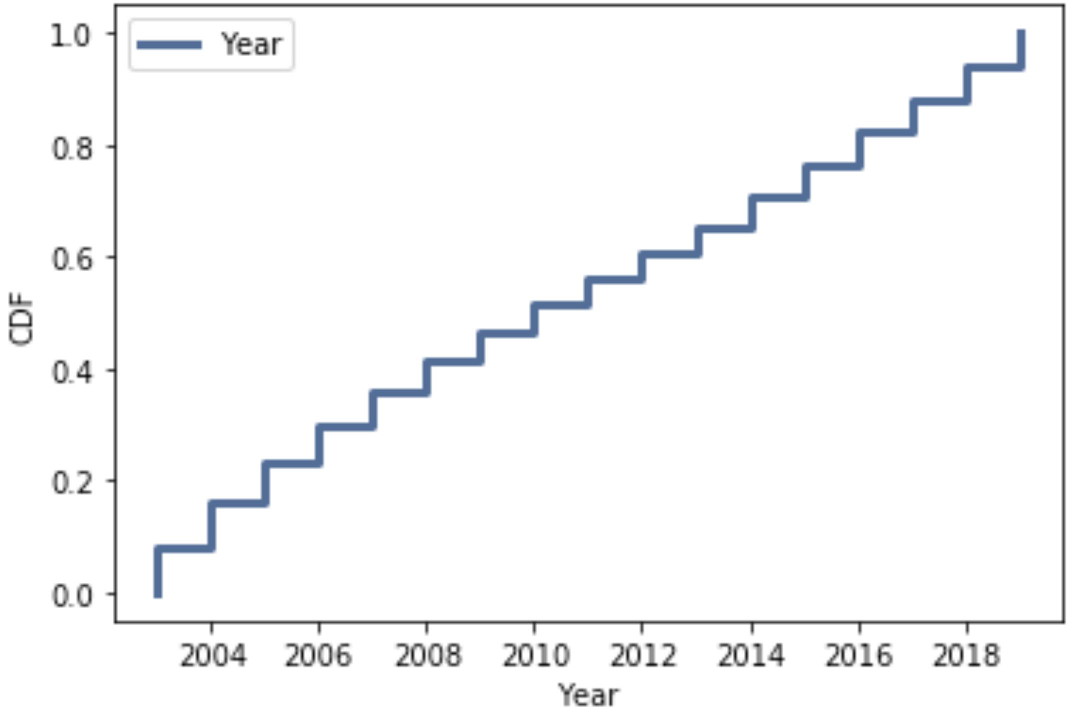
PMF- Second Graph

Comparing 2003 to 2019 by crime types to see if a crime became more prevalent



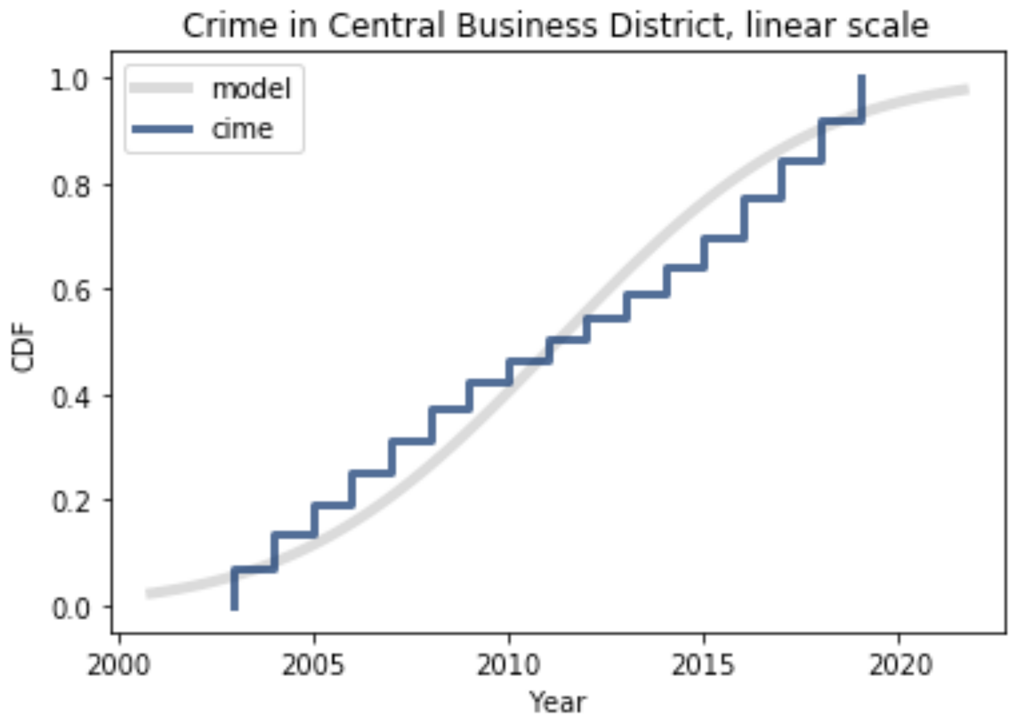
CDF

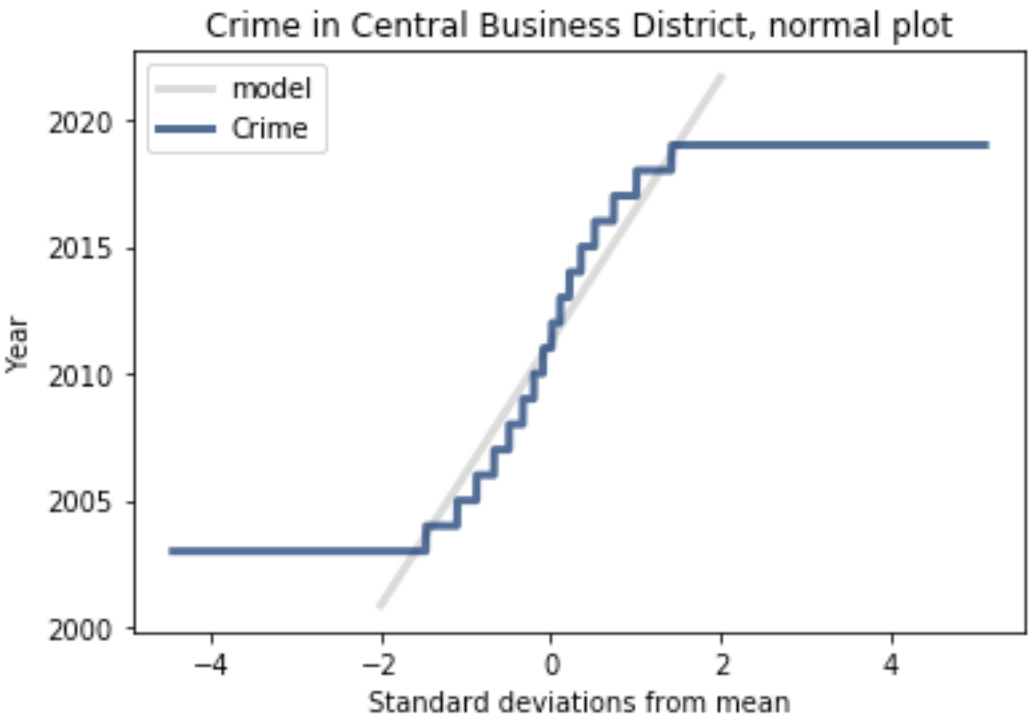
This CDF helps me answer my one of my question which is, Is crime increasing throughout the years in Vancouver, CA? Looking at the step size at the beginning the steps were larger so crimes were for frequent, but then in the middle show a slow down of crime because the steps got smaller, and then towards the end the steps are starting to pick up again.



Analytical Distribution

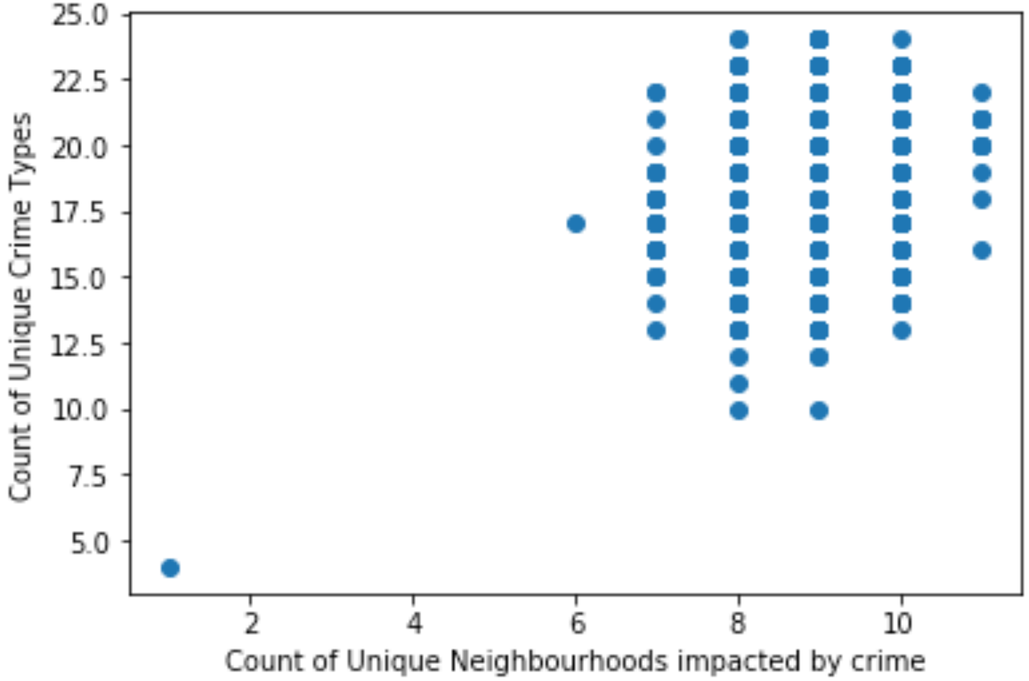
The histogram shown earlier had the neighborhood of Central Business District with the most crime. Therefore, I decided to plot a CDF to a normal model. Also generated a normal probability plot of Crime in Central Business District. The normal plot model seems to follow the data, whereas the linear scale model doesn’t quite fit.

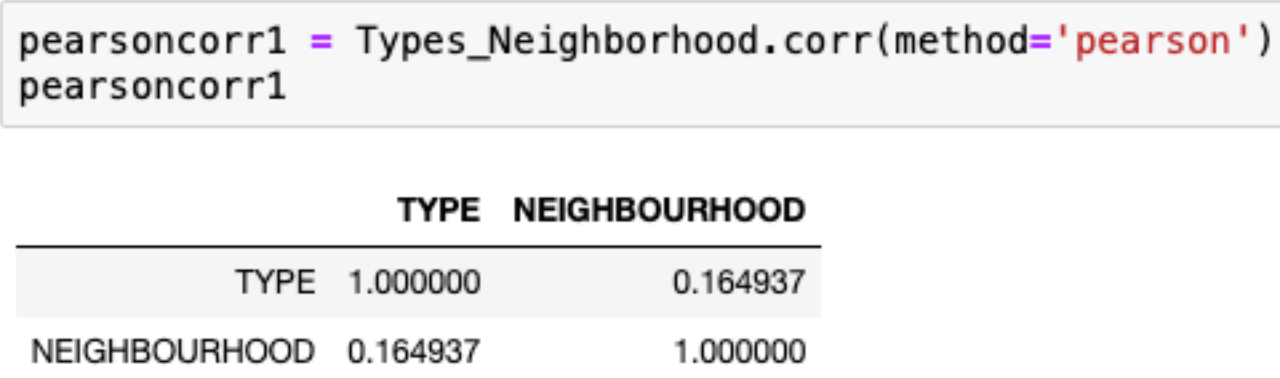

Crime in Central Business District, linear scale


Crime in Central Business District, normal plot

Relationships between Variables

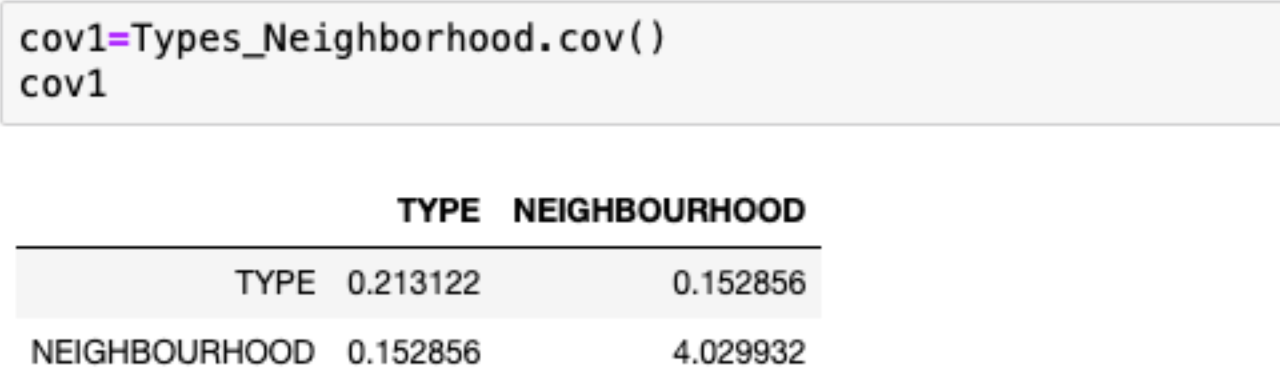
I needed to generate my own quantitative data to analyze. The following scatter plot looks at if there is a correlation between number of types of crimes committed and unique neighborhoods impacted. The correlation of 0.16 shows to be a weak positive correlation. The covariance shows to be 0.15 which means that there is a weak relationship





```
pearsoncorr1 = Types_Neighborhood.corr(method='pearson')
pearsoncorr1
```

	TYPE	NEIGHBOURHOOD
TYPE	1.000000	0.164937
NEIGHBOURHOOD	0.164937	1.000000

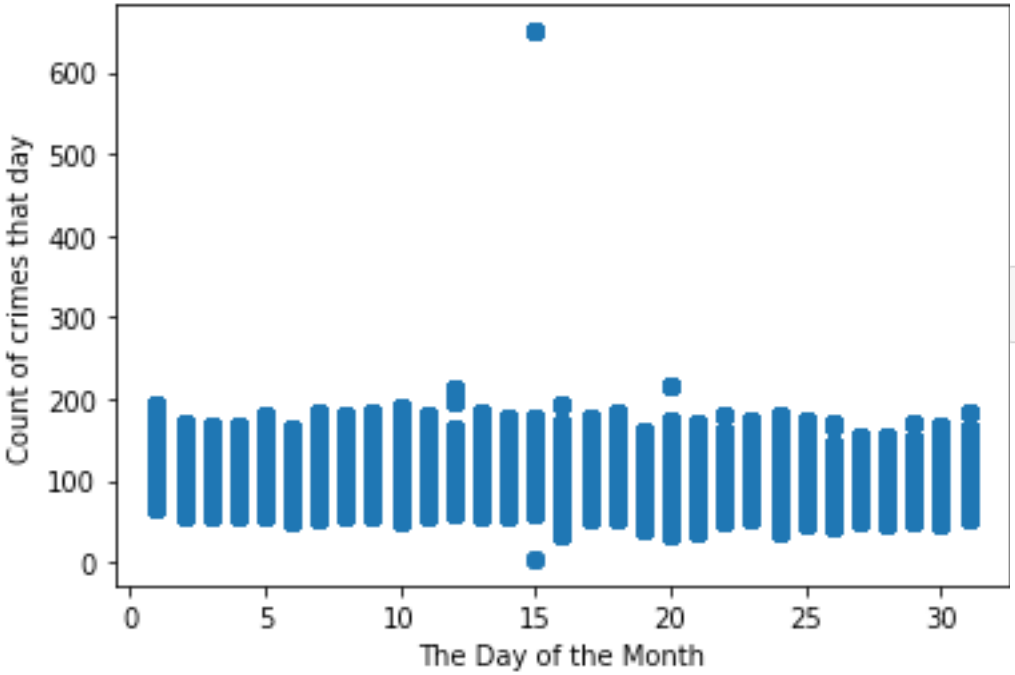


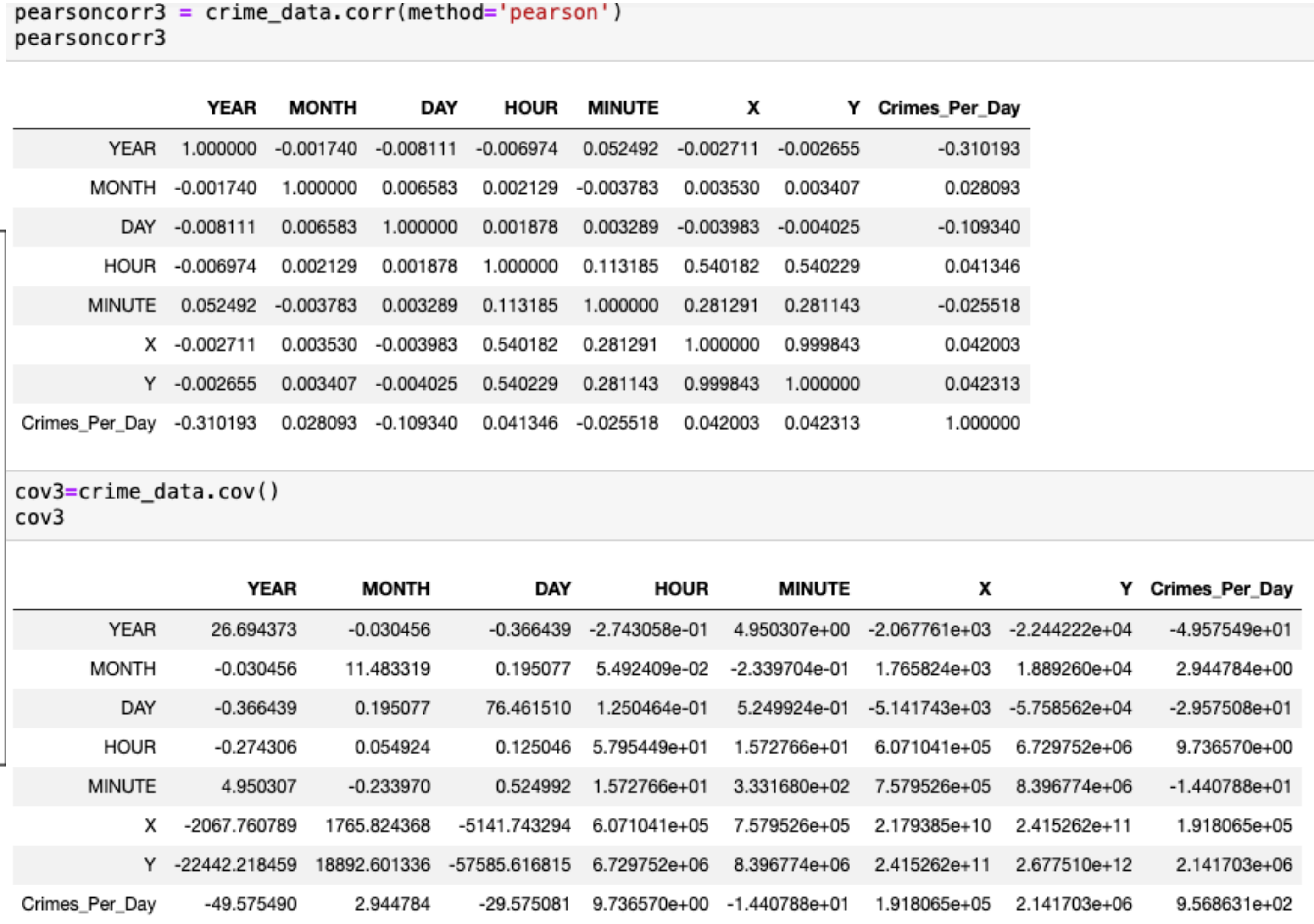
```
cov1=Types_Neighborhood.cov()
cov1
```

	TYPE	NEIGHBOURHOOD
TYPE	0.213122	0.152856
NEIGHBOURHOOD	0.152856	4.029932

Relationships between Variables

I needed to generate my own quantitative data to analyze. The following scatter plot looks at if there is a correlation between the number of crimes and the day of the month. The correlation of -0.109 shows to be a weak negative correlation. The covariance shows to be -29.5 which means that there is a strong inverse relationship among number of crimes count and the day of the month.





```
pearsoncorr3 = crime_data.corr(method='pearson')
pearsoncorr3
```

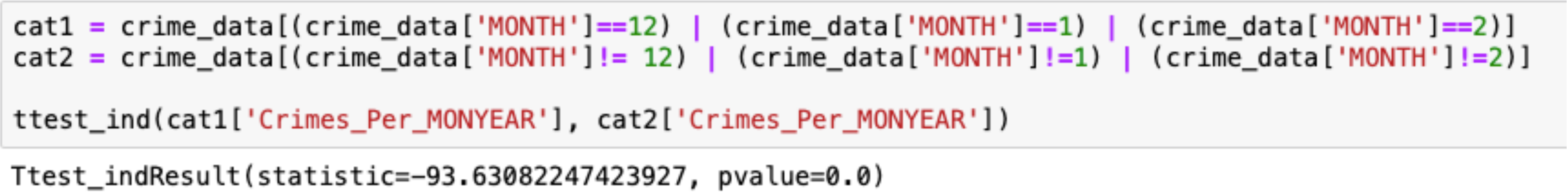
	YEAR	MONTH	DAY	HOUR	MINUTE	X	Y	Crimes_Per_Day
YEAR	1.000000	-0.001740	-0.008111	-0.006974	0.052492	-0.002711	-0.002655	-0.310193
MONTH	-0.001740	1.000000	0.006583	0.002129	-0.003783	0.003530	0.003407	0.028093
DAY	-0.008111	0.006583	1.000000	0.001878	0.003289	-0.003983	-0.004025	-0.109340
HOUR	-0.006974	0.002129	0.001878	1.000000	0.113185	0.540182	0.540229	0.041346
MINUTE	0.052492	-0.003783	0.003289	0.113185	1.000000	0.281291	0.281143	-0.025518
X	-0.002711	0.003530	-0.003983	0.540182	0.281291	1.000000	0.999843	0.042003
Y	-0.002655	0.003407	-0.004025	0.540229	0.281143	0.999843	1.000000	0.042313
Crimes_Per_Day	-0.310193	0.028093	-0.109340	0.041346	-0.025518	0.042003	0.042313	1.000000

```
cov3=crime_data.cov()
cov3
```

	YEAR	MONTH	DAY	HOUR	MINUTE	X	Y	Crimes_Per_Day
YEAR	26.694373	-0.030456	-0.366439	-2.743058e-01	4.950307e+00	-2.067761e+03	-2.244222e+04	-4.957549e+01
MONTH	-0.030456	11.483319	0.195077	5.492409e-02	-2.339704e-01	1.765824e+03	1.889260e+04	2.944784e+00
DAY	-0.366439	0.195077	76.461510	1.250464e-01	5.249924e-01	-5.141743e+03	-5.758562e+04	-2.957508e+01
HOUR	-0.274306	0.054924	0.125046	5.795449e+01	1.572766e+01	6.071041e+05	6.729752e+06	9.736570e+00
MINUTE	4.950307	-0.233970	0.524992	1.572766e+01	3.331680e+02	7.579526e+05	8.396774e+06	-1.440788e+01
X	-2067.760789	1765.824368	-5141.743294	6.071041e+05	7.579526e+05	2.179385e+10	2.415262e+11	1.918065e+05
Y	-22442.218459	18892.601336	-57585.616815	6.729752e+06	8.396774e+06	2.415262e+11	2.677510e+12	2.141703e+06
Crimes_Per_Day	-49.575490	2.944784	-29.575081	9.736570e+00	-1.440788e+01	1.918065e+05	2.141703e+06	9.568631e+02

Hypothesis Test

For my hypothesis test I wanted to figure out if colder months have less crime prevalence. To do this I created two new column, one of the column combined month and year to provide an unique index for each month, and the second column was used to calculate the count of how many crimes occurred that month. Next I split out the data into two categories, one for colder months which included December, January and February, and the second category had the remainder of the months in the year. I used two methods to test and get a p-value, and both showed the P-value to be 0. Therefore, we can reject the null hypothesis of that there is no link between colder months and crime prevalence in Vancouver. It appears that colder months have less crime prevalence.



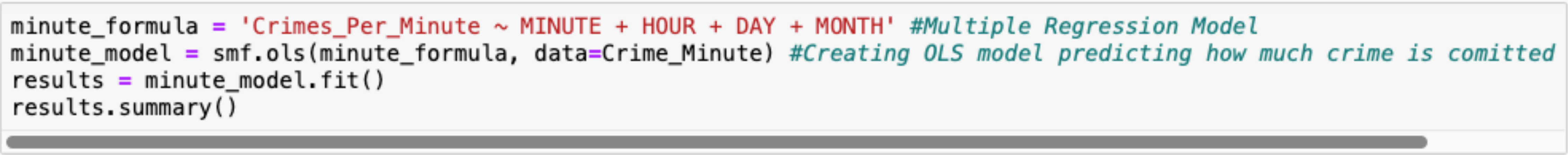
```
cat1 = crime_data[(crime_data['MONTH']==12) | (crime_data['MONTH']==1) | (crime_data['MONTH']==2)]
cat2 = crime_data[(crime_data['MONTH']!= 12) | (crime_data['MONTH']!=1) | (crime_data['MONTH']!=2)]

ttest_ind(cat1['Crimes_Per_MONYEAR'], cat2['Crimes_Per_MONYEAR'])
```

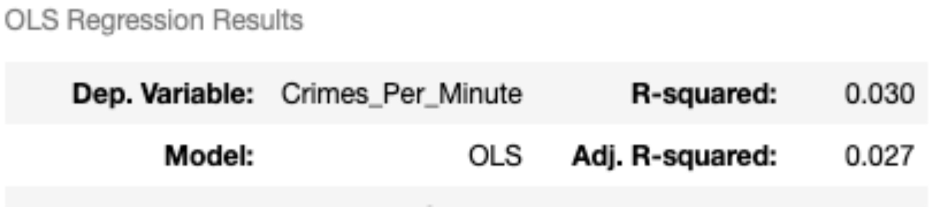
```
Ttest_indResult(statistic=-93.63082247423927, pvalue=0.0)
```

Regression Analysis

For my Regression Analysis test, I decided to do a multiple regression test to answer the question of can we predict when a crime is going to occur. I did this by creating two new columns, the first column that I created was a time string consisting of year, month, day, hour, and minute and another that counted how many crimes occurred each minute. Since I had a large dataset, I had filtered down to the most prevalent crime type Theft from Vehicle. I also filtered down to the Central Business district as this neighborhood had the most crime and the year 2019. I plotted the scatterplot to check the data and saw a few outliers, so I then filtered down to a reasonable limit of 50 crimes per minute. Once I had my data set up I ran an OLS multiple regression analysis using Number of Crimes per minute as the dependent variable and month, day, hour, and minute as the explanatory variables. The regression analysis gave me an R^2 value of 0.03 which meant that the model was not a very good fit to the data. Based on this outcome I would say it would be difficult to predict when a crime was going to occur.



```
minute_formula = 'Crimes_Per_Minute ~ MINUTE + HOUR + DAY + MONTH' #Multiple Regression Model
minute_model = smf.ols(minute_formula, data=Crime_Minute) #Creating OLS model predicting how much crime is comitted
results = minute_model.fit()
results.summary()
```



OLS Regression Results

Dep. Variable:	Crimes_Per_Minute	**R-squared:**	0.030
Model:	OLS	**Adj. R-squared:**	0.027

Conclusions from Project

I have investigated these following questions, and this is what I can conclude:

- Is there a particular crime in Vancouver, CA that has become more prevalent?
  - According to the histograms, Theft from Vehicle shows to be most prevalent
- Is crime increasing throughout the years in Vancouver, CA?
  - According to the CDF, and histogram overall crime has gone down between 2003 and 2019, but crime has started to pick up again slowly starting 2014 and now its leveling off.
- Do colder months have less crime prevalence?
  - The hypothesis test showed that we can reject the null hypothesis, and therefore conclude that colder months have less crime prevalence in Vancouver, CA.
- Which neighborhoods have the most crime?
  - According to the histogram, Central Business District had the most crime incident occurrence in Vancouver, CA.
- Can we predict when crimes are most likely to occur?
  - The regression analysis gave me an R^2 value of 0.03 which meant that the model was not a very good fit to the data. Based on this outcome I would say it would be difficult to predict when a crime was going to occur.

References:

- Downey, A. B. (2015). Think Stats. Sebastopol, CA: O'Reilly Media, Inc.
- Lu, k. (2019, November 18). Vancouver Crime Report, Version 2. Retrieved from https://www.kaggle.com/agilesifaka/vancouver-crime-report
- Lynch, D. (2019, June 3). Weather in Vancouver, B.C.: Climate, Seasons, and Average Monthly Temperature. Retrieved from https://www.tripsavvy.com/vancouver-average-monthly-temperatures-3371376
- Wikipedia. (2020, January 14). Vancouver. Retrieved from https://en.wikipedia.org/wiki/Vancouver