Introduction

I chose the Crime Data dataset that represents crime reported to the Seattle Police Department. I got the data from data.gov (link: https://catalog.data.gov/dataset/crime-data-76bd0). The reason why I chose this data is that my family was considering relocating to Seattle due to the high housing price in the Bay area, and therefore I would like to better understand Seattle, especially the security/safety aspect of the city.

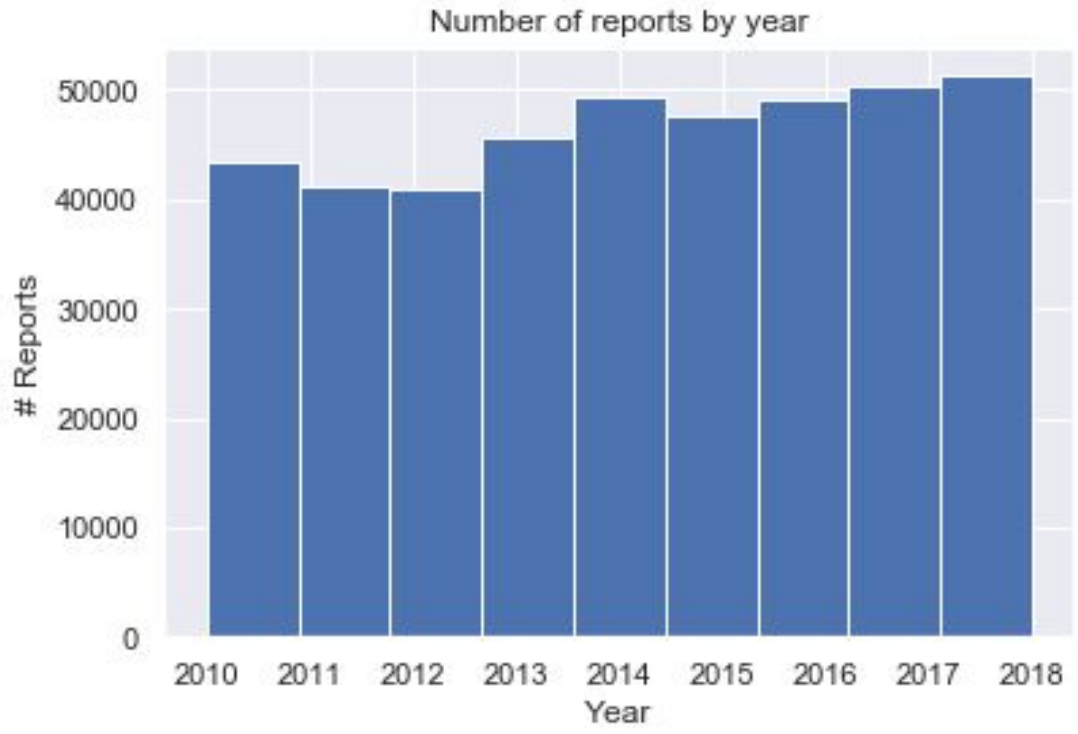
I was hoping to answer the following questions by exploring the data:

- At high level, what is the trend of quantity of crime reports over years? What neighborhoods/precincts are safer?
- At low level, what are the common categories of crime? How is the distribution changed over years?
- What are the most dangerous times in the day to get out?
- Is there any correlation between common crime categories?

Summary of data

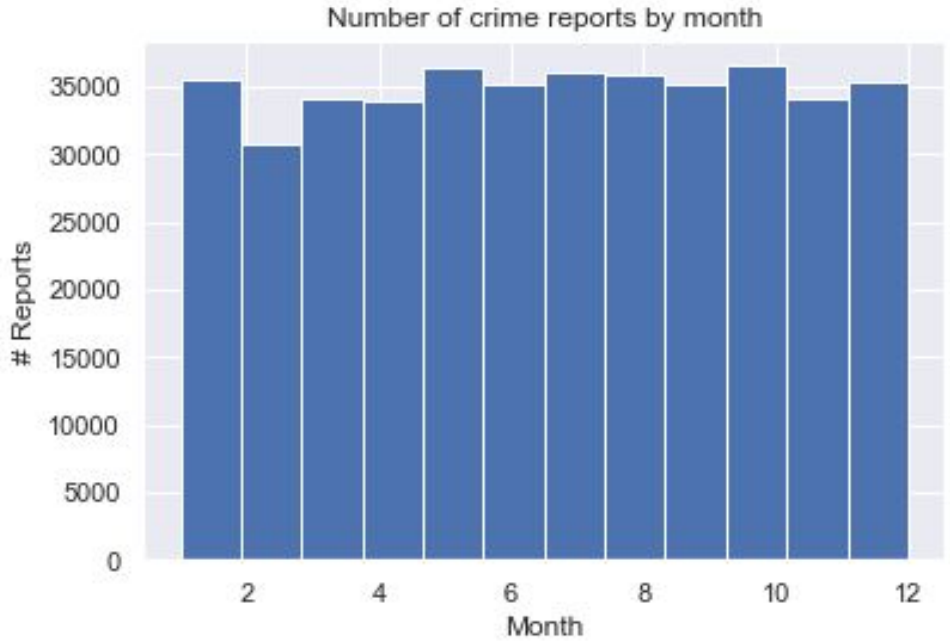
Quantity of crime reports over recent years

The histogram below shows the number of crime reports in each year between 2010 - 2018.



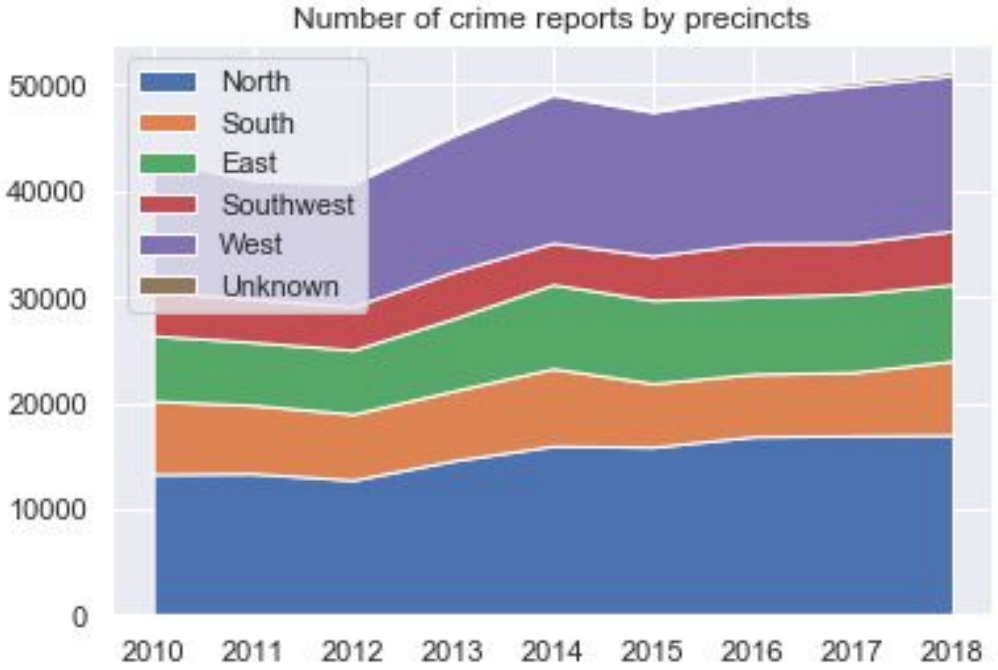
Quantity of crime reports in each month

The histogram below shows the number of crime reports in each month.

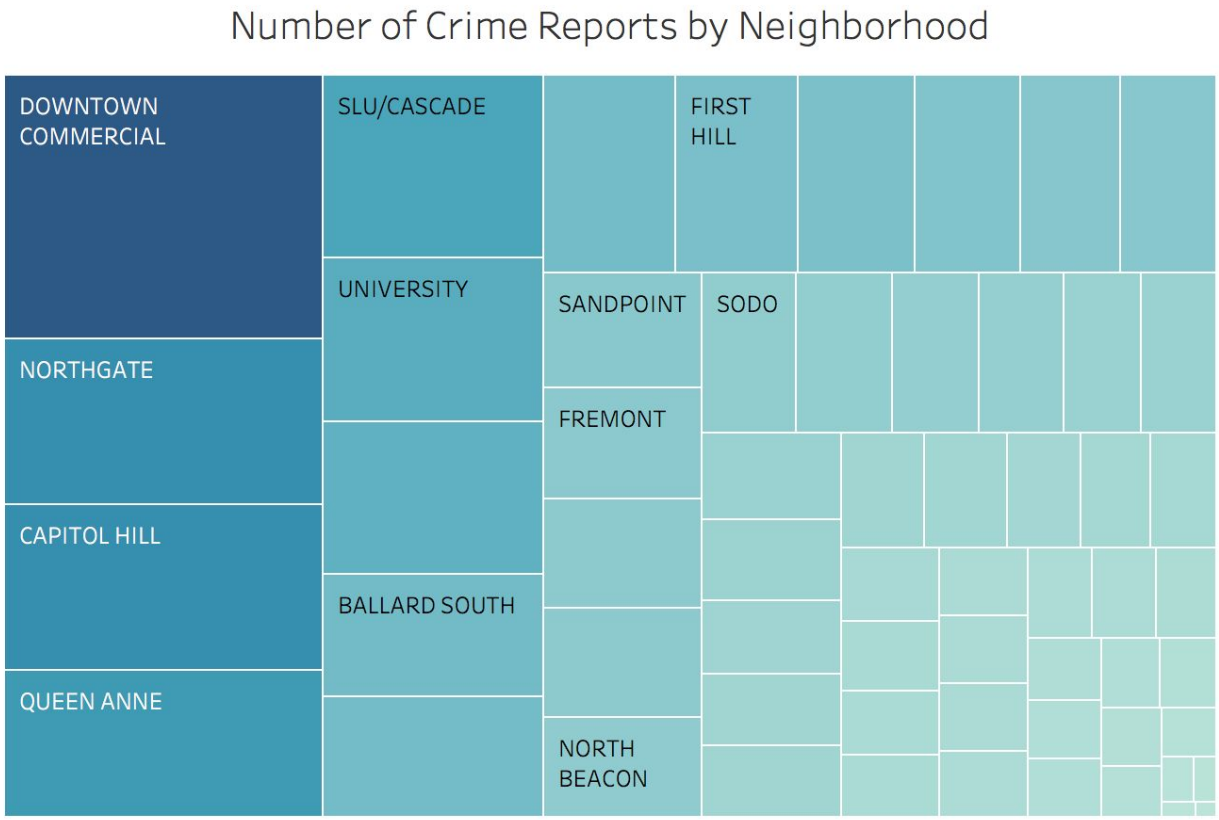


Quantity of crime reports by precincts over years

The stacked area plot below shows the number of crime reports in each precinct and how it changes over years.

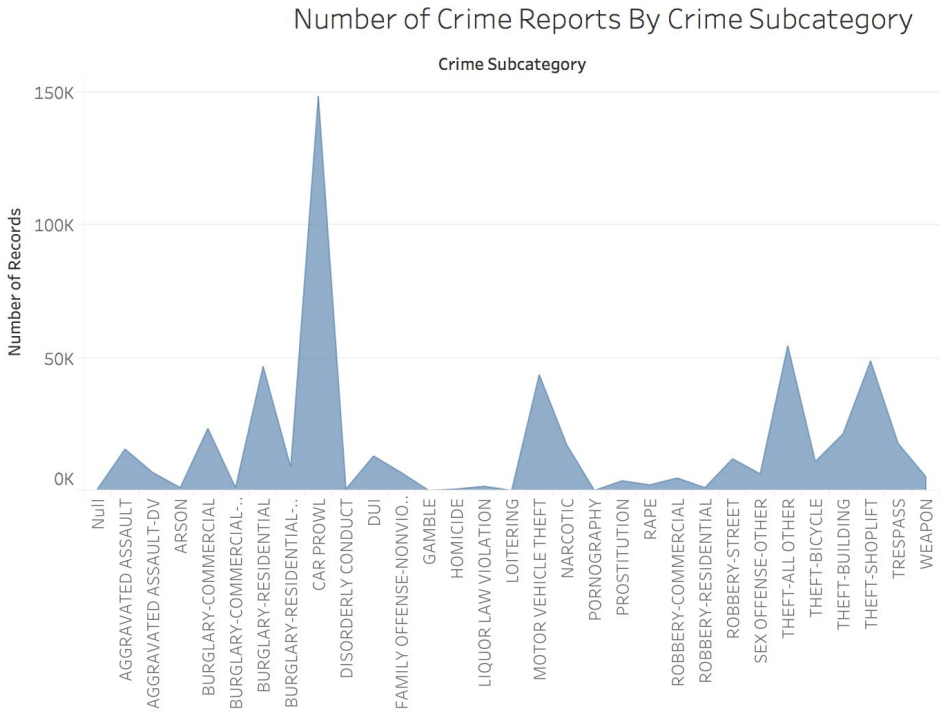


The tree map below visualizes the number of crime reports in each neighborhood in 2010-2018.

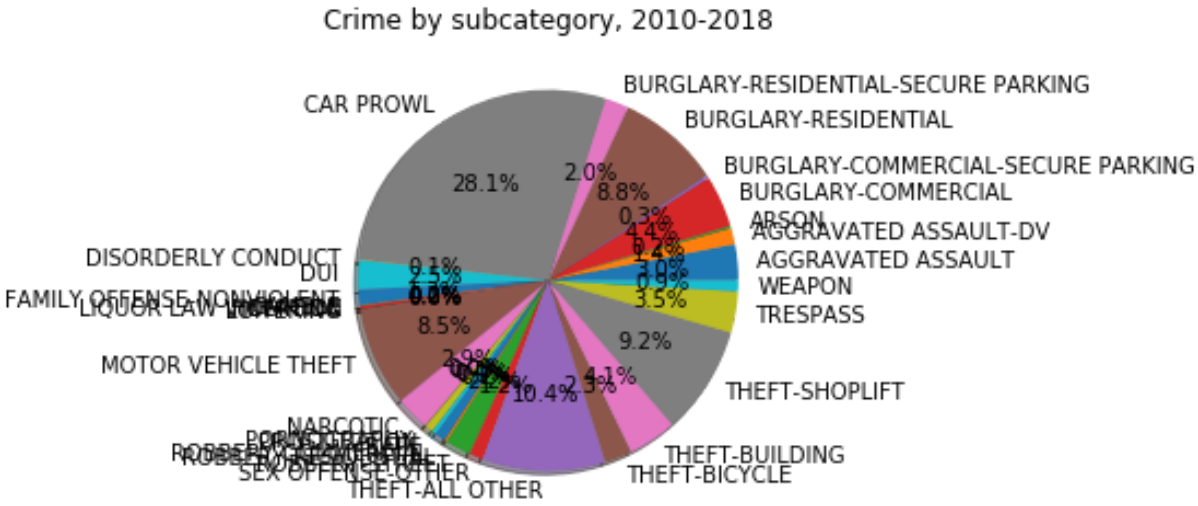


Quantity of crime reports by subcategory

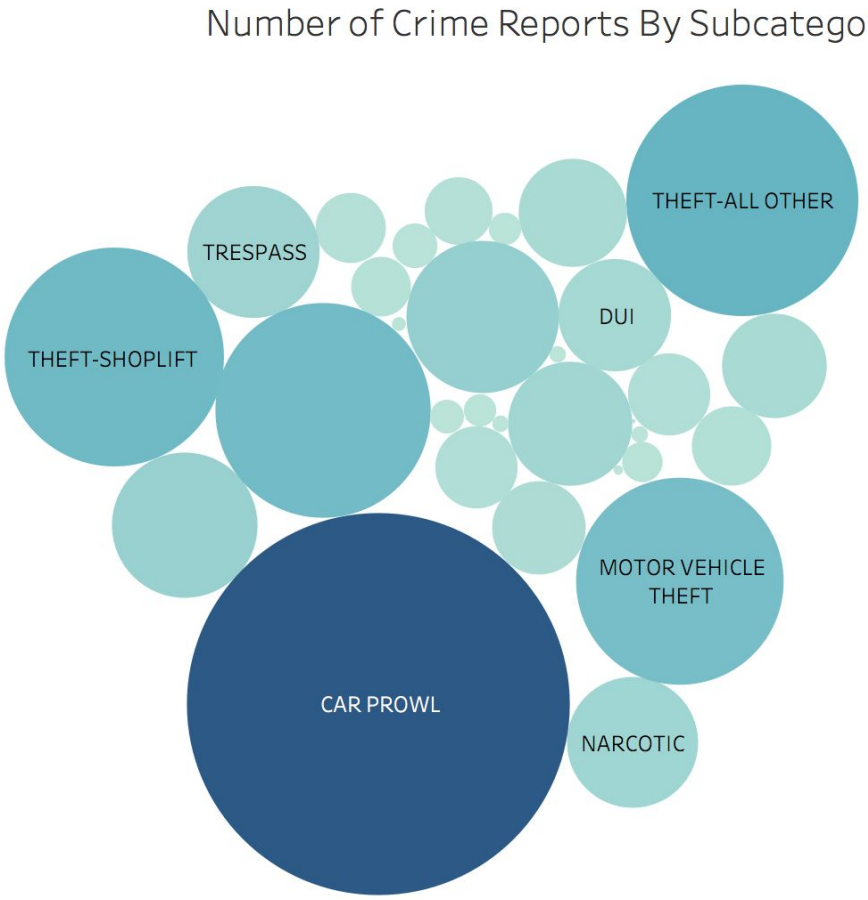
The plot below shows the number of crime reports in each crime subcategory.



The pie chart below visualizes the number of crime reports in each crime category.

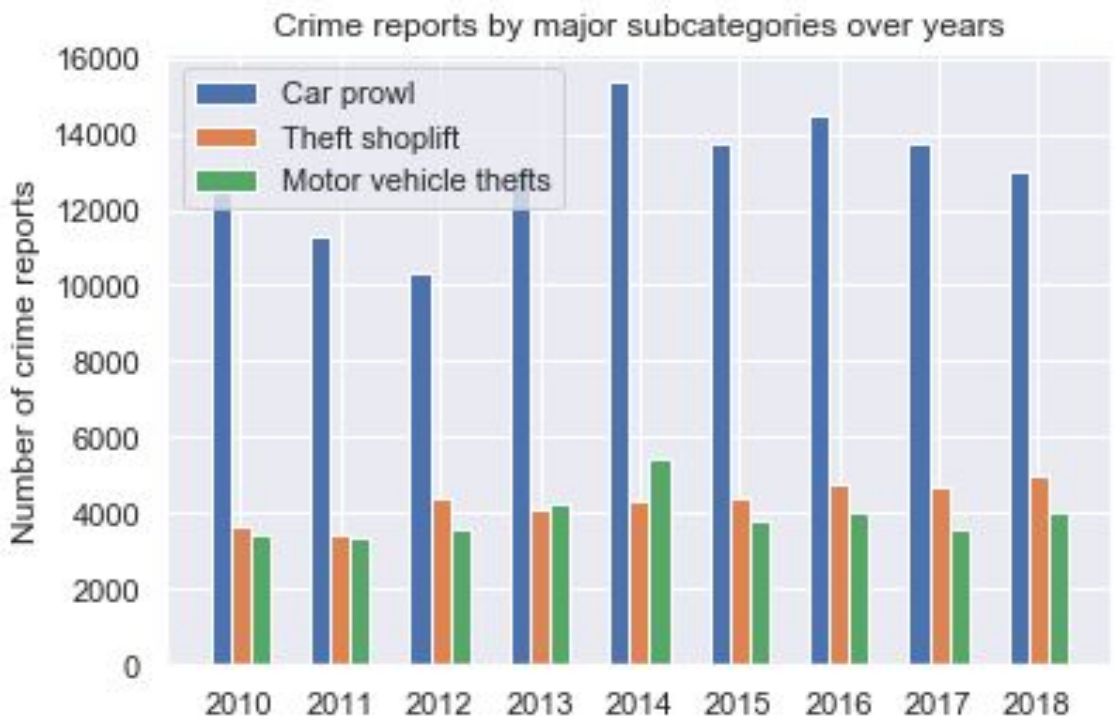


The bubble map below visualizes the number of crime reports in each subcategory.



Quantity of crime reports of major subcategories over years

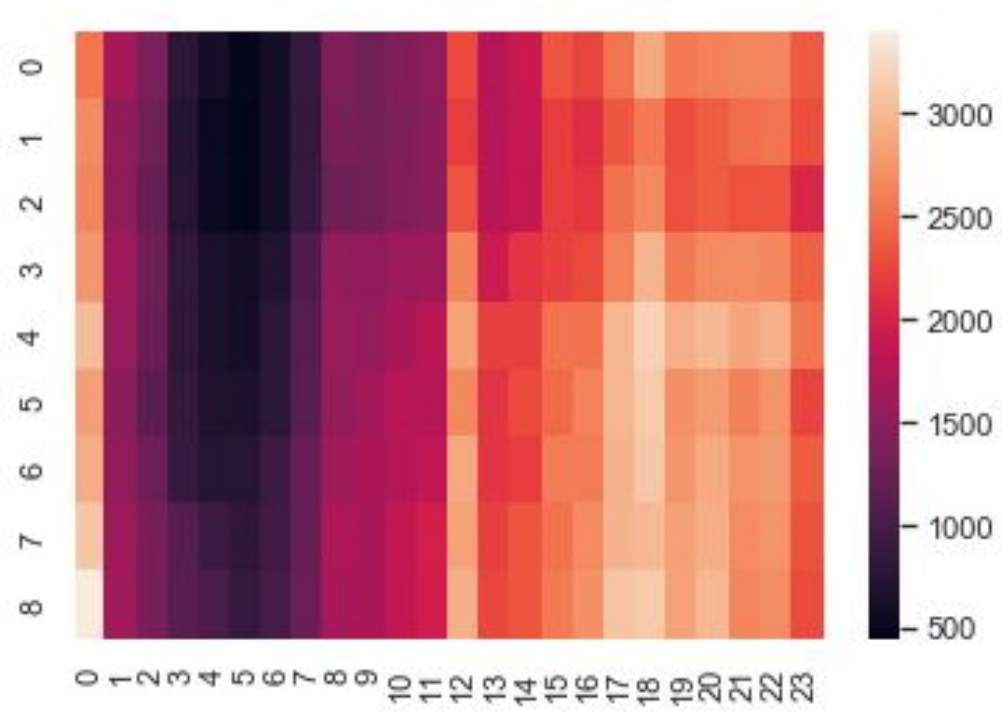
The barplot below visualizes the number of crime reports by three major subcategories over year.



Mean and variance of number crime reports of major subcategories 2010-2018
The boxplot below visualizes the distribution of number of crime reports for each major subcategory in 2010-2018.

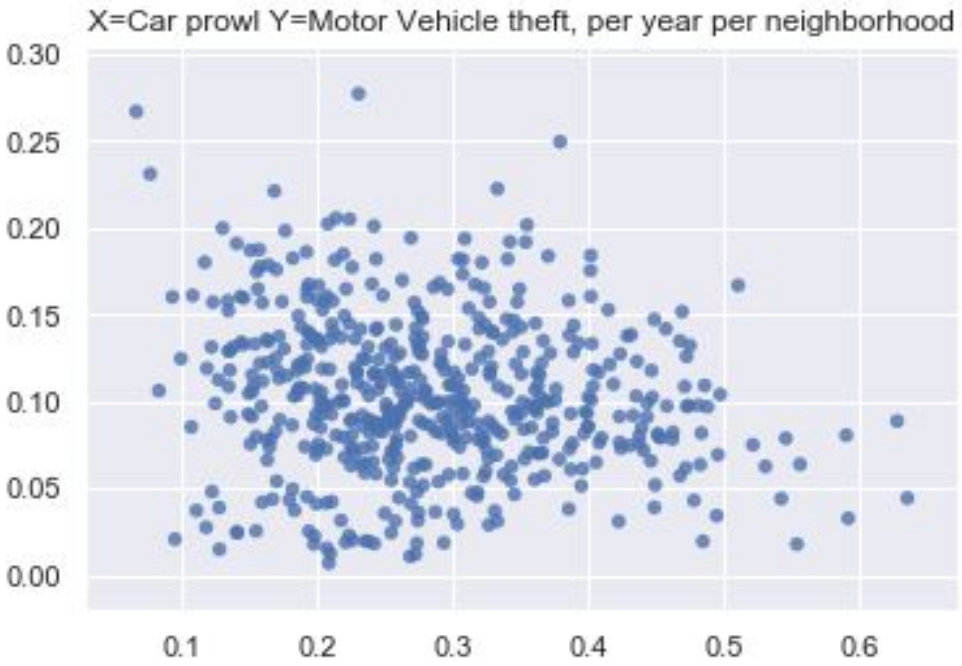
Quantity of crime reports over years in each hour in the day

The heat map below shows the number of crime reports in each hour (x-axis) in each year (2010-2018).



Correlation between number of car prowls and number of motor vehicle thefts

The scatterplot shows the correlation between the fraction of car prowl over all crime reports and the fraction of motor vehicle theft over all crime reports. Each sample point represents the data of a neighbor in a certain year.



Storyline

Overall the number of crime reports has been steadily increasing over years (see the histogram in **Quantity of crime reports over recent years**), especially in the less safe areas (see the stacked area plot in **Quantity of crime reports by precincts over years**). On the other hand, the number of crime reports of the major crime subcategories (e.g. car prowl, theft shoplift, motor vehicle theft) doesn't go up proportionally (see the **Quantity of crime reports of major subcategories over years**), which probably means the distribution of crime subcategories becomes diversified over years. The number of crime reports becomes more evenly distributed during the day (see the heat map in **Quantity of crime reports over years in each hour in the day,** especially the fact that the dark area becomes less and less over years)

Conclusion

All the observations made in 'storyline' indicate that Seattle has become a less safe city over years, especially

- The number of crime reports steadily goes over years, and this trend is more prominent in areas that were already dangerous.
- The crime types become diversified over years, which probably means it is harder for average people to avoid being affected by the crimes.
- The occurrence time of crimes become diversified over year, which may further make it hard for average people to stay immune to crime activities.

Appendix containing all code

Please reference to jupyter notebook.

Link to github page with this analysis

Crime data visualization

May 15, 2019

In [1]:
```
import pandas as pd
import numpy as np

# Seattle crime data.
# https://catalog.data.gov/dataset/crime-data-76bd0
df = pd.read_csv("Crime_Data.csv")
df.columns
```

Out[1]:
```
Index(['Report Number', 'Occurred Date', 'Occurred Time', 'Reported Date',
       'Reported Time', 'Crime Subcategory', 'Primary Offense Description',
       'Precinct', 'Sector', 'Beat', 'Neighborhood'],
      dtype='object')
```

In [2]:
```
# Histogram of crime report vs year
```

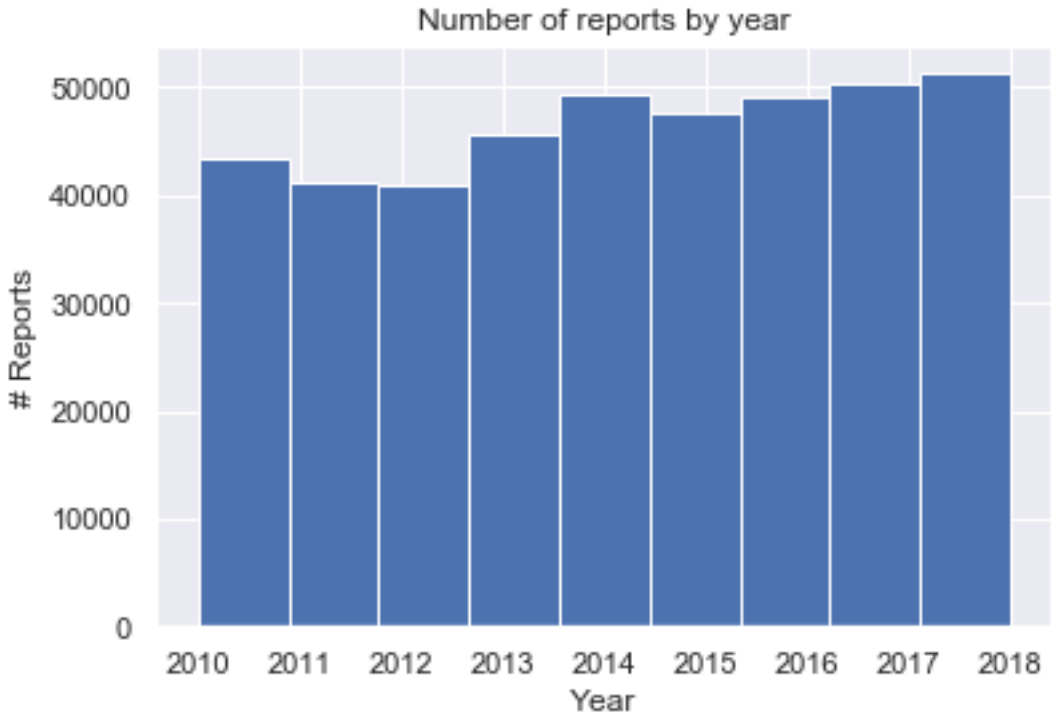
In [3]:
```
df['Occurred Date'] = pd.to_datetime(df['Occurred Date'])
```

In [4]:
```
df = df.loc[(df['Occurred Date'].dt.year >= 2010) & (df['Occurred Date'].dt.year <= 201
df.head()
```

Out[4]:

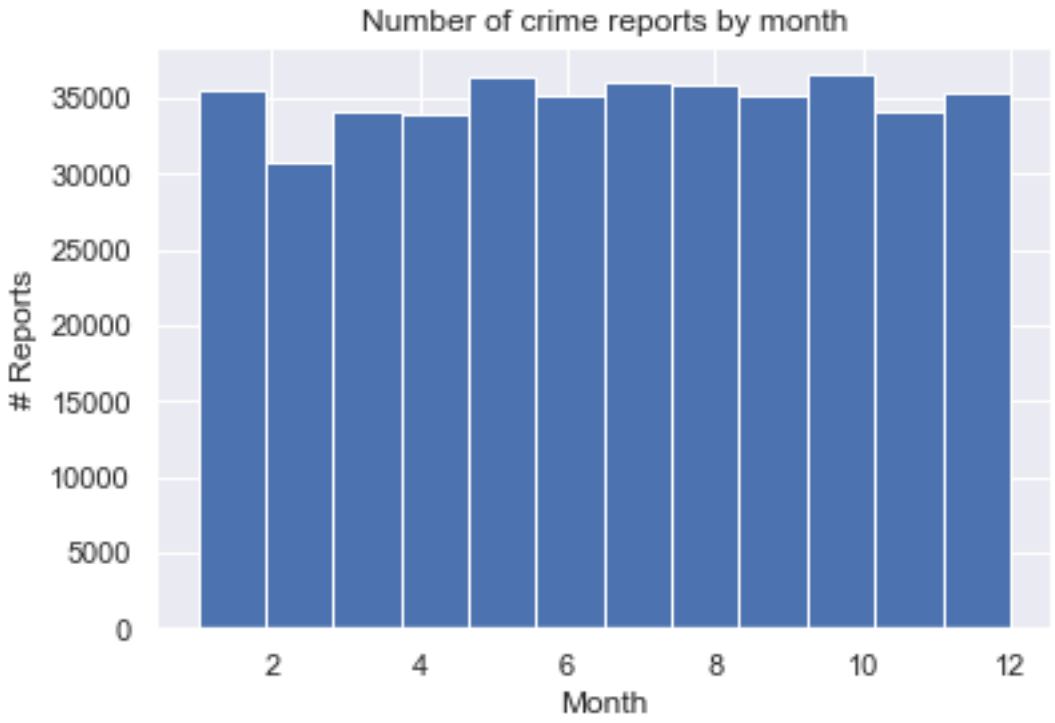
	Report Number	Occurred Date	Occurred Time	Reported Date	Reported Time	Crime Subcategory	Primary Offense Description	Precinct	Sector	Beat	Neighborhood
88079	20100000100029	2010-03-27	239.0	03/27/2010	310.0	NARCOTIC	NARC-POSSESS-HALLUCINOGEN	EAST	G	G2	CENTRAL AREA/SQUIRE PARK
88080	20100000100052	2010-03-27	1.0	03/27/2010	501.0	CAR PROWL	THEFT-CARPROWL	WEST	K	K1	DOWNTOWN COMMERCIAL
88081	20100000100057	2010-03-27	348.0	03/27/2010	402.0	DUI	DUI-LIQUOR	UNKNOWN	NaN	NaN	UNKNOWN
88082	20100000100076	2010-03-27	300.0	03/27/2010	450.0	ROBBERY-STREET	ROBBERY-STREET-BODYFORCE	NORTH	B	B3	WALLINGFORD
88083	20100000100104	2010-03-27	508.0	03/27/2010	508.0	DUI	DUI-DRUGS	EAST	E	E1	CAPITOL HILL

```
In [59]: import matplotlib.pyplot as plt

         df['Occurred Date'].dt.year.dropna().hist(bins=9)
         plt.ylabel('# Reports')
         plt.xlabel('Year')
         plt.title('Number of reports by year')
         plt.show()
```



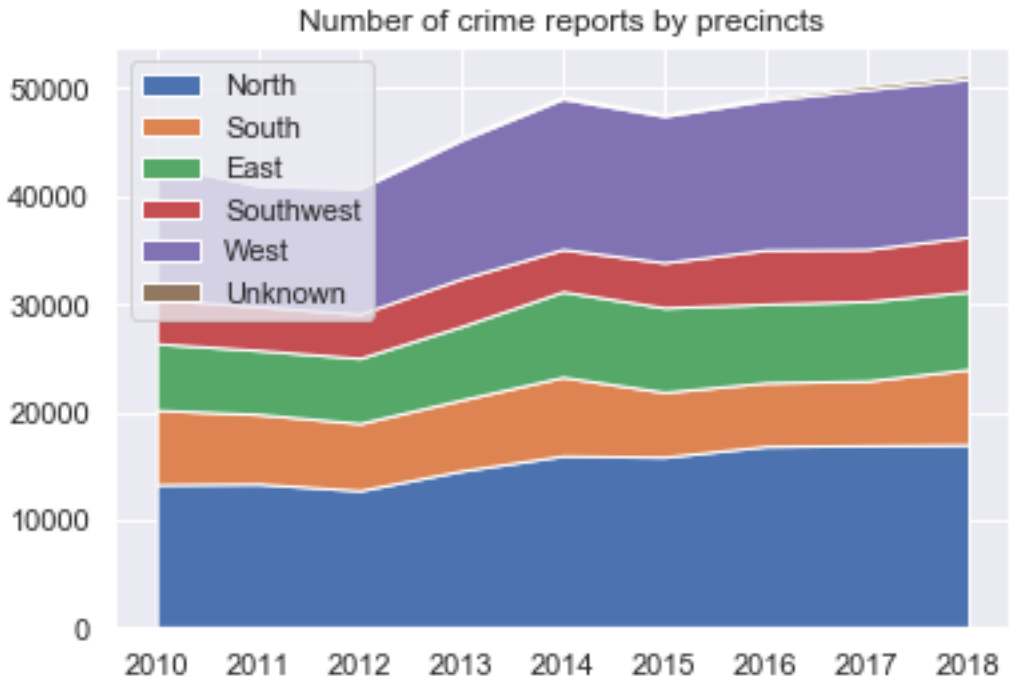
The histogram above shows the number of crime reports over time. The number of crime reports steadily goes up over years.

```
In [63]: df['Occurred Date'].dt.month.dropna().hist(bins=12)
         plt.ylabel('# Reports')
         plt.xlabel('Month')
         plt.title('Number of crime reports by month')
         plt.show()
```



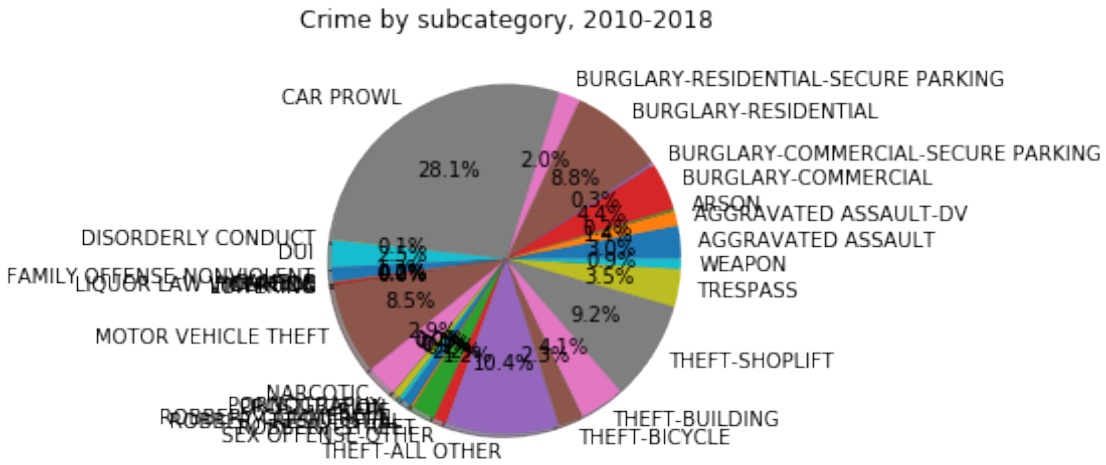
The histogram above shows the number of reports for each month in the year. The distribution is almost uniform, except for January, probably because it is the beginning of the year/right after the holiday season.

```
In [8]: # Crime report by precinct in 2010

In [9]: north = df.loc[(df['Precinct'] == 'NORTH')].groupby(df['Occurred Date'].dt.year).size()
        south = df.loc[(df['Precinct'] == 'SOUTH')].groupby(df['Occurred Date'].dt.year).size()
        east = df.loc[(df['Precinct'] == 'EAST')].groupby(df['Occurred Date'].dt.year).size().
        southwest = df.loc[(df['Precinct'] == 'SOUTHWEST')].groupby(df['Occurred Date'].dt.yea
        west = df.loc[(df['Precinct'] == 'WEST')].groupby(df['Occurred Date'].dt.year).size().
        unknown = df.loc[(df['Precinct'] == 'UNKNOWN')].groupby(df['Occurred Date'].dt.year).si

In [61]: x=range(2010,2019)
         y=[ north, south, east, southwest, west, unknown ]
         plt.stackplot(x,y, labels=['North','South','East', 'Southwest', 'West', 'Unknown'])
         plt.legend(loc='upper left')
         plt.title('Number of crime reports by precincts')
         plt.show()
```

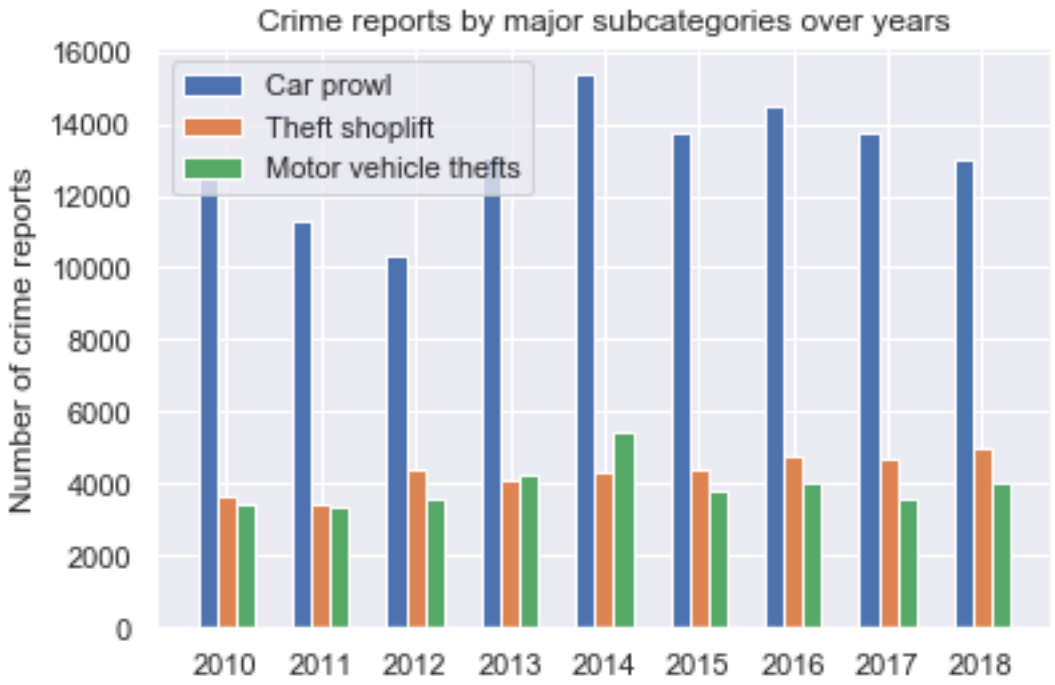


Observations: * North takes one third of all crime reports and its fraction goes up over years. * West has the second largest number of crime reports. * No significant change in terms of distribution is observed over years.

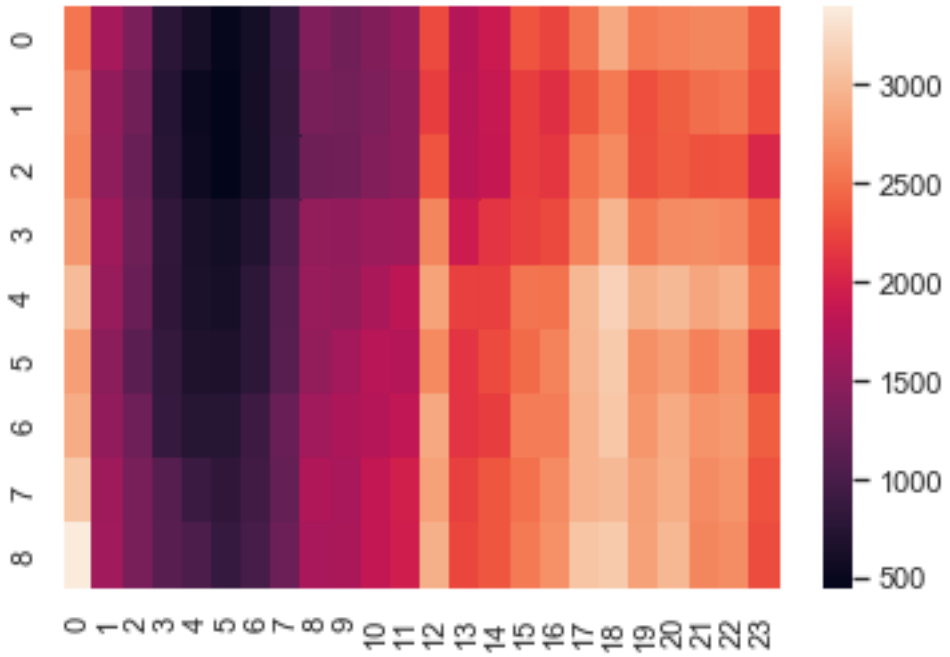
```
In [11]: # Crime Subcategory
         num_reports_by_crime_subcategory = df.groupby("Crime Subcategory").size()

         def piechart(series, title):
             labels = series.index.tolist()
             sizes = series.tolist()
             plt.pie(sizes, labels=labels, autopct='%1.1f%%', shadow=True)
             plt.title(title)
             plt.show()


         piechart(num_reports_by_crime_subcategory, "Crime by subcategory, 2010-2018")
```



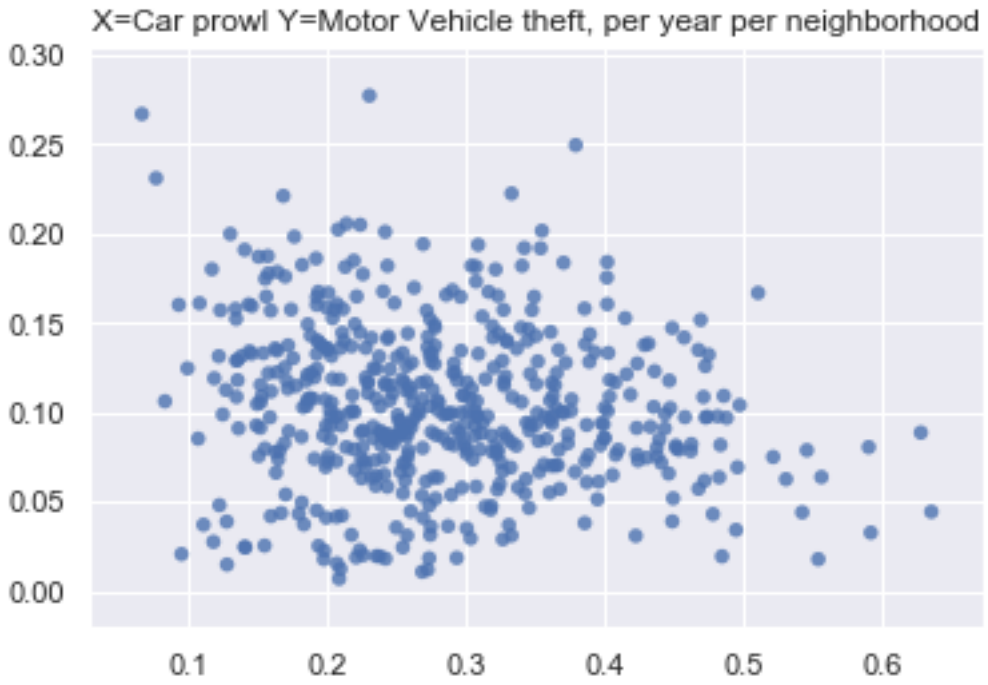

```
In [12]: # Let's look into the three major subcategories: car prowl (28.1%), theft-shoplift (9
         car_prowls = df.loc[(df['Crime Subcategory'] == 'CAR PROWL')]
         theft_shoplifts = df.loc[(df['Crime Subcategory'] == 'THEFT-SHOPLIFT')]
         motor_vehicle_thefts = df.loc[(df['Crime Subcategory'] == 'MOTOR VEHICLE THEFT')]

In [13]: # By year
         car_prowls_by_years = car_prowls.groupby(df['Occurred Date'].dt.year).size()
         theft_shoplifts_by_years = theft_shoplifts.groupby(df['Occurred Date'].dt.year).size()
         motor_vehicle_thefts_by_years = motor_vehicle_thefts.groupby(df['Occurred Date'].dt.y

In [64]: ind = np.arange(9)
         width = 0.2
         plt.bar(ind, car_prowls_by_years.tolist(), width, label='Car prowl')
         plt.bar(ind + width, theft_shoplifts_by_years.tolist(), width, label='Theft shoplift'
         plt.bar(ind + width*2, motor_vehicle_thefts_by_years.tolist(), width, label='Motor ve

         plt.legend(loc='upper left')

         plt.ylabel('Number of crime reports')
         plt.title('Crime reports by major subcategories over years')

         plt.xticks(ind + width, car_prowls_by_years.index.tolist())
         plt.show()
```



In [23]:
```
# Heat map between x = hours (0-23), y = years (2010 to 2018)
years = range(2010,2019)
matrix = np.empty([9, 24])
for year in years:
    matrix[year-2010,:] = df.loc[df['Occurred Date'].dt.year == year].groupby((df['Oc

import seaborn as sns; sns.set()
ax = sns.heatmap(matrix)
```



In [62]:
```
# Scatter plot between x = #number of NARCOTICS, y = #number of DUI, per year, per ne

x = []
y = []

category1 = 'CAR PROWL'
category2 = 'MOTOR VEHICLE THEFT'
for year in range(2010, 2019):
    category1_by_neighborhood = df.loc[(df['Crime Subcategory'] == category1) & (df['
    category2_by_neighborhood = df.loc[(df['Crime Subcategory'] == category2) & (df['
    all_by_neighborhood = df.loc[df['Occurred Date'].dt.year == year].groupby(df['Neig
    for index in category1_by_neighborhood.index:
        if index not in category2_by_neighborhood.index:
            continue
        if index not in all_by_neighborhood.index:
            continue
        x.append(float(category1_by_neighborhood[index]) / all_by_neighborhood[index])
        y.append(float(category2_by_neighborhood[index]) / all_by_neighborhood[index])

plt.scatter(x, y, alpha=0.8, edgecolors='none', s=30)

plt.title('X=Car prowl Y=Motor Vehicle theft, per year per neighborhood')
plt.show()
```



In [50]:
```
# boxplot,  car prowl, theft-shoplift, motor vehicle theft, per year
data_to_plot = [car_prowls_by_years, theft_shoplifts_by_years, motor_vehicle_thefts_by
plt.boxplot(data_to_plot)
plt.xticks(np.arange(1, 4), ('Car prowl', 'Theft-shoplift', 'Motor vehicle theft'))
plt.show()
```