Garrett Yamane

Investigation of San Francisco Crime Data 2003-2019

1. Introduction

San Francisco is a beautiful city that attracts a wide variety of visitors and residents. From Mission District to Chinatown, San Francisco is an amazing city full of different cultures. With a population of 880,000 and an area of 46.87 mi^2, San Francisco's density brings together quite a unique network of people each day. Safety is important for anywhere you travel or live, and being aware of crime and potential risks is always beneficial for one's own safety.

2. Problem Statement

San Francisco crime and safety are concerns for anyone wanting to travel, visit, or live in the city. These concerns highlight the importance of staying educated regarding police reports that have occurred in the past. Are certain areas more dangerous than others? What times of the day do more crimes tend to occur, and do additional factors such as the time of the year (Christmas, Black Friday, etc.) influence the risk of crime? These types of questions help people in San Francisco keep safe and aware of the potential dangers that could pose a direct threat to them. Through observing past crime data and police reports, I can illustrate key features and suggest patterns that contribute to common crime in San Francisco. I will use my findings to model San Francisco crime data in order to decipher important factors that contribute to San Francisco crime patterns and ideally predict future potential crime location/classification. The City of San Francisco could use this to better prepare for certain events when crime rates tend to spike and take appropriate precautionary measures and target areas that are unsafe to keep visitors and residents protected from harm.

3. Target Client

The San Francisco Police Department has to be prepared at all times for any potential report of crime. They must act accordingly and it is essential that they are readily on duty at peak crime times and areas. The more informed they are about the patterns and conditions that trigger spikes in crime rates, the better it will be for both the Police Department and public. My client (SF Police Department) can use the models I build to better understand patterns in past crimes and understand conditions that may lead to more in the future. They can

prepare for certain events and be aware of certain conditions that tend to lead to higher rates of crime, making sure that police units are readily available and onsite.

4. Data Wrangling

The San Francisco Government keeps records of police reports filed hosted at https://datasf.org/, dating from 2003-present. From 2008-2013, "the district attorney's domestic violence team dropped about 72 percent of all domestic violence criminal cases before they reached court" (https://sfpublicpress.org/news/2013-06/domestic-violence-case-that-spurred-san-francisco-reforms-comes-to-a-close). Stacks of hardcopy police report files made record-keeping difficult and cases like these could easily fall through. The San Francisco Government took steps toward improving their access to important records and made moves toward updating to the new Crime Data Warehouse data system, which made accessibility easy for its users.

a) Gathering the Data

Until May 2018, San Francisco crime data was maintained through CABLE, the legacy mainframe used to store their crime reports. However, CABLE was extremely prone to issues relating to delays and data accessibility and was discontinued. At the start of 2018, the San Francisco Police Department announced its release of the newly updated Crime Data Warehouse that clearly organizes SF crime data and is quick and responsive.

Because of this update, my dataset contains a union of two datasets: one for CABLE police reports dating from 2003-2018 and the other for Crime Data Warehouse police reports from 2018-2019.

- CABLE mainframe 2003-2018 police reports
- Crime Data Warehouse 2018-2019 police reports

b) Consolidating the Data

Prior to merging the two datasets, I wanted to prepare each of the datasets to make merging as easy as possible. I renamed columns from the CABLE mainframe to match their corresponding column names in the Crime Data Warehouse, ordered each crime report by increasing date, and reformatted the joint columns to have the same value format to keep consistency.

There was also some overlap with reports in 2018, so in order to keep all 2018 reports consistent I took 2003-2017 reports from CABLE and 2018-2019 reports from the Crime Data Warehouse to create a merged dataframe. There were some columns that were unique to each of the datasets, leaving many columns in the merged dataframe having a high percentage of missing values.

Incident Number	0.00
Incident Category	0.00
Incident Description	0.00
Incident Day of Week	0.00
Incident Date	0.00
Incident Time	0.00
Police District	0.00
Resolution	0.00
Intersection	0.59
Longitude	0.59
Latitude	0.59
point	0.59
Row ID	0.00
SF Find Neighborhoods	1.04
Current Police Districts	0.65
Current Supervisor Districts	0.62
Analysis Neighborhoods	0.64
Incident Year	0.00
Incident Datetime	89.78
Report Datetime	89.78
Incident ID	89.78
CAD Number	92.11
Report Type Code	89.78
Report Type Description	89.78
Filed Online	97.83
Incident Code	89.78
Incident Subcategory	89.78
CNN	90.37
Analysis Neighborhood	90.37
Supervisor District	90.37
HSOC Zones as of 2018-06-05	97.70
OWED Public Spaces	99.48
Central Market/Tenderloin Boundary Polygon - Updated	98.62
Parks Alliance CPSI (27+TL sites)	99.87

Figure 1: Merged dataframe's columns and percentage of missing values

The merged dataframe has a total of 2,415,298 rows and 34 columns. Imputing over columns that are missing such a large number of values would not be realistic. To account for the vast number of missing values, I focused my attention on the shared columns between the two datasets and combined features that contained similar information such as "Analysis Neighborhood" and "Analysis Neighborhoods". My finalized merged dataframe has a shape of 2,415,298 rows (each representing a separate crime report) and 15 crime feature columns.

Incident Number	0.00
Incident Category	0.00
Incident Description	0.00
Incident Day of Week	0.00
Incident Date	0.00
Incident Time	0.00
Police District	0.00
Resolution	0.00
Intersection	0.59
Longitude	0.59
Latitude	0.59
point	0.59
Row ID	0.00
Incident Year	0.00
Analysis Neighborhood	0.64

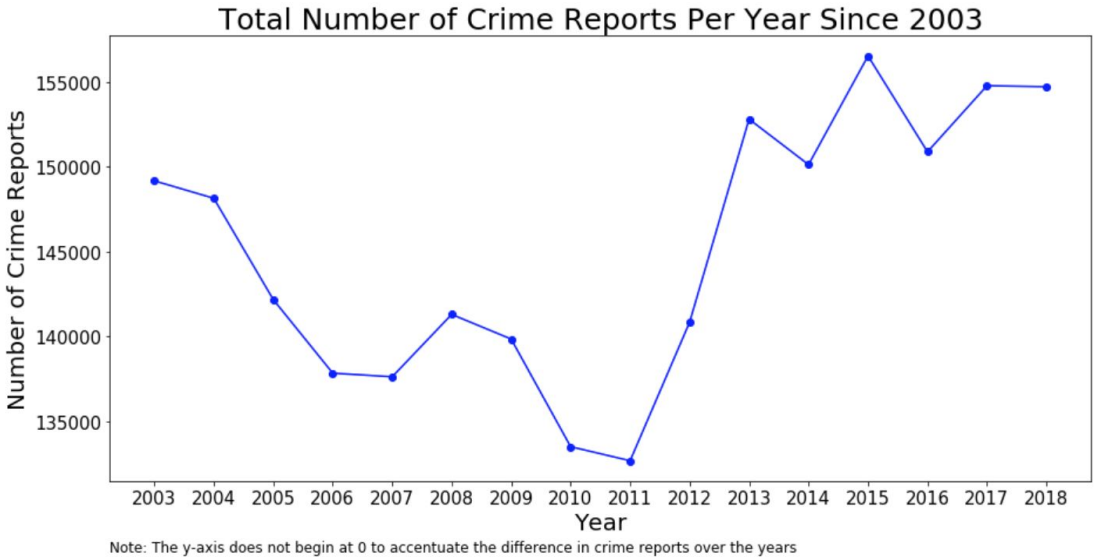
Figure 2: Merged dataframe's columns and percentage of missing values after modifying

5. Exploratory Data Analysis and Initial Findings

With my merged data frame, all data from 2003-2019 that has been cleaned and is consolidated into one place. By digging deeper into the data, my goal was to reveal underlying stories behind crimes being reported. Having all crime reports in one place allows me to easily access all available data to answer questions such as "what time of the year are crimes most likely to happen?" and "are certain police districts more at risk than others?"

a) How has crime report rate changed per year since 2003?

I began by examining the number of crimes being reported per year from 2003-2018. In order to not show any biases in the graph, I left out 2019 since data had only been logged up through August.

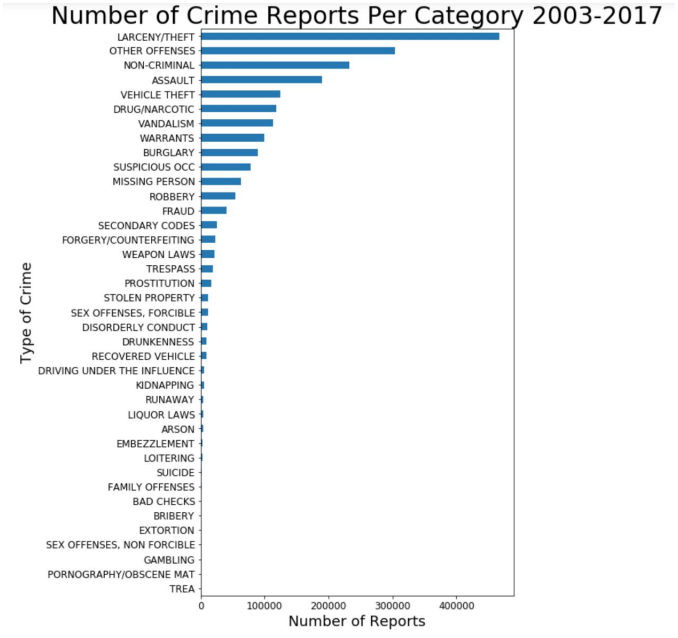


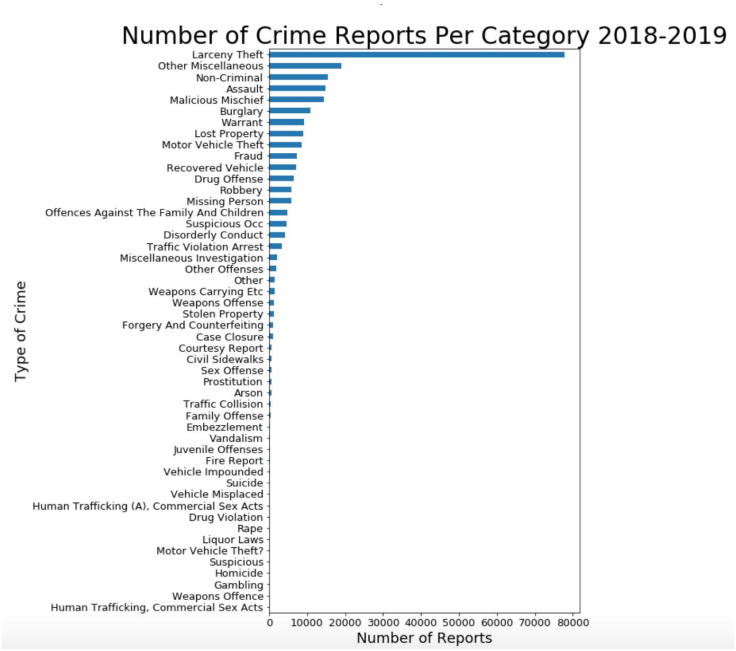
There are a couple of interesting takeaways from this graph that are worth noting. Beginning in 2003, the number of crime reports per year was high relative to the total average of 2003-2018. However, the steady

decline until the dip in 2011 is interesting. Observing patterns that could have caused this and comparing this year to the peak in 2015 motivate opportunities for further exploration.

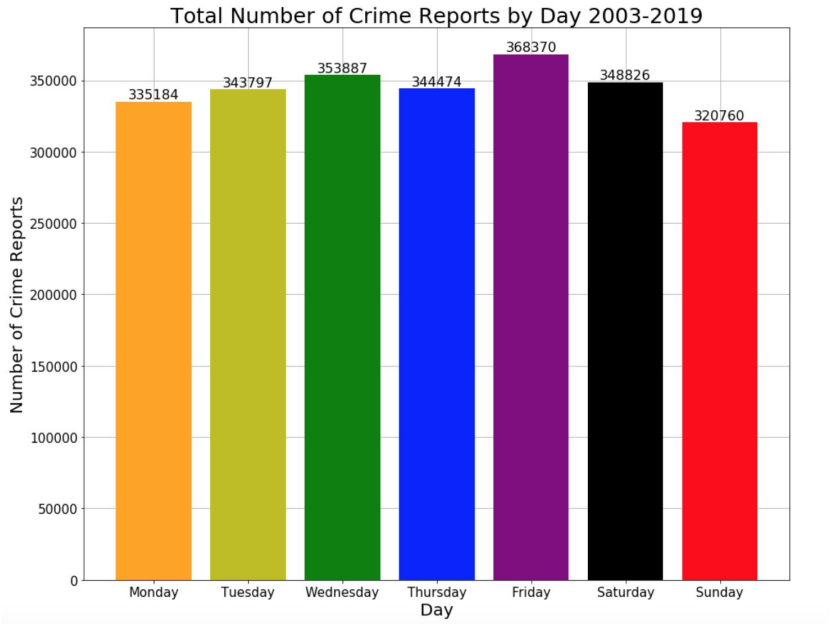
b) What classification of crimes is reported the most?

Going back to my merged data frame, there were a couple of things I needed to consider before answering this question. For each crime report, the classification of crime was listed under the "Incident Category" column. However, 2003-2017 crimes from the CABLE mainframe were labeled sightly different compared to the 2018-2019 Crime Data Warehouse crimes. For example, "LARCENY/THEFT" and "Larceny Theft" were two respective classifications for each data set that represented the same type of report. I did not want to treat these two differently, so I split up the reports into 2003-2017 reports and 2018-2019 reports. By doing this, I could examine the original classifications for each report and compared the results to each other.



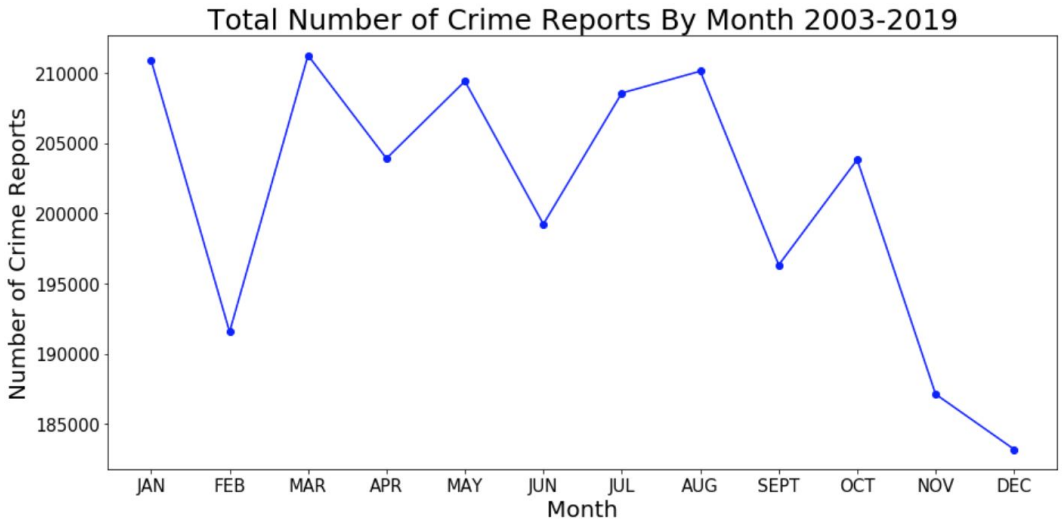


c) What days or months are crimes most likely?



From the graph, it appeared that Fridays have been the most likely for a crime to be reported. This seems reasonable since it marks the beginning of the weekend when more activity during the night outside of work is common. Sunday, the end of the weekend, sees the lowest number of crime reports. The drop from the beginning to the end of the weekend, though small relative to the entire data set (47,610 reports compared to the 2,415,298 total reports), could still affect the likelihood of crimes being reported.

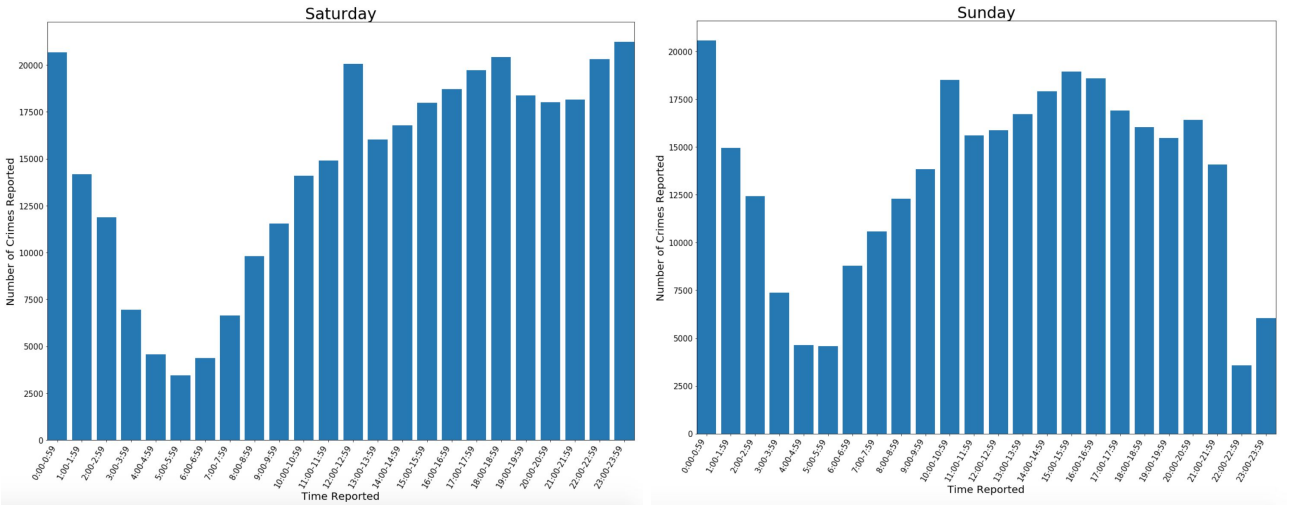
I also did a similar comparison between months to visualize how the number of crime reports change from month to month.



Crime reports appear to be more common at the beginning of the year, with the lowest numbers occurring in November and December. The dip in February between January and March (which are the two highest months) could be because there are less days but would be interesting to look further into.

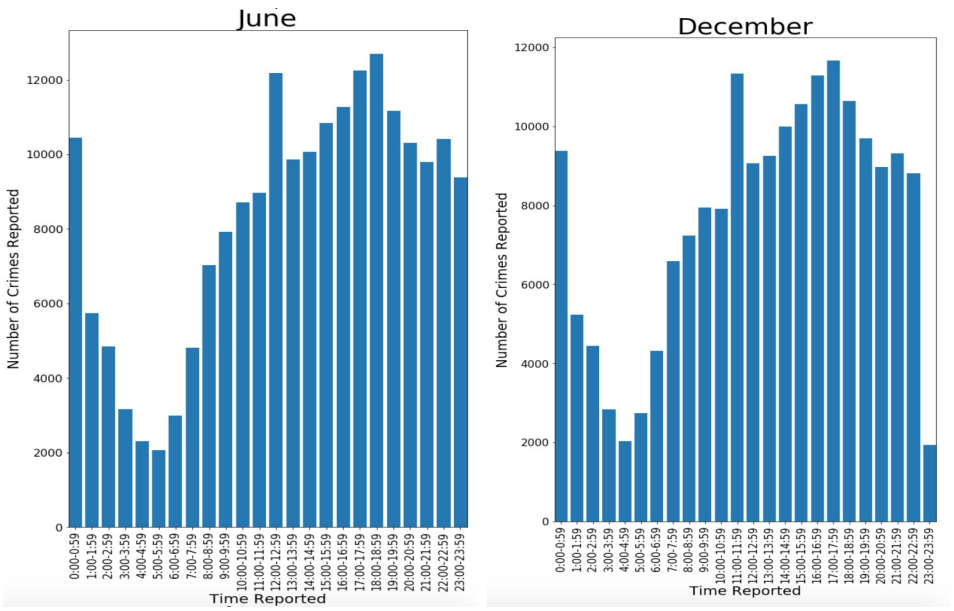
d) Are crimes more likely at certain times of the day or year?

Next, I wanted to isolate each crime by the time it was reported to see if time had any effects on how many crimes were reported. First, I formed buckets representing crimes by the hour. Crimes reported between 12:00am-12:59am were in one bucket, 1:00am-1:59am were in another bucket, and so on. I then took each of these buckets and split them by day reported to get individual graphs representing times crimes were reported for a specific day.



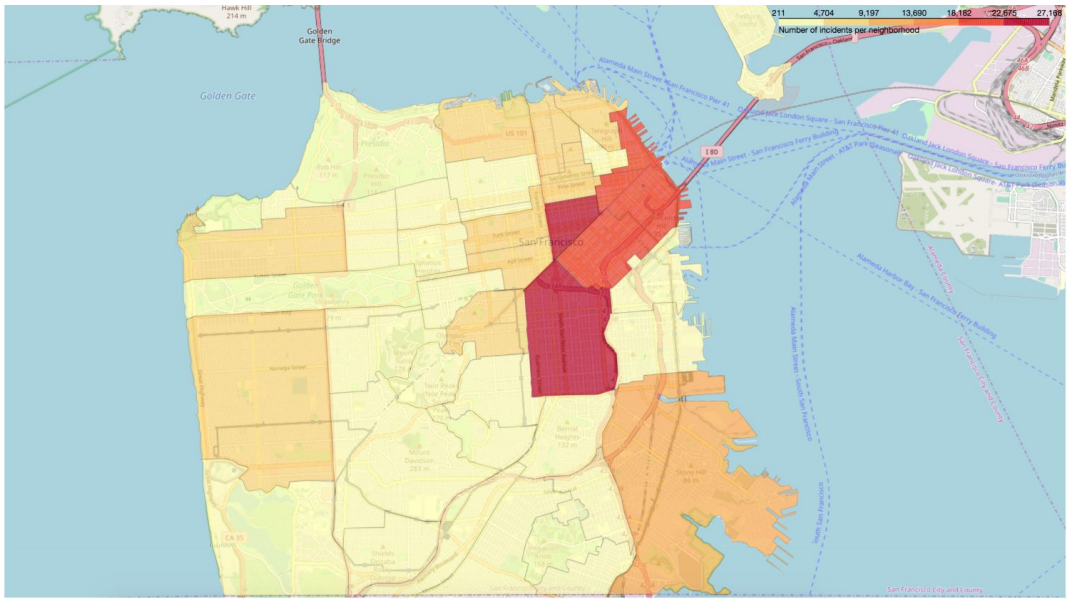
Above is an example for Saturday and Sunday, where each bin represented the hour in which a crime was reported. Binning the crimes by hour made visualizing patterns relating to time much more clear. Hours 10:00pm-11:59pm highlighted an interesting distinction between the two days. Saturday crimes are more likely compared to Sunday, and there is a large decline in late night crimes toward the end of the weekend.

I decided to take this a step further and look at times of reported crimes per month as well. I was curious to see if the time of the year had any significant effect on the times that crimes were being reported. Below are graphs for June and December representing the total number of crimes reported for each time bin.



e) Which neighborhoods are most crimes reported?

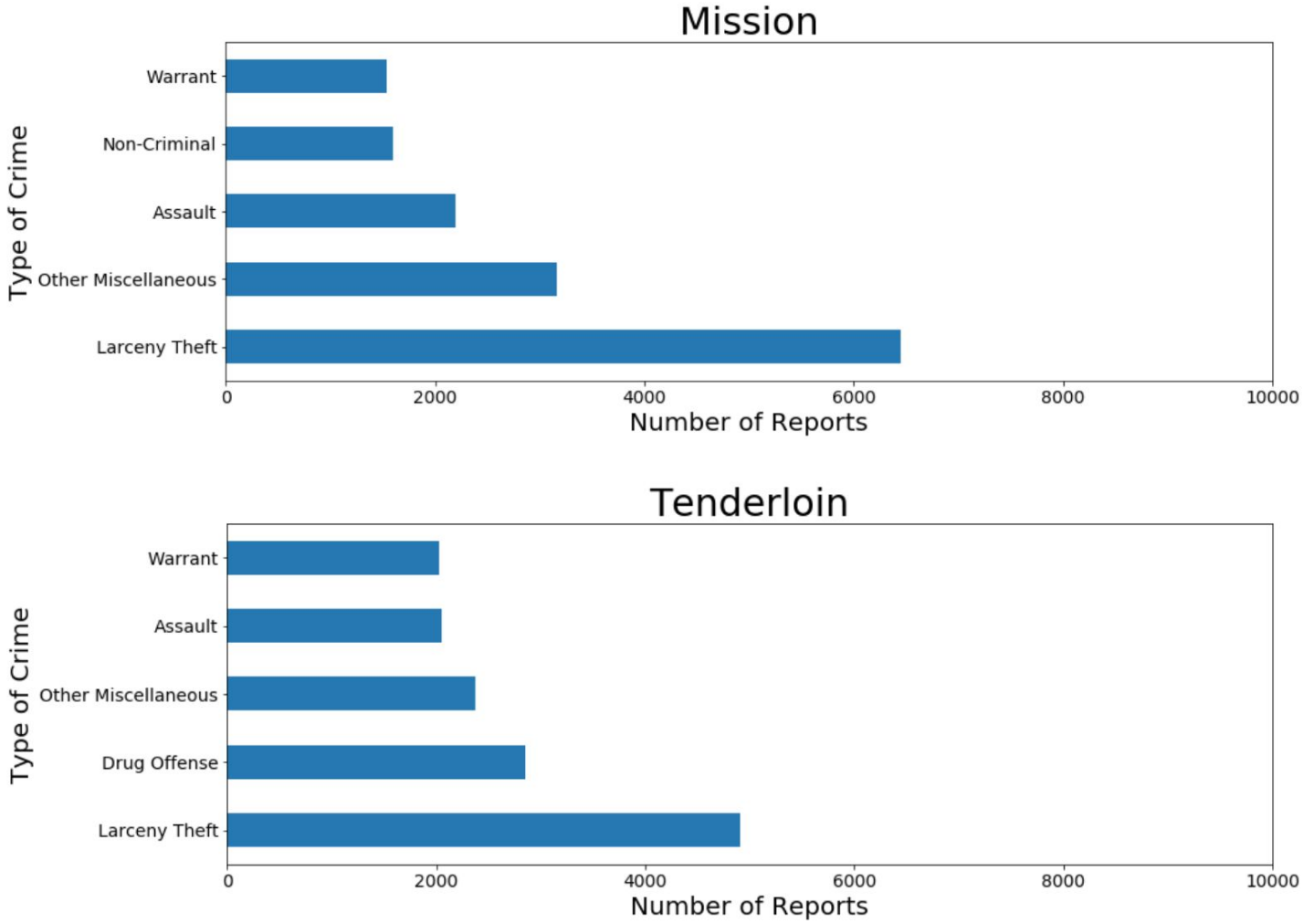
San Francisco consists of 41 different neighborhoods created by The Department of Public Health and the Mayor's Office of Housing and Community Development for the purpose of providing consistency in the analysis and reporting of socio-economic, demographic, and environmental data (https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783). Each crime report was assigned one of these 41 neighborhoods, and visually representing the crime reports via a choropleth map allowed for an easily interpretable visualization of areas of San Francisco that were most dangerous.



	Neighborhood	Number of Reports
0	Mission	27168
1	Tenderloin	24788
2	Financial District/South Beach	22052
3	South of Market	20614
4	Bayview Hunters Point	13385
5	North Beach	7746
6	Western Addition	7475
7	Castro/Upper Market	7114
8	Sunset/Parkside	6935
9	Nob Hill	6572

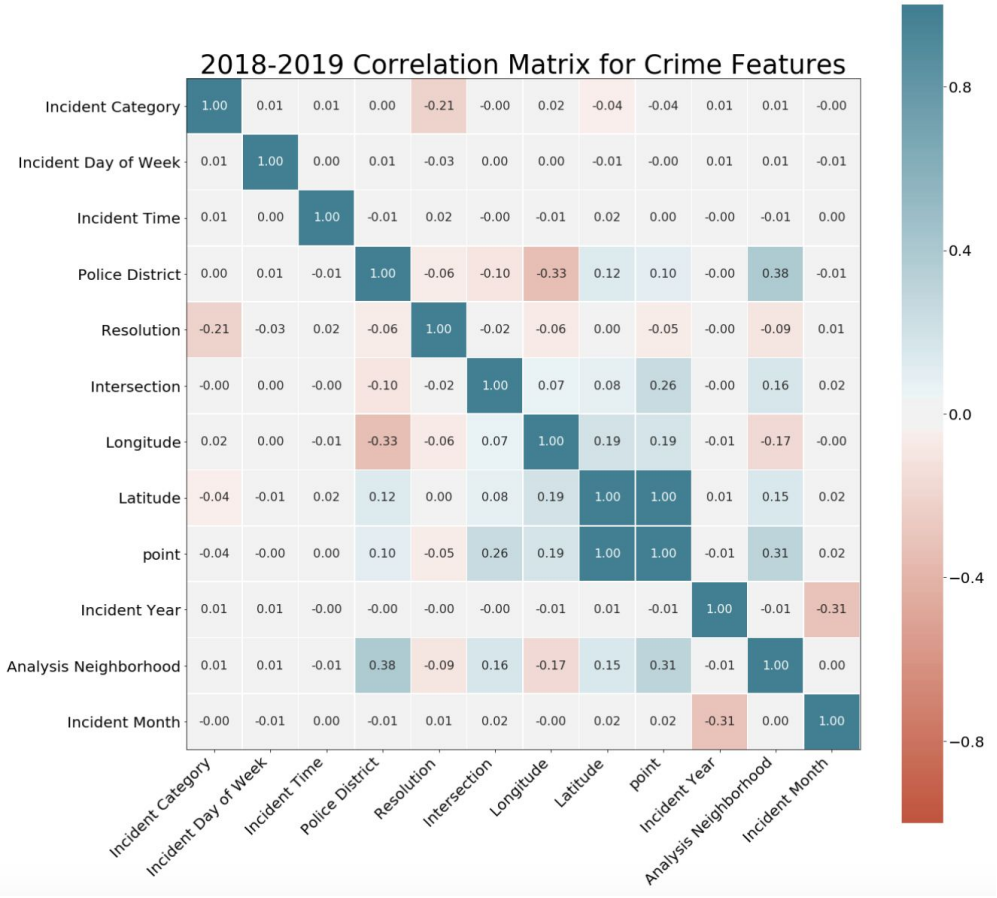
Figure 3: Chloropleth Map of Crimes per Neighborhood for 2018-2019 Reports

I decided to look at 2018-2019 reports to examine the current most dangerous neighborhoods. A direction I wanted to take this information was to see if certain classifications of crimes were different among the most dangerous neighborhoods. Was theft more common in Mission than Tenderloin? Or were drug offenses more of a concern in certain neighborhoods? These questions motivated me to hone in on the five most dangerous neighborhoods to see the types of crimes they struggle with.



Above are the two neighborhoods with the most crime reports from 2018-2019. Drug offense crimes ranked in the top five most reported crimes in Tenderloin, yet didn't show up for Mission. This could help the San Francisco Police Department target areas when dealing with specific crimes by knowing which areas are most likely to have those crime classifications reported.

f) Heatmap and Correlation Matrix for Crime Features



Initially, each crime report contained some categorical information that was stored as a String:

- Incident Category
- Incident Day of Week
- Incident Time
- Police District
- Resolution
- Intersection
- Analysis Neighborhood
- point

String columns initially couldn't be compared for their correlation. Because of this, I converted the above string columns to 'categorical codes' using Pandas' built-in *Series.cat.codes*. This allowed me to create a correlation matrix between the different crime features to visualize how certain features change in relation to others. For example, it appears that there is some correlation between "Resolution" and "Incident Category". Even though

the negative correlation doesn't tell us much at first, there is some relationship between the type of crime that is reported and the way it was dealt with, which could be interesting to look even further into.

6. Statistical Data Analysis

a) Crime Report Rate: Is there a statistical significance between the average number of crimes per year between different police districts?

San Francisco is split into 11 different police districts. In 2015, the boundaries for each police district changed. If I were to take all reports from 2003-2018, some reports would be cross-referenced due to the boundaries being different after 2015. Due to the change, I examined the average number of crimes reports per year for each police district starting in 2015 until 2018, leaving me with a sample of 4 data points per police district.

Police District	Mean Per Year	Variance	Total Reports
southern	26848.50	1.949013e+07	107394
mission	20342.50	2.423115e+06	81370
northern	20130.25	5.389582e+05	80521
central	20011.50	1.041427e+07	80046
bayview	13818.50	1.223255e+06	55274
ingleside	11874.50	1.143297e+06	47498
taraval	11501.00	5.272353e+05	46004
tenderloin	11223.75	8.588913e+06	44895
richmond	8993.25	1.796556e+05	35973
park	8479.75	7.493216e+05	33919

i) T-test of difference of means for police districts example: Southern vs. Park

For each comparison between two police districts, I tested the following hypotheses:

- H_0: The true mean crime rate between the two police districts are the same
- H_1: The true mean crime rate between the two police districts are not the same
    - For each, I assumed α-level of 0.05

```
# Compute the test statistic
test_stat = df_districts.loc["southern", "Mean Per Year"] - df_districts.loc["park", "Mean Per Year"]
print("Difference of means for Southern and Park:", test_stat)
```

```
Difference of means for Southern and Park: 18368.75
```

```
sample_size = 4

# Calculate MSE
mse = (df_districts.loc["southern", "Variance"] + df_districts.loc["park", "Variance"]) / 2

# Calculate standard error of test statistic
standard_err = np.sqrt(2*mse/sample_size)

# Compute t-statistic
t_stat = test_stat / standard_err

# Degrees of freedom
dof = 2*sample_size - 2

# Compute p-value
p_val = 1 - stats.t.cdf(t_stat,df=dof)

print("p-value:", p_val)
```

```
p-value: 9.075809058534112e-05
```

Since my p-value was less than my assumed α-level of 0.05, I could reject my null hypothesis that the true means of crimes of "Southern" and "Park" are the same.

b) Chi-square test of independence to identify important features associated with the type of crime committed

My target feature that I wanted to run chi-square tests against was *"Incident Category*", which gives the classification of each crime (i.e. theft, assault). Running chi-square tests against other crime features allows me to test the following hypotheses:

- H_0: There is no association between "*Incident Category*" and the comparison variable
- H_1: There is evidence to suggest that there is an association between "*Incident Category*" and the comparison variable
    - For each, I assumed α-level of 0.05

The five features I tested *"Incident Category"* against were:

- *Incident Hour*
- *Incident Day of Week*
- *Incident Month*
- *Police District*
- *Resolution*

Incident Hour and *Incident Month* contained interval data (1, 2, 3...). Before running the chi-square tests, I prepped them into more usable categorical data:

- *Incident Hour* = {morning, afternoon, evening, night}
- *Incident Month* = {spring, summer, fall, winter}

```
# Run chi-square independence test against comparison columns
chi_square = ChiSquare(df_2018_2019)
for var in comparison_cols:
    chi_square.TestIndependence(colX=var,colY="Incident Category" )
```

```
Time Category is IMPORTANT for Prediction
Incident Day of Week is IMPORTANT for Prediction
Season is IMPORTANT for Prediction
Police District is IMPORTANT for Prediction
Resolution is IMPORTANT for Prediction
```

Figure 4: Loop running chi-square test against each feature

Each test that I ran resulted in a p-value less than 0.05, meaning that I could reject my null hypothesis and accept that there is evidence to suggest that there is an association between each of these crime features with "Incident Category" (the type of crime that is committed). Lastly, to test that my chi-square test worked with irrelevant information, I tested it with"*Row ID*". "*Row ID*" is a unique identifier for each crime report in the dataset, which has no relationship to the type of crime.

```
chi_square.TestIndependence(colX="Row ID",colY="Incident Category")
```

```
Row ID is NOT an important predictor. (Discard Row ID from model)
```

Figure 5: Running chi-square test against irrelevant crime feature

7. Future Work

With the guidance of my mentor, I plan on building a baseline K-means model to cluster my dataset. I will explore different K-values using the elbow and silhouette methods and compare the results, followed by a post-analysis where I will try to extract what my K-means model is telling me. I will replicate some of the analysis done from my data storytelling and initial findings to each cluster to reveal any underlying patterns within my crime dataset.

Following my initial analysis, I will discuss with my mentor possible extensions for my model. For example: implementing different distance functions, using other clustering algorithms, or adding/engineering other features to be used when clustering.