San Francisco Crime Analysis: 2003-2017

SEAL OF THE CITY AND COUNTY OF SAN FRANCISCO

ORO EN PAZ FIERRO EN GUERRA

By: Nathan Nguyen

Table of Contents

Table of Figures

Introduction

San Francisco is located in the northern part of California with a population of 870,000 people and covering a land area of roughly about 46.87 square miles. SF is filled with beautiful attractions such as the Golden Gate Bridge, Alcatraz, and even the Golden Gate park. This beautiful city is surrounding by water and is blessed with the beautiful California weather, thus making San Francisco a top choice for tourism and a place you can call home. However, not everything is always as it seems. There are areas of San Francisco that is infested with crimes of all types of nature, which leads us to our purpose of this study.

Purpose

The main purpose of this study is to get a deeper understanding of crimes located in different districts of San Francisco. We want to be able to answers questions such as when and where do crimes occur? What types of crime do residence of San Francisco experience depending on a certain area? We want to see if there's any trends or triggers that cause crimes. This is all valuable information to help spread awareness. Not only will residence of San Francisco be more knowledgeable about the types of crimes that's occurring out there, but this study will ultimately also give better insight for the local SF Police Department. Knowing our purpose and target audience, we need to take the right approach in order to successfully complete the project.

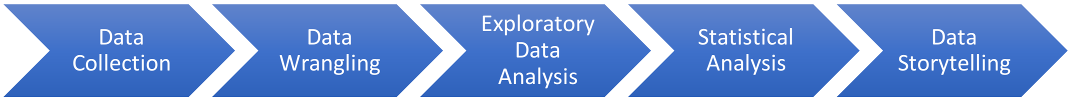


Figure 1: Approach Flow Diagram

Data Collection

Thanks to the help of open data sources such as DataSF, we are able to find valuable information that will help contribute to this study. Combining all sources of information, we are able to accumulate a dataset consisting of over two million rows spanning across 13 columns. The datasets can be found listed below:

- DataSF: Police Department Incidents 2017:
    - https://data.sfgov.org/Public-Safety/Police-Department-Incidents-Current-Year-2017-/9v2m-8wqu

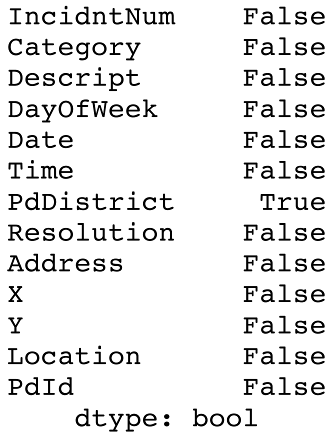
- DataSF: Police Department Incidents 2016:
    - https://data.sfgov.org/Public-Safety/Police-Department-Incidents-Previous-Year-2016-/ritf-b9ki
- DataSF: Police Department Incidents 2015-2003:
    - https://data.sfgov.org/Public-Safety/Police-Department-Incidents/tmnf-yvry/data

However, not all the data was easily accessible. Some data needed to be scrape in order to obtain the information. Web scrape data is listed below:

- Statistical Atlas:
    - https://statisticalatlas.com/neighborhood/California/San-Francisco/Mission/Population#figure/neighborhood-in-san-francisco/total-population

Data Wrangling

After collecting and compiling down all of our datasets and data source into one repository, the data needs to be examine and cleaned. Knowing that are dataset consist of over 2 million rows, we use pandas to locate missing information from the dataset.



```
IncidntNum    False
Category      False
Descript      False
DayOfWeek     False
Date          False
Time          False
PdDistrict     True
Resolution    False
Address       False
X             False
Y             False
Location      False
PdId          False
    dtype: bool
```

Figure 2: Dataset Columns

We can see that our data has missing information in the PdDistrict column and that our dataset isn’t perfect to start with. With the use of other building functions, the data was successfully cleaned and rendered down to a total of 2,123,167 rows and 13 columns and spanning over the years of 2003 to 2017. Now that our data is clean, it’s ready to be put into good use, which leads us to the start of our exploratory data analysis.

Crime Rate 2003 - 2017

How has crime change from 2003 to 2017?

Since our dataset is based on incidents of crimes reported throughout San Francisco, a good way to help visualize how the crime changed from 2003 to 2017 is to use a heat map.

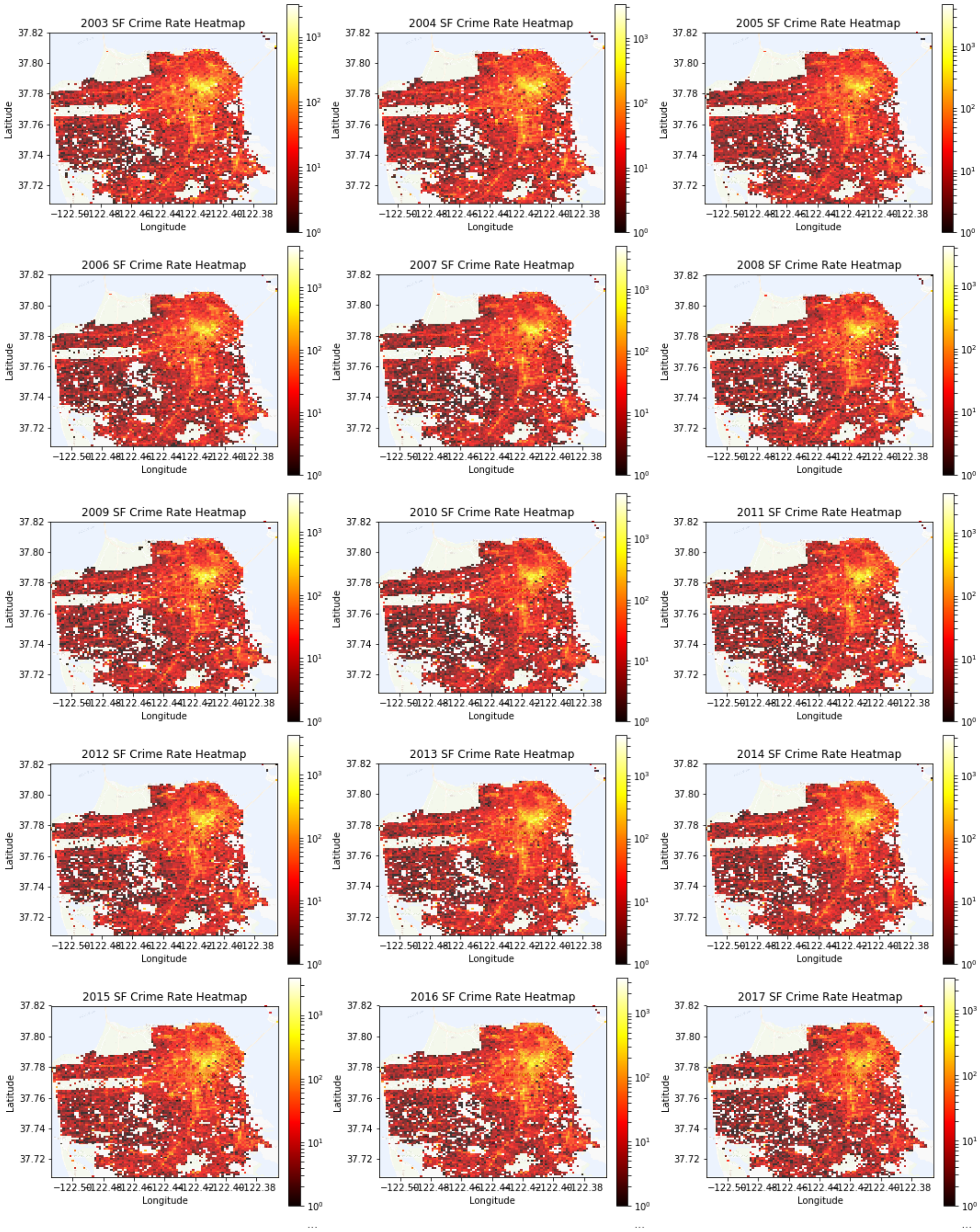

2003 SF Crime Rate Heatmap
2004 SF Crime Rate Heatmap
2005 SF Crime Rate Heatmap
2006 SF Crime Rate Heatmap
2007 SF Crime Rate Heatmap
2008 SF Crime Rate Heatmap
2009 SF Crime Rate Heatmap
2010 SF Crime Rate Heatmap
2011 SF Crime Rate Heatmap
2012 SF Crime Rate Heatmap
2013 SF Crime Rate Heatmap
2014 SF Crime Rate Heatmap
2015 SF Crime Rate Heatmap
2016 SF Crime Rate Heatmap
2017 SF Crime Rate Heatmap
Latitude
37.82
37.80
37.78
37.76
37.74
37.72
Longitude
−122.50122.48122.46122.44122.42122.40122.38
10^0
10^1
10^2
10^3

Figure 3: 2003 - 2017 Crime Rate Heat map

* Crime rate data for 2017 is not fully completed from Jan-Dec

There is clearly a reoccurring trend. There is a major hot spot around the Tenderloin area and the hot spot is consistent for all 15 years. We also see that there is a high crime rate running along the Bart station that leads use from South San Francisco into the heart of SF. Starting off in 2003, we can see that the hot spot is clearly having a wider spread. As we progress through the years this hot spot became tamer. However, starting in 2011, we can see that the crime rate starts to ramp up once again.

Is crime really increasing in 2011 to 2016?

Heat maps helps us identify the bigger picture and point out any trends or frequency of crimes rates for a particular year. The question to ask is how can we verify that crime is indeed decreasing from 2003 to 2011 and then increasing from 2011 to 2016? A more statistical approach would be to formulate the data into a line plot.

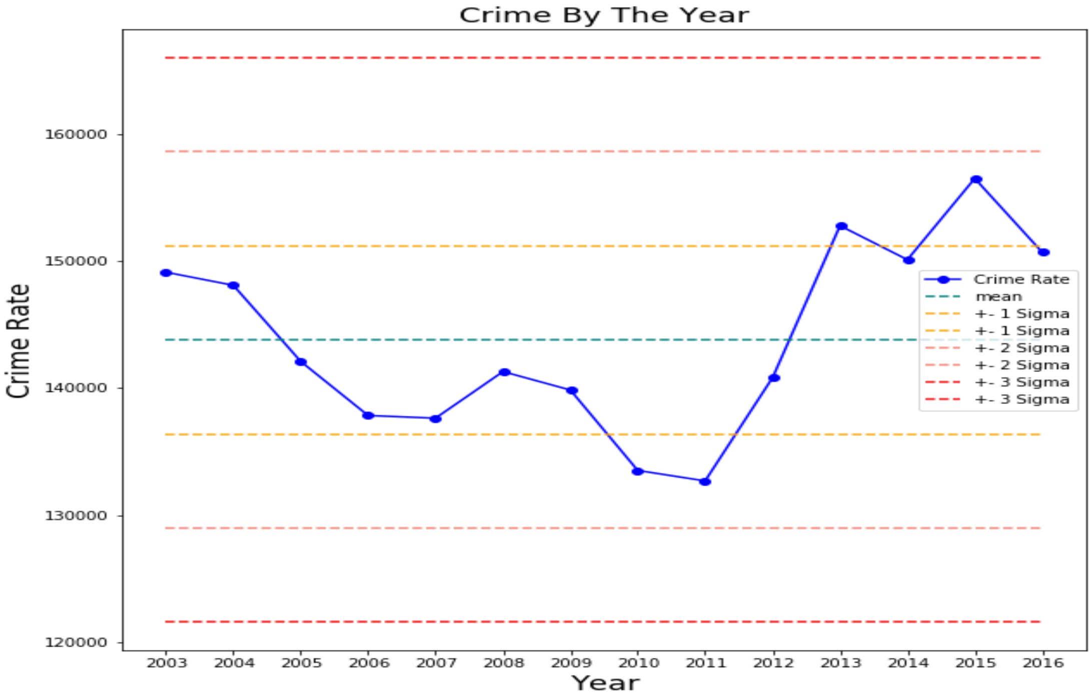
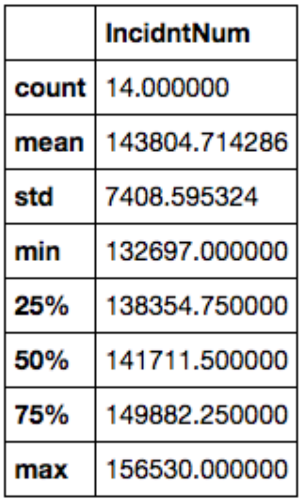


Figure 4: Crime Rate vs. Year

A line plot helps us map out the year and the amount of crimes reported for that particular year. On the plot above, we have the year on our X axis and the incident count on the Y axis. There are three main lines within our plot which consist of the crime rate line (blue), mean crime rate (teal, and multiple sigma limit lines. A numerical representation of the graph can be found below:



	IncidntNum
count	14.000000
mean	143804.714286
std	7408.595324
min	132697.000000
25%	138354.750000
50%	141711.500000
75%	149882.250000
max	156530.000000

Figure 5: Crime Rate vs. Year Descriptive Statistic

Talk about descriptive statistic and the error margins

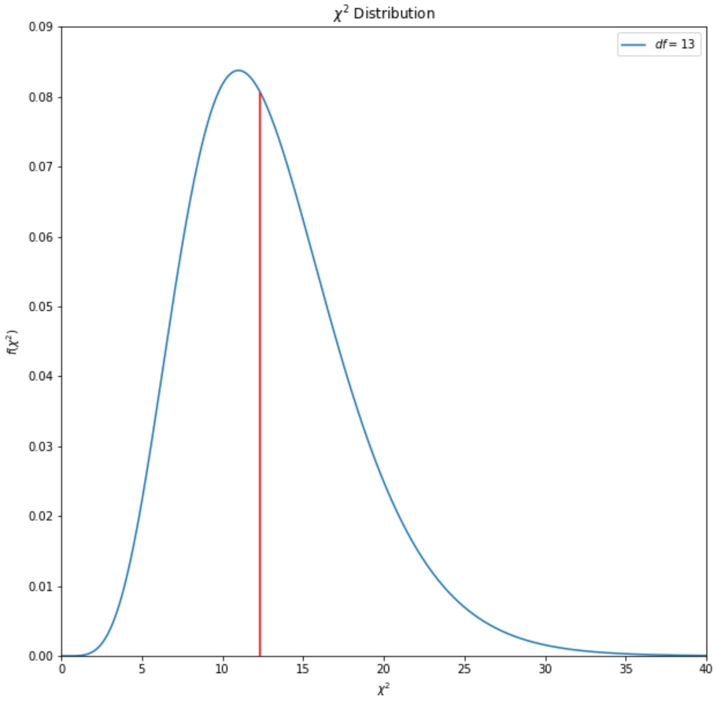
How can we statistically prove that crime is increasing or decreasing?

For use to know if crime is decreasing or increasing, we will use a chi square test to evaluate our observed data, which is the data that we used to graph the plot against our expected value, which will be our mean. We will test if there is any difference between our observed and expected value. In order to do this, we will need to come up with a hypothesis.

H_0: Crime rate is constant
H_1: Crime rate is not constant and is decreasing
a=.05
n=14
DF=13

Figure 6: Hypothesis Testing Variables

For our expected value, we will be using a mean crime rate value from 2003 to 2016. This gives us a flat line across our plot at 143805 crimes per year. Also, we want to note that we will be using an alpha level of .05 for our hypothesis testing. We also have 14 observations as well.



For a Chi2 distribution with a degree of freedom consisting of 13 and level of significance of .05, we have a critical value of 12.340 (Chi2 Table). If we find that our Chi2 value falls to the right of the rejection zone then we can reject the null hypothesis and conclude that crime is indeed reducing. In order to find the Chi2 test statistic, we will be using the equation below:

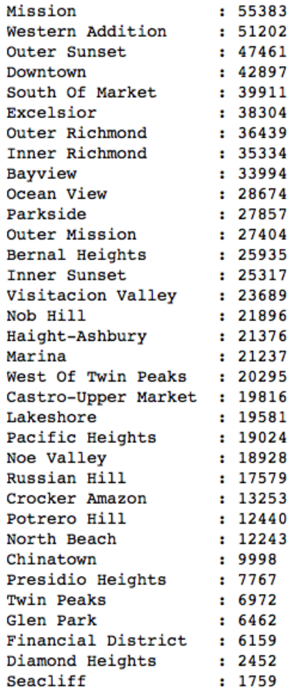
$$X^2 = \sum \left[\frac{(Obsereved - Expected)}{Expected} \right]^2$$

Figure 7: Chi-Squared Equation

This equation states that we must find the difference of the observed value and the expected value squared, then dividing that with the expected value and summing them all up. We can see that our Chi2 test statistic value is .0345, which is much less than our critical value of 12.340. Since, our value is less, then it must mean that it is to the left of our Chi2 distribution graph. Since it falls outside the rejection zone, we can conclude that we don't have enough evidence, so we fail to reject the null hypothesis, which states that the mean crime rate is 14,3805.

San Francisco District Population

There's 10 police district in San Francisco consisting of, Richmond, Taraval, Park, Northern, Central, Tenderloin, Southern, Mission, Ingleside, and Bayview. In order for me to find the population in each district, I first had to find out what part of the city belongs in what district. After doing so, the next biggest issue was to find data that had such information. San Francisco has roughly 800,000 residences living there and is disburse all throughout the city. I was able to find a website called Statistical Atlas, which contain population data for the district. In order for me to extract the information I had to use a few of python's request library to scrape the website.



```
Mission              : 55383
Western Addition     : 51202
Outer Sunset         : 47461
Downtown             : 42897
South Of Market      : 39911
Excelsior            : 38304
Outer Richmond       : 36439
Inner Richmond       : 35334
Bayview              : 33994
Ocean View           : 28674
Parkside             : 27857
Outer Mission        : 27404
Bernal Heights       : 25935
Inner Sunset         : 25317
Visitacion Valley    : 23689
Nob Hill             : 21896
Haight-Ashbury       : 21376
Marina               : 21237
West Of Twin Peaks   : 20295
Castro-Upper Market  : 19816
Lakeshore            : 19581
Pacific Heights      : 19024
Noe Valley           : 18928
Russian Hill         : 17579
Crocker Amazon       : 13253
Potrero Hill         : 12440
North Beach          : 12243
Chinatown            : 9998
Presidio Heights     : 7767
Twin Peaks           : 6972
Glen Park            : 6462
Financial District   : 6159
Diamond Heights      : 2452
Seacliff             : 1759
```

Figure 8: District Population

What’s the probability of crimes per district?

The list above shows us the population size for each part of the city. We will need to parse the list apart and store them in their appropriate district to find the population size for that district. Knowing this information will help us determine the probability of crime rate per district.

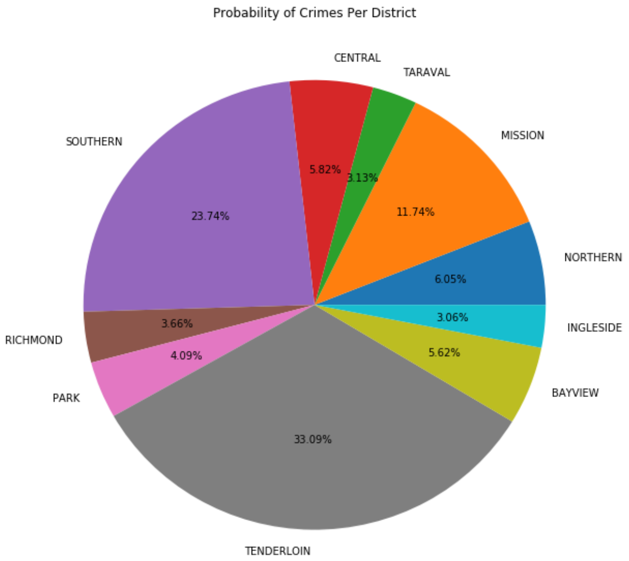
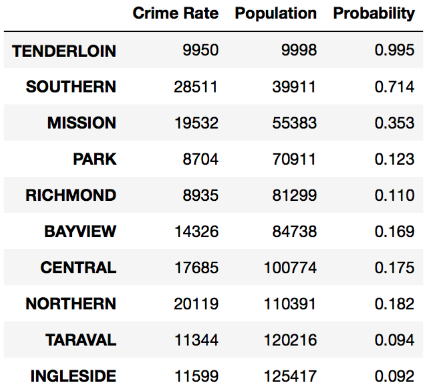


Figure 9: Probability of Crime Per District



	Crime Rate	Population	Probability
TENDERLOIN	9950	9998	0.995
SOUTHERN	28511	39911	0.714
MISSION	19532	55383	0.353
PARK	8704	70911	0.123
RICHMOND	8935	81299	0.110
BAYVIEW	14326	84738	0.169
CENTRAL	17685	100774	0.175
NORTHERN	20119	110391	0.182
TARAVAL	11344	120216	0.094
INGLESIDE	11599	125417	0.092

Figure 10: Probability of Crime Per District Table

After being combine the data and normalizing it. We can see that even though Ingleside has the highest population, there isn't as much crime happening in that area. The probability for crime to actually happen is roughly 9%. However, something to point out and notice is the probability for Tenderloin and Southern District. These two district has the smallest amount of population, but highest probability of crimes happening. Tenderloin comes in at an amazing 99%, while Southern District follows behind at 71%.

How has the crime rate changed per district between 2003 to 2017?

From the analysis above, we can determine which district experience the highest probability of crimes occurring per person. With this in mind, are crimes actually changing in those districts? Is crime decreasing or is it actually increasing? By knowing this major point, we will know what district needs be more secured and have more resources to help decrease the crime rate.