**Battle Against Crime in San Francisco**

1. Introduction & Problem  
     
   San Francisco crime statistics report an overall upward trend in crime based on data from 17 years with violent crime increasing and property crime increasing. Based on this trend, the crime rate in San Francisco for 2019 is expected to be higher than in 2016.

The city violent crime rate for San Francisco in 2016 was higher than the [national violent crime rate](https://www.cityrating.com/crime-statistics/) average by 78.95% and the city property crime rate in San Francisco was higher than the [national property crime rate](https://www.cityrating.com/crime-statistics/) average by 122.03%.  
  
In 2016 the city violent crime rate in San Francisco was higher than the [violent crime rate in California](https://www.cityrating.com/crime-statistics/california/) by 59.55% and the city property crime rate in San Francisco was higher than the [property crime rate in California](https://www.cityrating.com/crime-statistics/california/) by 113.13%. This is as per the report published on <https://www.cityrating.com/crime-statistics/california/san-francisco.html>.

1. Interest  
     
   The Police Department of San Francisco would be very much interested in what the trend was during 2018 so that they can focus on the top crime categories and the top neighborhoods in 2018 with the most incidents and make action plans to address the problem. They could also can create task forces to battle crime for the top neighborhoods in crime.
2. Data Sources  
     
   I have used the public data available at https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783 for the crime data of 2018.
3. Data Pre-processing  
     
   The dataset was then filtered to gather data for 2018 alone.
4. Feature Selection  
    Only the useful columns were selected into a new data frame that I could use for analysis. These are :  
    'analysis\_neighborhood','incident\_category', 'incident\_code',

'incident\_date', 'incident\_datetime', 'incident\_day\_of\_week',

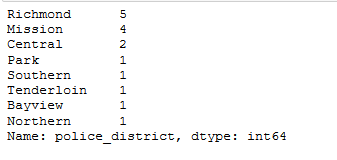
'incident\_description', 'incident\_id', 'incident\_number',

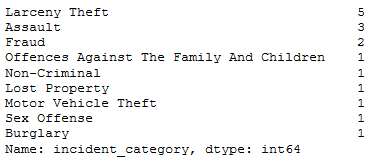
'incident\_subcategory', 'incident\_time', 'incident\_year',

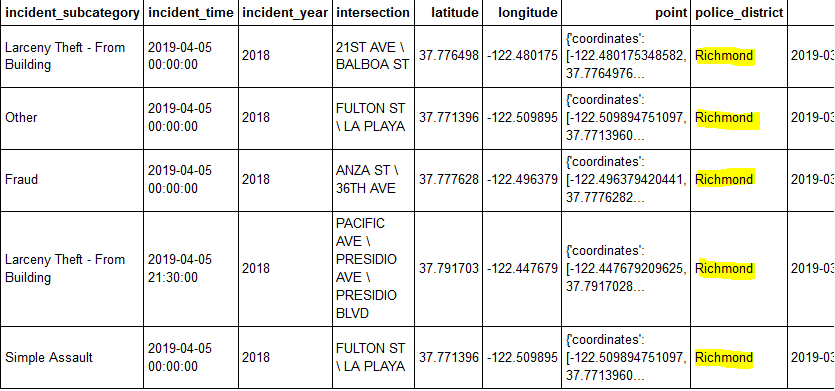
'intersection', 'latitude', 'longitude', 'point', 'police\_district',

'report\_datetime', 'report\_type\_code', 'report\_type\_description',

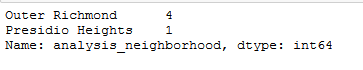
'resolution', 'row\_id', 'supervisor\_district'  
  
 Columns with NaN values in analysis\_neighborhood, latitude and longitude columns were also removed.

1. Data Analysis  
     
   Next, I looked up police district has the most incidents. The Richmond Police District shows to be on the top.  
   

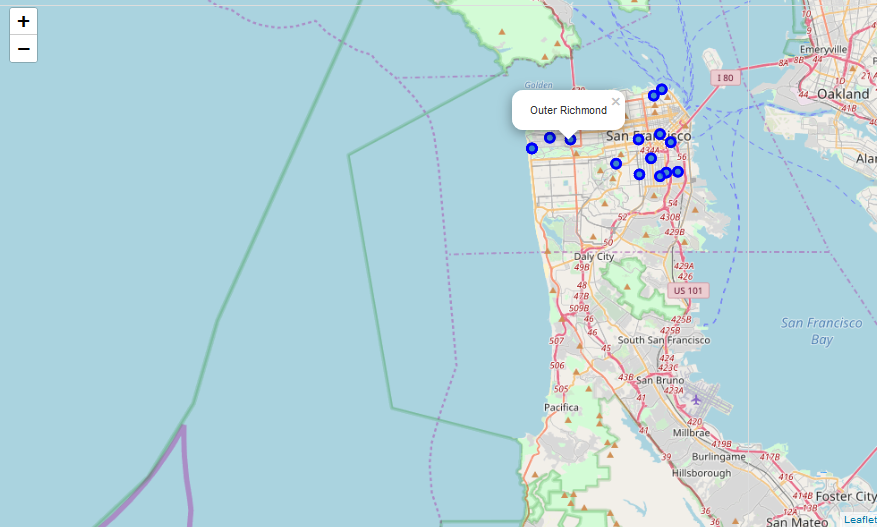
Also, looked up which incident\_category had the most occurrences. Larceny Theft came on top.  


Looked up data for all neighborhoods in the police district of Richmond by creating a new data frame for Richmond police district alone.  


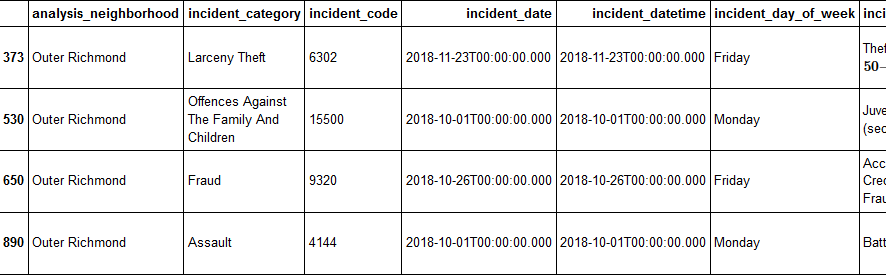
Next step was to find out which neighborhood in the Richmond police district had the most incidents. Outer Richmond neighborhood was on top.

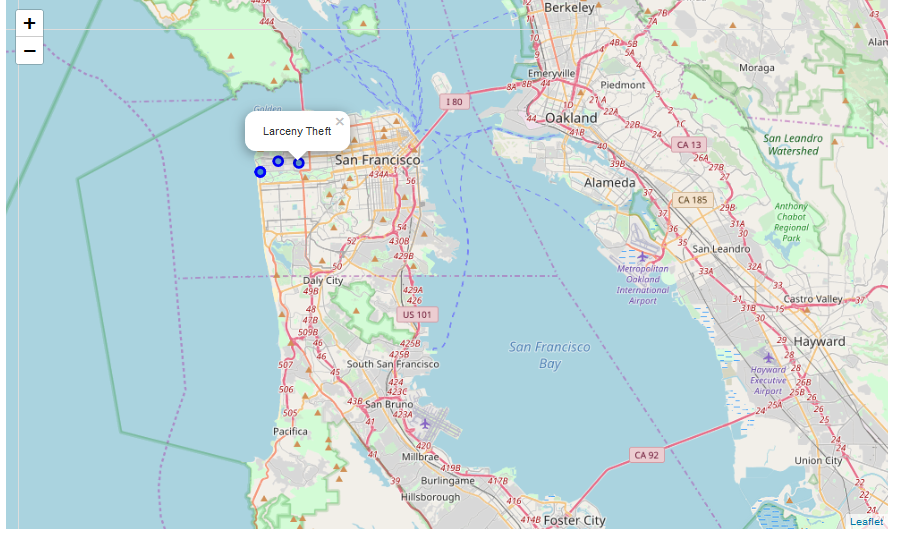


Mapped San Francisco showing all the analysis neighborhoods on it listed on the Report for 2018 using Folium to visualize.

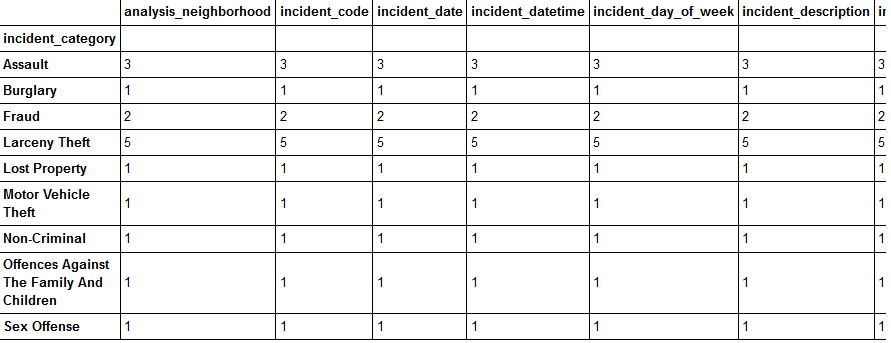


Also gathered data for incident categories in Outer Richmond and used Folium to visualize them.

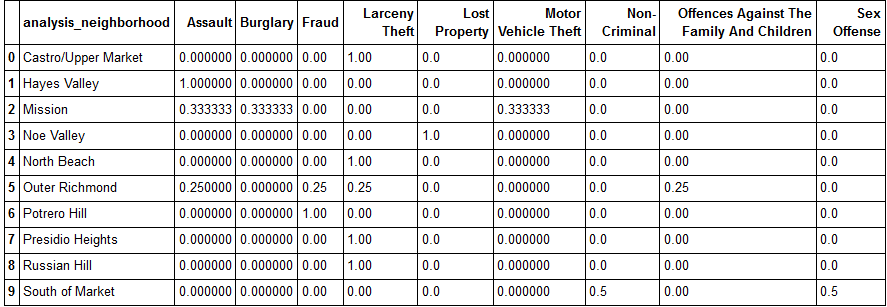




Now that we knew which neighborhood and crime category came on top for 2018 next step was to get a count of the various types of incident categories in the City of San Francisco for all neighborhoods for clustering.



Grouped rows by neighborhood and by taking the mean of the frequency of occurrence of each category



Pulled info on each neighborhood along with the top 5 most common incidents:

----Castro/Upper Market----

Incident freq

0 Larceny Theft 1.0

1 Assault 0.0

2 Burglary 0.0

3 Fraud 0.0

4 Lost Property 0.0

----Hayes Valley----

Incident freq

0 Assault 1.0

1 Burglary 0.0

2 Fraud 0.0

3 Larceny Theft 0.0

4 Lost Property 0.0

----Mission----

Incident freq

0 Assault 0.33

1 Burglary 0.33

2 Motor Vehicle Theft 0.33

3 Fraud 0.00

4 Larceny Theft 0.00

----Noe Valley----

Incident freq

0 Lost Property 1.0

1 Assault 0.0

2 Burglary 0.0

3 Fraud 0.0

4 Larceny Theft 0.0

----North Beach----

Incident freq

0 Larceny Theft 1.0

1 Assault 0.0

2 Burglary 0.0

3 Fraud 0.0

4 Lost Property 0.0

----Outer Richmond----

Incident freq

0 Assault 0.25

1 Fraud 0.25

2 Larceny Theft 0.25

3 Offences Against The Family And Children 0.25

4 Burglary 0.00

----Potrero Hill----

Incident freq

0 Fraud 1.0

1 Assault 0.0

2 Burglary 0.0

3 Larceny Theft 0.0

4 Lost Property 0.0

----Presidio Heights----

Incident freq

0 Larceny Theft 1.0

1 Assault 0.0

2 Burglary 0.0

3 Fraud 0.0

4 Lost Property 0.0

----Russian Hill----

Incident freq

0 Larceny Theft 1.0

1 Assault 0.0

2 Burglary 0.0

3 Fraud 0.0

4 Lost Property 0.0

----South of Market----

Incident freq

0 Non-Criminal 0.5

1 Sex Offense 0.5

2 Assault 0.0

3 Burglary 0.0

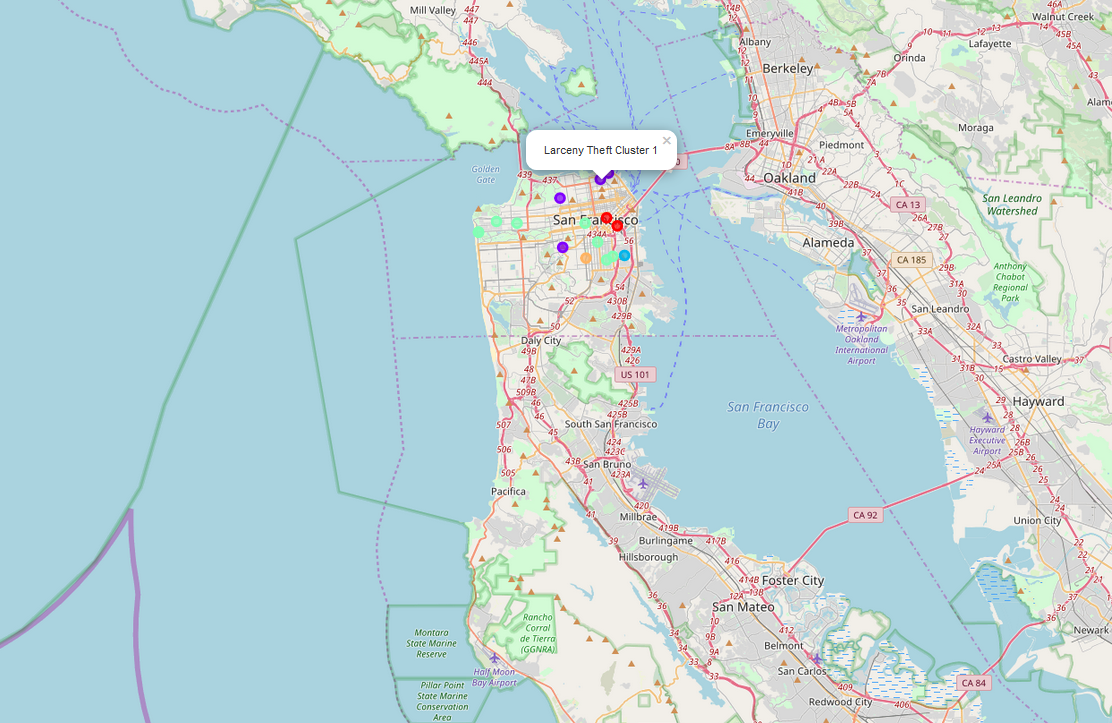
4 Fraud 0.0

Next created a new data frame that displayed the top 5 incidents for each neighborhood:



To create clusters showing various clusters of the incident categories, ran k-means to cluster the neighborhood into 5 clusters and created a new data frame that includes the cluster as well as the top 10 incidents for each neighborhood.

Finally visualized the various clusters that show the various incident categories using Folium Map.



**The Purple cluster shows for Larceny Theft which has the most occurrences in San Francisco. Also the top neighborhood for it is Outer Richmond in the Richmond police district of San Francisco as per this data analysis.**