# Investigation of San Francisco Crime Data 2003-2019

By: Garrett Yamane

# Problem Statement

- **Target Client**: San Francisco Police Department

- **Goal**: Identify important crime features and use them to develop unsupervised machine learning models for suggesting patterns that contribute to crimes in San Francisco

*https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwihgoXMy9rlAhV8HzQIHQvBANcQjRx6BAgBEAQ&url=https%3A%2F%2Fwww.aarp.org%2Ftravel%2Fdestinations%2Funited-states%2Fsan-francisco%2F&psig=AOvVaw2j0Ncqi3xC617wHsUwMGMy&ust=1573301668804179

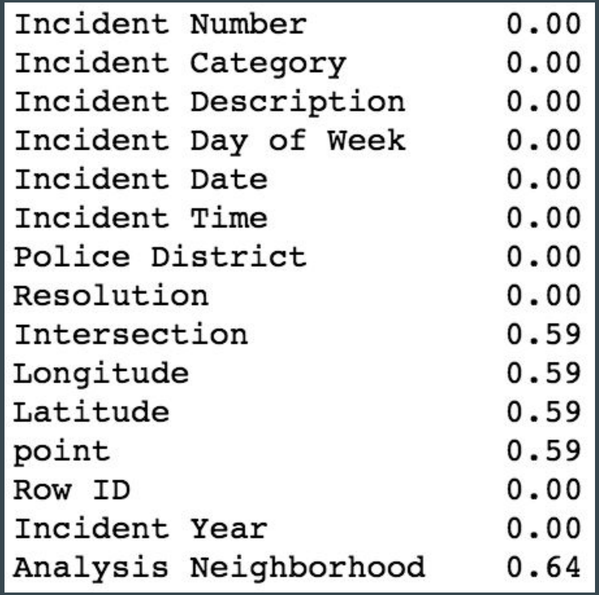
# Data Acquisition and Wrangling

- **Sources**:
  - CABLE mainframe 2003-2018 police reports
  - Crime Data Warehouse 2018-2019 police reports (update in 2018)

- **Final Data Frame**:
  - 2,415,298 rows (*independent crime reports*)
  - 15 crime feature columns

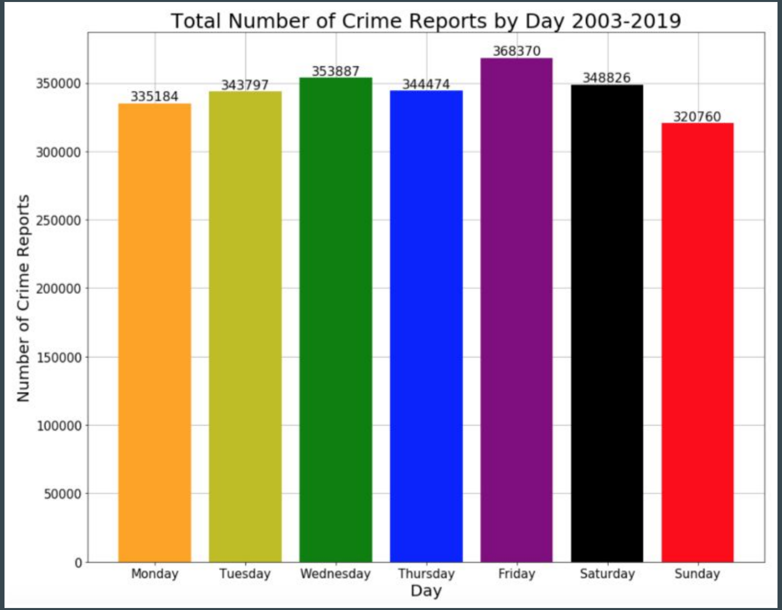
- **2018**: update made from CABLE mainframe to Crime Data Warehouse to make crime data more accessible

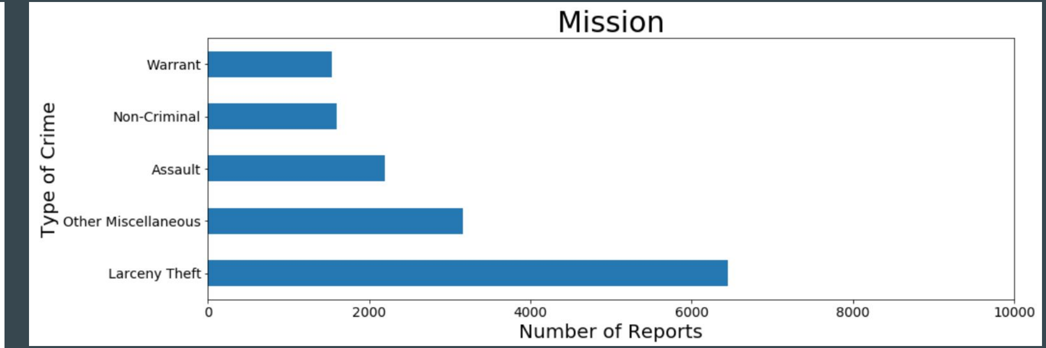
Crime Features for Final Data Frame
(% missing values)



```
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
Incident Year             0.00
Analysis Neighborhood     0.64
```

# Initial Findings

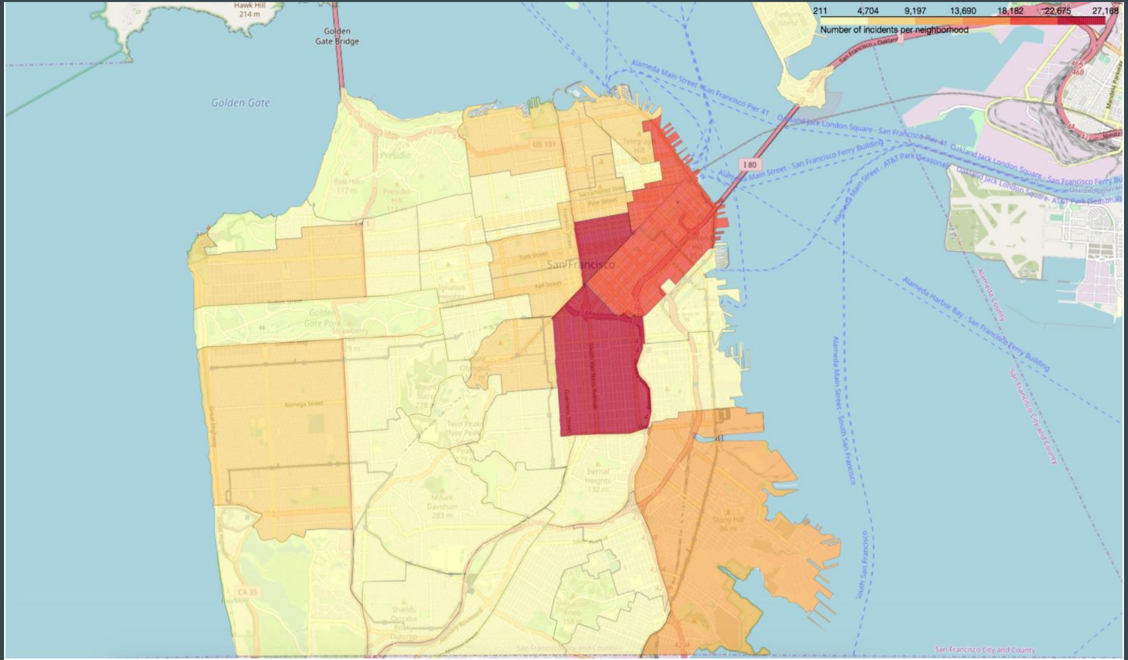


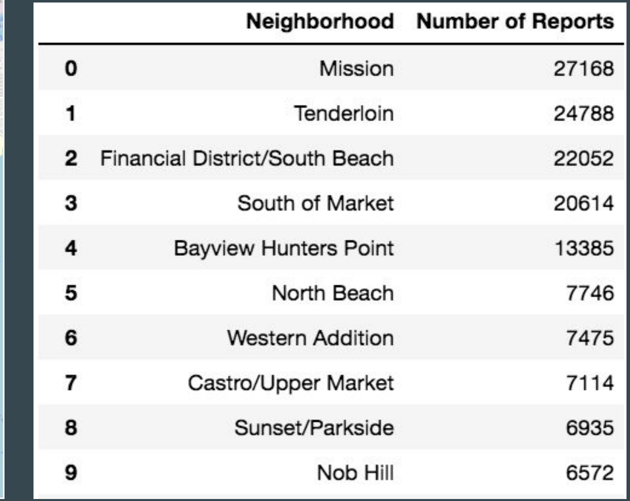


## Questions for Data Set

- Yearly crime rate change?
- What types of crimes are reported the most?
- Are crimes more likely to happen on certain days? Months? Times of the day?
- What neighborhoods are most dangerous?

# 2018-2019 Choropleth Map of Crimes per Neighborhood



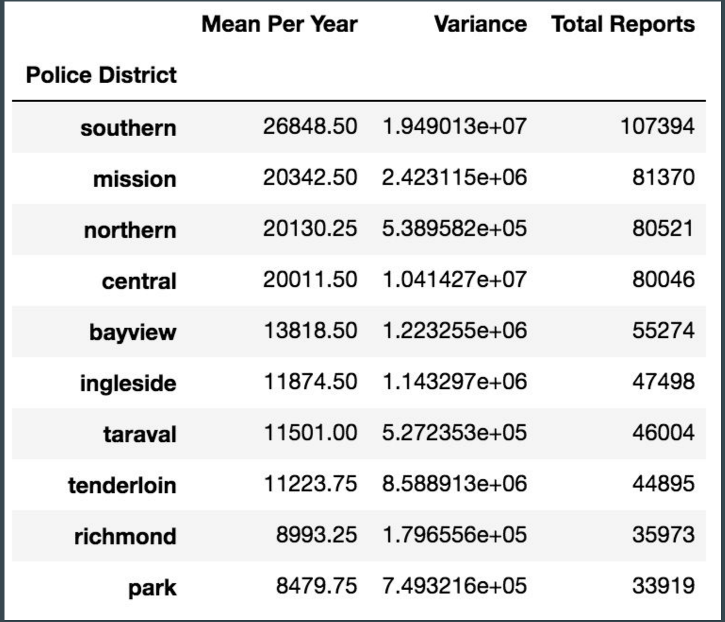


| | Neighborhood | Number of Reports |
|---|---|---|
| **0** | Mission | 27168 |
| **1** | Tenderloin | 24788 |
| **2** | Financial District/South Beach | 22052 |
| **3** | South of Market | 20614 |
| **4** | Bayview Hunters Point | 13385 |
| **5** | North Beach | 7746 |
| **6** | Western Addition |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

# Application of Inferential Statistics

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

- $H_0$: 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

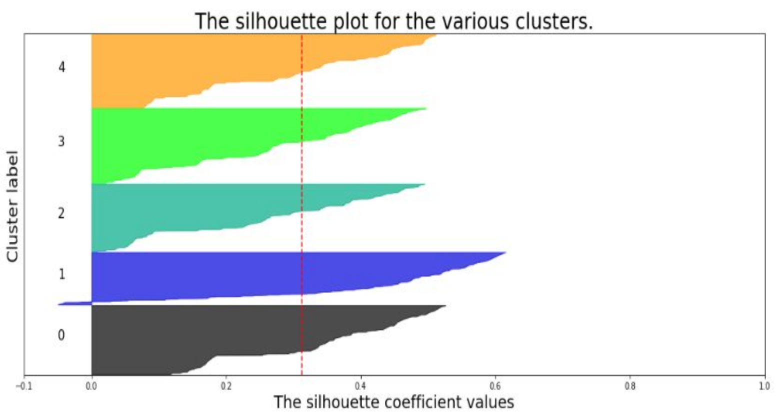


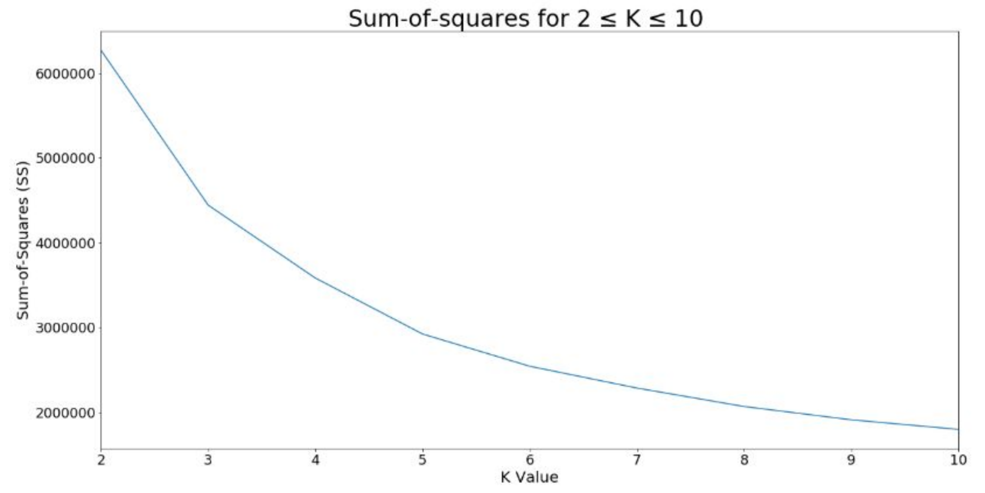
| | Mean Per Year | Variance | Total Reports |
|---|---|---|---|
| **Police District** | | | |
| **southern** | 26848.50 | 1.949013e+07 | 107394 |
| **mission** | 20342.50 | 2.423115e+06 | 81370 |
| **northern** | 20130.25 | 5.389582e+05 | 80521 |
| **central** | 20011.50 | 1.041427e+07 | 80046 |
| **bayview** | 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 |

# Application of Inferential Statistics: Feature Selection

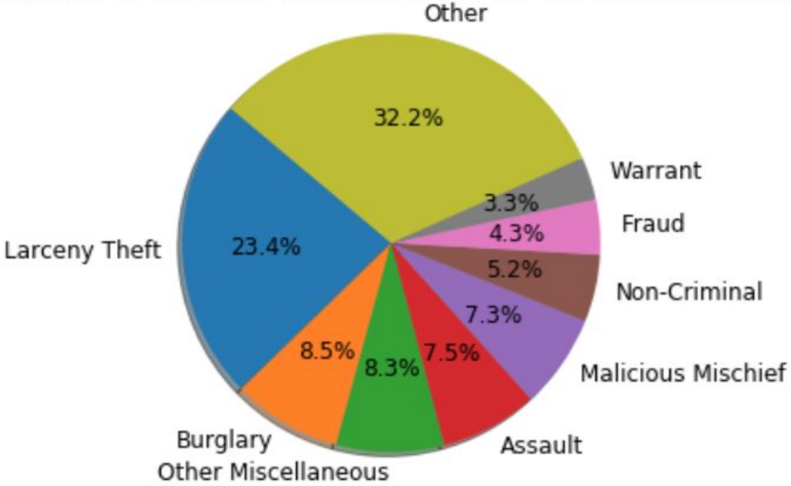
- **Chi-Square Test**
  - **Target Feature:** *Incident Category*
  - **Test Features:**
    - *Incident Hour*
    - *Incident Day of Week*
    - *Incident Month*
    - *Police District*
    - *Resolution*
- **Goal**
  - What features are useful determining the type of crime to be committed?

# Baseline Clustering Model: K-Means

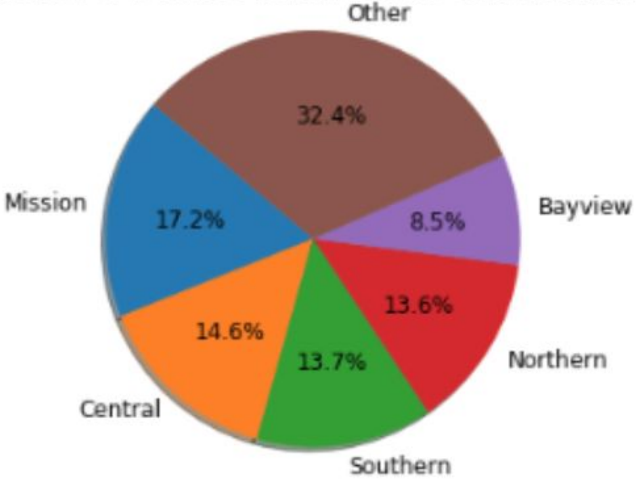




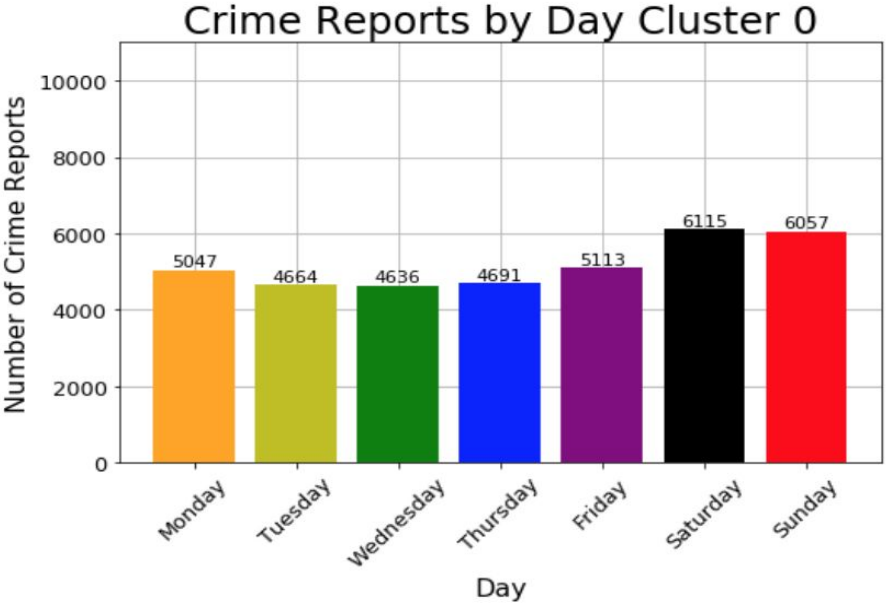
**Baseline k-means model built with k = 5*
**5 features used to cluster: Incident Hour, Incident Day of Week, Incident Month, Police District, Incident Category*



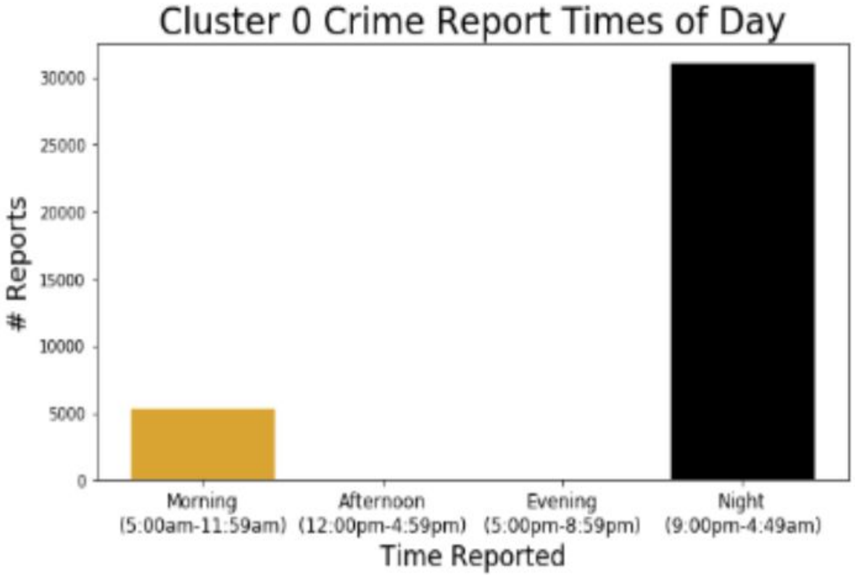
Cluster 0 Crime Classification % Distribution



Cluster 0 Police District % Distribution

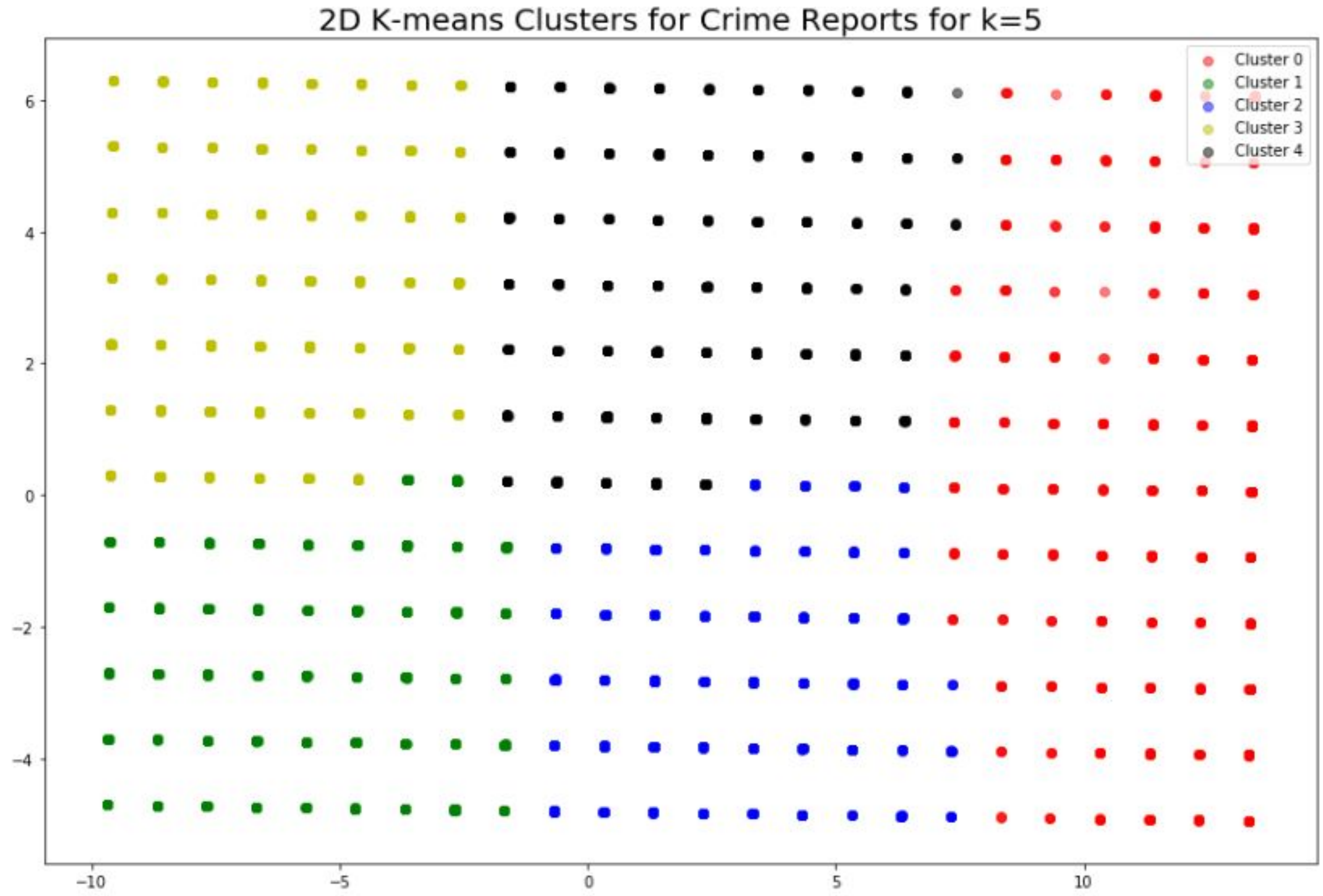


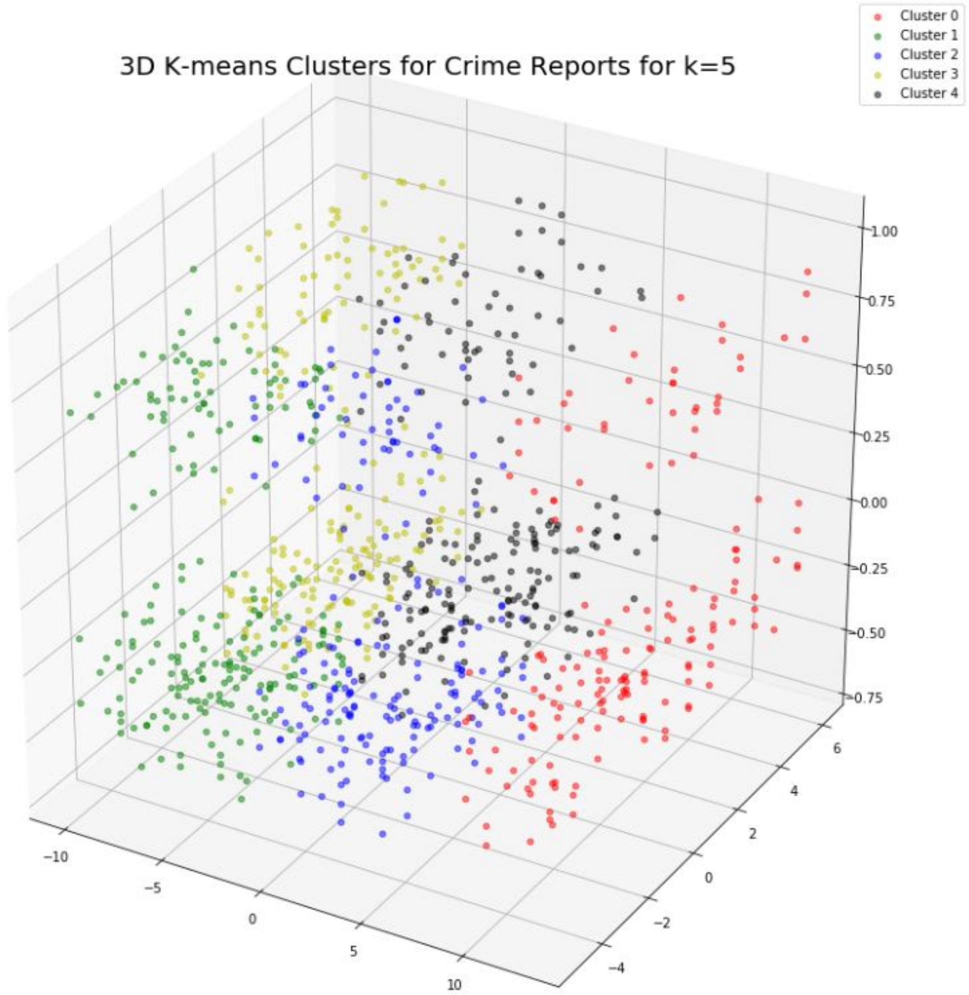
Crime Reports by Day Cluster 0



Cluster 0 Crime Report Times of Day

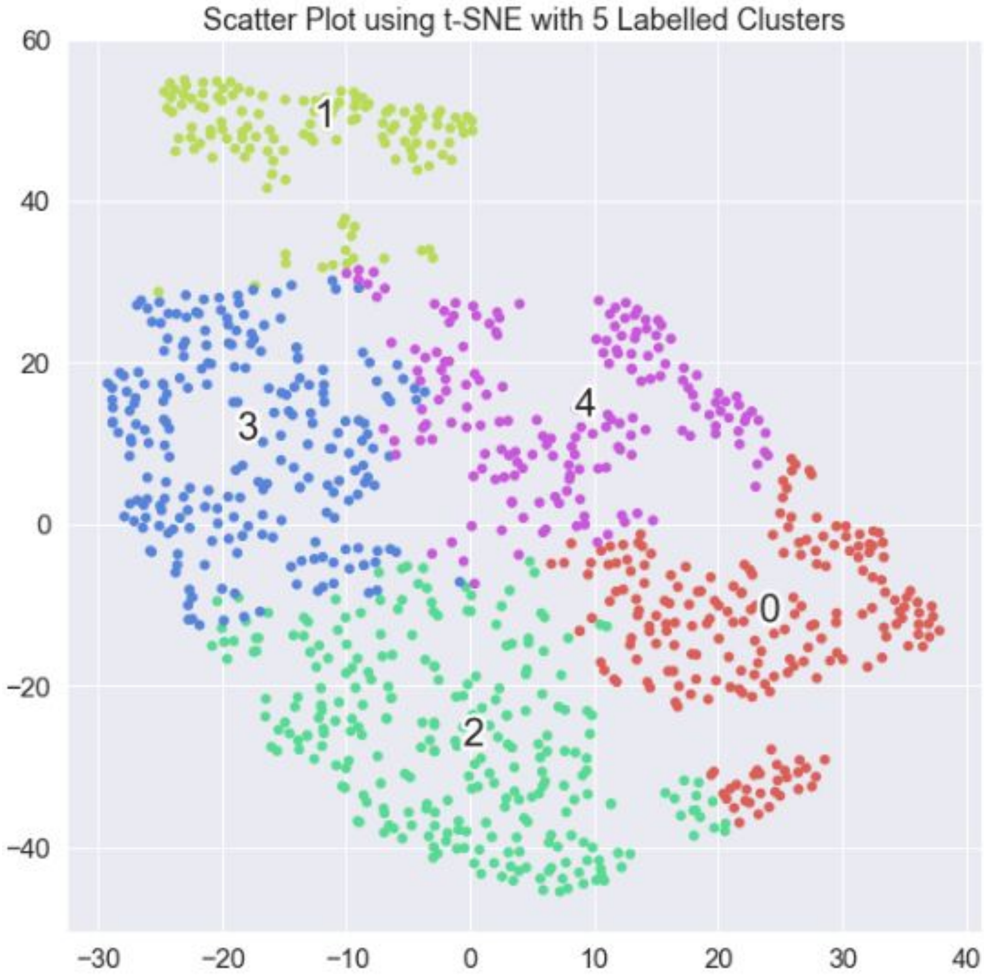
# Visualizing K-Means Clusters: PCA


2D K-means Clusters for Crime Reports for k=5
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
6
4
2
0
−2
−4
−10
−5
0
5
10



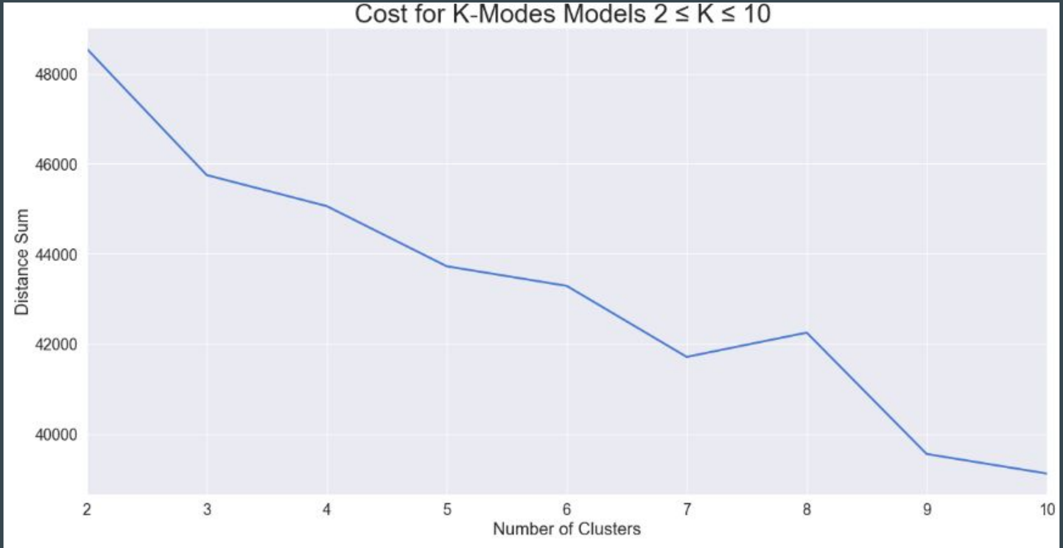
3D K-means Clusters for Crime Reports for k=5
● Cluster 0
● Cluster 1
● Cluster 2
● Cluster 3
● Cluster 4
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−10
−5
0
5
10
6
4
2
0
−2
−4

# Model Extension: Application of t-distributed stochastic neighbor embedding (t-SNE)


Scatter Plot using t-SNE with 5 Labelled Clusters
60
40
20
0
−20
−40
−30
−20
−10
0
10
20
30
40
1
3
4
0
2

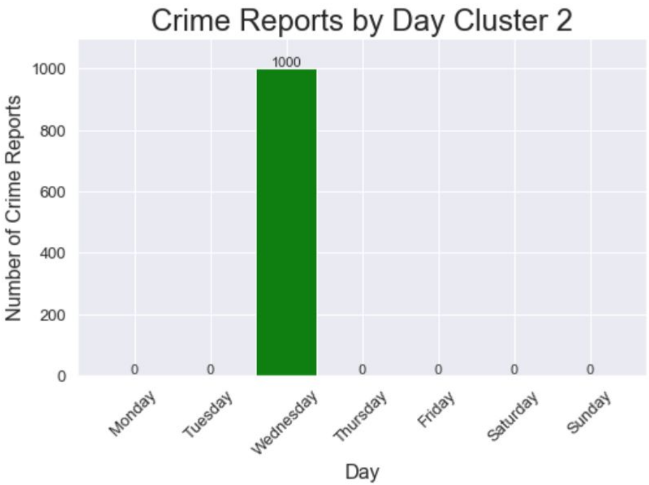
# Model Extension: K-Modes

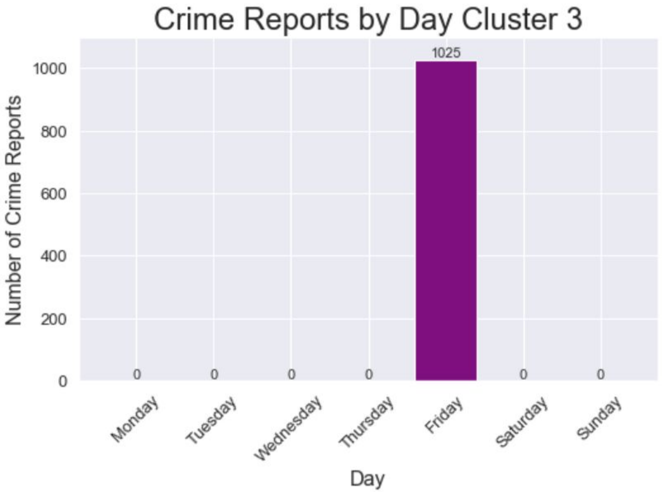
- Better for handling categorical data
- Distance Function: dissimilarity between objects rather than Euclidean distance

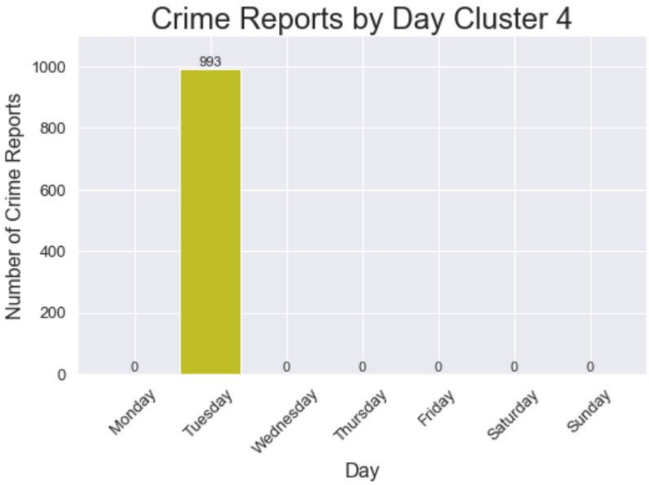


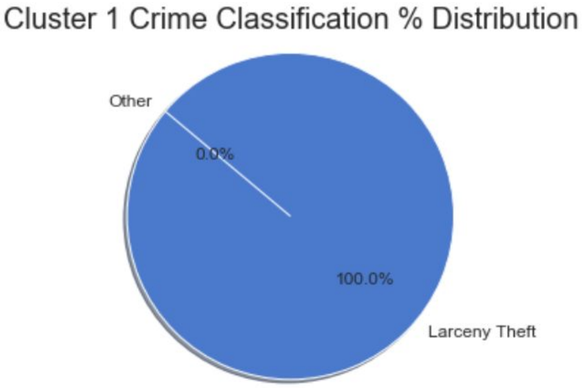
**Used same features to cluster as k-means model*

# K-Modes: Key Differences


Crime Reports by Day Cluster 2


Crime Reports by Day Cluster 3


Crime Reports by Day Cluster 4


Cluster 1 Crime Classification % Distribution

# Conclusions and Future Work

**- Big takeaways**

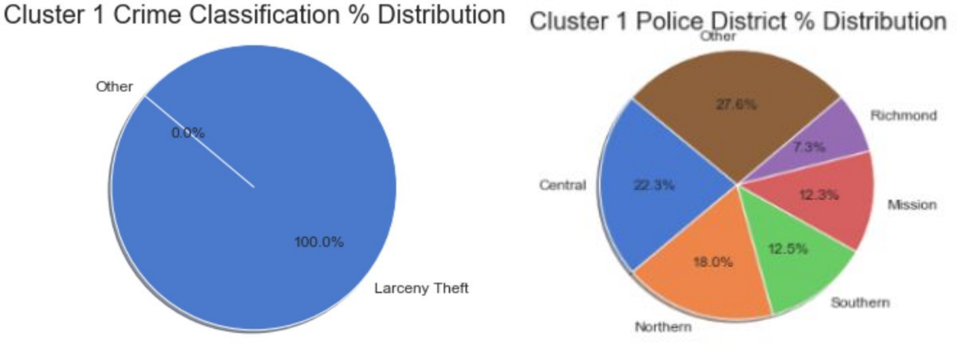
- No single algorithm can reveal all the underlying patterns of my crime data
- K-modes specifically targets categorical data -> stronger representation of SF crime data

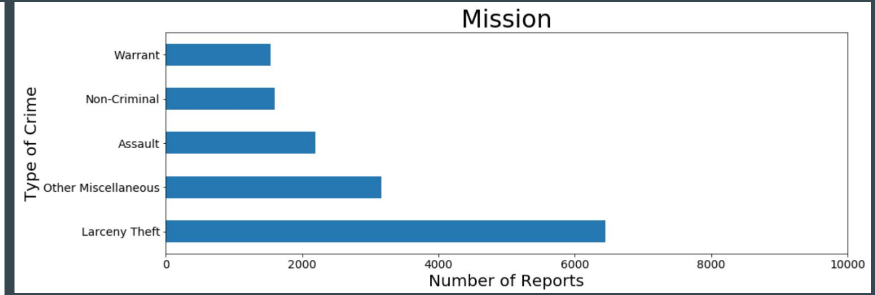
**- Future Work**

- Configure own crime classifications rather than dealing with every unique category
- Supervised Learning Perspective
    - Predict the type of crime that will happen given certain conditions (time of day, location, etc.)

# Recommendations to Client

- "Larceny Theft" by far the most reported crime
  - Large impact on crimes that could be connected to each other
- Target police districts such as Central and Northern with high priority
- Allocate more resources to districts where crimes are most likely may help reduce and prevent future crimes


Cluster 1 Crime Classification % Distribution
Other
Larceny Theft
Cluster 1 Police District % Distribution
Other
Richmond
Central
Mission
Southern
Northern


Mission
Type of Crime
Warrant
Non-Criminal
Assault
Other Miscellaneous
Larceny Theft
Number of Reports