Project Proposal

2020 Vision

Mihir Patel, Tina Xia, Leah Okamura, Kyra Cooperman

Section 1

The subject matter we're investigating is information about crime in San Francisco. In recent years, San Francisco hasn't been the safest place to live; the overall crime rate in San Francisco is 151% higher than the national average. According to SFChronicle, "homicides increased by 21.4% in San Francisco from March to June of this year," compared to 2019 (https://www.sfchronicle.com/bayarea/article/Which-crimes-are-up-down-in-SF-during-15408485.php). There is a 1 in 15 chance of becoming a victim of any crime. We wanted to use this dataset to obtain conclusions about specific factors that correlate to higher levels of crime, which will hopefully inform us of some key insights we can keep during future travels.

Research Question: What factors can the general population associate with local crime in order to be the safest while in San Francisco?

Hypotheses: A later time (e.g. nighttime hours) correlates to a higher level or rate of crime. Location is correlated to levels of crime.

We are interested in these two hypotheses because we believe they can then lead to other interesting relationships between variables within this dataset. For example, if there is a strong correlation between night and rate of crime, then is there a correlation between which night of the week (ex. Sunday night) and rate of crime? With location, are there certain districts that have a specific crime that is common there? By delving further and examining these relationships, we will be able to understand if crime has any specific pattern in San Francisco.

Section 2

The observations in the dataset are of crime data in San Francisco from 2016. We found our dataset at https://www.kaggle.com/roshansharma/sanfranciso-crime-dataset. Each observation in this datase is a crime whose various aspects have been recorded. There were originally 150,500 individual crimes/observations in this dataset. However, because of the nature of R Studio through OIT, we will be taking a random and reproducible sample from the larger dataset. We created this sample by using the function sample_n() on sanfrancrimeBIG to randomly select 15,000 observations. We chose 15,000 because it is still large enough to get an accurate portrayal of the total data set, yet is much more manageable to process.

There are 13 variables in the dataset.

IncidntNum (double): gives the Incident Number of the crime

Category (character): gives category of crime

Description (character): gives description of crime

DayofWeek (character): gives day of week the crime occurred on

Date (character): gives date (day, month, and year) of crime

Time (double): gives time of crime (in military time)

PdDistrict (character): gives police district crime occurred in

Resolution (character): gives kind of punishment given to the criminal to resolve the case

Address (character): gives address where the crime happened

X (double): gives latitude of crime location

Y (double): gives longitude of crime location

Location (character): exact location using latitude and longitude

PdId (double): ID of police officer

The curator of the dataset got it from the final assignment for Coursera and IBM's Data Visualization Course. The information in this dataset is most likely directly from the San Francisco Police Department for their reported crimes during 2016. This dataset was originally used to practice analyzing and visualizing data through geo spatial mapping by using folium maps for geographical understanding.

Section 3

```
sanfrancrime <- sanfrancrimeBIG%>%
  sample_n(15000)
  glimpse(sanfrancrime)
```

```
## Rows: 15,000
## Columns: 13
## $ IncidntNum <chr> "160915165", "160707558", "166035351", "160474133", "160...
## $ Category   <chr> "BURGLARY", "VANDALISM", "LARCENY/THEFT", "ASSAULT", "OT...
## $ Descript   <chr> "BURGLARY OF STORE, FORCIBLE ENTRY", "MALICIOUS MISCHIEF...
## $ DayOfWeek  <chr> "Thursday", "Wednesday", "Wednesday", "Saturday", "Wedne...
## $ Date       <chr> "11/10/2016 12:00:00 AM", "08/31/2016 12:00:00 AM", "02/...
## $ Time       <time> 03:54:00, 20:15:00, 15:00:00, 01:45:00, 14:34:00, 09:17...
## $ PdDistrict <chr> "TARAVAL", "RICHMOND", "SOUTHERN", "SOUTHERN", "SOUTHERN...
## $ Resolution <chr> "NONE", "NONE", "NONE", "ARREST, BOOKED", "ARREST, BOOKE...
## $ Address    <chr> "1100 Block of OCEAN AV", "500 Block of 7TH AV", "200 Bl...
## $ X          <dbl> -122.4547, -122.4652, -122.3656, -122.4090, -122.4034, -...
## $ Y          <dbl> 37.72349, 37.77809, 37.80967, 37.78113, 37.77542, 37.782...
## $ Location   <chr> "(37.7234871655198, -122.454662311927)", "(37.7780927956...
## $ PdId       <dbl> 1.609152e+13, 1.607076e+13, 1.660354e+13, 1.604741e+13, ...
```