Project Proposal

2020 Vision

Mihir Patel, Tina Xia, Leah Okamura, Kyra Cooperman

Section 1

The subject matter we're investigating is information about the crime in San Francisco. We're planning on analyzing the data and visualizing it to draw conclusions. Dataset: https://www.kaggle.com/roshansharma/sanfranciso-crime-dataset

Research Question: What factors will allow the general population of San Francisco to be safest from local crime?

Motivation: 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.

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

Section 2

In this section, you will describe the data set you wish to explore. This includes a description of the observations in the data set, a general description of the variables in the data set, and a description of how the data was originally collected (not how you found the data but how the original curator of the data collected it).

The observations in the dataset are of crime data in San Francisco from 2016. Each observation is a crime whose various aspects have been recorded. All observations are not missing any data. **only one observation is missing a recording, the other ones have everything**

There are 12 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): gives the ID of the police officer? **not sure what exactly what this variable is**

The curator of the dataset got it from the final assignment for Coursera and IBM's Data Visualization Course. I am not sure of where exactly how the course creators created it. **not sure where it came from**

Section 3

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