Project Draft

2020 Vision

Mihir Patel, Tina Xia, Leah Okamura, Kyra Cooperman

Introduction and data and methodology

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 (or in other cities with similar characteristics)?

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.

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.

introduce and justify the statistical method(s) that you believe will be useful in answering your research question.

The statistical methods we believe will be useful in answering our research question include the CLT, simulated null distributions, bootstrapping, etc. We're using these because:

Grouping violence based on violent vs nonviolent.

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

```
## Rows: 15,000
## Columns: 13
## $ IncidntNum <chr> "160201243", "160443683", "161062428", "160187994", "160...
## $ Category   <chr> "LARCENY/THEFT", "LARCENY/THEFT", "FRAUD", "NON-CRIMINAL...
## $ Descript   <chr> "GRAND THEFT FROM PERSON", "GRAND THEFT FROM LOCKED AUTO...
## $ DayOfWeek  <chr> "Tuesday", "Monday", "Saturday", "Friday", "Sunday", "Mo...
## $ Date       <chr> "03/08/2016 12:00:00 AM", "05/30/2016 12:00:00 AM", "12/...
## $ Time       <time> 22:00:00, 21:00:00, 12:05:00, 02:30:00, 18:03:00, 07:43...
## $ PdDistrict <chr> "SOUTHERN", "INGLESIDE", "MISSION", "CENTRAL", "NORTHERN...
## $ Resolution <chr> "NONE", "NONE", "ARREST, BOOKED", "NONE", "NONE", "ARRES...
## $ Address    <chr> "900 Block of MISSION ST", "100 Block of BROOKDALE AV", ...
## $ X          <dbl> -122.4081, -122.4209, -122.4067, -122.4076, -122.4347, -...
## $ Y          <dbl> 37.78157, 37.71209, 37.75534, 37.78834, 37.78885, 37.781...
## $ Location   <chr> "(37.7815668300024, -122.40805253847)", "(37.71209206756...
## $ PdId       <dbl> 1.602012e+13, 1.604437e+13, 1.610624e+13, 1.601880e+13, ...
```

Variables we're considering: Category Day of Week Date Time PdDistrict Resolution

Results

Showcase how you arrived at answers to your question using any techniques we have learned in this class (and some beyond, if you're feeling adventurous). Provide the main results from your analysis. The goal is not to do an exhaustive data analysis (i.e., do not calculate every statistic and procedure you have learned for every variable), but rather let me know that you are proficient at asking meaningful questions and answering them with results of data analysis, that you are proficient in using R, and that you are proficient at interpreting and presenting the results. Focus on methods that help you begin to answer your research questions.

Turn time into a set of ranges.

Research Question: What factors can the general population associate with local crime in order to be the safest while in San Francisco (or in other cities with similar characteristics)?

Relationship between category and time? Mihir Which PD has the highest proportion of violent crime? Kyra Relationship between time and crime? Tina Day of the week and category? Leah

ADD etc

```
sanfrancrime%>%
  group_by(Category)%>%
  count()
```

```
## # A tibble: 39 x 2
## # Groups:   Category [39]
##    Category                      n
##    <chr>                     <int>
##  1 ARSON                        31
##  2 ASSAULT                    1344
##  3 BAD CHECKS                    3
##  4 BRIBERY                       9
##  5 BURGLARY                    576
##  6 DISORDERLY CONDUCT           72
##  7 DRIVING UNDER THE INFLUENCE  39
##  8 DRUG/NARCOTIC               407
##  9 DRUNKENNESS                  43
## 10 EMBEZZLEMENT                 11
## # ... with 29 more rows
```

```
sanfrancrime<- sanfrancrime%>%
  mutate(violent_crime = case_when(
    Category == "ASSAULT" | Category == "SEX OFFENSES FORCIBLE" |
      Category == "ROBBERY" | Category == "KIDNAPPING" ~ "YES",
    Category != "ASSAULT" | Category != "SEX OFFENSES, FORCIBLE" |
      Category != "ROBBERY" | Category !="KIDNAPPING" ~ "NO"))
sanfrancrime%>%
  group_by(PdDistrict)%>%
  mutate(yes_violent = violent_crime == "YES")%>%
  count(yes_violent)%>%
  mutate(perc = (n/sum(n) *100))%>%
  filter(yes_violent == TRUE)%>%
  arrange(desc(perc))
```

```
## # A tibble: 10 x 4
## # Groups:   PdDistrict [10]
##    PdDistrict yes_violent     n  perc
##    <chr>      <lgl>       <int> <dbl>
##  1 TENDERLOIN TRUE          152 15.5
##  2 BAYVIEW    TRUE          220 15.4
##  3 INGLESIDE  TRUE          162 14.3
##  4 MISSION    TRUE          250 12.9
##  5 NORTHERN   TRUE          214 10.6
##  6 SOUTHERN   TRUE          293 10.5
##  7 TARAVAL    TRUE          110  9.94
##  8 CENTRAL    TRUE          168  9.43
##  9 PARK       TRUE           65  7.42
## 10 RICHMOND   TRUE           59  6.18
```

Discussion

This section is a conclusion and discussion. This will require a summary of what you have learned about your research question along with statistical arguments supporting your conclusions. Also, critique your own methods and provide suggestions for improving your analysis. Issues pertaining to the reliability and validity of your data and appropriateness of the statistical analysis should also be discussed here. A paragraph on

what you would do differently if you were able to start over with the project or what you would do next if you were going to continue work on the project should also be included.