Staying Safe: Analyzing Crime in San Francisco

2020 Vision

Mihir Patel, Tina Xia, Leah Okamura, Kyra Cooperman

INTRODUCTION AND DATA

San Francisco is a city known for its strong economy and booming tech industry. In addition to Silicon Valley and San Jose, the Bay Area is home to many powerful companies such as Google, Tesla, Apple, and Cisco. Because of these many benefits, San Francisco is a popular destination for college graduates. In May 2020, San Francisco was ranked second as the best metro area for recent graduates. This especially took into consideration the "high wages, work from home ability, and a (mainly) pandemic-resilient economy" that many recent graduates worry about during this time [1].

However, with an overall crime rate in San Francisco that is 151% higher than the national average, is it also important to note that in recent years, San Francisco has not been the safest place to live. The SFChronicle reported that compared to 2019, "homicides increased by 21.4% in San Francisco from March to June of this year,"[2]. There is a 1 in 15 chance of becoming a victim of any type of crime. A quick search about travel in San Francisco includes many articles listing the "Places to Avoid After Dark" or "Most Dangerous Neighborhoods in SF."With the a high possibility of any of us moving to San Francisco after our time at Duke,and the recent popularity with college graduates, we wanted to analyze this dataset to obtain conclusions about specific factors that correlate to higher levels of crime, which will could then inform us of some key insights we can keep during future travels or moves.

Through our research, we plan to investigate what factors the general population can associate with local crime in order to be the safest while in San Francisco. Our main hypotheses are 1) a later time (e.g. nighttime hours) correlates to a higher level or rate of crime and 2) Location is correlated to levels of crime. We believe it is important to investigate this question because there likely are policy changes that can be implemented to increase safety throughout the city. Our investigation will shine light on potential patterns of crime.

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.

In order to assess these hypotheses, we will look at the following relationships: 1. Relationship between crime type and time 2. Relationship between crime and time 3. Relationship between violent crimes and police district 4. Relationship between days of the week and crime

```
## Rows: 15,000
## Columns: 13
## $ IncidntNum <chr> "160074818", "166163532", "160697272", "160666750", "160...
## $ Category   <chr> "ASSAULT", "LARCENY/THEFT", "NON-CRIMINAL", "NON-CRIMINA...
## $ Descript   <chr> "THREATS AGAINST LIFE", "GRAND THEFT FROM LOCKED AUTO", ...
## $ DayOfWeek  <chr> "Tuesday", "Wednesday", "Sunday", "Tuesday", "Wednesday"...
## $ Date       <chr> "01/26/2016 12:00:00 AM", "06/15/2016 12:00:00 AM", "08/...
## $ Time       <time> 13:45:00, 08:06:00, 12:55:00, 16:00:00, 06:30:00, 15:55...
## $ PdDistrict <chr> "NORTHERN", "BAYVIEW", "SOUTHERN", "CENTRAL", "NORTHERN"...
## $ Resolution <chr> "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", ...
## $ Address    <chr> "FRANKLIN ST / PACIFIC AV", "CESAR CHAVEZ ST / ILLINOIS ...
```

```
## $ X          <dbl> -122.4249, -122.3866, -122.4136, -122.4065, -122.4197, -...
## $ Y          <dbl> 37.79461, 37.75033, 37.77951, 37.79515, 37.78967, 37.719...
## $ Location   <chr> "(37.7946072650051, -122.424873688619)", "(37.7503255046...
## $ PdId       <dbl> 1.600748e+13, 1.661635e+13, 1.606973e+13, 1.606668e+13, ...
```

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 dataset 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.

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.

METHODOLOGY

Variables

We will analyze the validity of our hypotheses using various statistical methods, including a Chi-square test, bootstrapping, and a logistic regression model, among others. Note: we plan on grouping violence based on violent vs nonviolent. The main variables we will be using in our analysis are Category, DayOfWeek, Date, Time, PdDistrict, and Resolution. We also created new variables to assist us in our data. This includes the variable timerange, that organizes the hour of the day into four times of day "night", "morning", "day", and "evening."

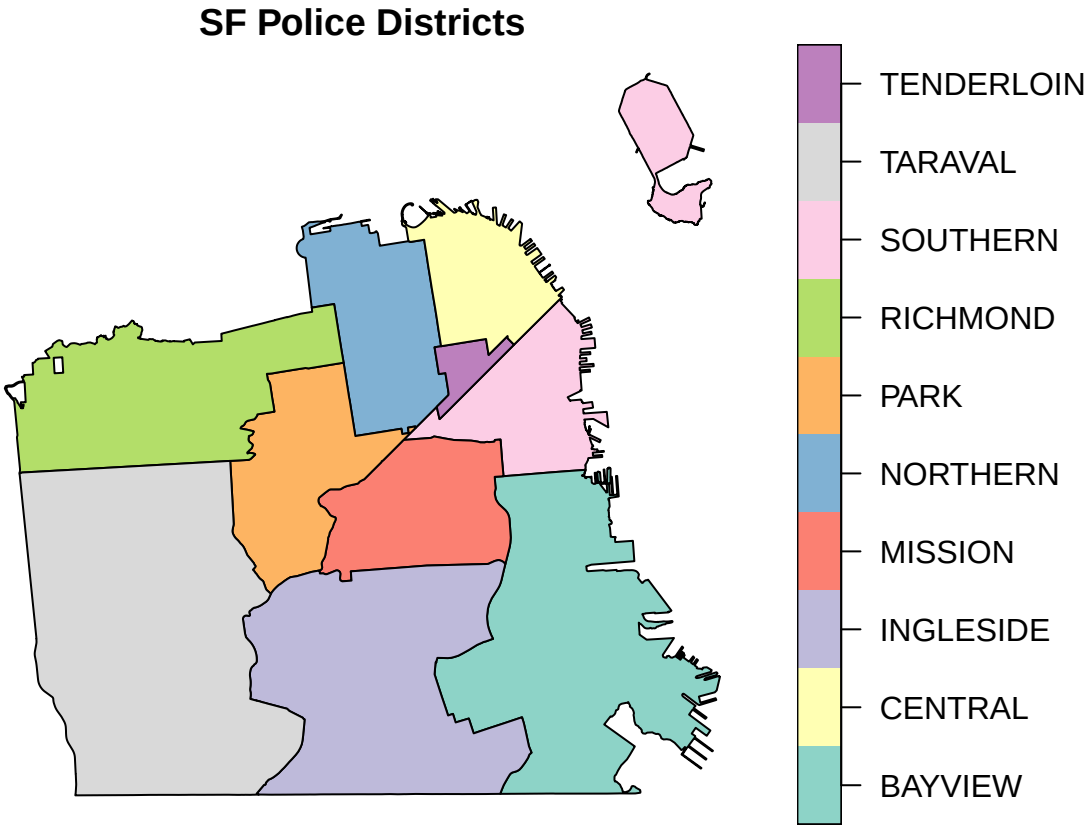
We also decided to categorize the all of the different types of crime that were reported. We organized the 39 types of crimes into varable crimetype, which consists of "Property", "Violent", "White Collar", "Drug/Alcohol", "Sex", "Suspicious", "Legal Violation", and "Miscellaneous".

```
## Rows: 15,000
## Columns: 8
## $ Category   <chr> "ASSAULT", "LARCENY/THEFT", "NON-CRIMINAL", "NON-CRIMINA...
## $ DayOfWeek  <chr> "Tuesday", "Wednesday", "Sunday", "Tuesday", "Wednesday"...
## $ Date       <chr> "01/26/2016 12:00:00 AM", "06/15/2016 12:00:00 AM", "08/...
## $ PdDistrict <chr> "NORTHERN", "BAYVIEW", "SOUTHERN", "CENTRAL", "NORTHERN"...
## $ Resolution <chr> "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", ...
## $ hour       <dbl> 13, 8, 12, 16, 6, 15, 8, 11, 22, 22, 23, 14, 0, 0, 2, 6,...
## $ timerange  <chr> "afternoon", "morning", "afternoon", "afternoon", "morni...
## $ crimetype  <chr> "Violent", "Property", "Miscellaneous", "Miscellaneous",...
```

Visualizations

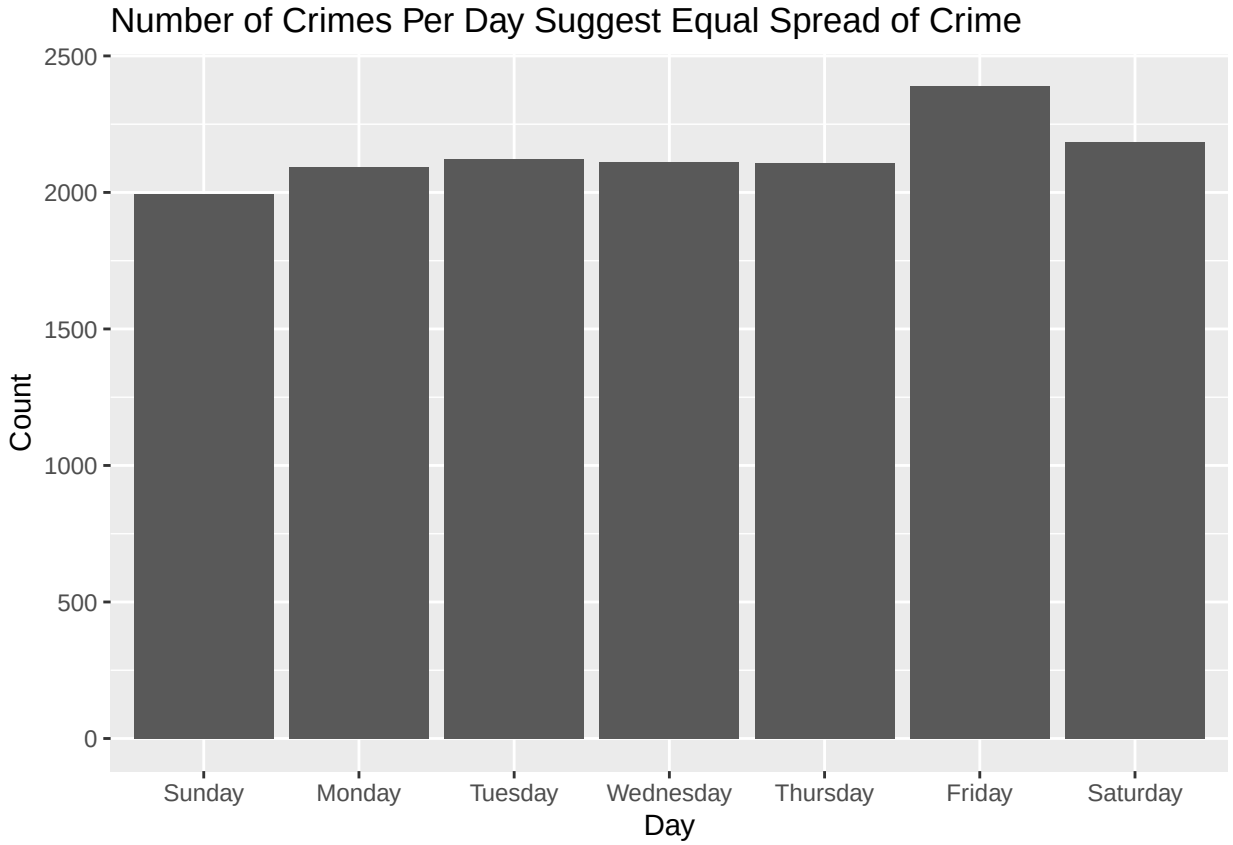
```
## Simple feature collection with 10 features and 5 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -122.5139 ymin: 37.7081 xmax: -122.357 ymax: 37.83329
## geographic CRS: WGS84(DD)
##    company   district shape_area shape_le_1 shape_leng
## 1        B   SOUTHERN   91344142  100231.35   87550.28
## 2        C    BAYVIEW  201384622  144143.48  163013.80
## 3        D    MISSION   80623840   40518.83   40152.78
## 4        E   NORTHERN   82781686   50608.31   56493.86
## 5        J TENDERLOIN   11072155   18796.78   12424.27
```

```
## 6        A    CENTRAL   55950269   67686.52   64025.13
## 7        F       PARK   84878956   50328.91   46307.78
## 8        G   RICHMOND  137964024   75188.63   69991.47
## 9        H  INGLESIDE  193580502   74474.18   74737.94
## 10       I    TARAVAL  284676678   73470.42   75350.22
##                          geometry
## 1  MULTIPOLYGON (((-122.3919 3...
## 2  MULTIPOLYGON (((-122.381 37...
## 3  MULTIPOLYGON (((-122.4095 3...
## 4  MULTIPOLYGON (((-122.4338 3...
## 5  MULTIPOLYGON (((-122.4022 3...
## 6  MULTIPOLYGON (((-122.4261 3...
## 7  MULTIPOLYGON (((-122.4396 3...
## 8  MULTIPOLYGON (((-122.4413 3...
## 9  MULTIPOLYGON (((-122.4045 3...
## 10 MULTIPOLYGON (((-122.4984 3...
```



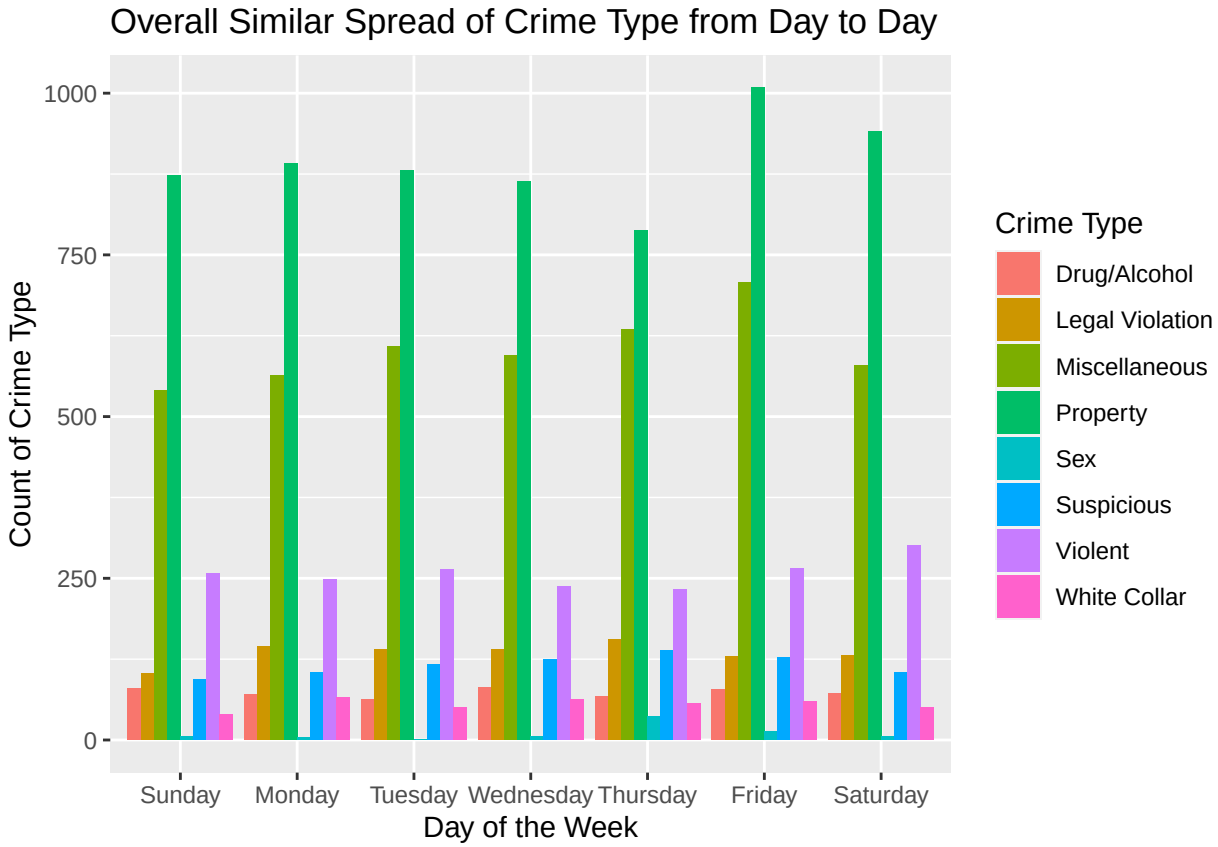

```
## # A tibble: 15,000 x 2
## # Groups:   DayOfWeek [7]
##    DayOfWeek cpday
##    <fct>     <int>
##  1 Tuesday    2124
##  2 Wednesday  2110
##  3 Sunday     1993
##  4 Tuesday    2124
##  5 Wednesday  2110
##  6 Monday     2093
##  7 Monday     2093
```

```
##  8 Thursday   2108
##  9 Wednesday  2110
## 10 Friday     2388
## # ... with 14,990 more rows
```



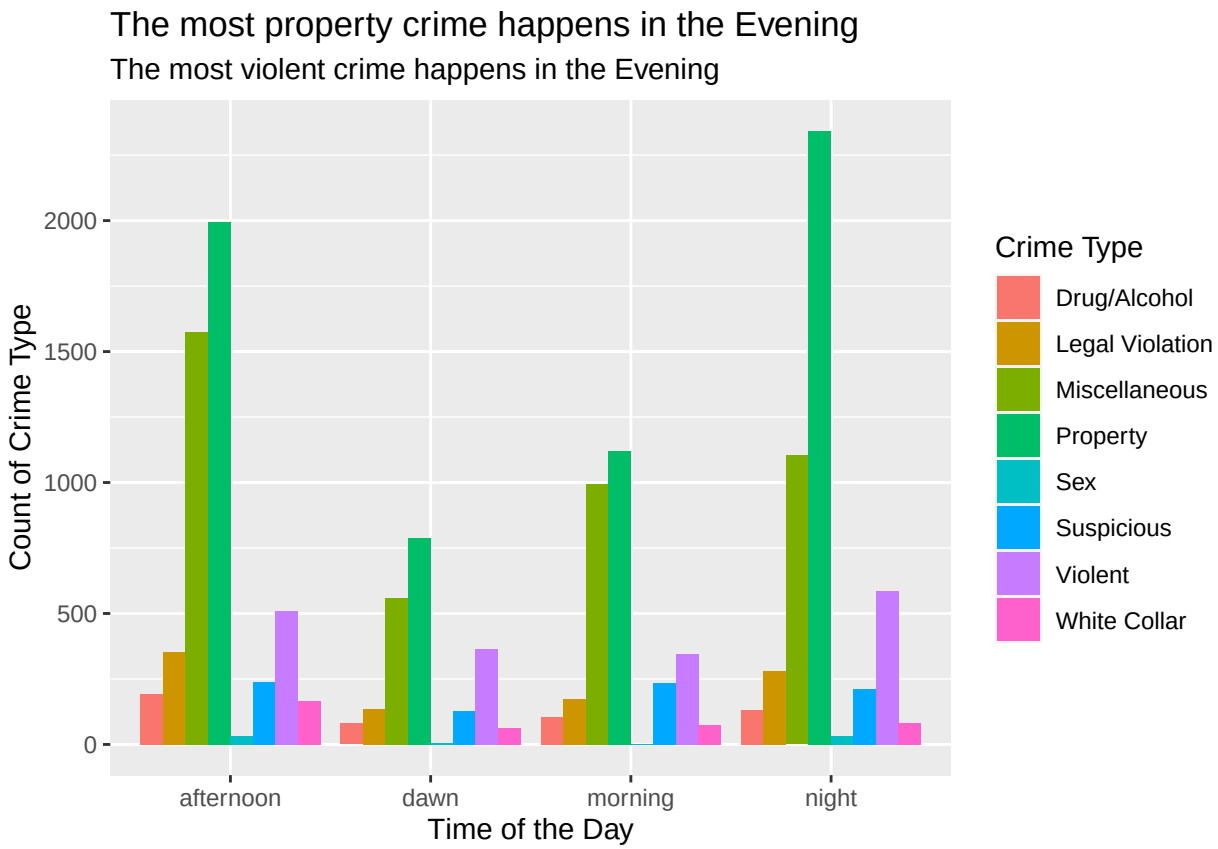
One relationship we were interested in was if certain days had a higher rates of crime. We visualized this relationship by creating a bar graph that compares the day of the week and number of crimes each day during this time period. By looking at the visual, we are able to see that each has a relatively similar crime count compared to the other. In addition to this, there is no significant pattern that sticks out as well.

```
## # A tibble: 15,000 x 9
## # Groups:   crimetype [8]
##    Category DayOfWeek Date  PdDistrict Resolution  hour timerange crimetype
##    <chr>    <fct>     <chr> <chr>      <chr>      <dbl> <chr>     <chr>
##  1 ASSAULT  Tuesday   01/2~ NORTHERN   NONE          13 afternoon Violent
##  2 LARCENY~ Wednesday 06/1~ BAYVIEW    NONE           8 morning   Property
##  3 NON-CRI~ Sunday    08/2~ SOUTHERN   NONE          12 afternoon Miscella~
##  4 NON-CRI~ Tuesday   08/1~ CENTRAL    NONE          16 afternoon Miscella~
##  5 NON-CRI~ Wednesday 02/0~ NORTHERN   NONE           6 morning   Miscella~
##  6 ROBBERY  Monday    03/2~ INGLESIDE  NONE          15 afternoon Violent
##  7 NON-CRI~ Monday    10/1~ SOUTHERN   NONE           8 morning   Miscella~
##  8 NON-CRI~ Thursday  02/0~ SOUTHERN   NONE          11 morning   Miscella~
##  9 WARRANTS Wednesday 05/0~ NORTHERN   ARREST, B~    22 night     Legal Vi~
## 10 VEHICLE~ Friday    04/0~ INGLESIDE  NONE          22 night     Property
## # ... with 14,990 more rows, and 1 more variable: ctcount <int>
```



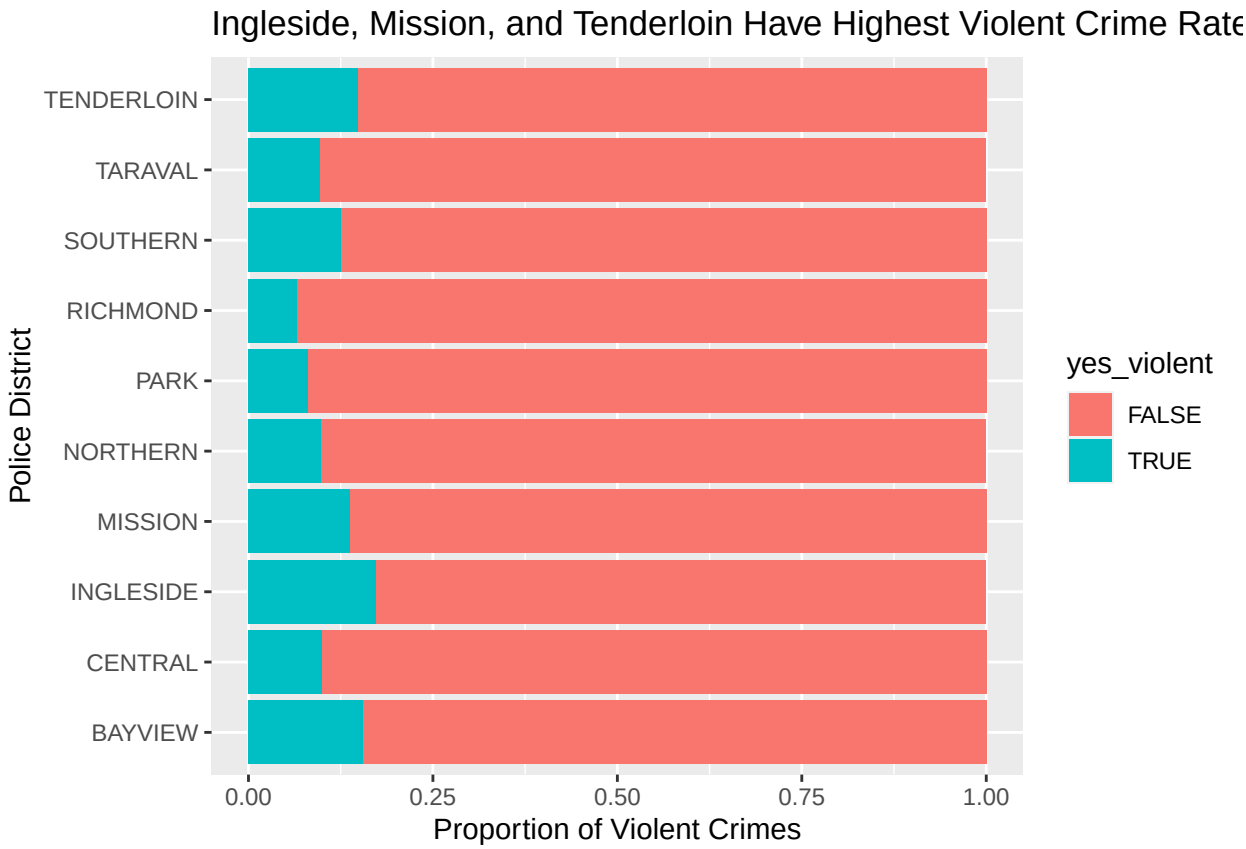
The faceted bar graph shows the frequency of each crime rate on a given day of the week. When looking at the visualization, it is easy to see the large difference between types of crime that exist. On each day, the number of property related crimes and miscellaneous crimes are significantly greater than the 5 other crime types. When looking at the frequency of crime types from day to day, every day has a similar pattern of frequency. This further supports the observation from the previous visualization where crime and day of the week do not necessarily have a relationship.

1- Relationship between crime type and time? Mihir



3- Which PD has the highest proportion of violent crime? Kyra

```
## # A tibble: 10 x 3
## # Groups:   PdDistrict [10]
##    PdDistrict     n  perc
##    <chr>      <int> <dbl>
##  1 INGLESIDE    199 17.2
##  2 BAYVIEW      223 15.6
##  3 TENDERLOIN   150 14.9
##  4 MISSION      257 13.7
##  5 SOUTHERN     367 12.6
##  6 CENTRAL      174  9.98
##  7 NORTHERN     200  9.88
##  8 TARAVAL      109  9.74
##  9 PARK          68  8.09
## 10 RICHMOND      58  6.60
```



By looking at the table and bar plot, it is clear that Ingleside, Mission, and Tenderloin have the highest rates of violent crime. However, Mission, Southern, and Bayview have the highest number of violent crimes. Park and Richmond both have the lowest rates and total numbers of violent crimes. For all police districts, the percentage of violent crimes is lower than 18%.

#Chi-Square Test

#Mihir We will be performing a Chi-Squared test between these crime types and categorical time of day to determine if there is the relationship between them is statistically significant.

H_0 : NO relationship between the crime types created above and categories for time of day created above.

H_a : There IS a relationship between the crime types created above and categories for time of day created above.

α of 0.05

```
##                  TimeOfDay
## CrimeCategory     afternoon dawn morning night
##   Drug/Alcohol          194   80     106   132
##   Legal Violation       354  135     173   280
##   Miscellaneous        1574  560     993  1103
##   Property             1994  788    1119  2343
##   Sex                    32    5       3    31
##   Suspicious            239  126     235   210
##   Violent               508  365     346   586
##   White Collar          165   63      75    83
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  table
## X-squared = 347.32, df = 21, p-value < 2.2e-16
```

#T Test

Given the significant p-value of our chi-square test statistic. We decided to conduct a hypothesis test to see how number of total crimes in San Francisco compares during the daylight and nightime.

We will now use the CLT to perform inference because the observations are independently selected and in this case, the sample size is large enough (n>30) for the CLT to apply. We are using t-distribution because we are testing a single sample's population mean and we don't know the true population SD.

H_0 : The true mean time of crimes committed during the week is equal to the true mean time of crimes committed during the weekend

H_a : The true mean time of crimes committed during the week is less than true mean time of crimes committed during the weekend

α of 0.05

#Logistic Regression

#Mihir

By using logistic regression, we hope to answer the question of how much more likely a violent crime is to occur depending on the time range of the crime committed. This model below shows the predicted proportion of crimes that are violent given the predictor of time range. The three time ranges used are evening, morning, and night. We hypothesize that the highest proportion of violent crimes will occur at night because there are typically fewer witnesses at these hours.

Before using logistic regression, we checked that all necessary conditions...

```
## # A tibble: 10 x 5
##    term               estimate std.error statistic   p.value
##    <chr>                 <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)         -2.13      0.0780   -27.3   2.14e-164
##  2 DayOfWeekMonday     -0.0714    0.0954    -0.749 4.54e-  1
##  3 DayOfWeekTuesday    -0.0215    0.0941    -0.228 8.19e-  1
##  4 DayOfWeekWednesday  -0.114     0.0964    -1.18  2.38e-  1
##  5 DayOfWeekThursday   -0.145     0.0969    -1.49  1.36e-  1
##  6 DayOfWeekFriday     -0.148     0.0937    -1.59  1.13e-  1
##  7 DayOfWeekSaturday    0.0792    0.0915     0.865 3.87e-  1
##  8 timerangedawn        0.608     0.0743     8.18  2.75e- 16
##  9 timerangemorning     0.137     0.0738     1.86  6.34e-  2
## 10 timerangenight       0.227     0.0643     3.53  4.14e-  4
```

```
## # A tibble: 10 x 5
##    term               estimate std.error statistic  p.value
##    <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)         -0.324     0.0512    -6.33  2.45e-10
##  2 DayOfWeekMonday     -0.0582    0.0636    -0.916 3.60e- 1
##  3 DayOfWeekTuesday    -0.107     0.0635    -1.69  9.13e- 2
##  4 DayOfWeekWednesday  -0.140     0.0637    -2.19  2.83e- 2
##  5 DayOfWeekThursday   -0.287     0.0642    -4.47  7.97e- 6
##  6 DayOfWeekFriday     -0.0934    0.0617    -1.51  1.30e- 1
##  7 DayOfWeekSaturday   -0.0480    0.0629    -0.764 4.45e- 1
##  8 timerangedawn       -0.109     0.0535    -2.04  4.10e- 2
##  9 timerangemorning    -0.115     0.0474    -2.42  1.53e- 2
## 10 timerangenight       0.396     0.0409     9.68  3.73e-22
```

RESULTS

#Chi-Square Test

The test statistic is 359.84, which has a chi squared distribution with 18 df under H_0. The p-value is < 2.2e-16 which is less than the α of 0.05. This means there is sufficient evidence to reject the null hypothesis. As a result, I conclude that there is sufficient evidence to suggest that at the 0.05 significance level that there is a relationship between the crime types created above and categories for time of day created above.

#Logistic Regression

#Mihir For Violent Crime:

Predicted logit(p) = -2.280 + 0.077* (Mon.) + 0.127* (Tues.) + 0.035* (Wed.) +
0.004* (Thur.) + 0.228* (Sat.) + 0.148* (Sun.) + 0.137* (morning) + 0.227* (evening) + 0.608* (night)

For Property Crime:

Predicted logit(p) = -0.417 + 0.035* (Mon.) - 0.013* (Tues.) - 0.046* (Wed.) -
0.193* (Thur.) + 0.045* (Sat.) + 0.093* (Sun.) - 0.114* (morning) + 0.396* (evening) - 0.109* (night)

#YUE: add quickly that we checked all conditions for regression, satisfied all conditions.

#Kyra Predicted logit(yes_violent) = 0.10383 + 0.01188(evening) + 0.00227(morning) +0.06939(night)

DISCUSSION

#leah's discussion

When trying to be safest in the busy city of San Francisco, we discovered through our analysis that certain measures can be taken to improve one's safety. This is proven by the multiple factors that influence where, what, and when crime is committed. However, it might first be important to discuss which factors do not play a substantial role in the act of a crime. For example, when looking at the bar graph comparing day of the week and number of crimes, it is clear that the difference from day to day is very minimal. Therefore, looking at just the day itself should not be a factor to whether it may be more dangerous or not. Looking at the visualization comparing type of crime and day of week furthers the point that the day of the week does not play a significant role on crime. Each day has a spread where Property crime is the highest, followed by Miscellaneous Crime, followed by Violent Crime. Something that could cause a possible influence in our data is the fact that the variable crime type was by us, so the organization of what crime fits into what category and the creation of categories is based on our research and knowledge.

#mihir's discussion: After creating the the categorical variables for time of day and also creating categorizing the types of crime within larger categories, we determined that there was a statistically significant relationship between the time of day and type of crime. As a result, I created a logistic model to calculate the log-odds of whether a violent crime occured with the predictors of day of the week and time of day. I also created a model for with the same predictors for a property crime.

This model cannot be applied to all cities but the base concept should remain the same. Most cities will have likely have crime peak during the night and during the weekends because more people will not be home.

I think if I were to continue to working the project I would likely break down the model to apply it specific districts By doing so, the characteristics of these individual districts could be identified. Hopefully, these characteristics allow us to potentially apply these models to districts with similar population density, population breakdown, etc.

#Tina's disc: After constructing the faceted visualization of crime count and time, we learned that certain categories of crime are far more prominent than others. For example, larceny/theft is more common, along with non-criminal crimes and assault. Most crimes happen during the afternoon and night, with the least happening in the hours from 0 to 6 (or in the early morning). Out of all the categories of crime listed, larceny/theft is mostly conducted during the evening, or between 6pm and 12am. This makes sense, as this is usually when night begins to set in, and it's a bit darker out, thus lending to increased obscurity and

decreased acuity and vision-related impairments. Overall, this visualization displayed more crimes occurring during afternoons and evenings.

From a policy standpoint, government leaders in San Francisco should consider having additional police on duty during the times when crime is more eminent (afternoon and night). Another course of action could be to simply hire more police trained in Arceny, Theft, and Assault, as they were the most prominent from the faceted graph.

We understand that we cannot extrapolate our analysis to every city; however, our conclusions will be generalizable to similar cities to a moderate degree. Other cities with similar infrastructure and economic conditions are more likely to utilize the analysis we've found. This analysis will not be applicable to Durham, NC, for example, because of the population density and overall difference in cities (SF is a bustling city, while Durham is a smaller, quaint town).

If we were to continue work on the project, we would add to our analysis by introducing data from different cities that are comparable to SF. It would be interesting to see the parallels in crime rates, as for many college students, traveling to their first job post-grad will be their first taste of independence and financial freedom – thus, safety is an important factor to take into consideration. Ultimately, expanding the population of interest to citizens in multiple cities would give a better picture of how cases of crime occur differently by region, state, country, or population density (urban vs. rural). Second, we would also adjust for additional potential confounding variables to improve the accuracy of our analysis and models. Finally, to learn more, we'd want to speak with current or past residents and police officers about their experiences. Data is a great way to create thoughtful questions but it may not provide the full or complete answer.

#Kyra's discussion:

The bar graph that shows crime rates and violent crime proportions that is faceted by police districts shows valuable insight as to which police districts are faced with the highest crime rates. The police districts of Tenderloin, Mission, and Ingleside have the highest percentages of violent crime (17.7%, 16.6%, 16% respectively). However, it is Bayview, Northern, and Southern that have the highest total number of violent crimes (239, 202, 291 respectively). Park and Richmond were both consistent in having the lowest numbers of total violent crimes as well as proportion of violent crimes. Noting the success of these districts in maintaining low levels of crimes, it could be beneficial to restructure other districts to mirror their practices.

Given that factors such as poverty level and unemployment rates are main drivers for crime[1], it would be valuable to assess these numbers for each police district. It would be valuable to know the differences in these factors for districts with more and less crime so that next steps can be taken to lower crime rates. For example, should a future study conclude that Park's public education system has higher test scores than that of Bayview, improving schools could be the best step for mitigating crime.

An important factor that this analysis is lacking is the populations of each police district. Having a larger population size would likely contribute to greater numbers of crime, even if per capita crime is lower. This information is not present in the dataset we used, but would be necessary to extrapolate a greater conclusion regarding which police district is most dangerous.

REFERENCES

[1]https://poetsandquantsforundergrads.com/2020/05/15/are-these-the-50-best-metro-areas-for-recent-college-grads/

[2] https://www.sfchronicle.com/bayarea/article/Which-crimes-are-up-down-in-SF-during-15408485.php

[3] https://www.sfchronicle.com/bayarea/philmatier/article/SF-ranks-high-in-property-crime-while-it-ranks-14439369.php

[4] https://ucr.fbi.gov/hate-crime/2011/resources/variables-affecting-crime

links to use for map: https://www.benjaminsorensen.me/project/sf_police/ https://data.sfgov.org/Public-Safety/Current-Police-Districts/wkhw-cjsf https://r-spatial.github.io/sf/articles/sf5.html#geometry-with-attributes-sf-1

NOTES

-notes from OH: -connect to the next level -interpret coefficients -final repo should look like a paper from poli sci -need to change order of days of the week -need to add README file -need to add coefficients

TO DO LIST

- t test MIHIR
- justifications methodology LEAH
- results add coefficient interpretations KYRA
- add regression model significance in discussion KYRA
- Discussion - organize it, add conclusion TINA
- clean up data sets
- References - LEAH
- SLIDES - LEAH, TINA