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

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

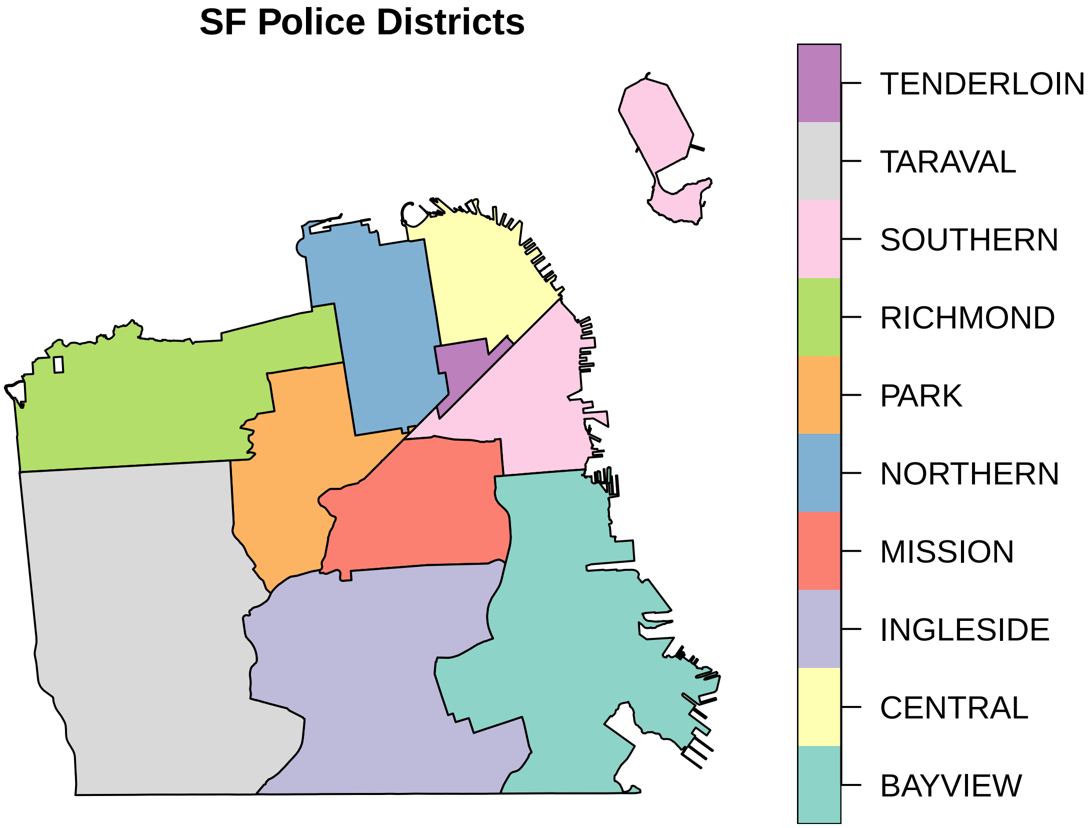
Variables

We will analyze the validity of our hypotheses using various statistical methods, including a Chi-square test, logistic regression model, and hypothesis tests. 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."

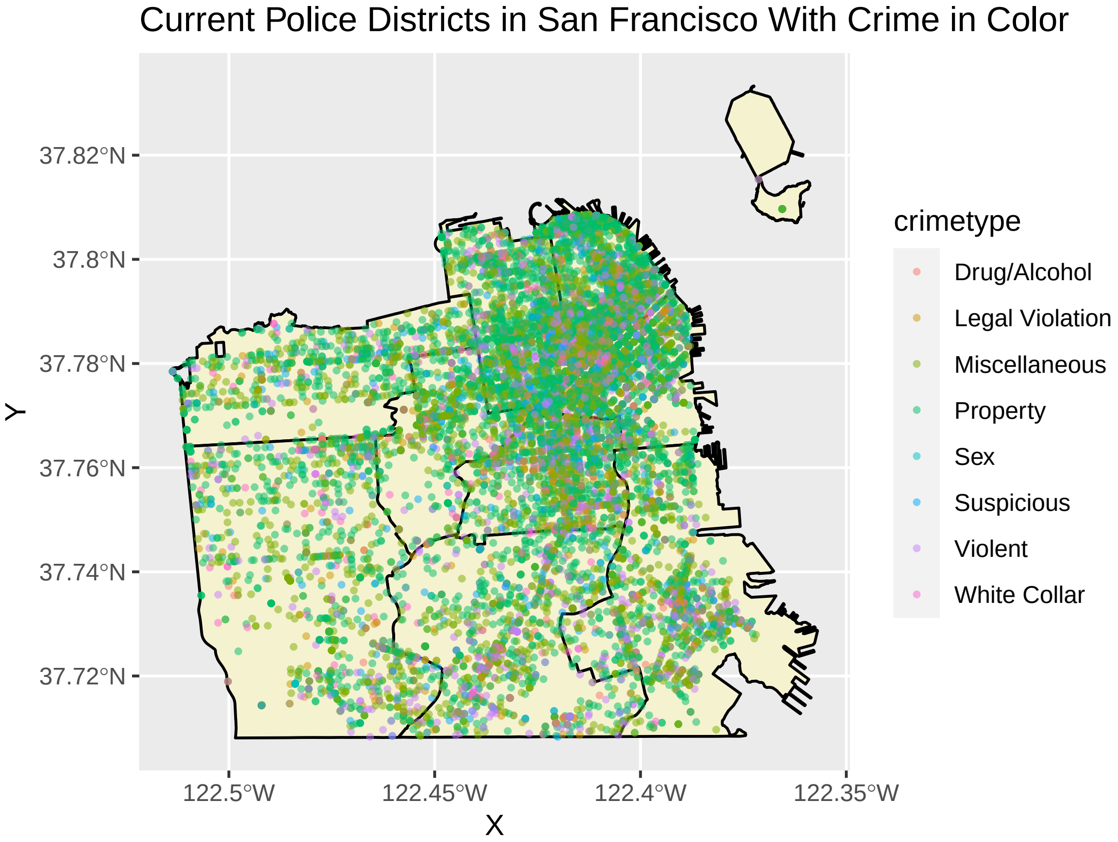
We also decided to categorize the all of the different types of crime that were reported. We organized the 39 types of crimes into variable crimetype, which consists of "Property", "Violent", "White Collar", "Drug/Alcohol", "Sex", "Suspicious", "Legal Violation", and "Miscellaneous". It is also important to note that we will also be analyzing crime in the context of whether it was violent or not.

Visualizations

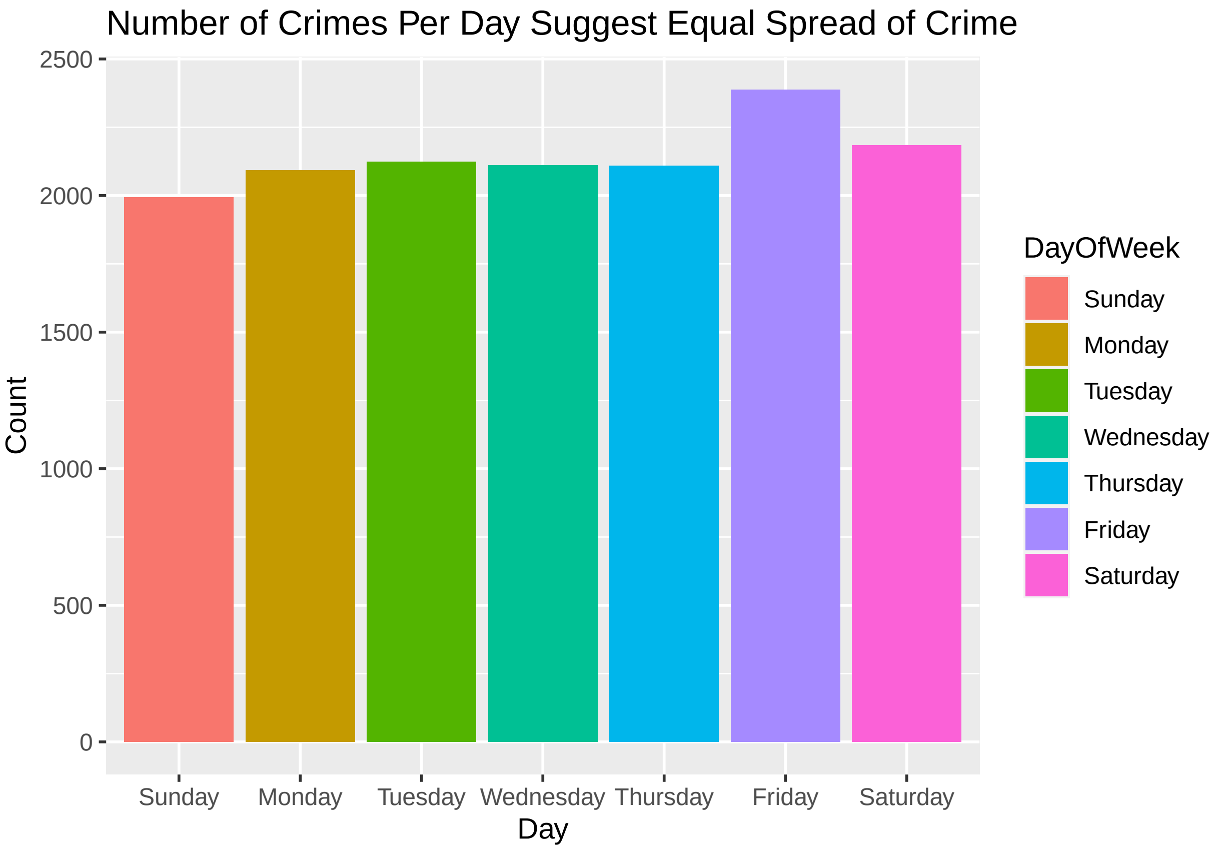
Because this dataset is directly from the San Francisco Police Department, Crime is not recorded by Neighborhood but Police District. To get a better understanding of how this may affect crime, we decided to visualize these districts as well as take a look at the number of crimes reported per district. Because there is not a clear pattern between area and crime count, we will be using proportions when analyzing crime rates from district to district.



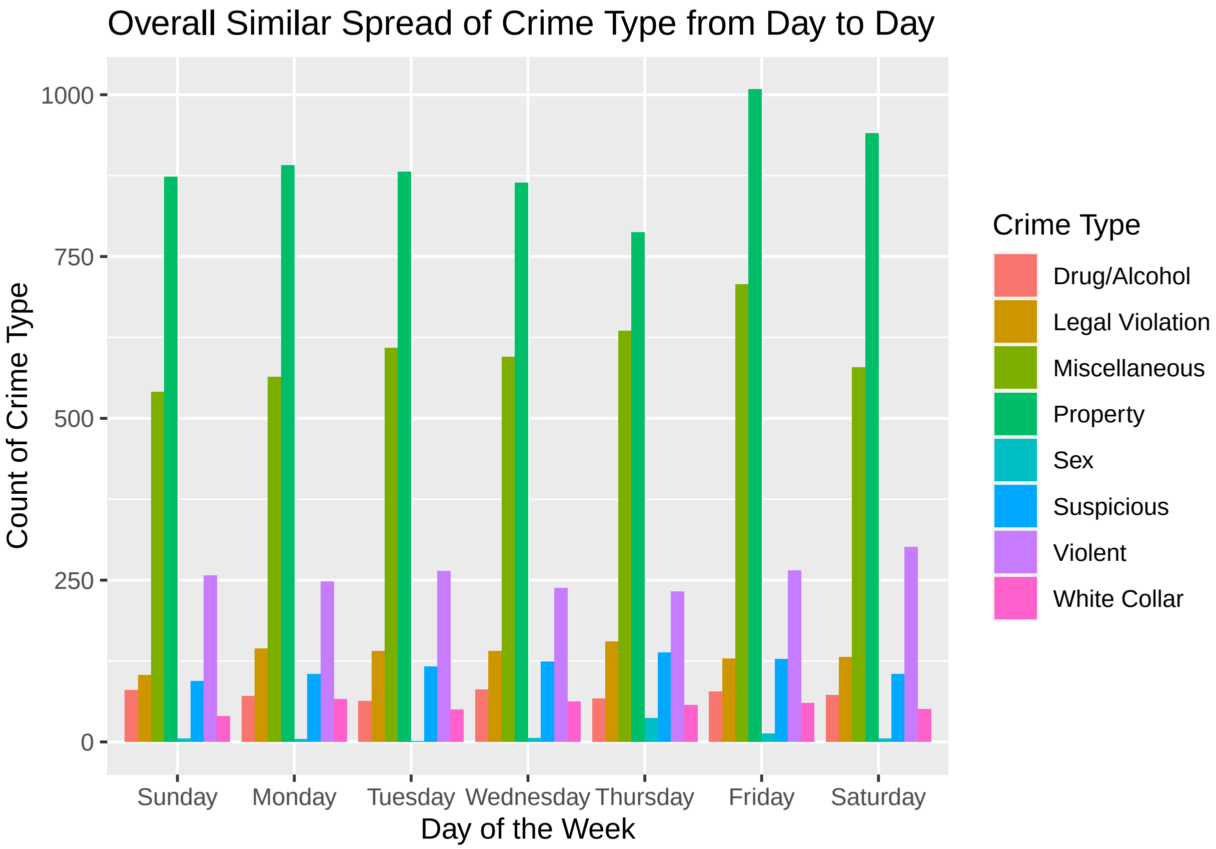
```
## # A tibble: 10 x 2
## # Groups:   PdDistrict [10]
##    PdDistrict     n
##    <chr>      <int>
##  1 BAYVIEW     1434
##  2 CENTRAL     1743
##  3 INGLESIDE   1156
##  4 MISSION     1872
##  5 NORTHERN    2025
##  6 PARK         841
##  7 RICHMOND     879
##  8 SOUTHERN    2921
##  9 TARAVAL     1119
## 10 TENDERLOIN  1010
```



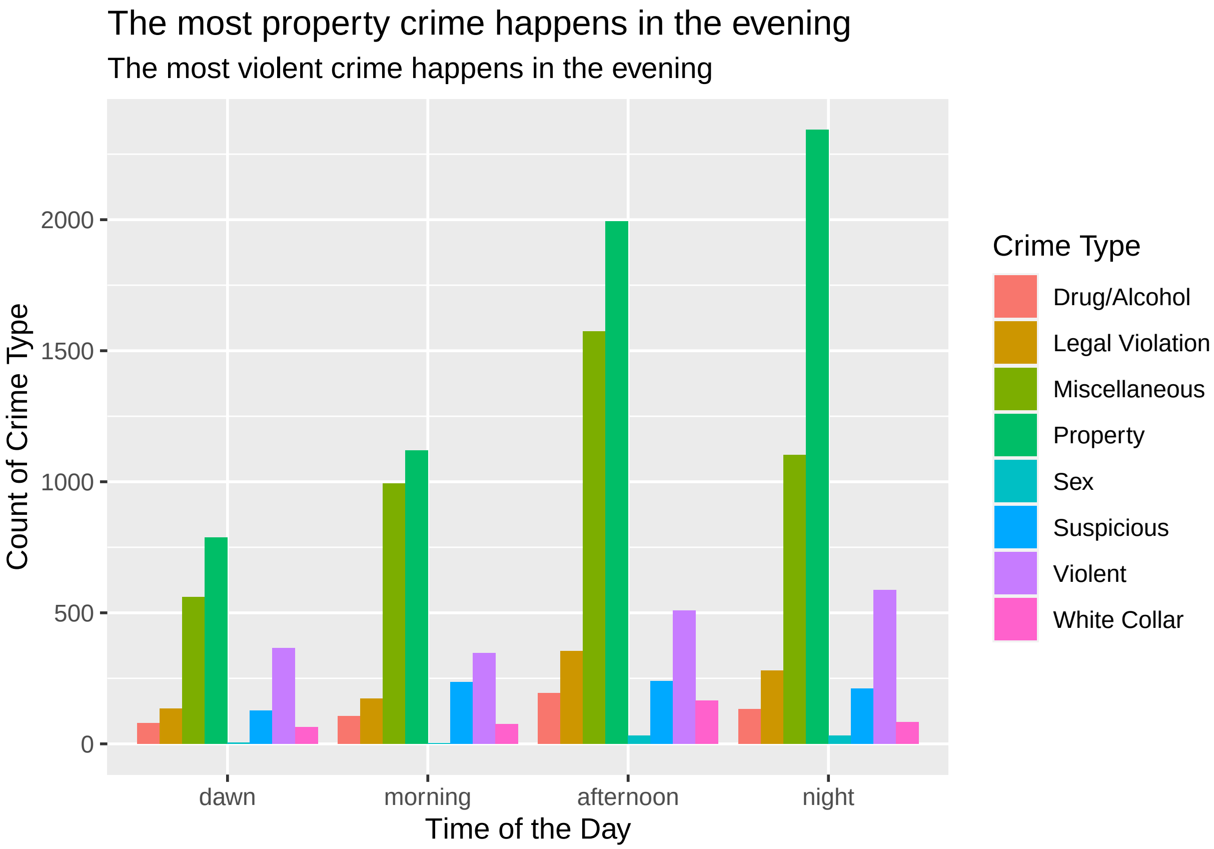
There's a lot of Property Crime concentrated in the Central, Northern, Southern, and Tenderloin.



The first 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. Because this recorded over a whole year, it may be important to identify that Friday has the highest number of crimes while Sunday has the least, but overall there is no significant pattern that sticks out.



Because of the fairly equal distribution of crime by day, we decided to delver further by comparing each crime type by day. 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 strong relationship.

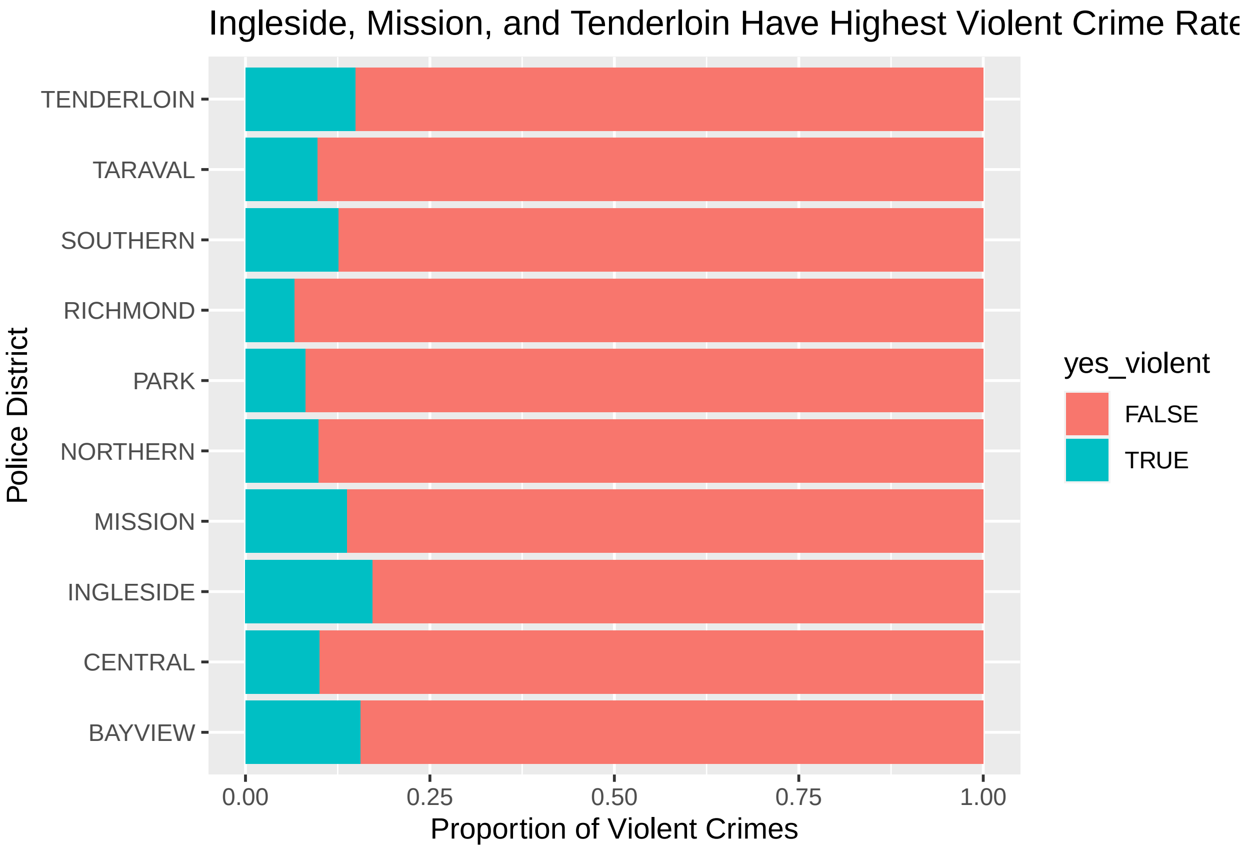


This then led us to the possibility that although day of the week may not have much of a relationship, this does not completely outrule time as a factor in how a crime takes places. We visualized this by creating a bar graph comparing frequency of crime and time of day. The visualization proves that there is a clear difference in what part of the day that crime is committed. The greatest amount of property crime occurs during the night, followed by the afternoon. Overall, the latter half of the day is when a higher amount of crime is committed.

###CHANGE DESCRIPTION

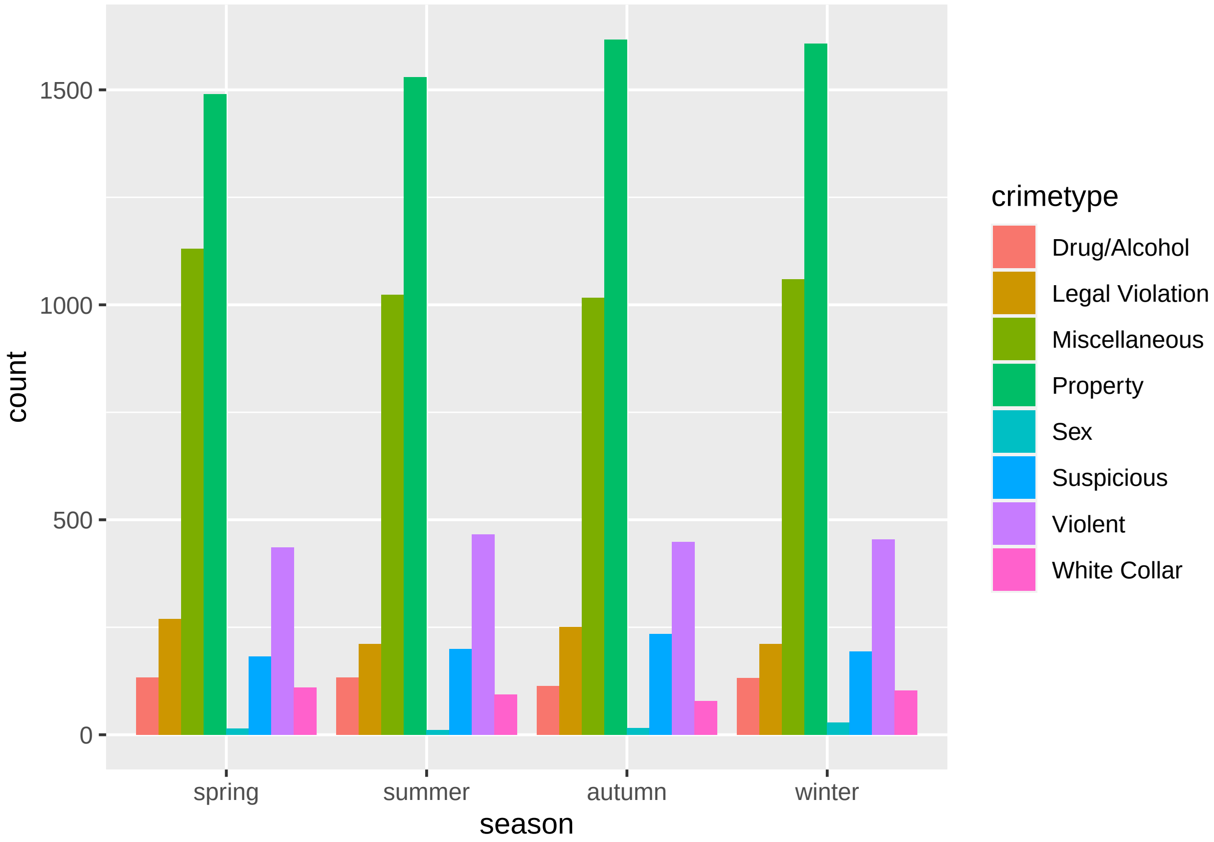
2 - Which PD has the highest proportion of violent crime?

```
## # 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        1    2    3    4
##   Drug/Alcohol      80  106  194  132
##   Legal Violation  135  173  354  280
##   Miscellaneous    560  993 1574 1103
##   Property         788 1119 1994 2343
##   Sex                5    3   32   31
##   Suspicious       126  235  239  210
##   Violent          365  346  508  586
##   White Collar      63   75  165   83
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  table
## X-squared = 347.32, df = 21, p-value < 2.2e-16
```

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 the necessary conditions of linearity, independence, normality, and equal variance were met.

```
## # A tibble: 10 x 5
##    term               estimate std.error statistic  p.value
##    <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)         -1.52      0.0828   -18.4   1.28e-75
##  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 timerangemorning    -0.471     0.0813    -5.80  6.65e- 9
##  9 timerangeafternoon  -0.608     0.0743    -8.18  2.75e-16
## 10 timerangenight      -0.381     0.0727    -5.24  1.58e- 7
```

```
## # A tibble: 10 x 5
##    term               estimate std.error statistic  p.value
##    <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)        -0.433      0.0604   -7.18   7.01e-13
##  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 timerangemorning   -0.00552    0.0587   -0.0941 9.25e- 1
##  9 timerangeafternoon  0.109      0.0535    2.04   4.10e- 2
## 10 timerangenight      0.505      0.0536    9.42   4.67e-21
```

Hypothesis Tests

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: Predicted target has the SAME likelihood of violent crime occuring than the sample population.

H_a: Predicted target has a GREATER likelihood of violent crime occuring than the sample population.

α of 0.05

```
##
##  Welch Two Sample t-test
##
## data:  is_Violent by targeted
## t = -5.0675, df = 381.07, p-value = 3.149e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.07524435
## sample estimates:
## mean in group 0 mean in group 1
##       0.1175747       0.2291105
```

H_0: Predicted target has the SAME likelihood of violent crime occuring than the sample population.

H_a: Predicted target has a GREATER likelihood of violent crime occuring than the sample population.

α of 0.05

```
##
##  Welch Two Sample t-test
##
## data:  is_Property by targeted
## t = -2.6675, df = 622.8, p-value = 0.00392
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.02160844
## sample estimates:
## mean in group 0 mean in group 1
##       0.4140896       0.4705882
```

H_0 : Property crime is equally likely to occur during colder Autumn and Winter than compared to other seasons

H_a : Property crime is more likely to occur during colder Autumn and Winter than other seasons

α of 0.05

```
##
##  Welch Two Sample t-test
##
## data:  is_Property by is_cold
## t = -2.4695, df = 14996, p-value = 0.006771
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.006636013
## sample estimates:
## mean in group 0 mean in group 1
##       0.4062416       0.4261168
```

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

For Violent Crime:

Predicted logit(p) = -2.131 - 0.071* (Mon.) - 0.021* (Tues.) - 0.114* (Wed.) - 0.145* (Thur.) - 0.148* (Fri.) + 0.079 * (Sat.) + 0.608* (dawn) + 0.137* (morning) + 0.227* (night)

While holding the day of the week constant, the log odds of a violent crime occurring increases by 0.114 if morning, 0.227 if night, and 0.608 if it is dawn.

While holding the time range of the day constant, the log odds of a violent crime occurring decreases by 0.071 if Monday, 0.021 if Tuesday, 0.114 if Wednesday, 0.145 if Thursday, and 0.148 if it is Friday. However, the log-odds increase by 0.079 if it is Saturday.

The reference level (not sure if that's the correct term) is Sunday at Afternoon.

According to the model, the log-odds of a violent crime occuring is greatest when it is Saturday at Dawn and the least when it is Friday at Afternoon.

For Property Crime:

Predicted logit(p) = -0.324 - 0.058* (Mon.) - 0.107* (Tues.) - 0.140* (Wed.) -
0.287* (Thur.) - 0.093* (Fri.) - 0.048* (Sat.) - 0.109* (dawn) - 0.115* (morning) + 0.396* (night)

While holding the day of the week constant, the log odds of a violent crime occurring decreases by 0.109 if night and 0.114 if it is dawn. However, the log-odds increase by 0.396 if it is morning.

While holding the time range of the day constant, the log odds of a violent crime occurring decreases by -0.058 if Monday, -0.107 if Tuesday, -0.140 if Wednesday, -0.287 if Thursday, -0.093 if Friday, and -0.048 if it is Saturday.

The reference level (not sure if that's the correct term) is Sunday at Afternoon.

According to the model, the log-odds of a violent crime occuring is greatest when it is Sunday at Night and the least when it is Thursday at Morning.

Hypothesis Tests

add descriptions for 3 t tests

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.

In addition, after creating categorical variables for time of day and also 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, we created a logistic model to calculate the log-odds of whether a violent crime occurred with the predictors of day of the week and time of day. We also created a model with the same predictors for property crime. However, this model cannot be applied to all cities; 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.

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.

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. 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 also be important 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.

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 Larceny, Theft, and Assault, as they were the most prominent from the graphs. Again, 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 San Francisco. 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 first-hand local experiences with crime. Data is a great way to create thoughtful questions but it may not provide the full or complete answer.

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 USED FOR SF 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

TO DO LIST

- t test results TINA
- justifications methodology LEAH
- results add coefficient interpretations KYRA & MIHIR
- add regression model significance in discussion KYRA
- Discussion - organize it, add conclusion TINA done

- References - LEAH
- SLIDES - LEAH, TINA