

# 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 [4].

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 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.

```
## 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/sanfrancisco-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.

There are 13 variables in the dataset: `IncidentNum` (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.

## 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 variable `crimetype`, which consists of "Property", "Violent", "White Collar", "Drug/Alcohol", "Sex", "Suspicious", "Legal Violation", and "Miscellaneous".

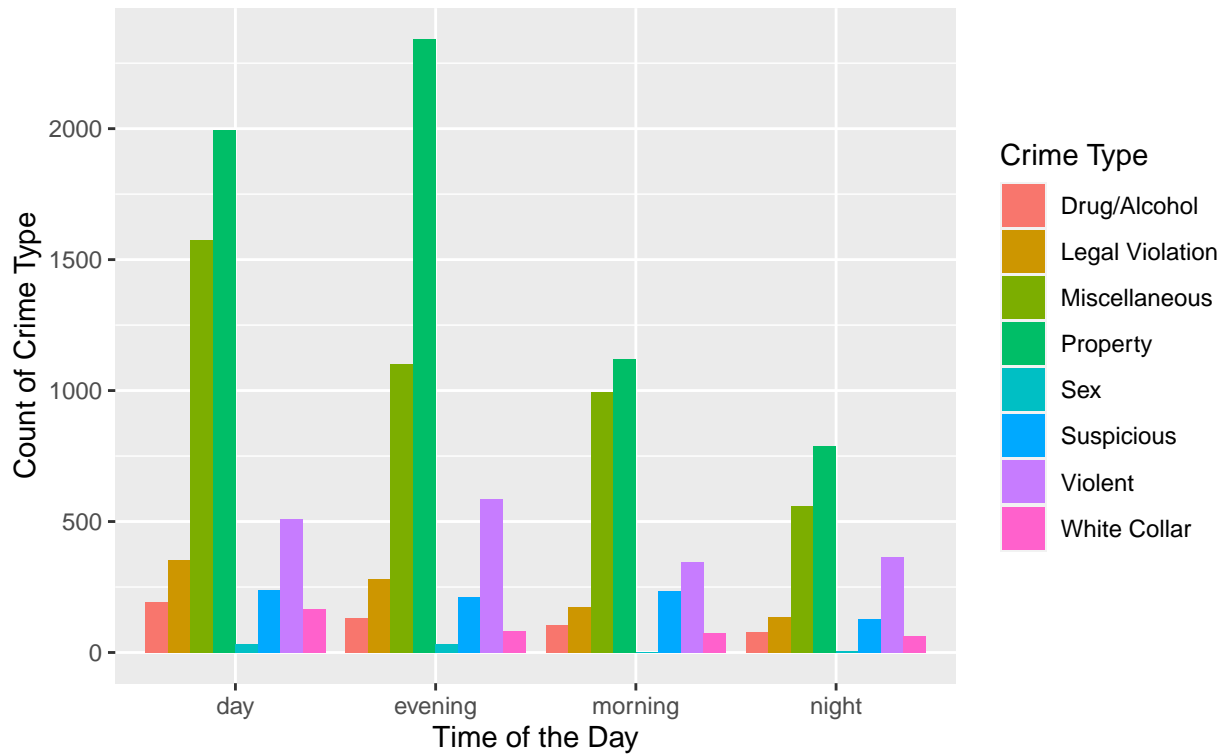
```
## Rows: 15,000
## Columns: 8
## $ Category   <chr> "ASSAULT", "LARCENY/THEFT", "NON-CRIMINAL", "NON-CRIMINA...
## $ DayOfWeek  <chr> "Tuesday", "Wednesday", "Sunday", "Tuesday", "Wednesday"...
## $ Date       <chr> "01/26/2016 12:00:00 AM", "06/15/2016 12:00:00 AM", "08/...
## $ PdDistrict <chr> "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> "day", "morning", "day", "day", "morning", "day", "morni...
## $ crimetype  <chr> "Violent", "Property", "Miscellaneous", "Miscellaneous",...
```

## Visualizations

### 1- Relationship between crime type and time? Mihir

The most property crime happens in the Evening

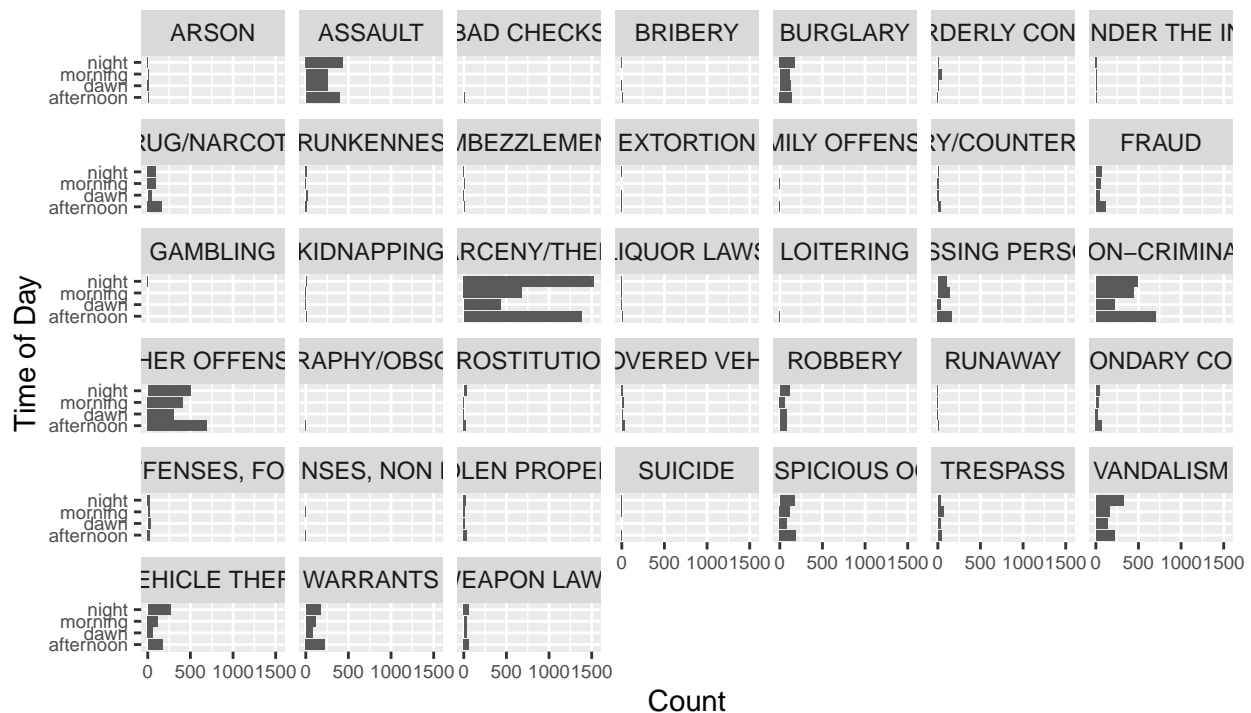
The most violent crime happens in the Evening



### ###2 - Relationship between time and crime? Tina

Question: Do more crimes generally occur at night in San Francisco? We will construct an effective, well-labeled visualization of the crime count and time.

## More Arceny/Theft, Assault and Non-Criminal Activity: the Relationship Between Category of Crime With Time of Day and Crime Faceted by Category of Crime



After constructing our visualization of crime count and time, a few things are clear: first, we can see that certain categories of crime are far more prominent than others. For example, larceny/theft is more common, along with non-criminal crimes, assault, and other crimes. Most crimes seem to 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 hours 18 & 24, ie 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 was quite interesting to dissect, as there does seem to be a correlation between crimes and their time of occurrence, as more crimes occur during afternoons and evenings.

tried this morning - issue is it can't read the .shp file.

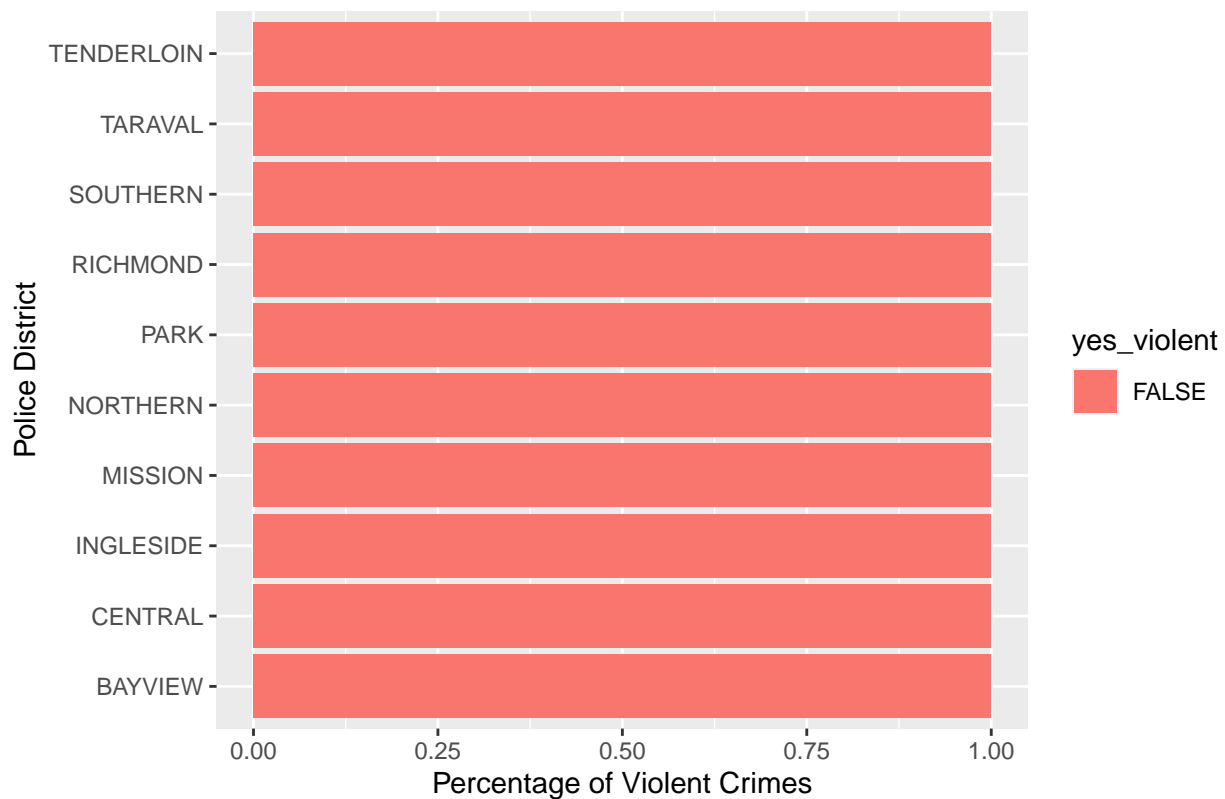
links to use for the map: [https://www.benjaminsorensen.me/project/sf\\_police/](https://www.benjaminsorensen.me/project/sf_police/) <https://data.sfgov.org/Public-Safety/Current-Police-Districts/wkhw-cjsf> <https://r-spatial.github.io/sf/articles/sf5.html#geometry-with-attributes-sf-1>

The purpose of this faceted barplot is to show which police districts have the highest rate of crime, as well as the highest proportion of violent crimes.

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

```
## # A tibble: 0 x 3
## # Groups:   PdDistrict [0]
## # ... with 3 variables: PdDistrict <chr>, n <int>, perc <dbl>
```

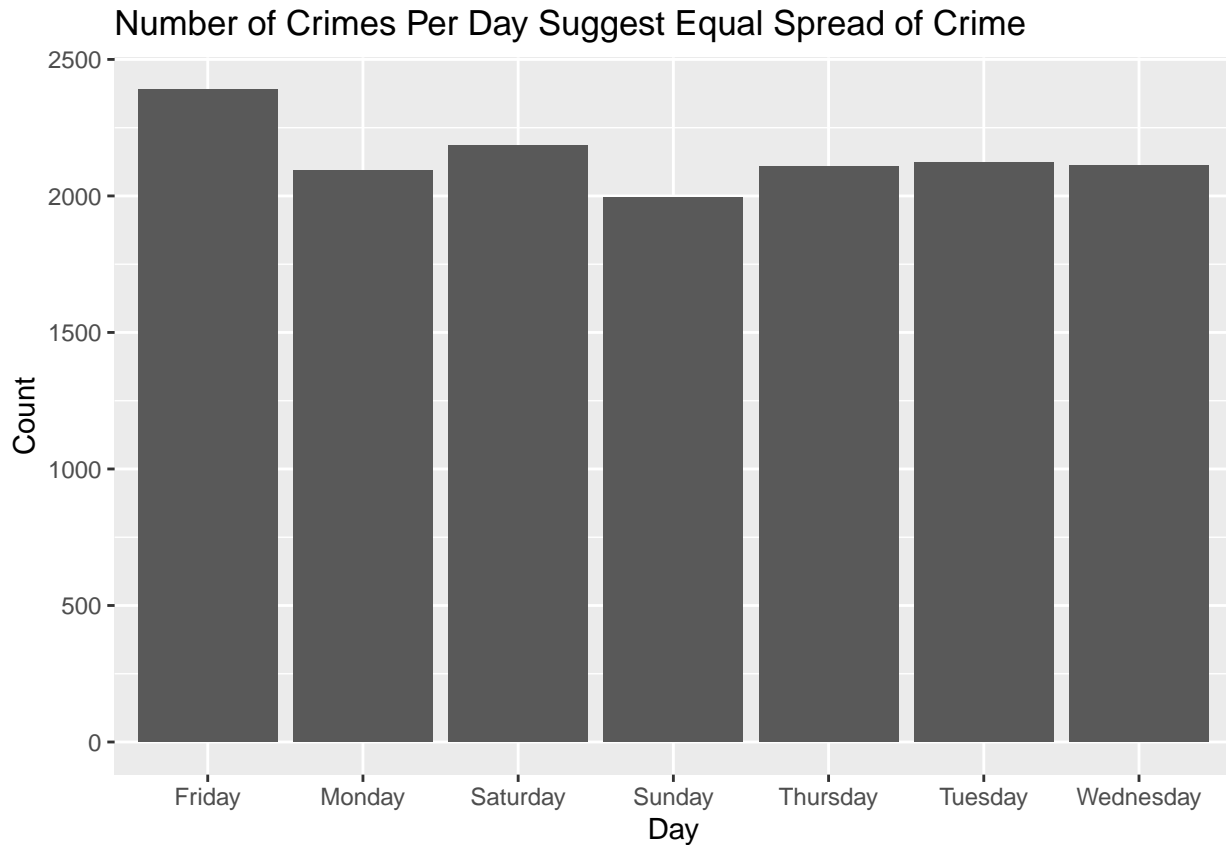
## Ingleside, Mission, and Tenderloin Have Highest Violent Crime Rate



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 16%.

## 5- Day of the week and category? Leah

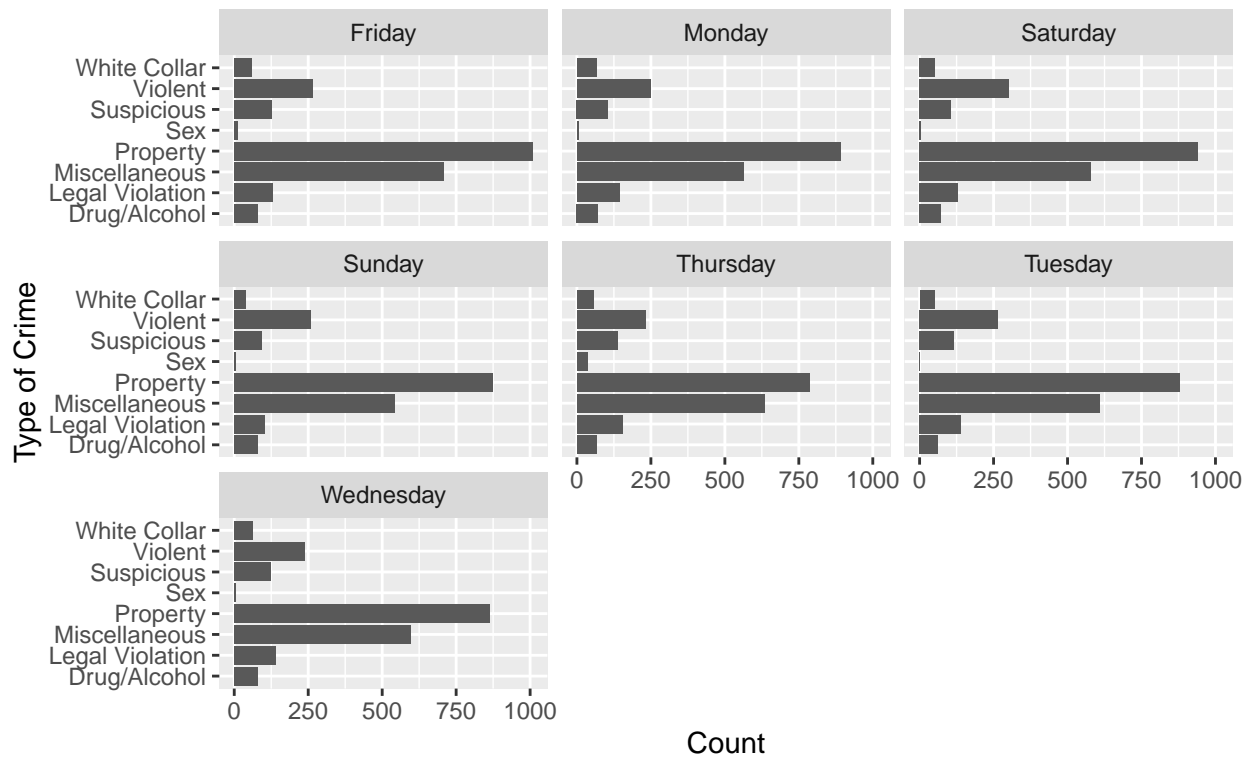
```
## # A tibble: 15,000 x 2
## # Groups:   DayOfWeek [7]
##   DayOfWeek cpday
##   <chr>      <int>
## 1 Tuesday    2124
## 2 Wednesday  2110
## 3 Sunday     1993
## 4 Tuesday    2124
## 5 Wednesday  2110
## 6 Monday     2093
## 7 Monday     2093
## 8 Thursday   2108
## 9 Wednesday  2110
## 10 Friday    2388
## # ... with 14,990 more rows
```



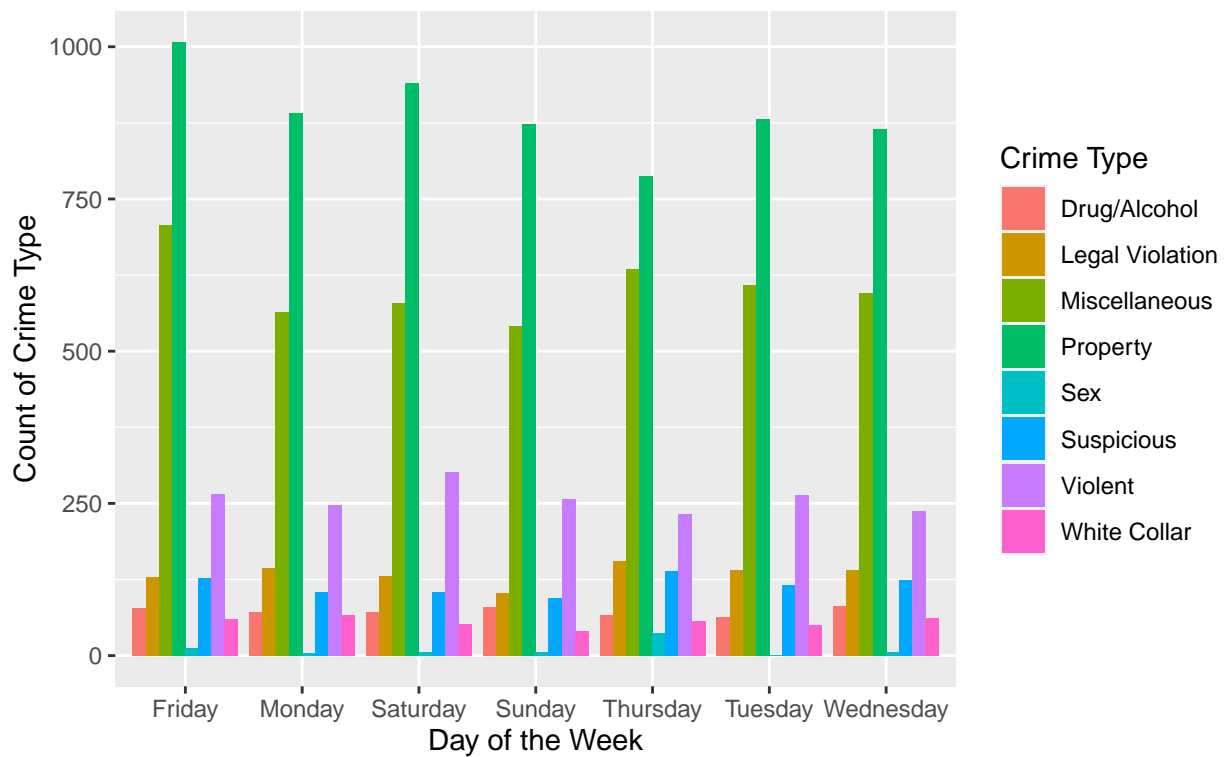
One relationship we were interested in was if certain days had a higher rates of crime. We visualized this relationship by creating a bar graph that compares the day of the week and number of crimes each day during this time period. By looking at the visual, we are able to see that each has a relatively 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> <chr> <chr> <chr> <chr> <dbl> <chr> <chr>
## 1 ASSAULT Tuesday 01/2~ NORTHERN NONE 13 day Violent
## 2 LARCENY~ Wednesday 06/1~ BAYVIEW NONE 8 morning Property
## 3 NON-CRI~ Sunday 08/2~ SOUTHERN NONE 12 day Miscella~
## 4 NON-CRI~ Tuesday 08/1~ CENTRAL NONE 16 day Miscella~
## 5 NON-CRI~ Wednesday 02/0~ NORTHERN NONE 6 morning Miscella~
## 6 ROBBERY Monday 03/2~ INGLESIDE NONE 15 day 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 evening Legal Vi~
## 10 VEHICLE~ Friday 04/0~ INGLESIDE NONE 22 evening Property
## # ... with 14,990 more rows, and 1 more variable: ctcount <int>
```

## Type of Crime by Day Shows A Large Proportion of Crimes as Property Related or Miscellaneous



The most property crime happens on Friday  
The most violent crime happens on Saturday



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.

#Chi-Square Test

#Mihir We will be performing a Chi-Squared test between these categorical variables 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.

$\alpha$  of 0.05

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

##
## 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.

```
## # A tibble: 15,000 x 4
##   DayOfWeek PdDistrict timerange isViolent
##   <chr>      <chr>      <chr>      <dbl>
## 1 Tuesday   NORTHERN    day         1
## 2 Wednesday BAYVIEW     morning      0
## 3 Sunday    SOUTHERN    day         0
## 4 Tuesday   CENTRAL     day         0
## 5 Wednesday NORTHERN    morning      0
## 6 Monday    INGLESIDE   day         1
## 7 Monday    SOUTHERN    morning      0
## 8 Thursday  SOUTHERN    morning      0
## 9 Wednesday NORTHERN    evening      0
## 10 Friday   INGLESIDE   evening      0
## # ... with 14,990 more rows

## # A tibble: 10 x 5
```



```
##      term                estimate std.error statistic    p.value
##      <chr>                <dbl>      <dbl>      <dbl>    <dbl>
##  1 (Intercept)          -2.28        0.0765   -29.8    5.49e-195
##  2 DayOfWeekMonday      0.0771        0.0941    0.818   4.13e-  1
##  3 DayOfWeekSaturday    0.228         0.0903    2.52    1.16e-  2
##  4 DayOfWeekSunday      0.148         0.0937    1.59    1.13e-  1
##  5 DayOfWeekThursday    0.00391       0.0955    0.0410  9.67e-  1
##  6 DayOfWeekTuesday     0.127         0.0928    1.37    1.71e-  1
##  7 DayOfWeekWednesday   0.0348        0.0950    0.367   7.14e-  1
##  8 timerangeevening     0.227         0.0643    3.53    4.14e-  4
##  9 timerangemorning     0.137         0.0738    1.86    6.34e-  2
## 10 timerangenight       0.608         0.0743    8.18    2.75e- 16
```

#4- How does time range affect whether crimes are violent? Kyra

```
## # A tibble: 4 x 5
##      term                estimate std.error statistic p.value
##      <chr>                <dbl>      <dbl>      <dbl>    <dbl>
##  1 (Intercept)              0          0         NaN      NaN
##  2 timerangeevening          0          0         NaN      NaN
##  3 timerangemorning          0          0         NaN      NaN
##  4 timerangenight            0          0         NaN      NaN
```

## 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  $\alpha$  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 Predicted logit(p) =  $-2.280 + 0.077^* (\text{Mon.}) + 0.127^* (\text{Tues.}) + 0.035^* (\text{Wed.}) + 0.004^* (\text{Thur.}) + 0.228^* (\text{Sat.}) + 0.148^* (\text{Sun.}) + 0.137^* (\text{morning}) + 0.227^* (\text{evening}) + 0.608^* (\text{night})$

#Kyra logit(yes\_violent) =  $0.10383 + 0.01188(\text{evening}) + 0.00227(\text{morning}) + 0.06939(\text{night})$

## 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.

#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.

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 crimes (239, 202, 291 respectively). Park and Richmond were both consistent in having the lowest numbers of total crimes as well as 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://ucr.fbi.gov/hate-crime/2011/resources/variables-affecting-crime>
- [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://poetsandquantsforundergrads.com/2020/05/15/are-these-the-50-best-metro-areas-for-recent-college-grads/>

## NOTES

-notes from OH: -connect to the next level -how can ur results inform policy decisions? -interpret coefficients  
-final repo should look like a paper from poli sci -need to get rid of daytime and crimetype duplicates -need to change order of days of the week