Project Draft

2020 Vision

Mihir Patel, Tina Xia, Leah Okamura, Kyra Cooperman

Introduction

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.

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

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.

Data

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

```
## Rows: 15,000
## Columns: 13
## $ IncidntNum <chr> "160277379", "160789794", "160341225", "160913090", "166...
## $ Category   <chr> "LARCENY/THEFT", "ROBBERY", "NON-CRIMINAL", "LARCENY/THE...
## $ Descript   <chr> "LOST PROPERTY, PETTY THEFT", "ROBBERY, BODILY FORCE", "...
## $ DayOfWeek  <chr> "Saturday", "Wednesday", "Monday", "Wednesday", "Sunday"...
## $ Date       <chr> "04/02/2016 12:00:00 AM", "09/28/2016 12:00:00 AM", "04/...
## $ Time       <time> 21:00:00, 23:09:00, 09:30:00, 09:50:00, 02:00:00, 00:01...
## $ PdDistrict <chr> "PARK", "MISSION", "CENTRAL", "SOUTHERN", "SOUTHERN", "I...
## $ Resolution <chr> "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", ...
## $ Address    <chr> "300 Block of CARL ST", "MISSION ST / 17TH ST", "300 Blo...
## $ X          <dbl> -122.4556, -122.4195, -122.4091, -122.4034, -122.4022, -...
## $ Y          <dbl> 37.76513, 37.76343, 37.78720, 37.77542, 37.78109, 37.711...
```

```
## $ Location  <chr> "(37.7651340492249, -122.455615084227)", "(37.7634292328...
## $ PdId      <dbl> 1.602774e+13, 1.607898e+13, 1.603412e+13, 1.609131e+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 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.

Methodology

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 certain variables are categorical, so they ETC ETC

Note: we plan on grouping violence based on violent vs nonviolent.

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, 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.

Relationship between category and time? Mihir

To determine the relationship between category and time, I have created 4 time intervals (morning, day, evening, and night) and categorized the crimes based on the type of crime. I will then be performing a Chi-Squared test between these categorical variables to determine if there is any relationship between them? (is that the right explanation).

```
important <- sanfrancrime %>%
  mutate(str = as.character(Time)) %>%
  mutate(hourstr = substr(str, 1, 2)) %>%
  mutate (hour = as.numeric(hourstr)) %>%
  select(Category, DayOfWeek, Date, PdDistrict, Resolution, hour)

important <- important %>%
```

```
  mutate(timerange = case_when( hour >= 0 & hour < 6 ~ "night",
                                hour >= 6 & hour < 12 ~ "morning",
                                hour >= 12 & hour < 18 ~ "day",
                                hour >= 18 & hour < 24 ~ "evening"))
glimpse(important)
```

```
## Rows: 15,000
## Columns: 7
## $ Category   <chr> "LARCENY/THEFT", "ROBBERY", "NON-CRIMINAL", "LARCENY/THE...
## $ DayOfWeek  <chr> "Saturday", "Wednesday", "Monday", "Wednesday", "Sunday"...
## $ Date       <chr> "04/02/2016 12:00:00 AM", "09/28/2016 12:00:00 AM", "04/...
## $ PdDistrict <chr> "PARK", "MISSION", "CENTRAL", "SOUTHERN", "SOUTHERN", "I...
## $ Resolution <chr> "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", ...
## $ hour       <dbl> 21, 23, 9, 9, 2, 0, 11, 13, 19, 18, 21, 17, 9, 12, 14, 1...
## $ timerange  <chr> "evening", "evening", "morning", "morning", "night", "ni...
```

```
important <- important %>%
  mutate(crimetype = case_when(

    Category == "BURGLARY" | Category == "LARCENY/THEFT" |
    Category == "STOLEN PROPERTY" | Category == "RECOVERED VEHICLE" |
    Category == "VEHICLE THEFT" | Category == "ARSON" |
    Category == "VANDALISM"  ~ "property related",

    Category == "ROBBERY" | Category == "ASSAULT" |
    Category == "KIDNAPPING" |
      Category == "SEX OFFENSES, FORCIBLE" ~ "violence related",

    Category == "BRIBERY" | Category == "BAD CHECKS" |
    Category == "EMBEZZLEMENT"| Category == "FORGERY/COUNTERFEITING" |
    Category == "FRAUD" | Category == "GAMBLING"|
    Category == "EXTORTION" ~ "money related",

    Category == "DRIVING UNDER THE INFLUENCE" | Category == "DRUG/NARCOTIC" |
    Category == "DRUNKENNESS"| Category == "LIQUOR LAWS" ~ "drug related",

    Category == "PORNOGRAPHY/OBSCENE MAT" | Category == "PROSTITUTION" |
    Category == "SEX OFFENSES, NON FORCIBLE" ~ "sex related",

    Category == "LOITERING" | Category == "TREA" |
    Category == "TRESPASS"| Category == "SUSPICIOUS OCC" |
    Category == "DISORDERLY CONDUCT" ~ "suss related",

    Category == "FAMILY OFFENSES" | Category == "MISSING PERSON" |
    Category == "NON-CRIMINAL"| Category == "OTHER OFFENSES" |
    Category == "TRESPASS"| Category == "SECONDARY CODES" |
    Category == "SUICIDE"| Category == "SECONDARY CODES" |
    Category == "WARRANTS"| Category == "WEAPON LAWS" |
    Category == "RUNAWAY" ~ "misc."))

#maybe we want to add more crime types?
glimpse(important)
```

```
## Rows: 15,000
```

```
## Columns: 8
## $ Category   <chr> "LARCENY/THEFT", "ROBBERY", "NON-CRIMINAL", "LARCENY/THE...
## $ DayOfWeek  <chr> "Saturday", "Wednesday", "Monday", "Wednesday", "Sunday"...
## $ Date       <chr> "04/02/2016 12:00:00 AM", "09/28/2016 12:00:00 AM", "04/...
## $ PdDistrict <chr> "PARK", "MISSION", "CENTRAL", "SOUTHERN", "SOUTHERN", "I...
## $ Resolution <chr> "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", "NONE", ...
## $ hour       <dbl> 21, 23, 9, 9, 2, 0, 11, 13, 19, 18, 21, 17, 9, 12, 14, 1...
## $ timerange  <chr> "evening", "evening", "morning", "morning", "night", "ni...
## $ crimetype  <chr> "property related", "violence related", "misc.", "proper...
```

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

```
crimecount <- important %>%
  count(crimetype)
crimecount
```

```
## # A tibble: 7 x 2
##   crimetype           n
##   <chr>           <int>
## 1 drug related      514
## 2 misc.            5180
## 3 money related     349
## 4 property related 6234
## 5 sex related        79
## 6 suss related      837
## 7 violence related 1807
```

```
test <- important %>%
  group_by(crimetype) %>%
  count(timerange)
test
```

```
## # A tibble: 28 x 3
## # Groups:   crimetype [7]
##    crimetype     timerange      n
##    <chr>         <chr>      <int>
##  1 drug related  day          204
##  2 drug related  evening      143
##  3 drug related  morning      112
##  4 drug related  night         55
##  5 misc.         day         1846
##  6 misc.         evening     1429
##  7 misc.         morning     1188
##  8 misc.         night        717
##  9 money related day          132
## 10 money related evening       87
## # ... with 18 more rows
```

```
crimestuff <- c(rep(crimecount$crimetype[1], crimecount$n[1]),
             rep(crimecount$crimetype[2], crimecount$n[2]),
             rep(crimecount$crimetype[3], crimecount$n[3]),
             rep(crimecount$crimetype[4], crimecount$n[4]),
             rep(crimecount$crimetype[5], crimecount$n[5]),
```

```
                rep(crimecount$crimetype[6], crimecount$n[6]),
                rep(crimecount$crimetype[7], crimecount$n[7]))

timestuff <- c(
rep(test$timerange[1], test$n[1]), rep(test$timerange[2], test$n[2]),
rep(test$timerange[3], test$n[3]), rep(test$timerange[4], test$n[4]),

rep(test$timerange[5], test$n[5]), rep(test$timerange[6], test$n[6]),
rep(test$timerange[7], test$n[7]), rep(test$timerange[8], test$n[8]),

rep(test$timerange[9], test$n[9]), rep(test$timerange[10], test$n[10]),
rep(test$timerange[11], test$n[11]), rep(test$timerange[12], test$n[12]),

rep(test$timerange[13], test$n[13]), rep(test$timerange[14], test$n[14]),
rep(test$timerange[15], test$n[15]), rep(test$timerange[16], test$n[16]),

rep(test$timerange[17], test$n[17]), rep(test$timerange[18], test$n[18]),
rep(test$timerange[19], test$n[19]), rep(test$timerange[20], test$n[20]),

rep(test$timerange[21], test$n[21]), rep(test$timerange[22], test$n[22]),
rep(test$timerange[23], test$n[23]), rep(test$timerange[24], test$n[24]),

rep(test$timerange[25], test$n[25]), rep(test$timerange[26], test$n[26]),
rep(test$timerange[27], test$n[27]), rep(test$timerange[28], test$n[28]))


table <- table(crimestuff, timestuff)

chisq.test(table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 321.86, df = 18, p-value < 2.2e-16
```

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.

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.

```
sanfrancrime <- sanfrancrime %>%
  mutate(str = as.character(Time)) %>%
  mutate(hourstr = substr(str, 1, 2)) %>%
  mutate (hour = as.numeric(hourstr))

sanfrancrime <- sanfrancrime %>%
  mutate(timerange = case_when( hour >= 0 & hour < 6 ~ "dawn",
                                hour >= 6 & hour < 12 ~ "morning",
                                hour >= 12 & hour < 18 ~ "afternoon",
```

```
                                        hour >= 18 & hour < 24 ~ "night"))


ggplot(sanfrancrime, mapping = aes(y = timerange)) +
  geom_bar() + facet_wrap(~ Category) + labs(
    x = "Count",
    y = "Time of Day",
    title = "Relationship Between Category of Crime With Time of Day and Crime Count", subtitle = "Facet
```

Relationship Between Category of Crime With Time of Day and Crime C
Faceted by Category of Crime
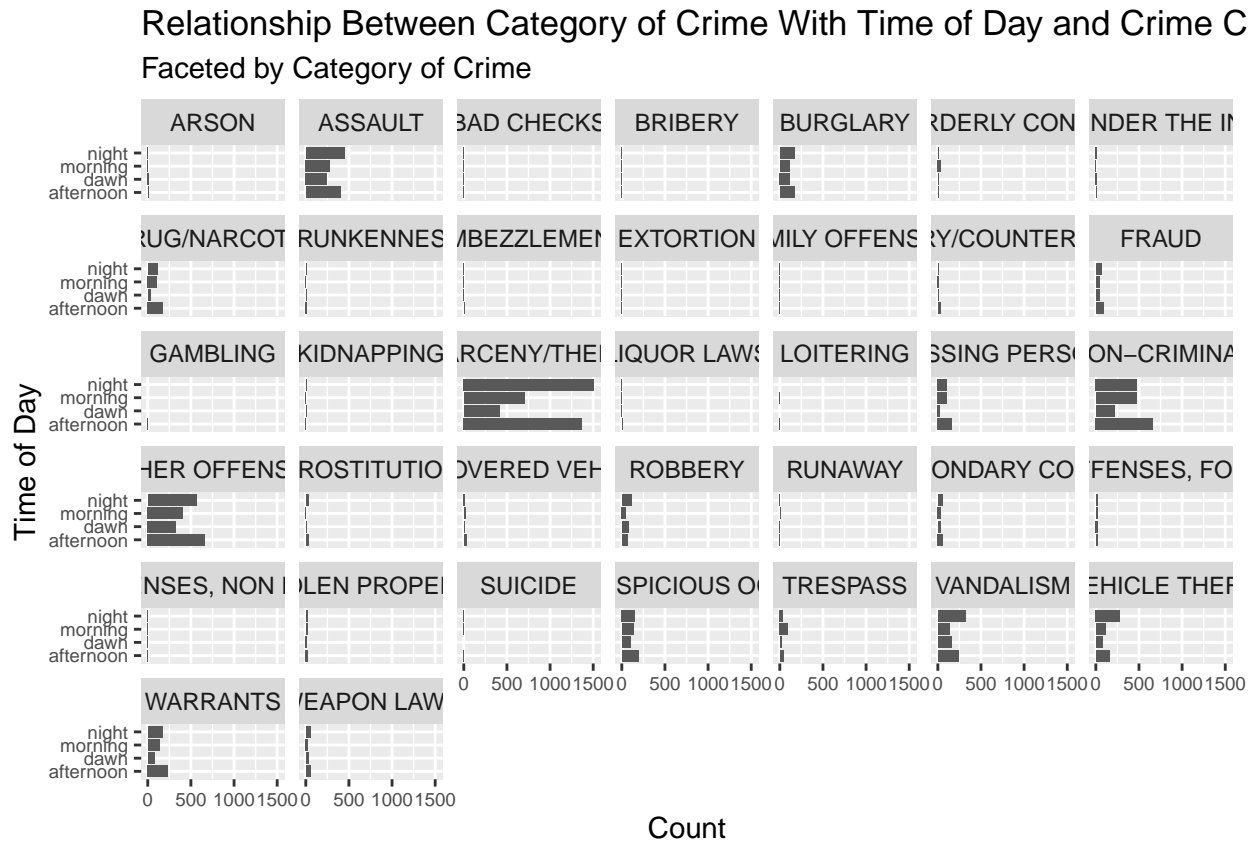

```
# sanfrancrime <- sanfrancrime %>%
#   group_by(Category) %>%
#   summarise(count = n())
# sanfrancrime
```

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.

```
# library(sf)
# data1 <- st_read("data/Police_Department_Incidents_-_Previous_Year__2016_.csv", quiet = TRUE)
# data1
```

```
# ggplot(data1) +
#   geom_sf(aes(fill = voted)) +
#   labs(title = "Higher population counties have more votes cast",
#         fill = "Total number of votes cast") +
#   theme_bw()
#
# ggplot(data1) +
#   geom_sf(color = "green", size = 1.5, fill = "orange", alpha = 0.50) +
#   labs(title = "SF data with theme and aesthetics") +
#   theme_bw()
```

Office hrs

Which PD has the highest proportion of violent crime? Kyra

```
pd_violent <- sanfrancrime%>%
  group_by(Category)%>%
  count()
pd_violent<- 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"))
important <- important%>%
  filter(PdDistrict!="NA")%>%
  group_by(PdDistrict)%>%
  mutate(yes_violent = crimetype == "violence related")%>%
  arrange(desc(yes_violent))


important%>%
  group_by(PdDistrict)%>%
  count(yes_violent)%>%
  mutate(perc = (n/sum(n)*100))%>%
  arrange(desc(perc))%>%
  filter(yes_violent=="TRUE")
```

```
## # A tibble: 10 x 4
## # Groups:   PdDistrict [10]
##     PdDistrict yes_violent     n  perc
##     <chr>        <lgl>       <int> <dbl>
##  1 MISSION     TRUE          320 16.2
##  2 TENDERLOIN  TRUE          153 16.0
##  3 INGLESIDE   TRUE          187 16.0
##  4 BAYVIEW     TRUE          224 15.6
##  5 TARAVAL     TRUE          123 10.7
##  6 SOUTHERN    TRUE          305 10.4
##  7 CENTRAL     TRUE          172 10.1
##  8 NORTHERN    TRUE          195 10.0
##  9 PARK        TRUE           74  8.57
## 10 RICHMOND    TRUE           54  6.27
```

```
ggplot(important, aes(x = PdDistrict, fill = yes_violent))+
  geom_bar(position = "fill") + coord_flip()+
  labs(title =
        "Ingleside, Mission, and Tenderloin Have Highest Violent Crime Rates ",
      y = "Percentage of Violent Crimes", x = "Police District")
```

Ingleside, Mission, and Tenderloin Have Highest Violent Crime Rate
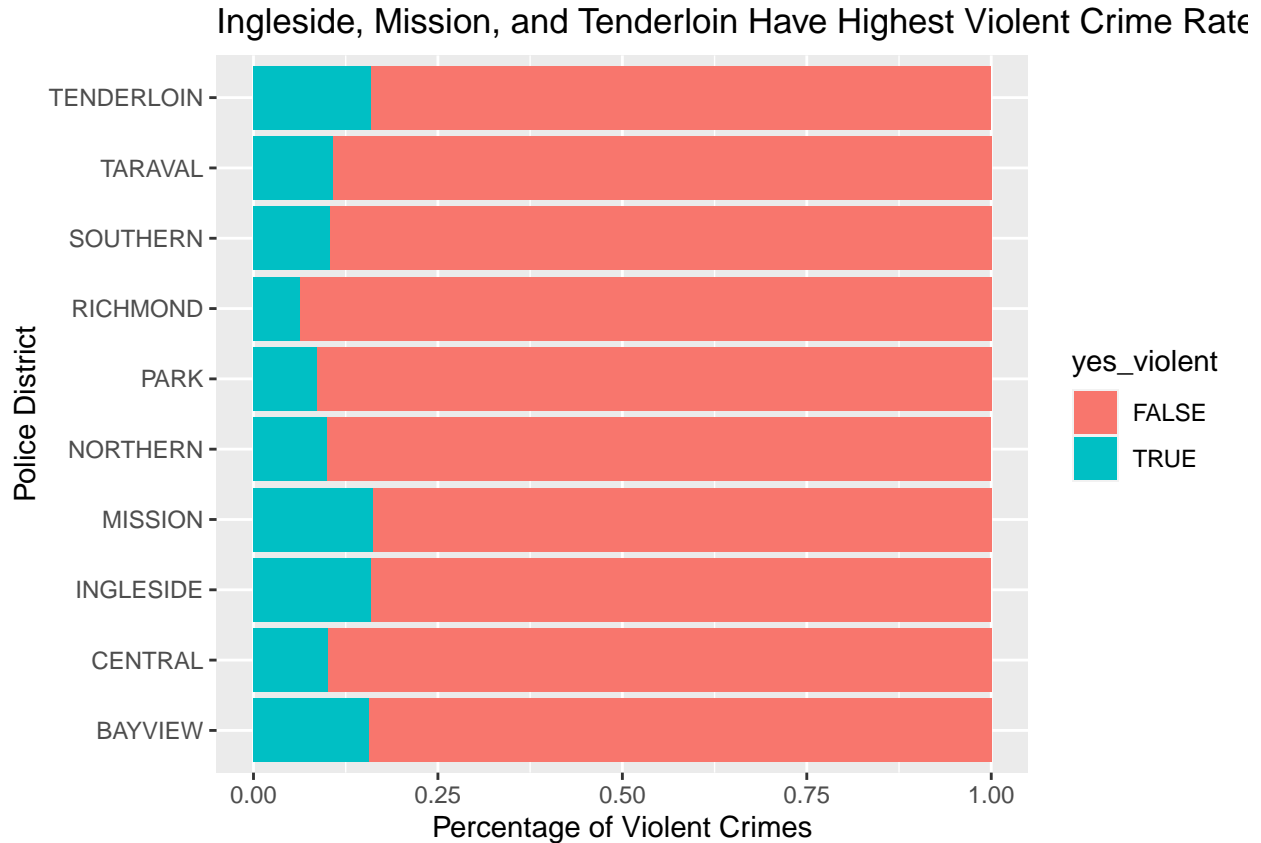
Ingleside, Mission, and Tenderloin have the highest rates of violent crime. However, Mission, Southern, and Bayview have the highest number of violent crimes. Park and Richmond both have the lowest rates and total numbers of violent crimes. For all police districts, the percentage of violent crimes is lower than 16%.

How does time range affect whether crimes are violent? Kyra

```
library(broom)
mod<- lm(yes_violent~timerange,
        data = important)
tidy(mod)
```

```
## # A tibble: 4 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       0.0998   0.00461     21.6  2.73e-102
## 2 timerangeevening  0.0243   0.00658      3.70 2.15e-  4
## 3 timerangemorning  0.0149   0.00743      2.00 4.55e-  2
## 4 timerangenight    0.0687   0.00841      8.17 3.20e- 16
```

$\text{logit}(p) = 0.10383 + 0.01188(\text{evening}) + 0.00227(\text{morning}) + 0.06939(\text{night})$

Day of the week and category? Leah

```
day <- sanfrancrime%>%
  group_by(DayOfWeek)%>%
  mutate(cpday = n())%>%
  select(DayOfWeek, cpday)
day
```

```
## # A tibble: 15,000 x 2
## # Groups:   DayOfWeek [7]
##    DayOfWeek cpday
##    <chr>     <int>
##  1 Saturday   2209
##  2 Wednesday  2112
##  3 Monday     2082
##  4 Wednesday  2112
##  5 Sunday     2062
##  6 Tuesday    2012
##  7 Monday     2082
##  8 Friday     2305
##  9 Sunday     2062
## 10 Monday     2082
## # ... with 14,990 more rows
```

```
ggplot(data = day, mapping = aes(x = DayOfWeek)) +
    geom_bar()  + labs(x = "Day", y = "Count",
      title = "Number of Crimes Per Day Suggest Equal Spread of Crime")
```


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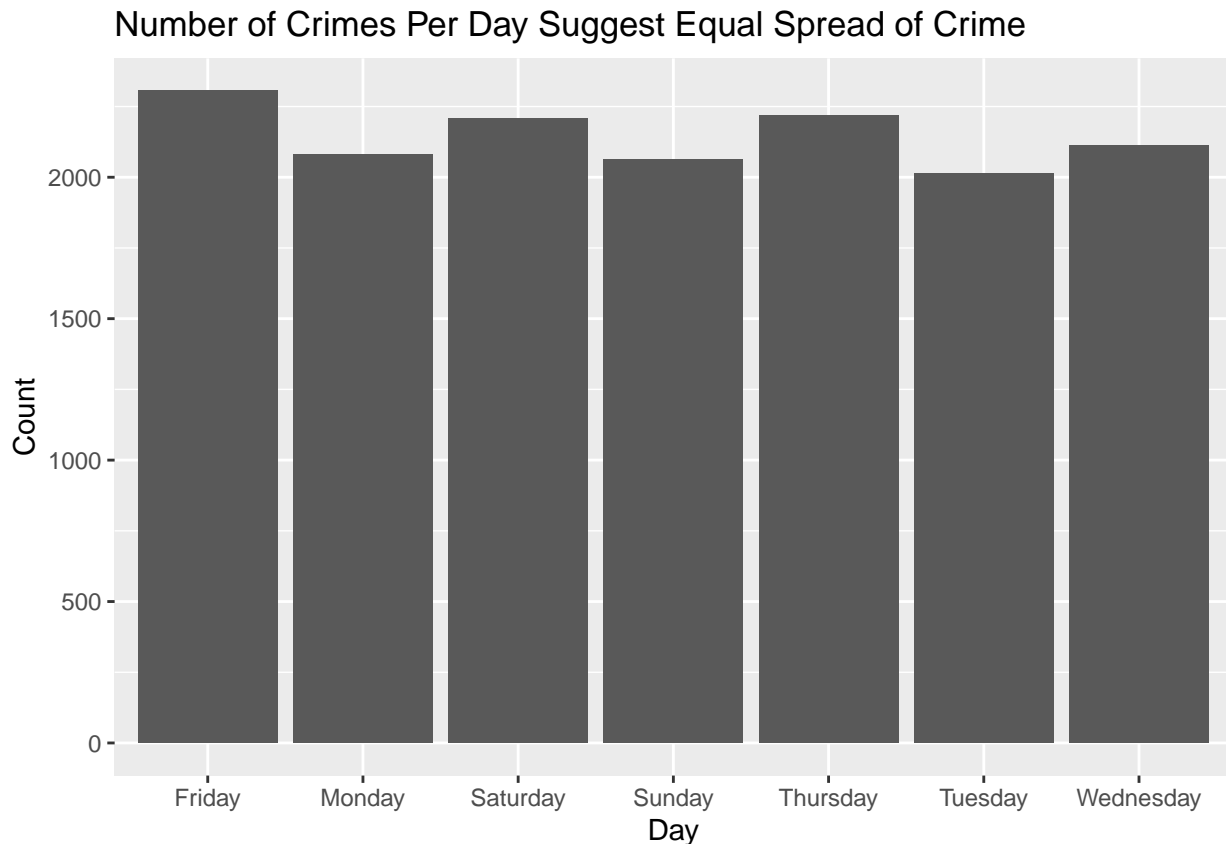
One relationship we were interested in was if certain days had a higher rates of crime. We visualized this relationship by creating a bar graph that compares the day of the week and number of crimes each day during this time period. By looking at the visual, we are able to see that each has a relatively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

```r
crimetypeday <-important%>%
  group_by(crimetype)%>%
  mutate(ctcount = n())
crimetypeday
```

```
## # A tibble: 14,999 x 10
## # Groups:   crimetype [7]
##    Category DayOfWeek Date  PdDistrict Resolution  hour timerange crimetype
##    <chr>    <chr>     <chr> <chr>      <chr>      <dbl> <chr>     <chr>
##  1 ROBBERY  Wednesday 09/2~ MISSION    NONE          23 evening   violence~
##  2 ASSAULT  Tuesday   02/1~ NORTHERN   NONE          18 evening   violence~
##  3 ASSAULT  Sunday    04/0~ MISSION    ARREST, B~     6 morning   violence~
##  4 ASSAULT  Thursday  01/2~ INGLESIDE  ARREST, B~    14 day       violence~
##  5 ASSAULT  Thursday  11/1~ SOUTHERN   NONE          21 evening   violence~
##  6 ASSAULT  Sunday    06/1~ BAYVIEW    NONE          12 day       violence~
##  7 ASSAULT  Friday    07/2~ NORTHERN   NONE          16 day       violence~
##  8 ASSAULT  Sunday    02/1~ TENDERLOIN NONE          15 day       violence~
##  9 ASSAULT  Sunday    02/2~ TARAVAL    ARREST, B~    10 morning   violence~
## 10 ASSAULT  Saturday  11/2~ BAYVIEW    NONE           1 night     violence~
## # ... with 14,989 more rows, and 2 more variables: yes_violent <lgl>,
## #   ctcount <int>
```

```r
ggplot(data = crimetypeday, mapping = aes(y = crimetype)) +
  geom_bar() + facet_wrap(~ DayOfWeek) +
  labs(
    x = "Count",
    y = "Type of Crime",
    title = "Type of Crime by Day Shows A Large Proportion of Crimes as Property Related or Miscellaneou
```

Type of Crime by Day Shows A Large Proportion of Crimes as Prop
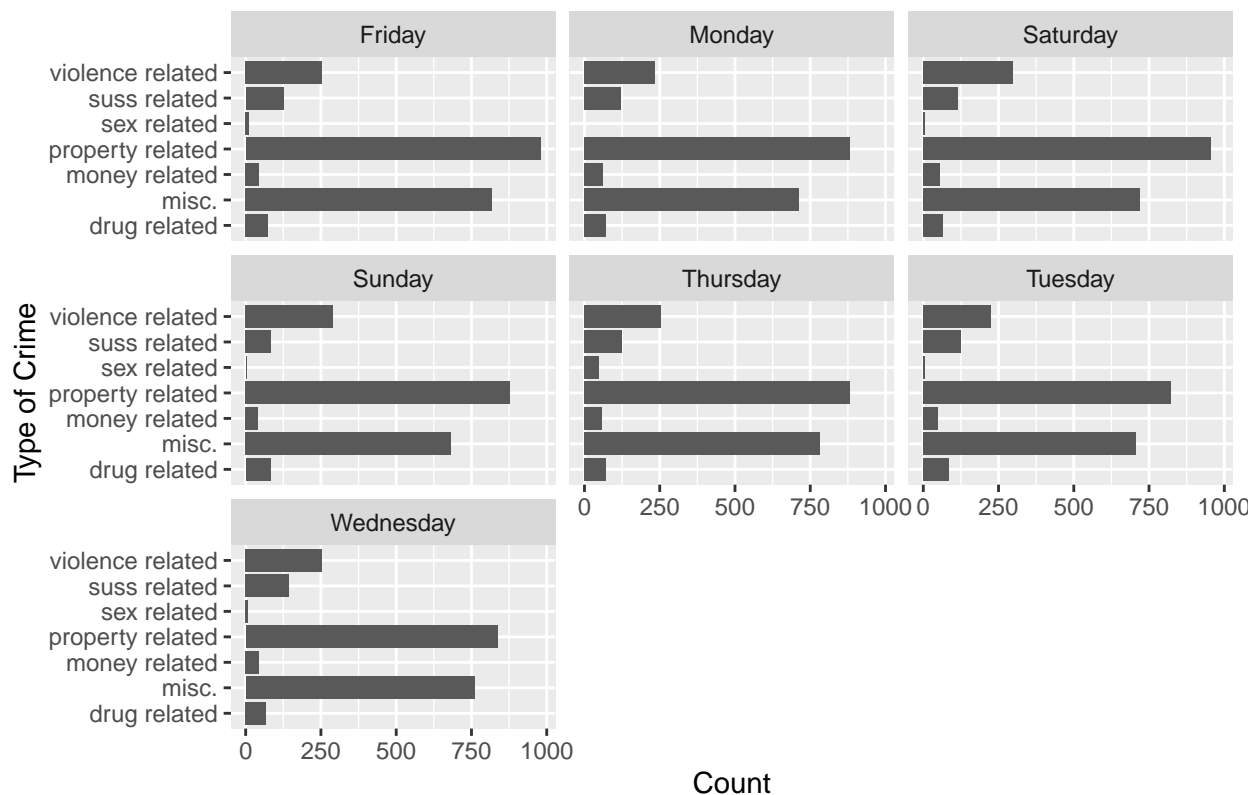

```
#need to fix crimetype names
# make miscellaneous crimes more specific?
```

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.

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.