# 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

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

```
## Rows: 15,000
## Columns: 13
## $ IncidntNum <chr> "160787544", "160289487", "160944233", "160427461", "160...
## $ Category   <chr> "LARCENY/THEFT", "NON-CRIMINAL", "ASSAULT", "SUSPICIOUS ...
## $ Descript   <chr> "GRAND THEFT FROM LOCKED AUTO", "LOST PROPERTY", "BATTER...
## $ DayOfWeek  <chr> "Wednesday", "Tuesday", "Saturday", "Monday", "Saturday"...
## $ Date       <chr> "09/28/2016 12:00:00 AM", "03/08/2016 12:00:00 AM", "11/...
## $ Time       <time> 10:45:00, 09:00:00, 21:30:00, 09:30:00, 05:29:00, 16:30...
## $ PdDistrict <chr> "RICHMOND", "INGLESIDE", "INGLESIDE", "SOUTHERN", "INGLE...
## $ Resolution <chr> "NONE", "NONE", "ARREST, BOOKED", "NONE", "NONE", "NONE"...
## $ Address    <chr> "2000 Block of POST ST", "100 Block of LISBON ST", "500 ...
## $ X          <dbl> -122.4356, -122.4308, -122.4354, -122.4017, -122.4124, -...
## $ Y          <dbl> 37.78489, 37.72584, 37.71968, 37.78417, 37.73406, 37.758...
## $ Location   <chr> "(37.784892281882, -122.435575820196)", "(37.72583961593...
## $ PdId       <dbl> 1.607875e+13, 1.602895e+13, 1.609442e+13, 1.604275e+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).

```r
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", "NON-CRIMINAL", "ASSAULT", "SUSPICIOUS ...
## $ DayOfWeek  <chr> "Wednesday", "Tuesday", "Saturday", "Monday", "Saturday"...
## $ Date       <chr> "09/28/2016 12:00:00 AM", "03/08/2016 12:00:00 AM", "11/...
## $ PdDistrict <chr> "RICHMOND", "INGLESIDE", "INGLESIDE", "SOUTHERN", "INGLE...
## $ Resolution <chr> "NONE", "NONE", "ARREST, BOOKED", "NONE", "NONE", "NONE"...
## $ hour       <dbl> 10, 9, 21, 9, 5, 16, 23, 10, 2, 11, 0, 17, 18, 0, 17, 19...
## $ timerange  <chr> "morning", "morning", "evening", "morning", "night", "da...
```

```
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", "NON-CRIMINAL", "ASSAULT", "SUSPICIOUS ...
## $ DayOfWeek  <chr> "Wednesday", "Tuesday", "Saturday", "Monday", "Saturday"...
## $ Date       <chr> "09/28/2016 12:00:00 AM", "03/08/2016 12:00:00 AM", "11/...
## $ PdDistrict <chr> "RICHMOND", "INGLESIDE", "INGLESIDE", "SOUTHERN", "INGLE...
## $ Resolution <chr> "NONE", "NONE", "ARREST, BOOKED", "NONE", "NONE", "NONE"...
## $ hour       <dbl> 10, 9, 21, 9, 5, 16, 23, 10, 2, 11, 0, 17, 18, 0, 17, 19...
## $ timerange  <chr> "morning", "morning", "evening", "morning", "night", "da...
## $ crimetype  <chr> "property related", "misc.", "violence related", "suss r...
```

$H_0$ : 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

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

```
## # A tibble: 7 x 2
##   crimetype            n
##   <chr>            <int>
## 1 drug related       515
## 2 misc.             5201
## 3 money related      357
## 4 property related  6291
## 5 sex related         58
## 6 suss related       790
## 7 violence related  1788
```

```
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         195
##  2 drug related  evening     131
##  3 drug related  morning     124
##  4 drug related  night        65
##  5 misc.         day        1908
##  6 misc.         evening    1421
##  7 misc.         morning    1203
##  8 misc.         night       669
##  9 money related day         141
## 10 money related evening      76
## # ... 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]),
```

```r
                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 = 372.17, 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 $\alpha$ 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.

### Relationship between time and crime (ie do crimes generally occur at night)? Tina

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

```r
# 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 ~ "night",
#                                 hour >= 6 & hour < 12 ~ "morning",
#                                 hour >= 12 & hour < 18 ~ "day",
#                                 hour >= 18 & hour < 24 ~ "evening"))
```

```
#
# sanfrancrime <- sanfrancrime %>%
#   group_by(Category) %>%
#   summarise(count = n())
# sanfrancrime
#
# ggplot(data = sanfrancrime, aes(x = timerange, y = count)) +
#   geom_bar() + facet_grid(~Category) + labs(x = "time", y = "Count of Crimes",
# title = "SSS")
```

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

ETC!

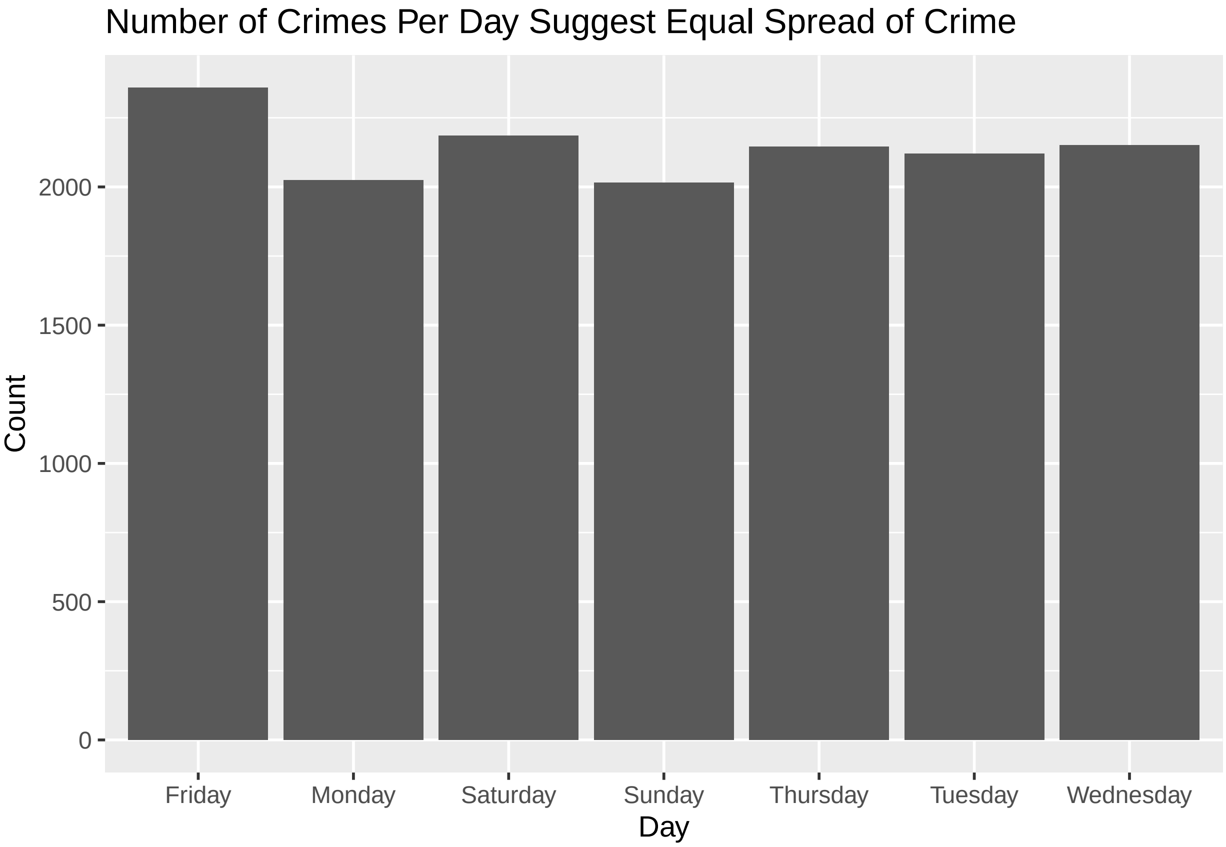
```
pd_violent <- sanfrancrime%>%
  group_by(Category)%>%
  count()
pd_violent<- sanfrancrime%>%
  mutate(violent_crime = case_when(
    Category == "ASSAULT" | Category == "SEX OFFENSES FORCIBLE" |
      Category == "ROBBERY" | Category == "KIDNAPPING" ~ "YES",
    Category != "ASSAULT" | Category != "SEX OFFENSES, FORCIBLE" |
      Category != "ROBBERY" | Category !="KIDNAPPING" ~ "NO"))
important2 <- important%>%
  group_by(PdDistrict)%>%
  mutate(yes_violent = crimetype == "violence related")%>%
  count(yes_violent)%>%
  mutate(perc = (n/sum(n) *100))%>%
  filter(yes_violent == TRUE)%>%
  arrange(desc(perc))


#make a ggplot with a stacked bar graph showing proportions##
```

### 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 Wednesday  2151
##  2 Tuesday    2121
##  3 Saturday   2185
##  4 Monday     2025
##  5 Saturday   2185
##  6 Monday     2025
##  7 Monday     2025
##  8 Wednesday  2151
##  9 Tuesday    2121
## 10 Monday     2025
## # ... 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")
```



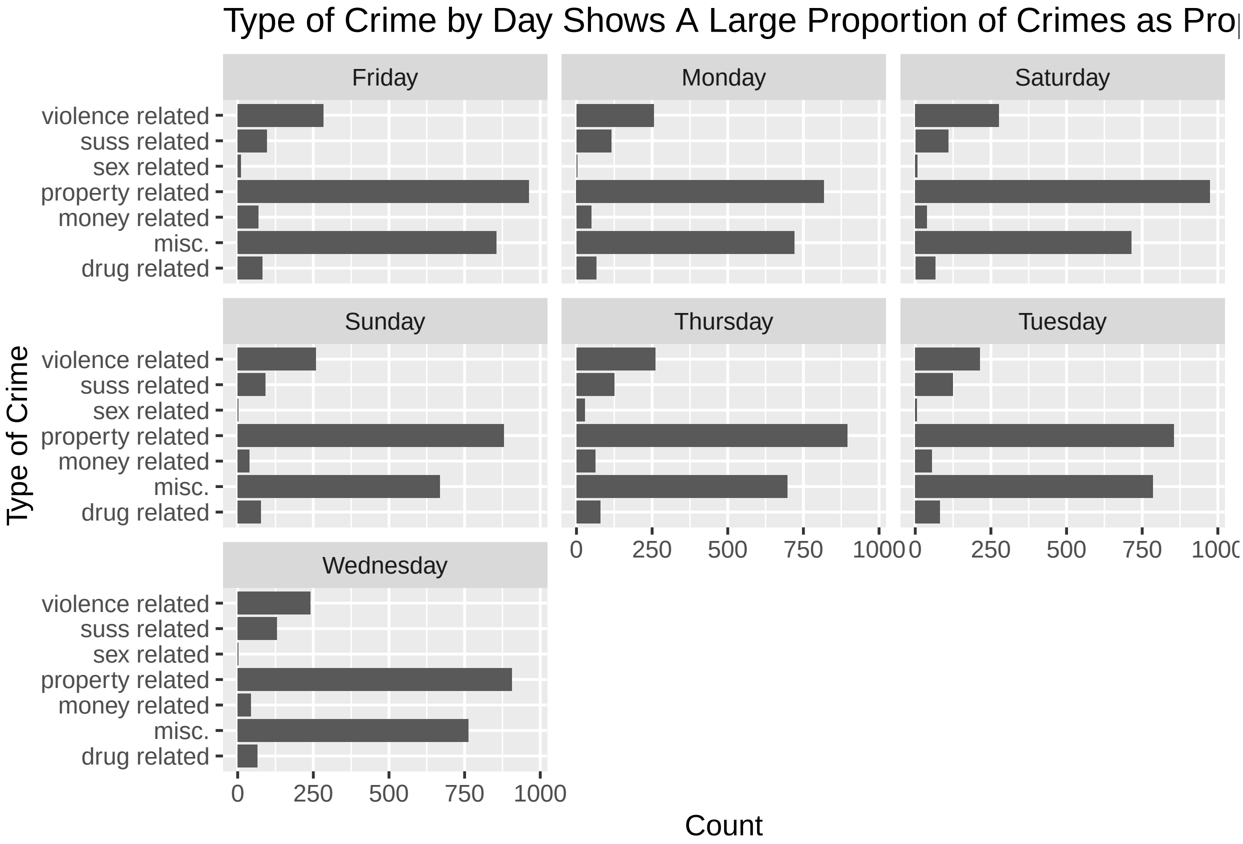
```
#need to change order of days of the week
```

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.

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

```
## # A tibble: 15,000 x 9
## # Groups:   crimetype [7]
##    Category DayOfWeek Date  PdDistrict Resolution  hour timerange crimetype
##    <chr>    <chr>     <chr> <chr>      <chr>      <dbl> <chr>     <chr>
##  1 LARCENY~ Wednesday 09/2~ RICHMOND   NONE          10 morning   property~
##  2 NON-CRI~ Tuesday   03/0~ INGLESIDE  NONE           9 morning   misc.
##  3 ASSAULT  Saturday  11/1~ INGLESIDE  ARREST, B~    21 evening   violence~
##  4 SUSPICI~ Monday    05/2~ SOUTHERN   NONE           9 morning   suss rel~
##  5 NON-CRI~ Saturday  03/0~ INGLESIDE  NONE           5 night     misc.
##  6 ASSAULT  Monday    07/1~ MISSION    NONE          16 day       violence~
##  7 WEAPON ~ Monday    05/0~ BAYVIEW    ARREST, B~    23 evening   misc.
##  8 OTHER O~ Wednesday 07/2~ PARK       NONE          10 morning   misc.
##  9 LARCENY~ Tuesday   08/0~ SOUTHERN   NONE           2 night     property~
## 10 SUSPICI~ Monday    01/2~ NORTHERN   NONE          11 morning   suss rel~
## # ... with 14,990 more rows, and 1 more variable: ctcount <int>
```

```
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
```



```
#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.