San Francisco Crime Analysis

Introduction

Nowadays we live in a data-driven society, with different problems and challenges. Therefore modern problems, required modern solutions. Gladly , today we have many tools that can help us to solve troubles. One of this tools is Data science, that can gather insights to help us understand a problematic and help us to find solutions to them.

The data generated each day is exponentially growing everyday it passes and it's begging for analysts to dive deep in it, and make use of the raw input in an impactful way. We can now fight crime as a vigilante, all from behind a computer by exploring and analyze crime data sets.

Objetive

In this project we will explore to the most deep and dubious streets of San Francisco in order to understand the relationship between reported crime incidents by civilians and police officers.

Data

This particular data set can be found at Kaggle and also at the San Francisco open data webpage, here, where you can find more interesting data about the city.

```
library(lubridate)
library(tidyverse)
library(ggplot2)
library(ggmap)

incidents <- read.csv("incidents.csv")
calls <- read_csv("calls.csv")

glimpse(calls)
```

```
Rows: 100,000
Columns: 15
$ ‘Crime Id‘        <dbl> 163003307, 180870423, 173510362, 163272811, 17281...
$ Descript          <chr> "Bicyclist", "586", "Suspicious Person", "911 Dro...
$ ‘Report Date‘     <dttm> 2016-10-26, 2018-03-28, 2017-12-17, 2016-11-22, ...
$ Date              <dttm> 2016-10-26, 2018-03-28, 2017-12-17, 2016-11-22, ...
$ ‘Offense Date‘    <dttm> 2016-10-26, 2018-03-28, 2017-12-17, 2016-11-22, ...
$ ‘Call Time‘       <dbl> 0.74097222, 0.24236111, 0.12500000, 0.73541667, 0...
$ ‘Call Date Time‘  <dttm> 2016-10-26 17:47:00, 2018-03-28 05:49:00, 2017-1...
$ Disposition       <chr> "GOA", "HAN", "ADV", "NOM", "GOA", "ADV", "REP", ...
$ Address           <chr> "The Embarcadero Nor/kearny St", "Ingalls St/van ...
$ City              <chr> "San Francisco", "San Francisco", "San Francisco"...
$ State             <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "...
$ ‘Agency Id‘       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ ‘Address Type‘    <chr> "Intersection", "Intersection", "Intersection", "...
$ ‘Common Location‘ <chr> NA, NA, NA, NA, NA, NA, "Midori Hotel Sro #612, S...
$ X15               <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
```

```
glimpse(incidents)
```

```
Rows: 84,000
Columns: 13
$ IncidntNum <int> 176122807, 160569314, 160362475, 160435298, 90543656, 18...
$ Category   <chr> "LARCENY/THEFT", "ASSAULT", "ROBBERY", "KIDNAPPING", "MI...
$ Descript   <chr> "GRAND THEFT FROM UNLOCKED AUTO", "BATTERY", "ROBBERY, B...
$ DayOfWeek  <chr> "Saturday", "Thursday", "Tuesday", "Friday", "Tuesday", ...
$ Date       <chr> "2017-05-13T00:00:00Z", "2016-07-14T00:00:00Z", "2016-05...
$ Time       <chr> "10:20:00", "16:00:00", "14:19:00", "23:57:00", "07:40:0...
$ PdDistrict <chr> "SOUTHERN", "MISSION", "NORTHERN", "SOUTHERN", "TARAVAL"...
$ Resolution <chr> "NONE", "NONE", "ARREST, BOOKED", "ARREST, BOOKED", "LOC...
$ Address    <chr> "800 Block of BRYANT ST", "MISSION ST / CESAR CHAVEZ ST"...
$ X          <dbl> -122.4034, -122.4182, -122.4299, -122.4050, -122.4612, -...
$ Y          <dbl> 37.77542, 37.74817, 37.77744, 37.78512, 37.71912, 37.806...
$ Location   <chr> "{'latitude': '37.775420706711', 'human_address': '{\"ad...
$ PdId       <dbl> 1.761228e+13, 1.605693e+13, 1.603625e+13, 1.604353e+13, ...
```

There has to be relationship between incidents reported by civilians and officers by the date on which the incidents were documented, so let's combine this information. By joining both datasets the structure preserves only days on which both civilians reported incidents and police encountered incidents.

```
daily_inc <- incidents %>%
    count(Date, sort = TRUE) %>%
    rename(n_incidents = n)
daily_cal <- calls %>%
    count(Date, sort = TRUE) %>%
    rename(n_calls = n)

daily_inc$Date <- as_datetime(daily_inc$Date)

df <- inner_join(daily_inc, daily_cal, by = c("Date" = "Date"))
glimpse(df)
```

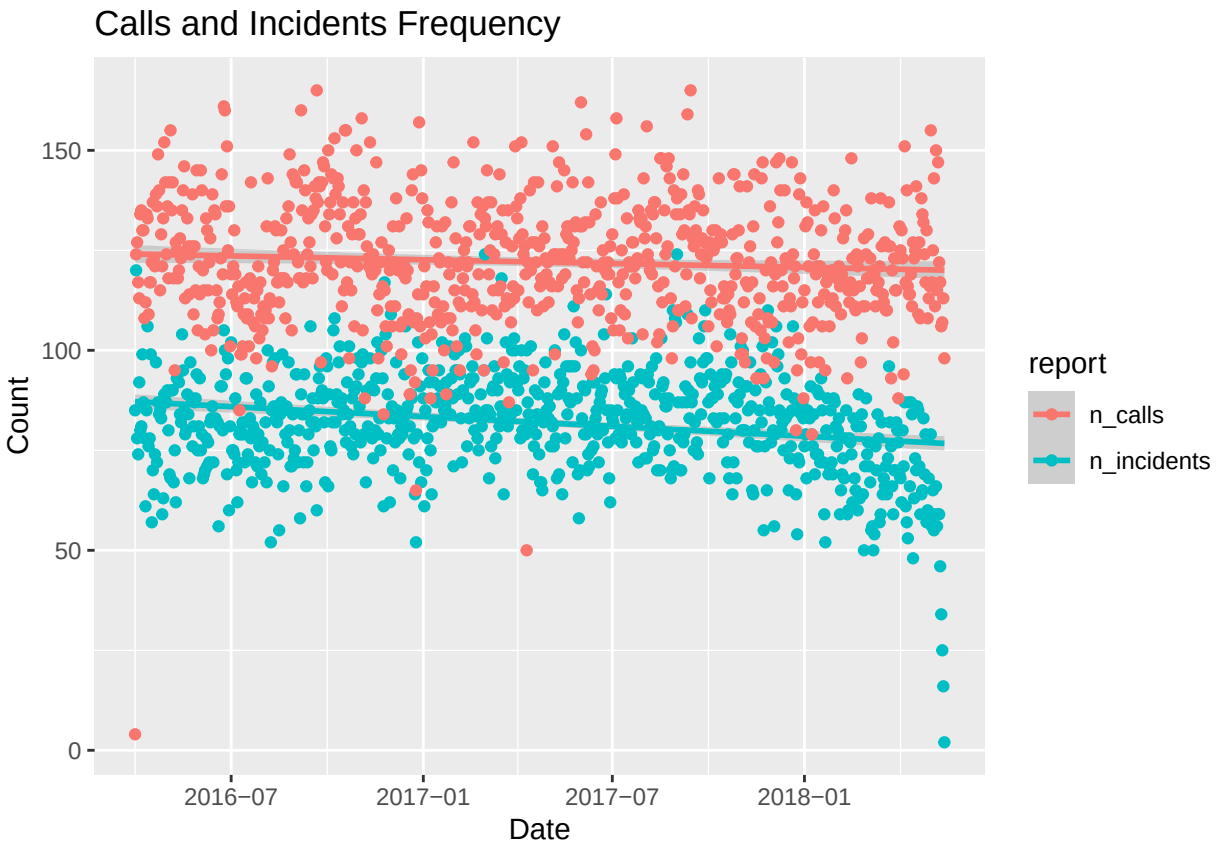
```
Rows: 776
Columns: 3
$ Date        <dttm> 2017-03-01, 2017-09-01, 2016-04-01, 2017-03-17, 2016-1...
$ n_incidents <int> 124, 124, 120, 118, 117, 114, 111, 110, 110, 110, 109, ...
$ n_calls     <int> 133, 138, 124, 129, 115, 120, 127, 106, 124, 108, 111, ...
```

Exploratory Data Analysis

We need to find a way to search and comprehend crime rates. We can do this by looking at the frequency of calls and incidents across time to help discern if there is a relationship between these variables.

```
plot <- df %>%
  gather(key = report, value = count, -Date)

ggplot(plot, aes(x = Date, y = count, color = report)) +
      labs(x='Date', y='Count', title='Calls and Incidents Frequency') +
      geom_smooth(method = "lm", formula = y ~ x) +
      geom_point()
```



A quantitative way to determine the relationship between 2 variables is to calculate the correlation coefficient between them, in other words, this number represents the linear dependence between two data sets. Firstly, let's calculate the daily count coefficient and subsequently take a broad view of the trends by summarizing the data into monthly counts and calculate the coefficient.

```
daily_cor <- cor(df$n_incidents, df$n_calls)

cor_df <- df %>%
  mutate(month = month(Date)) %>%
  group_by(month) %>%
  summarize(n_incidents = sum(n_incidents),
            n_calls = sum(n_calls))

monthly_cor <- cor(cor_df$n_incidents, cor_df$n_calls)

print('Correlation coefficient between daily frequencies:')
```

```
[1] "Correlation coefficient between daily frequencies:"
```

```
daily_cor
```

```
[1] 0.1469688
```

```
print('Correlation coefficient between monthly frequencies:')
```

```
[1] "Correlation coefficient between monthly frequencies:"
```

```
monthly_cor
```

```
[1] 0.970683
```

It will be helpful to have all the information from each police reported incident and each civilian call on their shared dates so we can calculate similar statistics from each dataset and compare results.

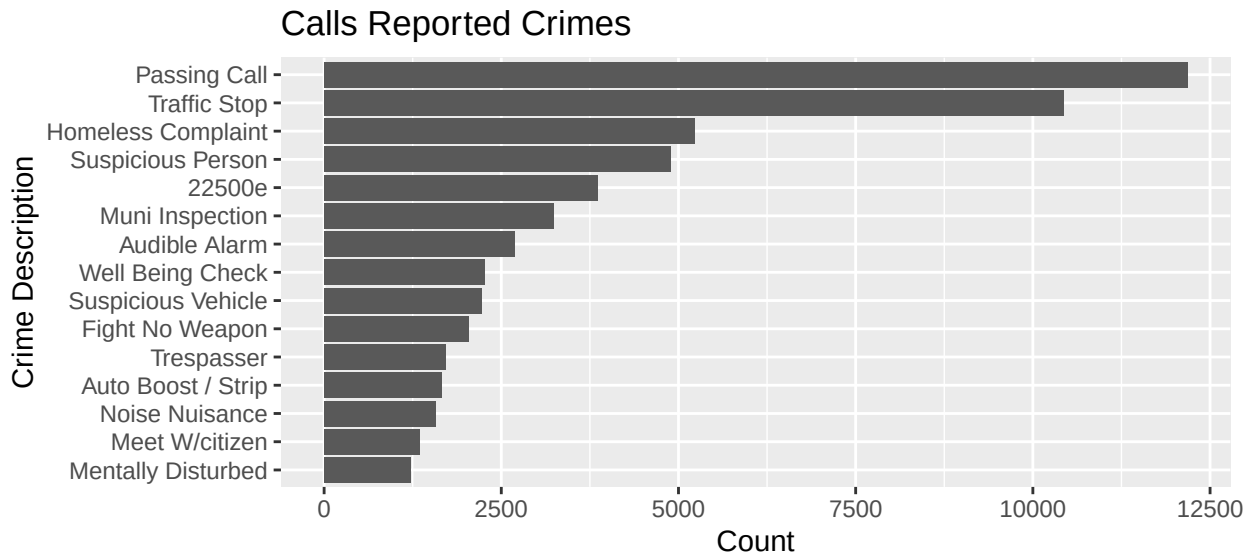
```
## [1] TRUE
```

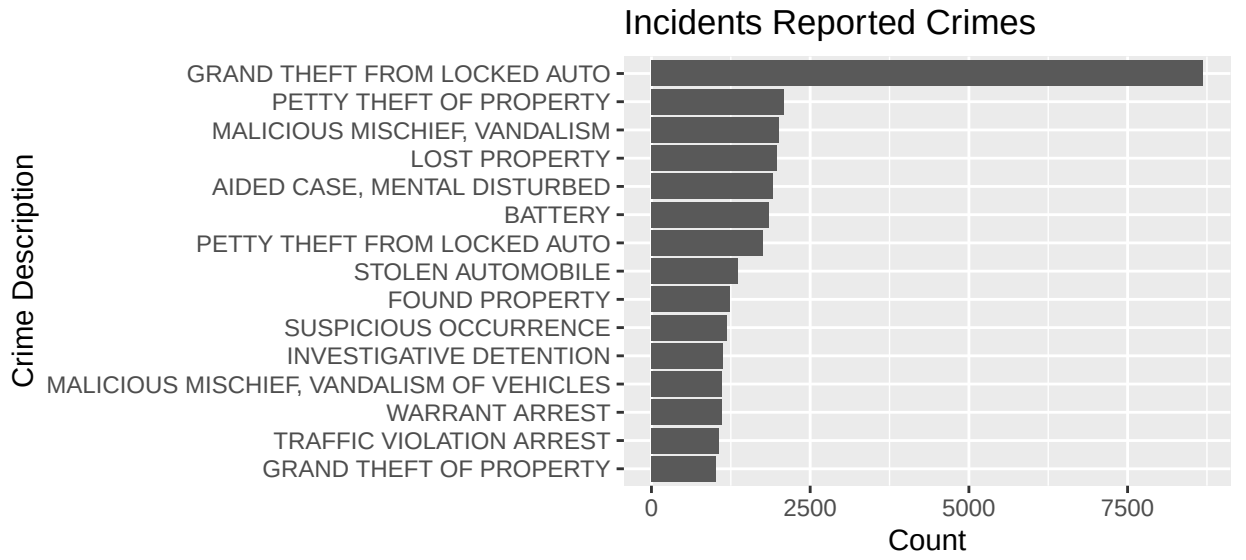
After some data manipulation we are ready to see some results and get interesting information. For searching trends in categorical data we need to understand the level of importance each category has. We can visualize this by ranking the order of the each category into a bar chart.

```
plot_calls_freq <- calls_dates %>%
  count(Descript) %>%
  top_n(15, n) %>%
  ggplot(aes(x = reorder(Descript, n), y = n)) +
  geom_bar(stat = 'identity') +
  ylab("Count") +
  xlab("Crime Description") +
  ggtitle("Calls Reported Crimes") +
  coord_flip()

plot_incidents_freq <- incidents_dates %>%
  count(Descript) %>%
  top_n(15, n)  %>%
  ggplot(aes(x = reorder(Descript, n), y = n)) +
  geom_bar(stat = 'identity') +
  ylab("Count") +
  xlab("Crime Description") +
  ggtitle("Incidents Reported Crimes") +
  coord_flip()
```

As you can visualize "Grand theft from locked auto" is the crime with highest incidence by far. However "Auto boost/Strip" falls way behind in the reported by civilian chart, which made me think that people are aware of the problem and try to help each other by preventing the crime. Yet, this is probably only the case where the location of a "called in crime" is similar to the crime incidence location. Let's check to see if the locations of the most frequent civilian reported crime and police reported crime are similar.





It appears the datasets share locations where auto crimes occur and are reported most frequently - such as on Point Lobos Avenue, Lyon Street, and Mission Street. It would be great to plot co-occurrence of these locations to visualize overlap, however we only have longitude and latitude data for police reported incidents. No matter, it will still be very valuable to inspect the frequency of auto crime occurrence on a map of San Francisco. This will give us immediate insight as to where auto crimes occur. Most importantly, this visualization will provide a powerful means of communication.

```
location_calls <- calls_dates %>%
  filter(Descript == "Auto Boost / Strip") %>%
  count(Address) %>%
  arrange(desc(n))%>%
  top_n(10, n)

location_incidents <- incidents_dates %>%
  filter(Descript == "GRAND THEFT FROM LOCKED AUTO") %>%
```

```
  count(Address) %>%
  arrange(desc(n))%>%
  top_n(10, n)

location_calls
```

```
# A tibble: 11 x 2
   Address                                n
   <chr>                              <int>
 1 1100 Block Of Point Lobos Av          21
 2 3600 Block Of Lyon St                 20
 3 100 Block Of Christmas Tree Point Rd  18
 4 1300 Block Of Webster St              12
 5 500 Block Of 6th Av                   12
 6 800 Block Of Vallejo St               10
 7 1000 Block Of Great Hy                 9
 8 100 Block Of Hagiwara Tea Garden Dr    7
 9 1100 Block Of Fillmore St              7
10 3300 Block Of 20th Av                  7
11 800 Block Of Mission St                7
```

```
location_incidents
```

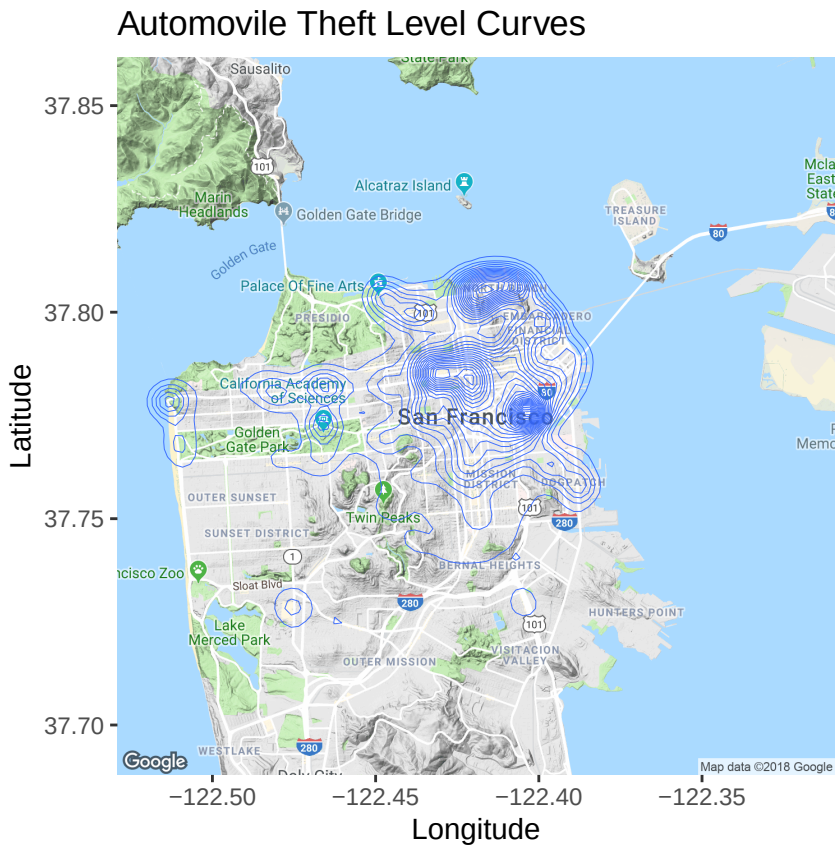
```
                       Address   n
1        800 Block of BRYANT ST 441
2  500 Block of JOHNFKENNEDY DR  89
3   1000 Block of POINTLOBOS AV  84
4       800 Block of MISSION ST  61
5        2600 Block of GEARY BL  38
6         3600 Block of LYON ST  36
7      1300 Block of WEBSTER ST  35
8     1100 Block of FILLMORE ST  34
9        22ND ST / ILLINOIS ST  33
10         400 Block of 6TH AV  30
```

It seems the datasets share locations where auto crimes occur and are reported most frequently, for example: Mission Street, Lyon Street and Point Lobos Avenue.

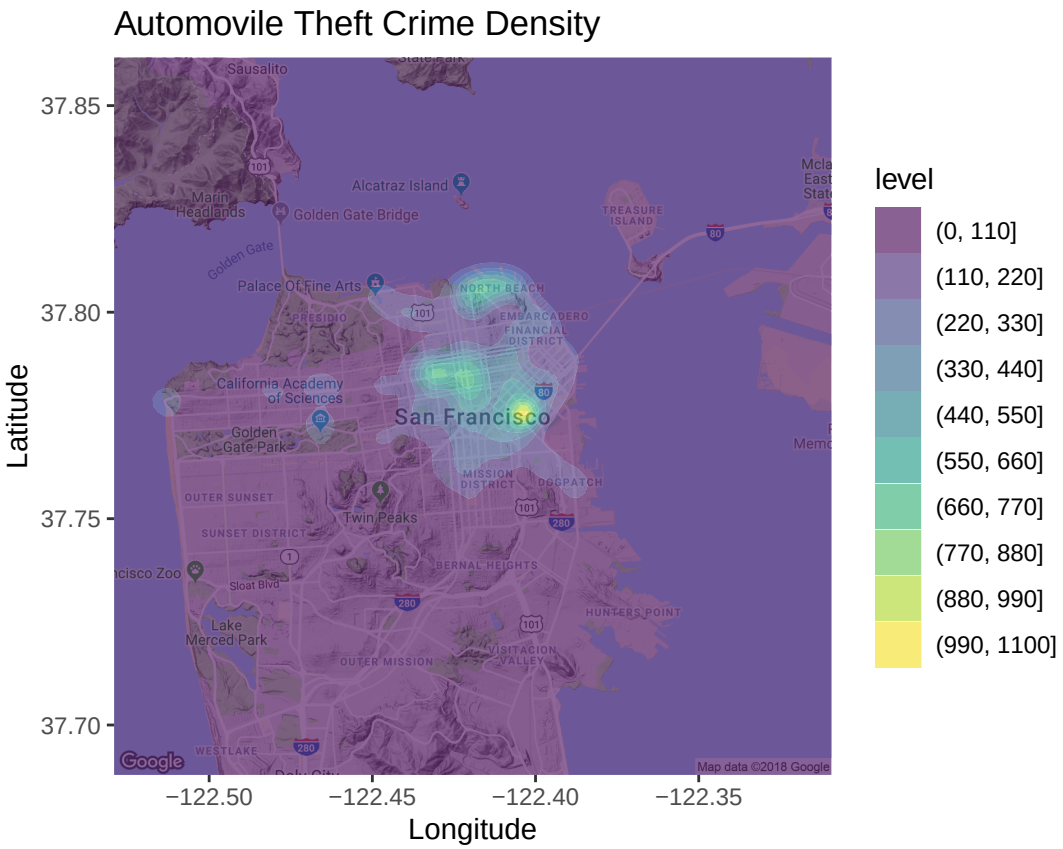
This is helpful information, however a visualization worths more than a thousand words. So, by plotting the occurrence of these locations to visualize the frequency of auto crime occurrence on a map of San Francisco. This will give us rapid insight of where auto crimes occur around San Francisco.

```
sf_map <- readRDS("sf_map.RDS")
auto_crime <- incidents_dates %>%
  filter(Descript == "GRAND THEFT FROM LOCKED AUTO")

ggmap(sf_map) +
  stat_density_2d(
    aes(x = X, y = Y), alpha = 0.9,
    size = 0.01, bins = 30, data = auto_crime,
    geom = "density_2d") + labs(x = 'Longitude', y = 'Latitude') +
    ggtitle('Automovile Theft Level Curves')
```




```
ggmap(sf_map) +
  geom_density_2d_filled(
    aes(x = X, y = Y, fill = ..level..), alpha = 0.60,
    size = 0.1, bins = 10, data = auto_crime,
    geom = "density_2d") + labs(x = 'Longitude', y = 'Latitude') +
    ggtitle('Automovile Theft Crime Density')
```



Conclusions and Recommendations

Data Science combines programming skills, statistics and math to extract meaningful insights from data. Whichever tool you prefer, it is often important for analysts to work with similar platforms so that they can share their insights.

These insigths can help the population to be informed about particular circumstances that surrounds them like crime rates around San Francisco. Our final visualization can provide powerful means of communication and make the difference in this world. There is still work to do, if you are curious enough, you can map other crime types and gather information about the crime arround the city.