# Crime Incidents reported in San Francisco over the years.

From the Police Departments' perspective.

## Abstract

Unfortunately, the occurrence of crime is a part of everyones' life. Various techniques have been used to understand about these crimes and come up with solutions which would increase the safety of the citizens. But, another way of understanding the data is through the perspective of the police department itself. The visualizations showed that the department was the busiest on Fridays as most of the incidents were reported on Fridays. Further Southern, Mission and Northern Police Districts got the highest number of cases reported to.  Finally over the period of time the most reported crimes were Theft, Vandalism and Burglary. Therefore assuming that the available data is accurate we can conclude that more officers should be working for the Southern police district. Also, the number of people in Friday's shift should be highest and a better plan should be used to reduce the top committed crimes around the city.

## Data Collection

My data is the collection of reported crime incidents in San Francisco City from the year 2003 till Present. The dataset satisfies two aspects of Big Data i.e, Volume and Velocity.

My dataset was the combination of two datasets provided by the Government of San Francisco, the first one being the data of crimes reported from 2003 to May, 2018. This dataset takes over 800 MB of storage and contains data spread across 35 different columns with a total of 2,160,953 records. The second dataset consists of data spread across 36 columns where the records are updated everyday at 10:00 AM Pacific time. I took the dataset updated till 10th Dec, 2020 which has a total  413,183 records in total and took over 300 MB of storage space. The number of records present in the datasets made it impossible for me to open it on my local machine which therefore fulfills the Volume aspect of Big Data and the frequency of data updation fulfills the Velocity aspect of it.

Even though I am merging two different datasets I don't consider it to satisfy the Variety aspect of Big Data as both the datasets are in structured format with more or less the same columns.

## Data Exploration, Processing, Cleaning and/or Integration

*What did you need to do to prepare the dataset(s) to create your graph/chart?*

In order to get the desired dataset, I cleaned the data first using OpenRefine where most of the name cleaning was done. To load the first dataset i.e, from 2003 to May, 2018 I had to divide it into two to successfully load it in OpenRefine. The division was done using the inbuilt function of the tool. The name cleaning included removal of whitespaces in the values, and converting the values from capital letters to small or vice versa to make them uniform across the dataset. Furthermore, in the dataset from 2018 to present there were records which had only null values or dummy variables in them, which I assumed were recorded to test the new reporting system therefore these records were also removed from the dataset. Both the dataset had the same values recorded in their columns but the names of the columns were different, I made changes to the column names to make it uniform for the across datasets. Final output of the cleaning process using the tool were three CSV files : two for data from 2003 to May 2018 and the other for data from 2018 till present.

### *How did you choose the attributes and data subset to visualise?*

The three CSV files obtained from the cleaning process were then further scrutinized to get the subset for further visualization.

For the first visualisation, I needed only the “DayOfWeek” column with the count of incidents reported on these days. Though the number of records were high, upon reading about the  first dataset I realised that to update about the incidents (Ex. Case closed, Arrested etc.) they made use of another record with the same incident number and updated the Description column of it. So I had to get only the first occurence of the incidents numbers to make sure I am getting new reported incidents. In the second dataset, the column “Row Type Description” had the information about the new incidents or the updated incidents. I filtered out the records which contained the “Initial” string in their “Row Type Description” which indicated the first reported incident. Finally, I combined the two datasets and created a new dataframe with two columns where one contained 7 days of week and the other had the count of incidents reported on that day.

Similarly for the next visualisation, I combined two datasets and calculated the number of incidents reported based on the different incident categories. From this I selected the top 5 Incident categories with highest reported incidents to be visualised and created a subset dataframe with only 4 columns i.e, Category, Latitude, Longitude and Incident Description.

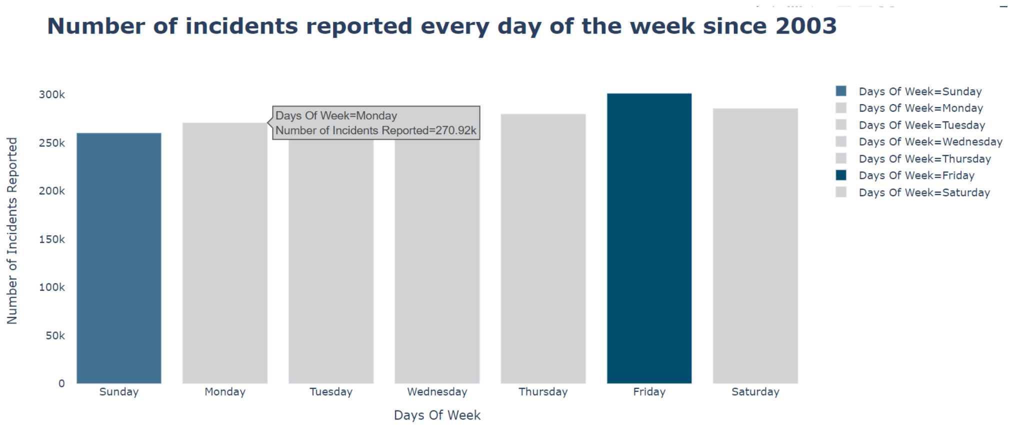
Finally for the final visualization, I again selected only the first occurence of the incident numbers and added a new column “Year” with the year in which the incident was reported. The year was extracted from the “Date and Time” columns. Further, I created a subset dataframe with three columns “Year”, “Police District” and “Count” which had the count of incidents reported to every police district for every year from 2003 till date. Finally, another column “Color” was added to represent different colors for each “Police District”.

## Visualisation

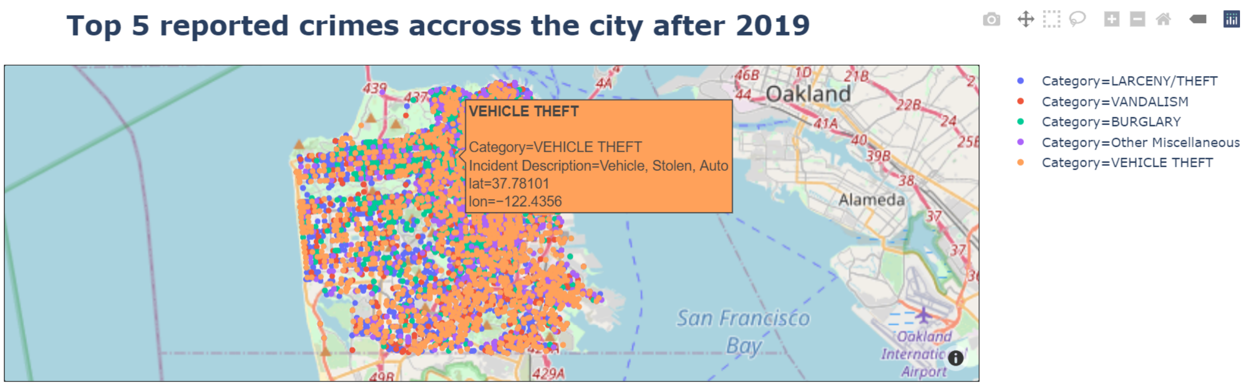
### *Graphs and Design choices*

The first visualization is an interactive graph which shows the number of cases reported each day of the week since 2003 till present. In my opinion, a bar chart is the best way of showing this visualisation as it provides a minimalistic approach in comparing different categories of data. It has a hover-over option where it gives you the number of incidents reported for any particular bar. Further, from the right side of the graph one can select the

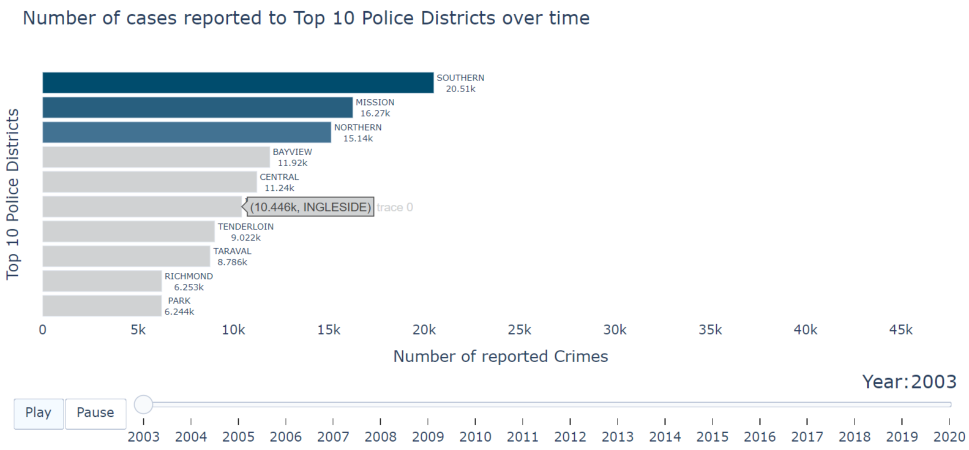
"Days of Week" upon which the data for that "Day" is hidden or shown from the graph. The *conclusion* that can be drawn from this graph is that more attention needs to be taken on Fridays and probably more force must be deployed on this day. Whereas a conclusion can be made that as most of the people are at home on Sundays less crimes are reported that day.



The next visualization is an interactive one where I wanted to show the top five types of crimes reported across the city. As I had the Latitude and the Longitude of the incidents I made use of a scatter plot on the map of the city for my visualization as it is easy to understand about various locations and the crimes committed in those areas. There is a hover-over option which upon pointing to any of the points will give the "Category", "Location" and "Description" of the crime. On the right, there are the top 5 crimes upon selecting which we can choose to "Show" or "Hide" those categories from the graph. As there are lots of records to be plotted, only the incidents reported after 2019 were considered for this visualization. The *conclusion* which can be drawn from this visualization is that extra preparation and planning needs to be done to control the top crimes reported and also the areas where which are not the safest can be found using this visualisation.



The final Visualization is an animated interactive graph which shows the number of cases reported to the "Top 10 Police Districts" since 2003 till date. The inspiration of using a bar chart for this task came from a YouTube Channel where I came across the usage of a Bar chart to visualize the data over a period of time. There is a hover-over option for this visualisation as well which will show the name of the police district and the count of the reported cases for the district for that particular year. I have placed a “Play” and “Pause” button upon clicking on them it starts or pauses the animation respectively. There is a slider in the bottom which can be used to select the year of choice which further will update the graph related to that year. And to emphasize the top three districts I have made use of shades of blue where the darker the color represents the higher value and to make sure that these values stand out I have used a lighter shade of grey to represent other districts. A *conclusion* that can be drawn from this graph would be that more number of officers are required to the top 3 districts as they have more number of cases to handle.



### Tools used

I made use of an open source Python called Plotly for all my visualizations. For the first two visualisations I made use of Plotly’s Express library and for the final racing bar chart I made use of Plotly’s Graph Object library.

# Conclusion

While overall I was satisfied with the visualization outcomes from the dataset I feel that there are still few aspects of it which could be improved upon to make the visualization much better.

In the first two visualizations the legends on the left side of the graphs can be minimised to only the categories and the extra string can be omitted. This would make the visualisation look much more neat than present.

In the second visualisation, initially I wanted to plot a map visualisation with a drop down menu for the categories from where one could select the category and the year and it would update the graph. My inspiration for doing such visualisation was this. But I was not able to do it as Plotly has a limit (applicable for its free version) to the number of points which could be plotted for such a design. Because of this reason I reduced the number of incidents and took only the incidents reported since 2019 because otherwise there would be too many points in the graph and it would make it difficult to understand it. Furthermore, there were few records with no locations, latitude or longitude values at all and I found it difficult to overcome this situation and finally had to drop these incidents.

# References

1. As I was new to Plotly I made use of their documentation to figure out it how it worked. (https://plotly.com/python/)
2. To understand the concept of colors to be used for visualisation I made use of these two articles:
    a. https://www.dataquest.io/blog/what-to-consider-when-choosing-colors-for-data-visualization/
    b. https://chartio.com/learn/charts/how-to-choose-colors-data-visualization/
3. For my last visualization I referred to this article and made appropriate changes according to my requirements to get the desired output. https://towardsdatascience.com/making-a-bar-chart-race-plot-using-plotly-made-easy-8dad3b1da955