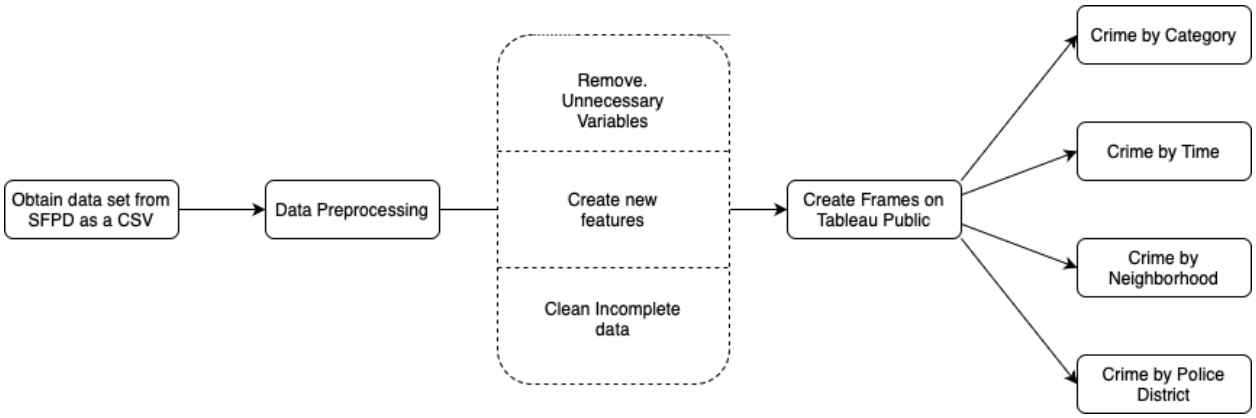
Temporal and Spatial Analysis of Crimes in San Francisco (2018-2023)

Gian Carlo L. Baldonado
Jarrett Zapata
Sanket Shah

Department of Computer Science
San Francisco State University
CSC 805 Data Visualization: Concepts , Tools, Techniques, and Paradigms Spring 2023
Prof. Shahrukh Humayoun
Phase 3 Project Implementation
May 11, 2023

I. System Architecture



The diagram above displays the architecture of our visualization system. We obtained "Police Department Incident Reports: 2018 to Present" data from DataSF as our dataset and saved it as a CSV file. Then we preprocessed the data for our key questions by removing unnecessary variables, creating new features and cleaning incomplete data. We utilized Tableau to develop four unique dashboards that address questions within their respective categories of crime by category, crime by time, crime by neighborhood and crime by police district.

II. Dataset Description

A. Removal of Unnecessary Columns

For data pre-processing, we have removed the following columns because we believe they do not provide information that can further our crime analysis. Most of these columns are also identifiers for the San Francisco Police Department's database, specific text description of crimes that can be substituted with the category or subcategory of crime, or simply the values of the columns are mostly null.

- CAD Number
- Incident ID
- Row ID
- Report Type Code
- Report Type Description
- Filed Online
- Incident Code
- Incident Description
- CNN
- Supervisor District
- Supervisor District 2012
- Neighborhoods

- ESNCAG - Boundary
- Current Supervisor Districts
- Current Police Districts

B. Creation of new columns

We created the following columns based on existing columns Incident Time and Report Time because we believe it can aid us in answering the questions we have above:

- Categorize whether a crime happened day or night using time by hour column. During this step, crimes committed between 10 PM to 6 AM are registered as crime committed at night
- Time taken to report a crime

C. Cleaning Incomplete Data

Some rows in the dataset had missing variables for some columns and hence, those rows were not suitable for data analysis. So, they were removed from the dataset.

III. System Description

Our goal for the San Francisco temporal and spatial crime analysis is to create four dashboards that provide insights into our key questions from Phase 1. These include identifying the most common times and locations for crimes, patterns of crime occurrence on specific days, and changes in crime rates and types over time. By creating these dashboards, we aim to gain insights into the temporal and spatial crime analysis in the city. These are our key questions below:

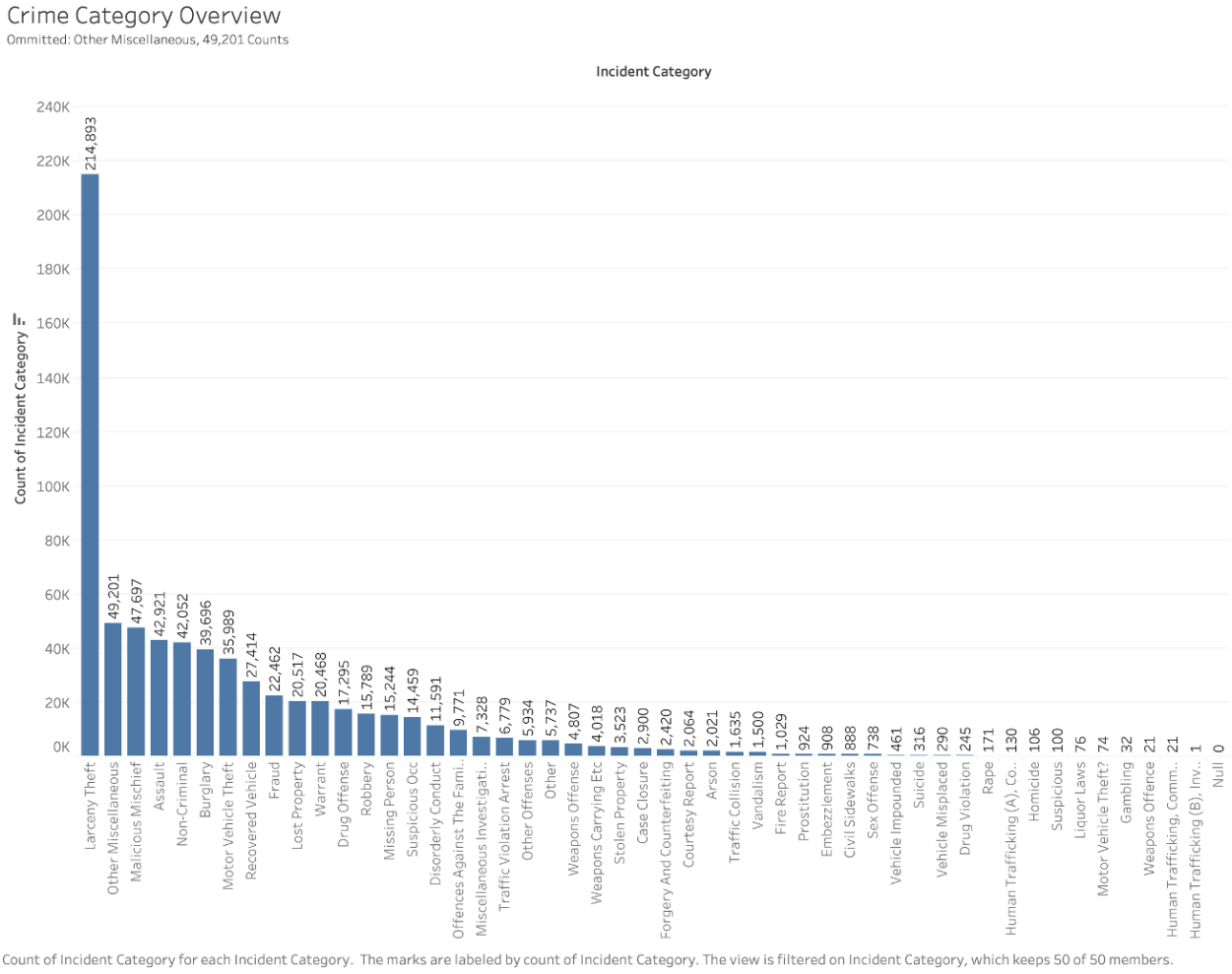
1. Is there a specific time when crimes typically happen?
2. Which neighborhoods are crimes most concentrated?
3. What days do crimes happen frequently?
4. Are crimes increasing or decreasing yearly?
5. What kinds of crimes are increasing or decreasing?

A. Technologies used

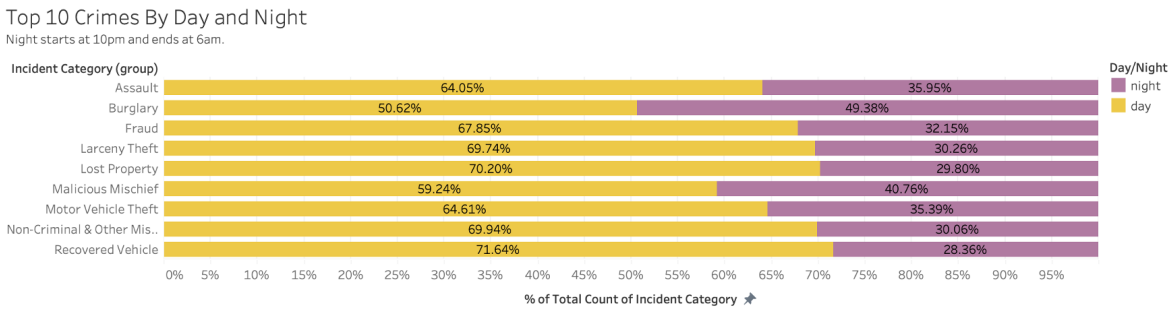
- Python: We preprocessed the data on python
- Jupyter: Jupyter notebooks were created to facilitate the preprocessing
- NumPy: Numpy is a library in python that we used for numerical manipulations
- Pandas: Pandas library in python handles the management of datasets in form of data frame variables
- Tableau: The visualizations were carried out on Tableau

IV. Screenshots Visualization System

a. Crime: Categorical Overview Dashboard

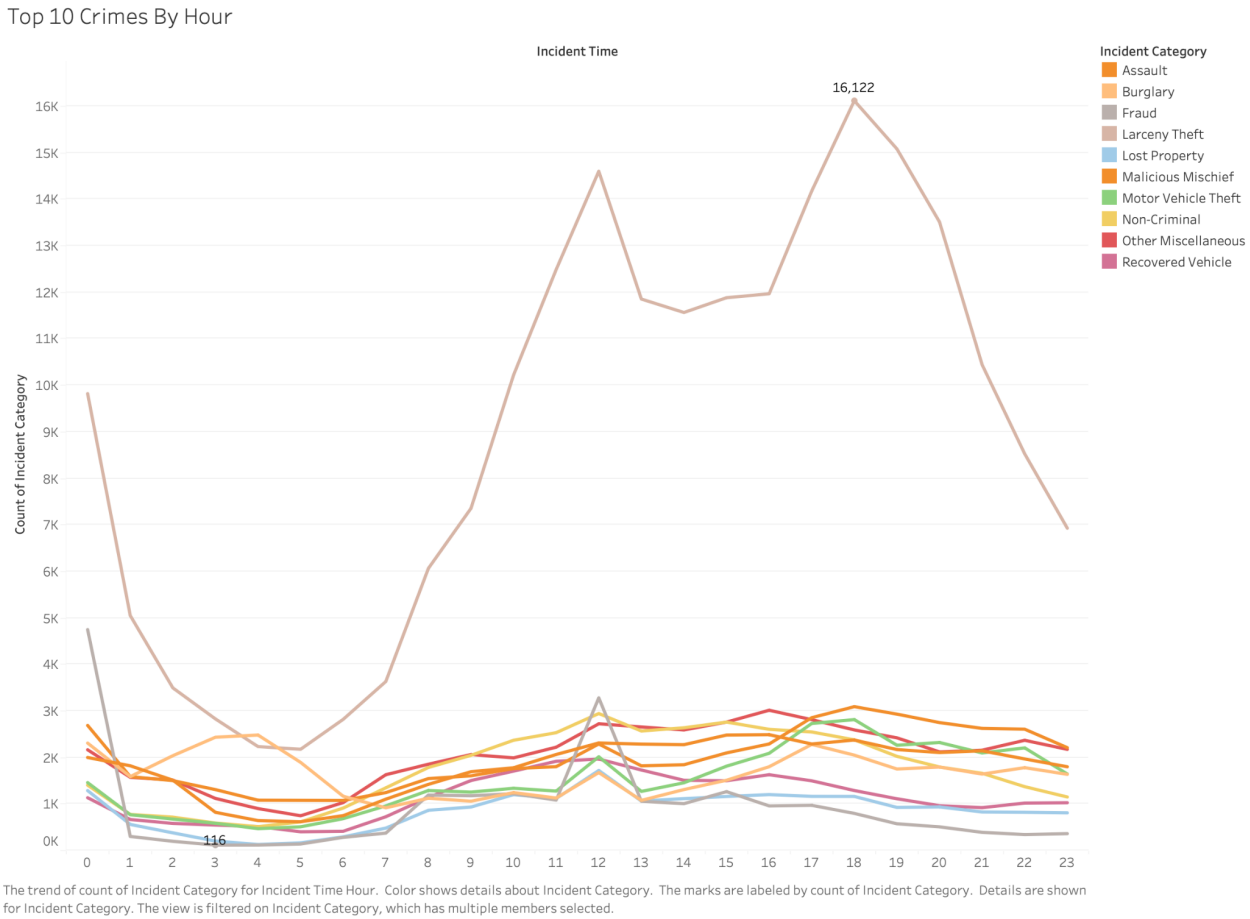


The Crime Category Overview is a bar chart that compares the crime counts by category. We can see from this chart that Larceny Theft is the top most crime in SF between 2018 and 2023, followed by Other Misc, Assault, Non-criminal, and Burglarly. Larceny Theft is such a massive crime in the city, that if you stack the next top 5 crimes together in one category, Larceny Theft would still be the top-most crime.



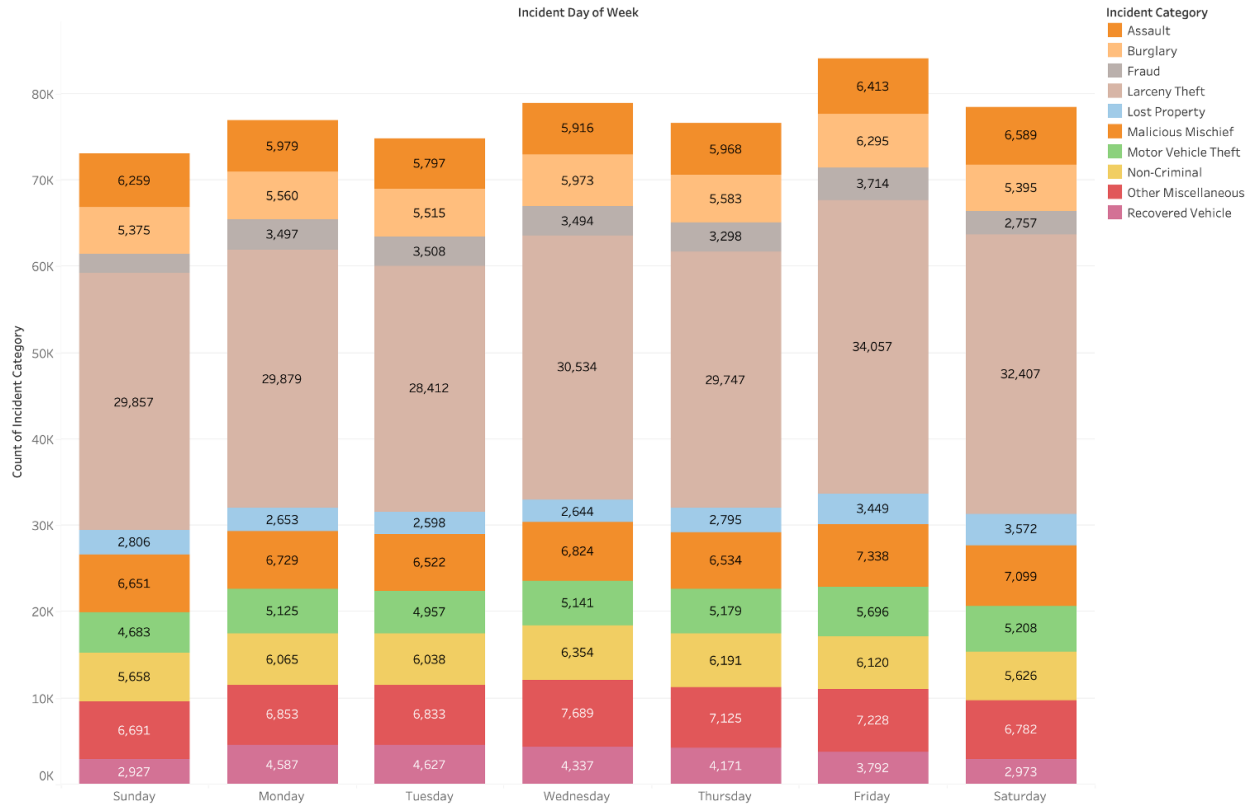
We define night as the time between 10pm to 6am. This visualization, which uses a 100% stacked column chart, tells you when do the top 10 crimes typically happen: day or night. From this visualization, we can say that 70 percent of larceny theft happened during the day, while for burglary it was equally likely to happen during the day and night.

b. Temporal Analysis of Crime



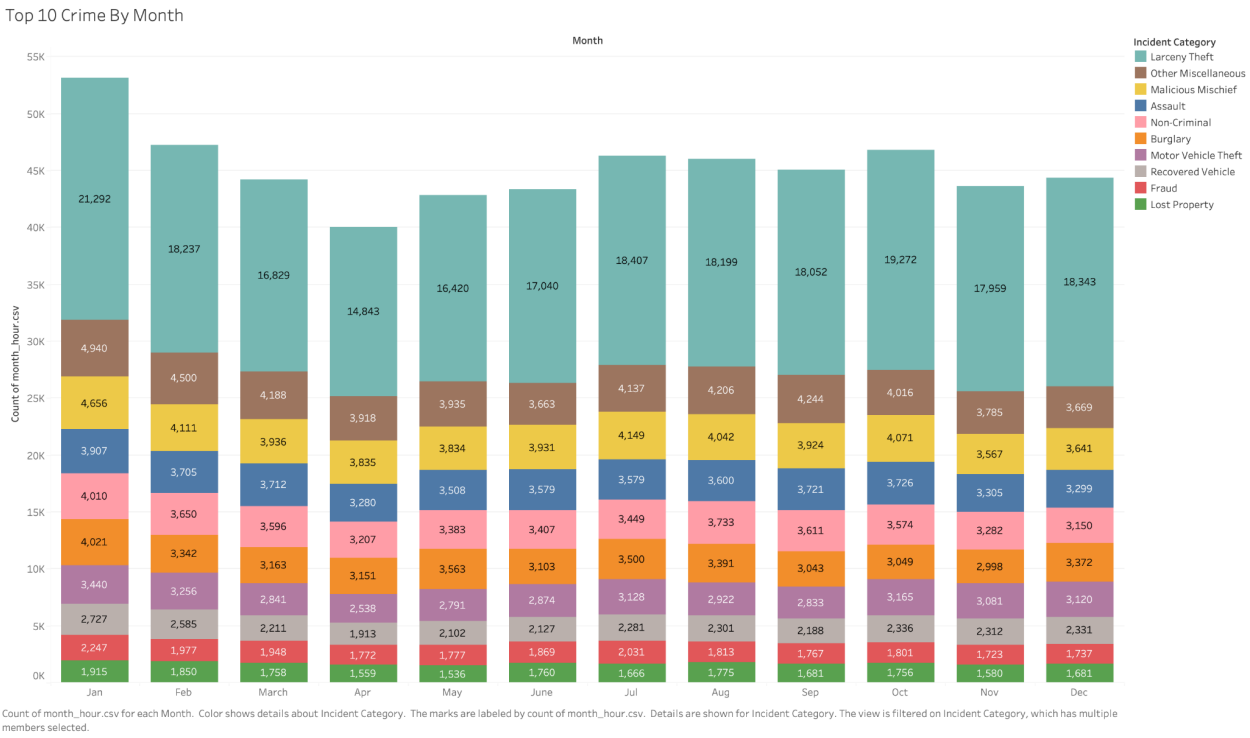
This is an hour by hour visual on different types of crimes observed. The axis starts with 0 (12 AM) to 23 (11 PM). Drastic variation in crime in general is not observed throughout the daytime however larceny dips around 4 AM and reaches a peak at 6 PM with the peak reaching a value 4 times higher than at 4 AM.

Top 10 Crime By Weekday

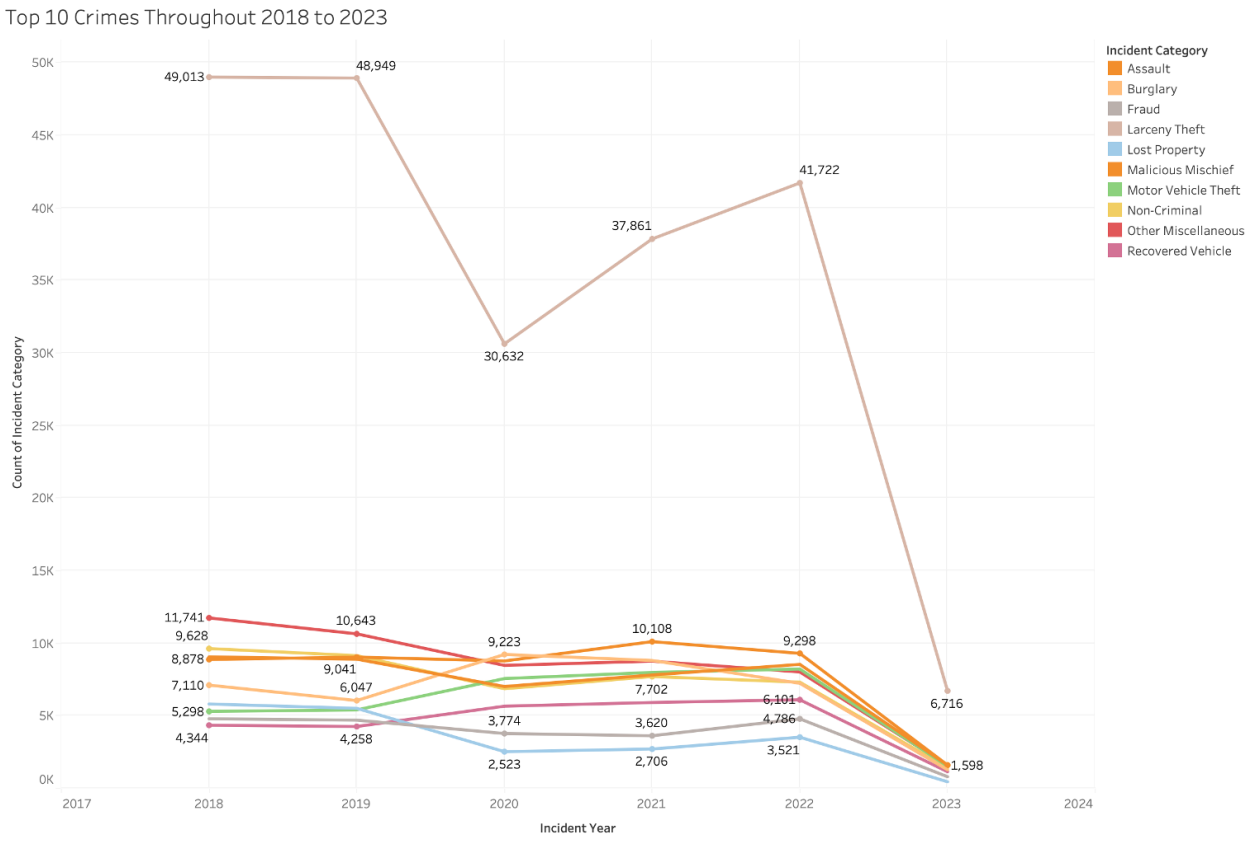


Count of Incident Category for each Incident Day of Week. Color shows details about Incident Category. The marks are labeled by count of Incident Category. Details are shown for Incident Category. The view is filtered on Incident Category, which keeps 10 of 50 members.

With this visualization we try to observe whether crimes varied with respect to day of the week. Much variation is not observed here either, the crime is at a low margin high in the beginning of the weekend i.e. Friday.

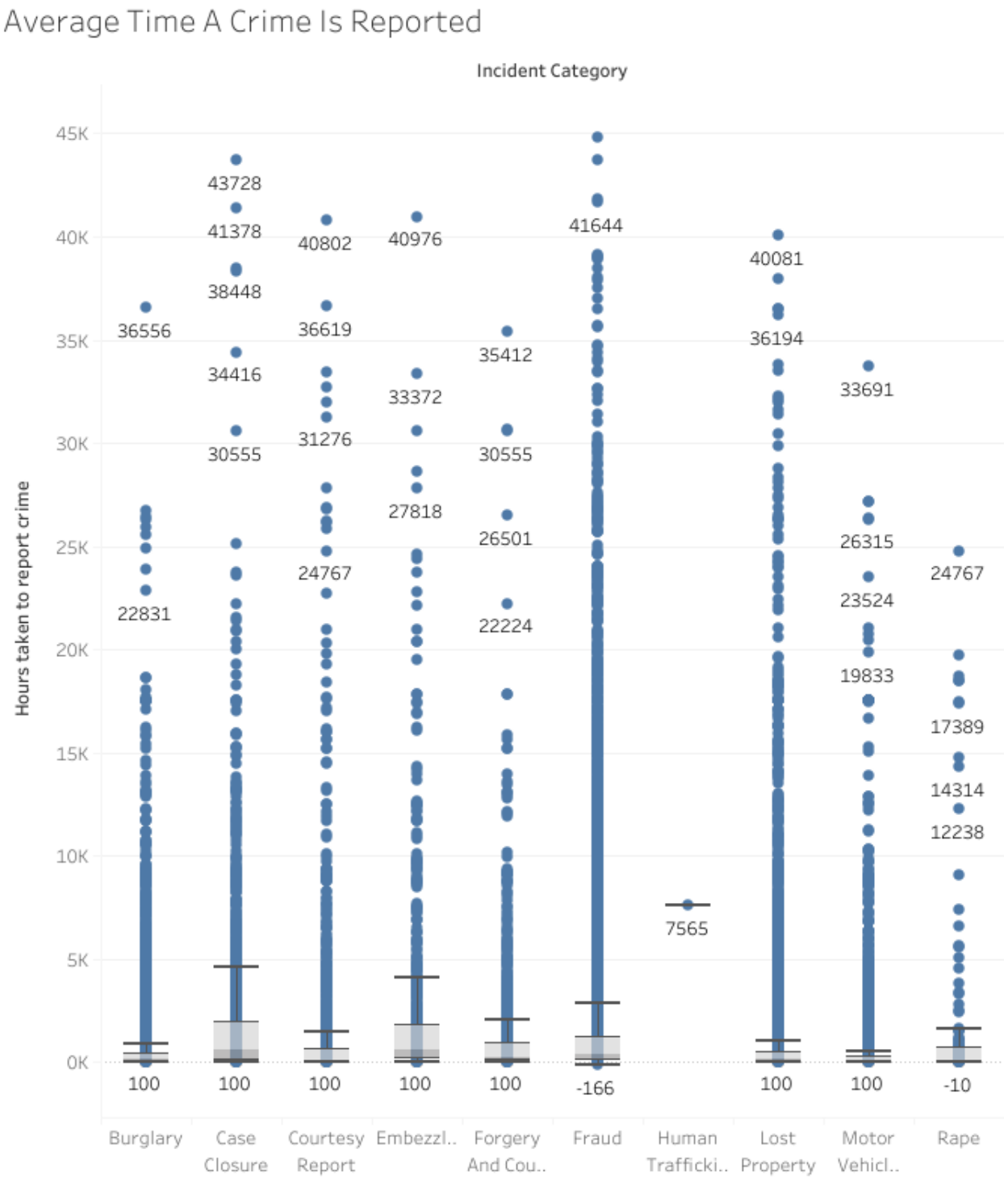


Here, we observe how much crime is reported in each month of the year. Maximum crime is seen in January where it is ~25% higher than in April, which is when minimum crime is observed.



The trend of count of Incident Category for Incident Year. Color shows details about Incident Category. The marks are labeled by count of Incident Category. Details are shown for Incident Category. The view is filtered on Incident Category, which keeps 10 of 50 members.

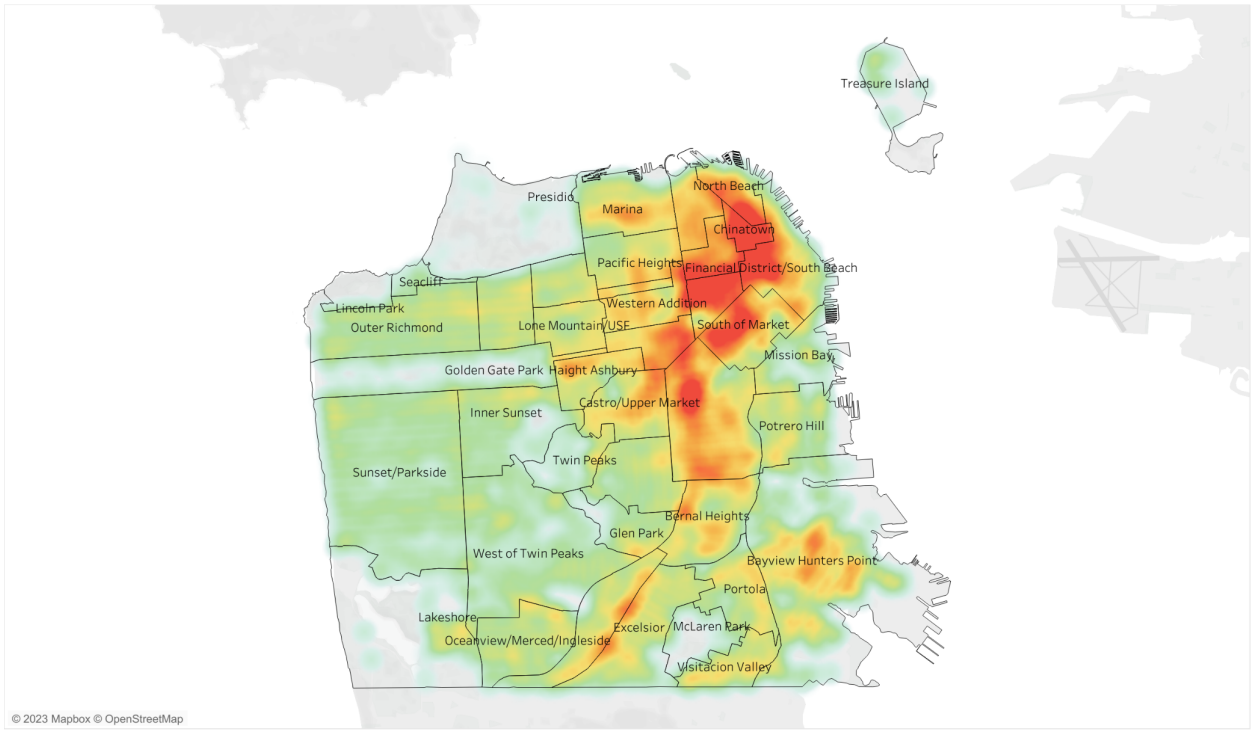
We plotted the year wise pattern of crimes occurred and observed that 2023 sees a sharp decline in the graph. However, this should not be understood as decrease in crimes for this year as the data only had incident reports from the first few months of 2023. There was a local minimum observed at 2020 in occurrence of larceny theft. This might have come as a result of lockdown restrictions during COVID-19 pandemic.



Next, we observe the time taken to report a crime: category-wise. Here, the "time taken to report a crime" is the difference of time between the occurrence of the crime and the time it was reported. The average of each category is highly skewed as we observed a lot of outliers with a high amount of time taken to report (the blue markers outside of the box in each column).

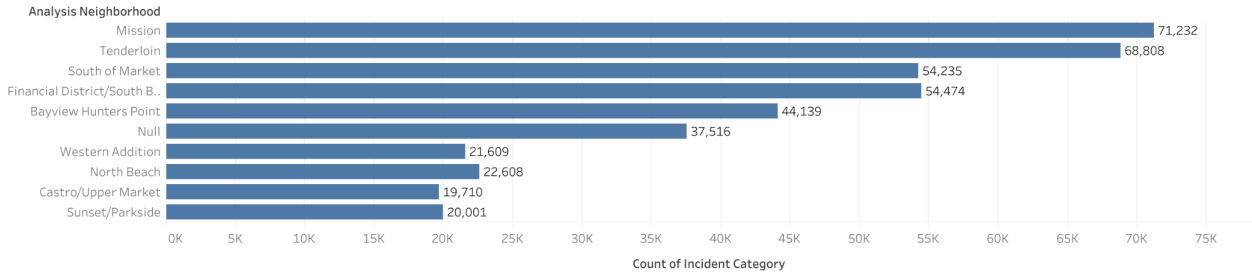
c. Geographic/Spatial Analysis of Crime

Crime Heat Map

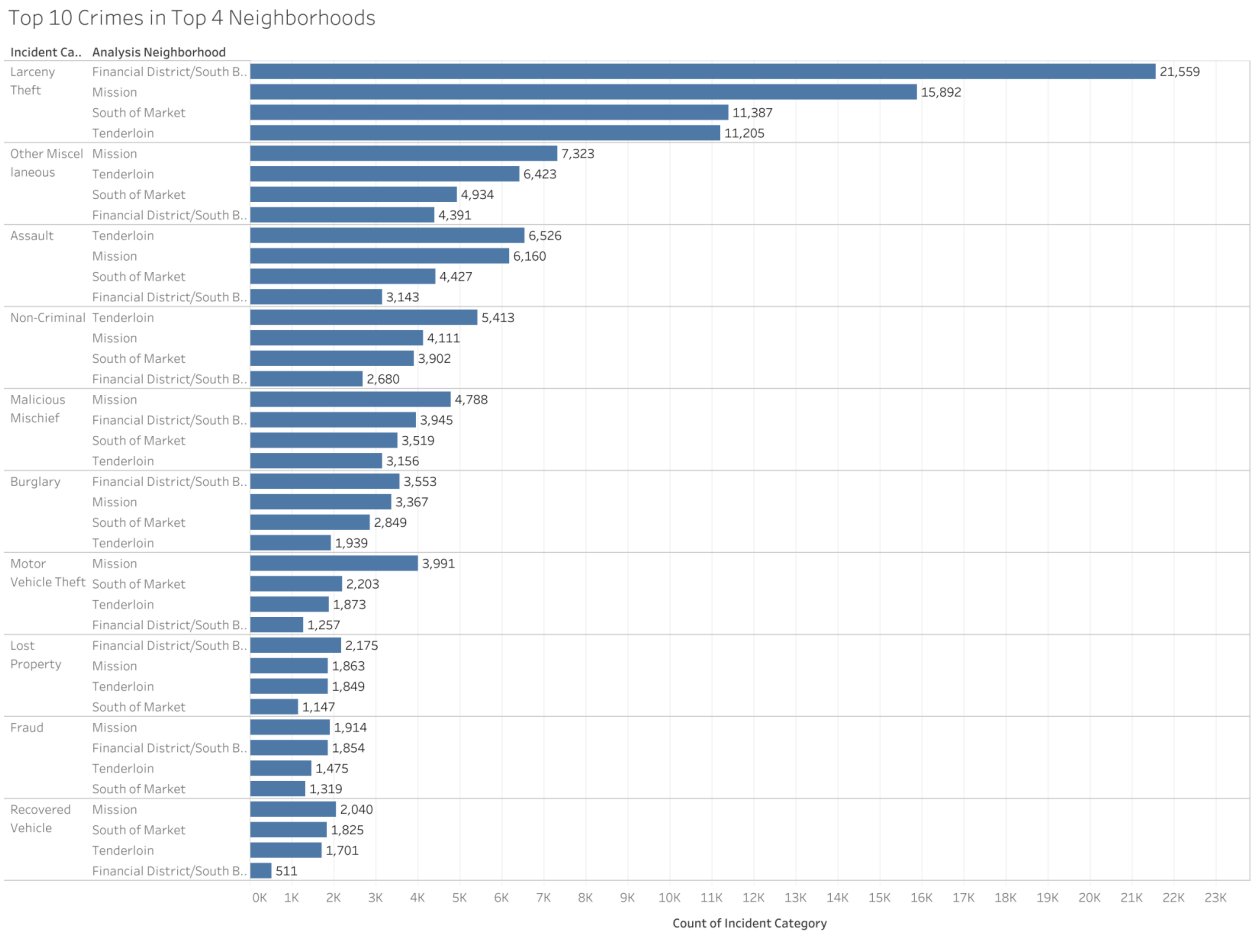


For this visualization, we wanted to answer the question, “In general, are crimes more concentrated in downtown SF?” Using the location data we have from the dataset and the intersection points of where the crimes happened, we were able to create a heatmap that shows the density of the crimes in San Francisco. Red means most concentration while green means least concentration. According to this visualization, there is more concentration of crimes happening in downtown SF compared to the outskirts like in Sunset and Outer Richmond. It can be seen here that the neighborhoods with the most “red” areas are Financial District, Tenderloin, South of Market, and Mission.

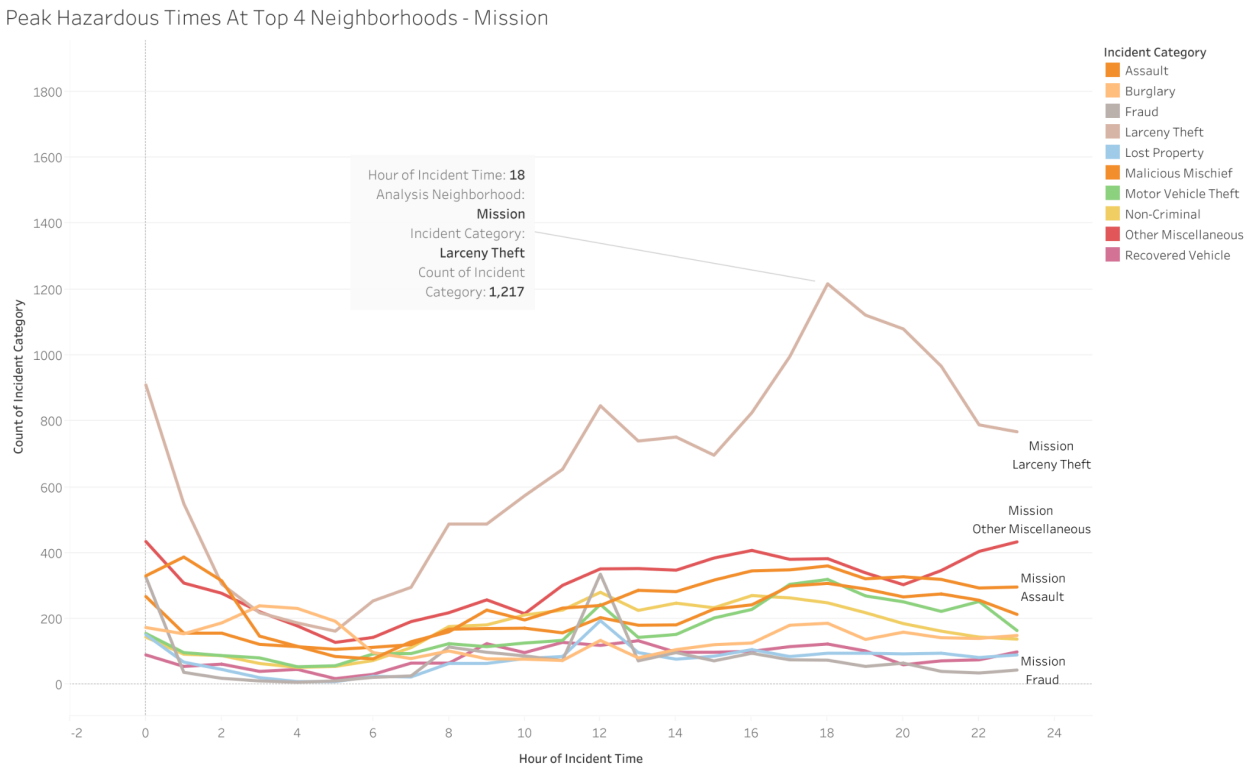
Top Neighborhoods with Crimes



This bar chart gives you a better picture of how the neighborhoods compare to each other when it comes to crime counts. We can see that the top neighborhood with most crimes is Mission, followed by Tenderloin, South of Market (SOMA), and then Financial District (FiDi).

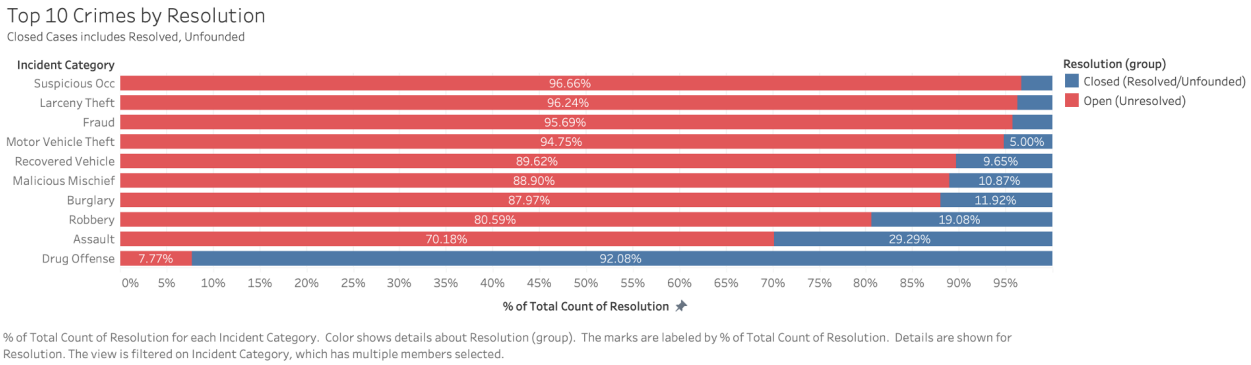


Given the top 4 neighbrhoods (Mission, Tenderloin, SOMA, and FiDi), we intevestigated further in which of these neighborhoods are the top 10 crimes most prevalent. For example, we can tell from this visualization that most of larceny theft happen in FiDi, while most assault happen in Tenderloin and the Mission.

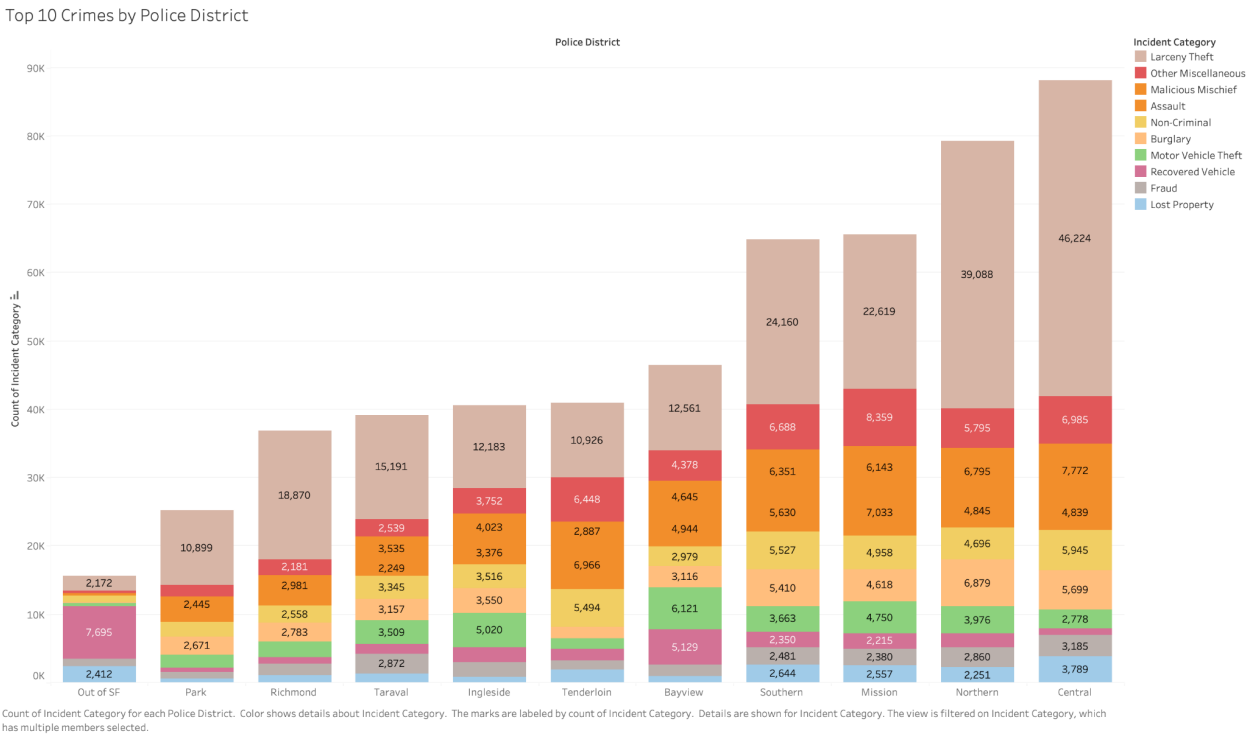


This visualization is a series or pages of visualization by neighborhood, which can be all viewed in the Tableau Public Link below. It tells us what are the top peak hazardous crimes at the top 4 neighbrhoods. Among all the neighborhoods, it's consistent that 7pm to 8pm is when peak of larceny crimes happen. A second peak can be observed around 12pm.

d. Crime by Police District



Here we have a percentage bar chart for the top 10 crimes by resolution. Our goal for this chart is to see which crime incident categories are difficult to resolve. We separated open by unresolved incidents and closed by resolved or unfounded incidents. We see that most crimes are open or unresolved except for drug offenses in which about 92% of cases are closed.



In this stacked bar chart we see the top 10 crimes per police district. Our goal for this chart is to examine what incident categories are most frequent in each police district. For each bar chart, we see that larceny theft is the largest incident category for each police district followed by other miscellaneous categories or malicious mischief.

V. Source Material

a. Demo Video: https://drive.google.com/file/d/1dp7C1YzoShrvkVrwpkPMfTirIkJO5r7t/view?usp=sharing

b. Tableau Public Link: Temporal and Spatial Analysis of Crimes in San Francisco (2018-2023) | Tableau Public