Trends in Thefts from Vehicles in San Francisco from January 2018 to November 2022

W200 Project 2, Dec 2022

Priya Reddy, Felicia Liu, Riya Shrestha, Allegra Simmons

Team GitHub Repository:
https://github.com/UC-Berkeley-I-School/Project2_Reddy_Shrestha_Liu_Simmons
Primary Dataset to Analyze:
Data downloaded on November 7, 2022.
https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783

Introduction

In recent years, there has been an increase in concern surrounding the frequency of crime in San Francisco. In a recent poll conducted by the SF Chronicle, 45% of participants reported that they had had an item stolen from them in the last 5 years.[1] In particular, there has been a lot of concern around car break ins and theft from cars in SF. Residents concerned about this perceived uptick in thefts from cars have turned to a number of different methods, from leaving notes to keeping the windows rolled down, to stop themselves from becoming a victim of this crime[2]. Understanding trends in this type of crime would be key in helping SF residents feel safe, and would give them insight into what steps they can take to protect themselves.

In this report, we examine the trends in larceny from vehicles from 2018 to 2022 to see if there have been recent shifts in crime from the pre-COVID to post-COVID era. The first trends we plan to inspect are the weekly and monthly trends in larceny from vehicles. In doing this, we can report which days of the week and which months or days of the month that this crime is more likely to occur. We will also look into the specific police districts in the city to see if there are any differences in the crime rates between districts that could be linked to the differences in socioeconomic status. Finally, to further examine the disparities among the districts, the resolution rates of larceny from vehicles will be explored from each district to see if the number of resolutions appears correlated with the average median income of the district. In our conclusion, we will take a look into possible confounding variables that could impact the numbers observed in the data.

The results from this report can be used for better crime prevention strategies, better allotment of resources for different districts, and to advise residents on steps they can take to prevent this crime. This data can be used by the local city government to improve the lives of their residents, ensure that citizens are well informed of the risks, and to prevent these crimes accordingly.

Data Structure

The dataset we analyzed came from the city of San Francisco's DataSF initiative. The dataset is titled Police Department Incident Reports: 2018 to Present and it compiles data on incident reports from the Crime Data Warehouse and the SFPD. The dataset contains reports from January 1, 2018 to Present

[1] Rachel Swan, "Here's How Many San Franciscans Say They've Been the Victim of a Crime, According to New Poll," San Francisco Chronicle, September 19, 2022, https://www.sfchronicle.com/sf/article/sfnext-poll-crime-sfpd-17439346.php.

[2] Pender, C. (2022, July 13). SF residents take creative steps to prevent car break-ins. KRON4. Retrieved December 3, 2022, from https://www.kron4.com/news/the-crazy-lengths-sf-drivers-are-going-to-prevent-car-break-ins/

and is updated daily. For our analysis we decided to use data collected up until November 7, 2022, the day we started to analyze the dataset.

Each row of the dataset represents an incident record and these incident records are filed by police officers or by members of the public. Once filed, the reports are then approved by a supervising Sergeant or Lieutenant.For this report, we will use Initial reports, the first report filed for an incident. The dataset does not include any identifiable information nor does it contain information about policing of the crime. We cannot consider the dataset complete for all police incidents, as all incidents that involved a Juvenile have been deleted from the dataset (account for about 2% of recorded incidents per year) In addition, some of the data is redacted for confidential reasons such as domestic violence reports (account for about 5% of recorded incidents). These omitted cases make up a small enough percentage of the total data (< ~7%), that we felt the data was complete enough to work with.

The original, full dataset had 658,728 rows and 34 columns. Although the dataset contained a lot of information on incidents in SF, we decided to filter for the observations that will help us answer our research questions. Most importantly, we filtered the data by *Incident Subcategory* as we were specifically interested in the “Larceny - From Vehicle” rows. Additional steps that we took to clean our data can be found in the data cleaning section of our report.

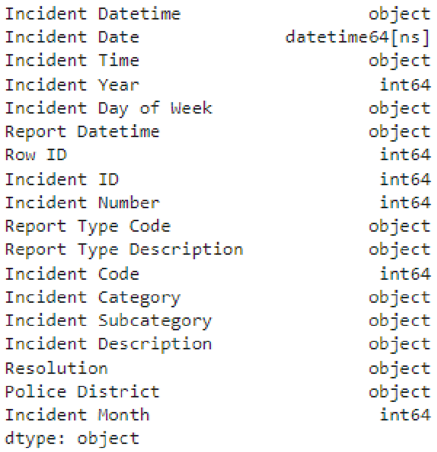
After all considerations the dataset contains only information about “Larceny - From Vehicle,” and the corresponding information about where and when it happened. Beyond this dataset, we did not use any supplemental datasets in our project.

Data Cleaning

After initial examination of the variables and observations, we performed several sanity checks on the dataset and have summarized our main findings below:

1. The information contained in *Incident Datetime* is separated into three other variables: *Incident Date*, *Incident Time*, and *Incident Year*
    a. *Incident Datetime* and *Incident Date* are in proper date format (YYYY/MM/DD)
    b. We added a separate *Incident Month* variable in order to make monthly average comparisons in our data analysis section
2. *Incident Day of Week* contains the day of the week that the incident occurred, which includes all days of the week from Monday through Sunday
3. *Report Datetime* is in proper formatting as well; however, this variable is not separated into three other variables like *Incident Datetime* is
    a. This is not too big of an issue, as we primarily focused on *Incident Datetime* rather than *Report Datetime*
4. 15 out of the original 34 columns have no null values and these variables include *Incident Datetime, Incident Date, Incident Time, Incident Year, Incident Day of Week, Report Datetime, Row ID, Incident ID, Incident Number, Report Type Code, Report Type Description, Incident Code, Incident Description, Resolution,* and *Police District*
5. Variables that have a particularly large number of null values include *CAD Number, Filed Online, Intersection, CNN, Analysis Neighborhood, Supervisor District, Latitude, Longitude, Point, Neighborhoods, ESNCAG - Boundary File, Central Market/Tenderloin Boundary Polygon - Updated, Civic Center Harm Reduction Project Boundary, HSOC Zones as of 2018-06-05, Invest In Neighborhoods (IIN) Areas, Current Supervisor Districts*, and *Current Police Districts*
    a. We dropped these columns from our analysis, as we didn’t need them to answer our questions and too many null values simply adds noise to the data

6. The variables *Incident Category* and *Incident Subcategory* contain 556 null values each, but we keep these columns because we need to filter the incidents by vehicle thefts
    a. We only kept rows where the *Incident Subcategory* was "Larceny - From Vehicle"
7. *Police District* contains 11 districts in the Bay Area, including “Bayview,” “Central,” “Ingleside,” “Mission,” “Northern,” “Out of SF,” “Park,” “Richmond,” “Southern,” “Taraval,” and “Tenderloin”
    a. We filtered out the “Out of SF”rows since the focus of our analysis was thefts in SF
8. There are 36 duplicates of *Incident ID* after filtering the data for "Larceny - From Vehicle" and excluding “Out of SF” police districts
    a. The data documentation states that multiple reports of the same incident (e.g. from different people) are different rows in the dataset but have the same *Incident ID*
    b. Similarly, files that have been updated are entered as new rows and share the same *Incident ID* as well
    c. We removed the 36 duplicate rows and kept only the initial incident filing, as it didn’t make too much sense to include the same incident multiple times in our analysis



```
Incident Datetime                   object
Incident Date               datetime64[ns]
Incident Time                       object
Incident Year                        int64
Incident Day of Week                object
Report Datetime                     object
Row ID                               int64
Incident ID                          int64
Incident Number                      int64
Report Type Code                    object
Report Type Description             object
Incident Code                        int64
Incident Category                   object
Incident Subcategory                object
Incident Description                object
Resolution                          object
Police District                     object
Incident Month                       int64
dtype: object
```

Table 1: Table of dataset columns and their data types

After the above process, we obtained a final dataframe with no null values and the data types shown in Table 1. Our cleaned dataset contained 113,400 rows and 18 columns.

For this analysis, to better understand the trend over the course of the COVID pandemic, we chose to sort the data into 3 time periods. “Pre-COVID” which we defined as the time period before March 17, 2020, which was the day SF implemented the stay-at-home orders. “Peak-COVID”, which we defined as the period between March 17, 2020 and December 31, 2021, when stay at home orders were in place, and “Post-COVID”, which we defined as the period of time beginning January 1, 2022 through the end of our data (Nov. 7, 2022), which was when the stay at home orders were beginning to be lifted and there was a general return to in-person activities. To be clear, these are designations we have come up with to better understand the COVID related trends in SF, and not a reflection of the actual end of the disease.

Data Analysis

Thefts from Vehicles By Month and COVID Period in SF

The first question that we explored was how the COVID-19 pandemic affected the number of thefts from vehicles in SF. More specifically, we examined the average monthly thefts for our three time periods: pre-COVID, peak-COVID, and post-COVID. Our hypothesis was that car thefts increased during the peak-COVID period and have decreased post-COVID, though not to pre-COVID levels.

We first created a graph to show the number of thefts from vehicles by month from January 1, 2018 to October 31, 2022. November of 2022 had the lowest number of thefts because we downloaded the original dataset early in the month (November 7, 2022). Thus, we removed these incidents when creating the graph, as the low number was likely due to the fact that we didn’t have the data, not because the number of thefts was actually lower than usual. Including incidents from November of 2022 would have given us a figure with a sharp drop on the far right, which would have been visually misleading.

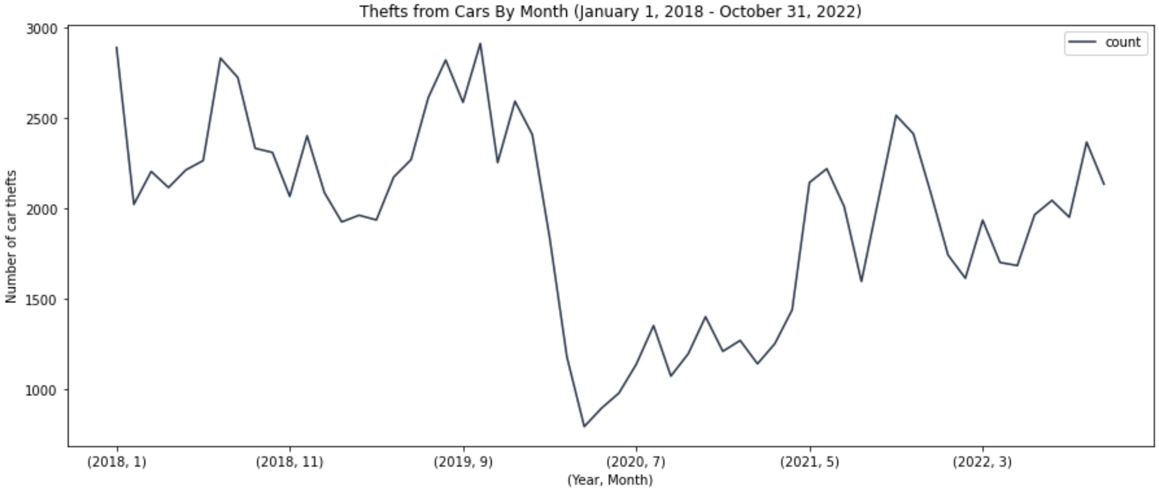


Figure 1. Thefts from Cars by Month in SF

As shown in Figure 1, the number of thefts from vehicles began to decline in February of 2020, and reached an all-time low in April of 2020. Thefts remained relatively low until around May of 2021, when there was a sudden, drastic increase. Most importantly, the figure clearly shows a big drop in thefts around July of 2020, as well as the fact that thefts are on the rise once again.

In order to compare the number of thefts from vehicles between the three time periods, we took the monthly average of each period. We first filtered our dataset by date and then divided it into three smaller datasets by period. We calculated the total number of thefts for each of these smaller datasets and divided those numbers by the total number of months within the time period. Pre-COVID lasted 26.5 months, peak-COVID lasted 21.5 months, and post-COVID was at 10.25 months at the time of our analysis.

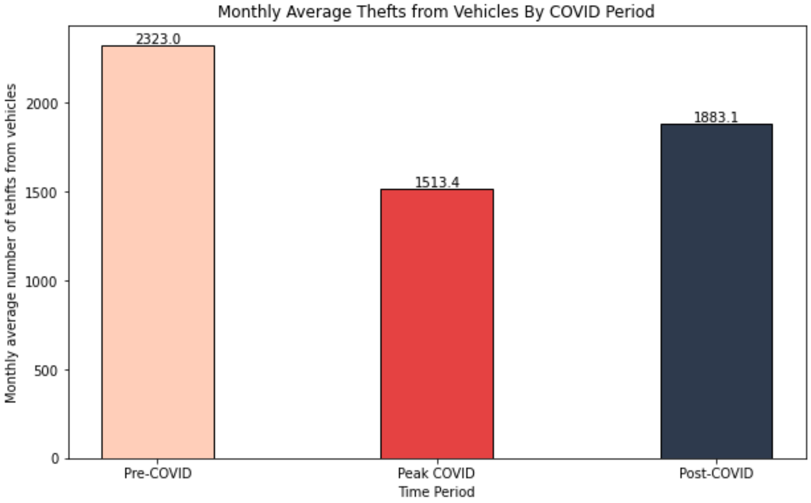


Figure 2. Monthly Average Thefts from Vehicles by COVID Period

As shown in Figure 2, for the pre-COVID period, the average monthly number of thefts from vehicles was 2323.0 thefts. For peak-COVID, the monthly average was 1513.4, and for post-COVID, the monthly average was 1883.1. Our analysis shows that the number of thefts during peak-COVID was considerably lower than that of the pre-COVID period. Thefts have increased again now that we are in the post-COVID period, but the average monthly number of thefts has not yet reached the pre-COVID average.

Weekly and Monthly Trends in Theft from Vehicle

We wanted to further analyze the monthly and weekly trends in car thefts in SF spanning the three COVID periods previously mentioned. First, we took a deeper look into the specific days of the month to observe which days of the month in each period that theft most commonly occurred. We originally hypothesized that car thefts would occur more towards the end of the month due to upcoming bills and rent that need to be paid leading to more thefts as well as increased thefts from cars during the peak months of the COVID-19 pandemic due to losses of jobs and financial insecurity.

To see the full range of thefts over different days of the month, we created a graph of all the days of the month faceted by COVID period. The graph was created by finding the number of thefts per day in each period and dividing by the total number of months in each period to account for the periods being different lengths.

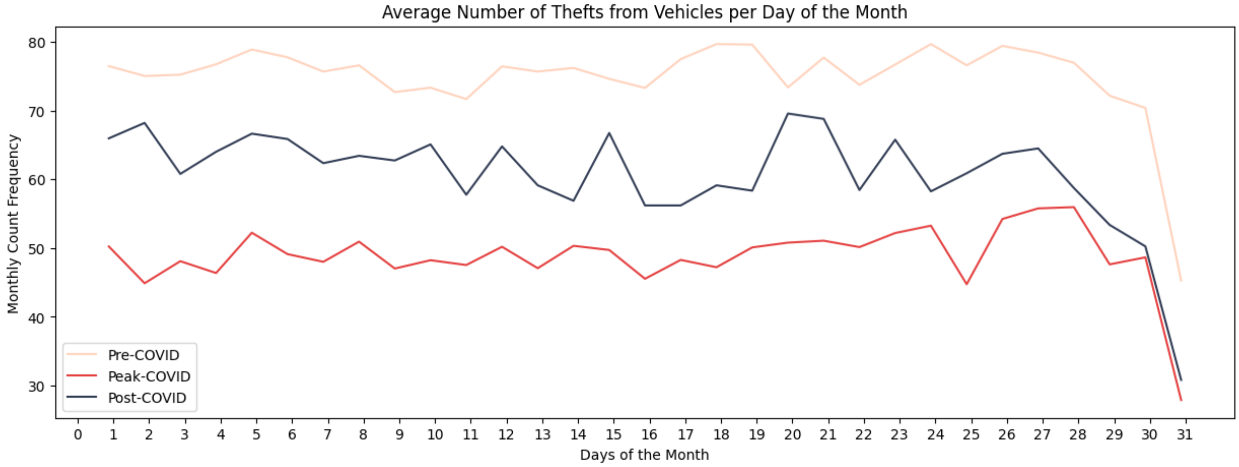


Figure 3. Average Number of Thefts from Vehicles per Days of the Month

As shown in Figure 3, the only time period that had the uptick in thefts towards the end of the month was the peak-COVID period. During this period, the highest rates of theft were on the 28th, 27th and 26th. This uptick is most likely due to the reason stated in the hypothesis that the end of the month is when bills and rent are due. With the financial strain of COVID, this could have played a bigger role in the peak-COVID era. Though pre-COVID did show a bit of a trend towards the end of the month, it was not as drastic as peak-COVID. Post-COVID actually showed even less of a trend to the end of the month, with many of the top average thefts occurring in the beginning of the month. There is a big drop off in thefts on the 31st for all periods which is due to the fact that only 7 months out of the year have a 31st day, so we are not counting it towards the result.

Next we delved into thefts on different days of the week. We hypothesized that thefts in cars would increase on the weekends as more people and cars are in SF leading to increased car thefts. We also hypothesized that during peak-COVID, the theft numbers may have decreased since people were not moving around, even on the weekends. In order to see the trends we have visualized the average monthly thefts from vehicles over the course of the days of the week, faceted by COVID period. This was done by finding the counts for each day of the week within the period specified. That number of overall thefts per day was then divided by the number of months in each period to find the average number of thefts per month for the days of the week in that period. As with the days of the month, this was done to standardize the numbers because each period was not the same length.

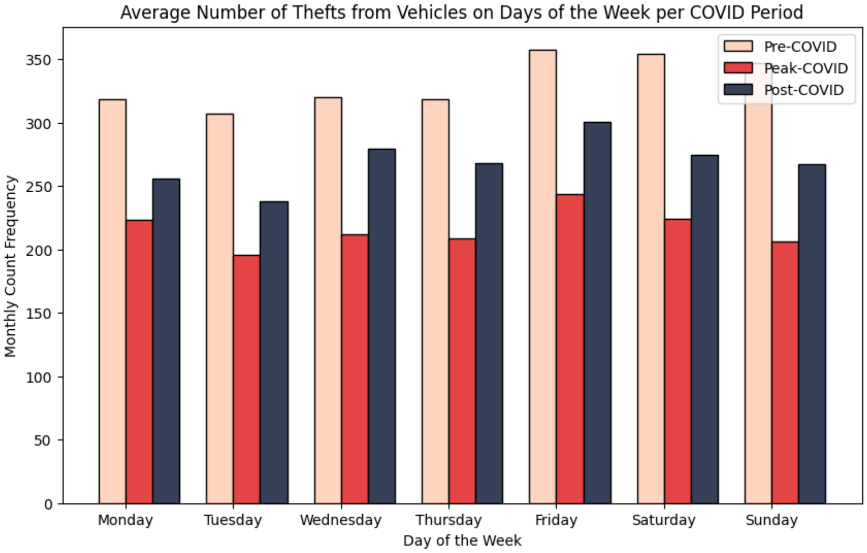


Figure 4. Average Number of Thefts from Cars per Days of the Week

In Figure 4, the pattern between days of the week and COVID periods can be seen. The highest monthly averages for thefts from vehicles happened in the pre-COVID time period, followed by post-COVID and finally peak-COVID. For all three periods, the day with the highest average number of thefts from cars was Friday. For pre-COVID, all three days of the weekend had the top 3 highest average number of thefts, which was what we predicted. However, for peak-COVID and post-COVID only Friday and Saturday were in the top 3 highest averages for the two periods. Instead the other highest dates were weekdays, Monday and Wednesday respectively.

District Trends in Theft from Vehicle

The third question we wanted to answer was where these car thefts were occurring, and whether these locations had changed over the course of the pandemic period. We expected to see that across police districts the number of thefts increased during the peak-COVID period and then decreased (though not to pre-COVID levels) during the post-COVID period. In addition, we hypothesized that districts that have a higher income level would be more likely to have a greater number of car thefts during the peak-COVID period, as these areas likely had more expensive cars (with theoretically more expensive contents) left outside during the shelter in place orders. Looking at the average median income across the zipcodes in each of the police districts we thus hypothesized that the police districts of Park and Tarval, (which had the highest average median incomes)[3] would have the highest number of car thefts during the peak-COVID period and that the police districts of the Tenderloin and Central (which had the lowest average median incomes)[4] would have the lowest.

To complete this analysis we looked at the average monthly number of car thefts in each recorded police district. That is, since each "larceny from car" entry had a recorded police district we were able to group incidents regionally by these districts. We chose to use the police districts instead of neighborhoods

[3] *Map of all ZIP codes in San Francisco, California - updated December 2022*. Zipdatamaps.com. (n.d.). Retrieved December 3, 2022, from https://www.zipdatamaps.com/zipcodes-san-francisco-ca

[4] *Ibid.*

or exact locations for this analysis, simply because every entry had a recorded police district but a majority of entries were missing an exact location or a neighborhood. We also chose to look at the average monthly number of car thefts since each time frame (pre-, peak-, post-) are different lengths, and we wanted to be able to compare the normalized numbers.

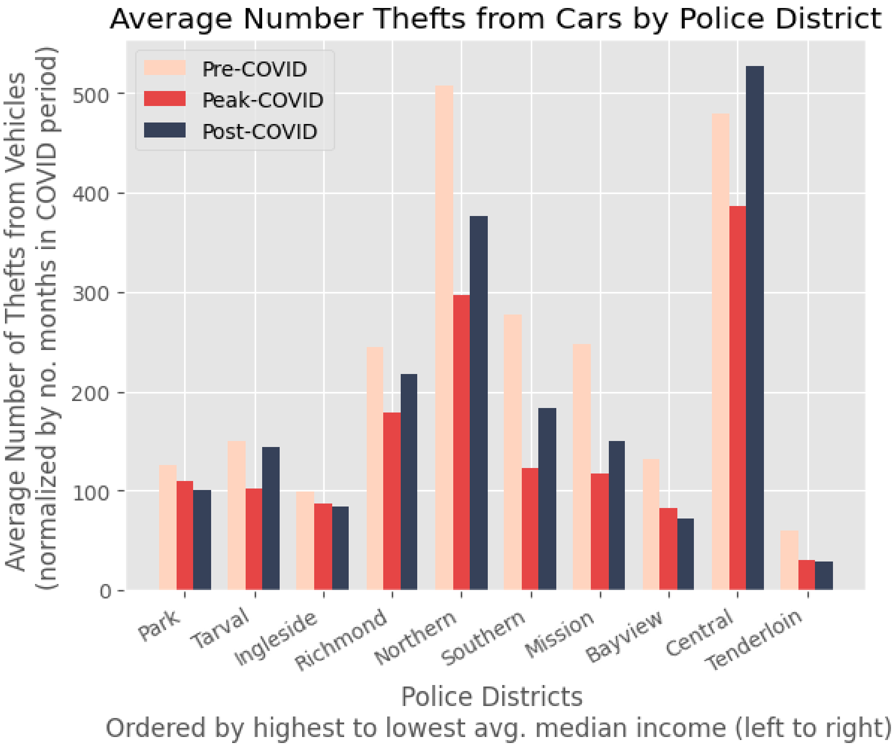
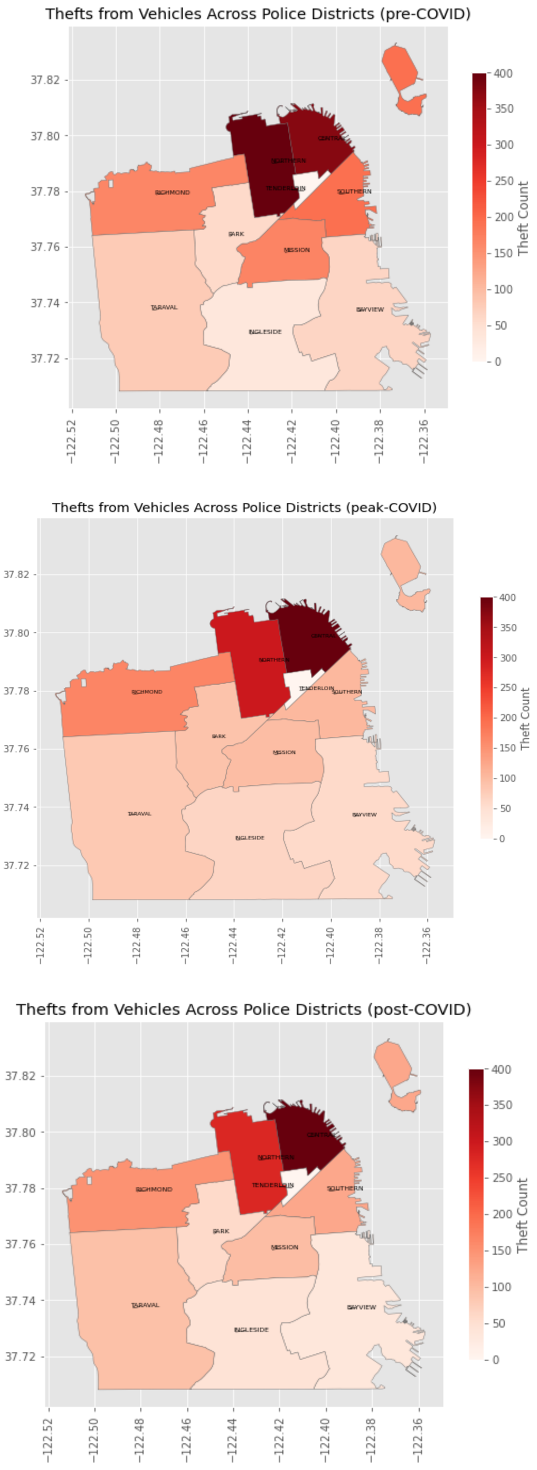


Figure 5. Average monthly car thefts in SF by police district across COVID periods. Police districts sorted from left to right in order of highest to lowest avg. median income. Avg. Median Income taken by average median incomes across zip codes contained in each police district.

In Figure 5, we can see a pattern that echoes the findings from our early analyses. Across police districts the number of cases of larceny from cars decreased from the pre-COVID period to the peak-COVID period. And from the peak-COVID to the post-COVID period, the number of incidents either increased but not to pre-COVID levels or the number of incidents decreased further. This is in direct opposition to our initial hypothesis that the number of car thefts increased during the peak-COVID period. We can also see that the districts with the greatest rate of car thefts across all three time periods are the Northern and Central districts, and the districts with the lowest rate of car thefts across the three time periods are the Tenderloin, Ingleside, and Bayview districts. As a result, we can see that this pattern does not follow our hypothesis that the higher income districts would have more thefts, instead we see that the districts in the middle of our graph (the districts with incomes in the middle of our range) are the ones that generally have the most thefts. In addition, looking at the overall patterns among the districts in Figures 6 a-c, it doesn't look as though the districts that had the most or the least thefts changed much across the peak- or post- COVID periods. Overall this analysis alone does not give us a clearer picture of *why* these particular districts have more car thefts, it just shows us that it is likely not income related. A future analysis might include plotting other factors along with the number of car thefts, for example, number of cars on patrol, number of cars registered to zipcodes in the districts, or size of the district.

There are some limitations to this analysis, namely that despite having normalized across time frames we did not normalize by the physical size of the police districts. Since some of the police districts


Thefts from Vehicles Across Police Districts (pre-COVID)

Thefts from Vehicles Across Police Districts (peak-COVID)

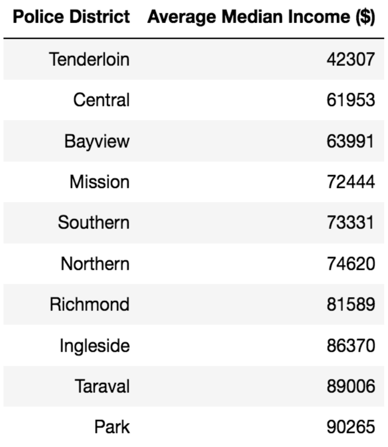
Thefts from Vehicles Across Police Districts (post-COVID)

Figure 6a, 6b, 6c. Heatmaps of theft from car trends in San Francisco. From top to bottom: Pre-COVID map, Peak-COVID map, Post-COVID map

are much larger or much smaller than the other districts it is possible that the differences in average number of thefts from cars can be attributed to the difference in size. In a future analysis it would be worthwhile to attempt to normalize the number of car thefts by the size of the district.

Trends in Resolution of Thefts from Vehicles

Lastly, we thought it was important to look at the trends of resolutions from vehicle theft. Specifically, we wanted to investigate if higher income districts, such as Park and Taraval[5] (as shown in Table 2 below) were associated with high percentages of resolutions.



Police District	Average Median Income ($)
Tenderloin	42307
Central	61953
Bayview	63991
Mission	72444
Southern	73331
Northern	74620
Richmond	81589
Ingleside	86370
Taraval	89006
Park	90265

Table 2 : Avg. Median Income of Police District in SF. Made by averaging median incomes of zip codes contained in each police district.

Incident reports can have multiple different types of conclusions. The case can be left “Open or Active,” or if a resolution was found they can be marked as “Cite or Adult Arrest". Other resolutions could conclude that the report was “Unfounded,” where the report was deemed false by an investigator, or it can be found as an “Exceptional Adult.” Both “Exceptional Adult” and “Unfounded” are seen as valid resolutions for incident reports.

In order to find the percentage of incident reports left “Open or Active” and “Resolved” for each district, we conducted a comparison of the two (shown by Figure 7 and Figure 8). The Tenderloin has the least number of cases “Open or Active,” at 98.46%, whereas Richmond had the

[5] *Map of all ZIP codes in San Francisco, California - updated December 2022*. Zipdatamaps.com. (n.d.). Retrieved December 3, 2022, from https://www.zipdatamaps.com/zipcodes-san-francisco-ca

highest percentage of “Open or Active” incident reports at 99.74%. As expected, the inverse is seen when looking at resolved incident reports. The Tenderloin and Bayview districts have the highest percentage of “Resolved” incident reports, and Richmond and Northern have the lowest .

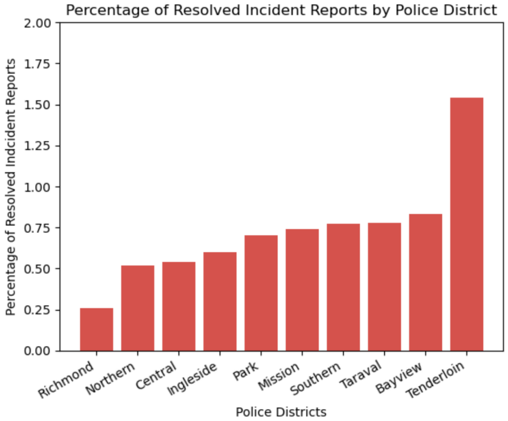


Figure 7. Percentage of Resolved Incident Reports by Police District

Next, Figure 9 looks at the percentage of “Resolved” incident reports and organizes the districts by their income from highest income to lowest income. From this graph, it is apparent that income is not a clear indicator for the percentage of “Resolved” incident reports by districts. This is made evident by the Tenderloin, with a median income of $42,307 (the lowest in the city)[6], having the highest percentage of resolved cases. On the other hand, Park, the district with the highest median income at $90,265[7] had a lower percentage of resolved cases compared to districts with lower median household incomes. More specifically, Bayview, Mission, and Southern have lower median household incomes, but higher percentages of resolved incident reports compared to Park.

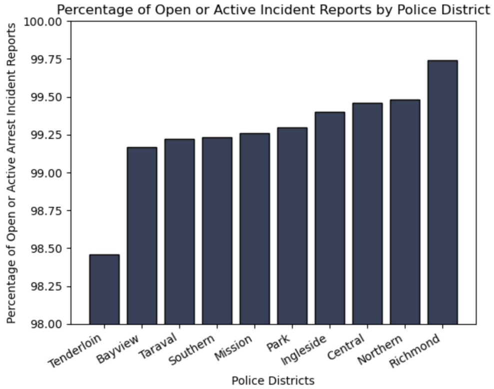


Figure 8. Percentage of Open or Active Incident Reports by Police District

From the investigation, it is hard to define the relationship between household median income in a police district and percentages of “Resolved” incident reports for that district. It is important to note that the percentages of “Resolved” cases between each police district for larceny from vehicle differ by less than one percent each, therefore it might not be meaningful to determine why one district has a higher resolved incident report percentage from the other, when the the difference might not be very large.

Lastly, we thought it was important to note that the percentage of overall arrests for larceny from vehicles cases for all of the SFPD is around 0.58%, whereas the Percentage of overall arrests for all crimes is around 19.32%. This could be due to the fact that larceny from vehicles is harder to resolve than most other crimes, as the discovery of the crime is delayed, which could explain why resolutions from district to district did not vary by a lot.

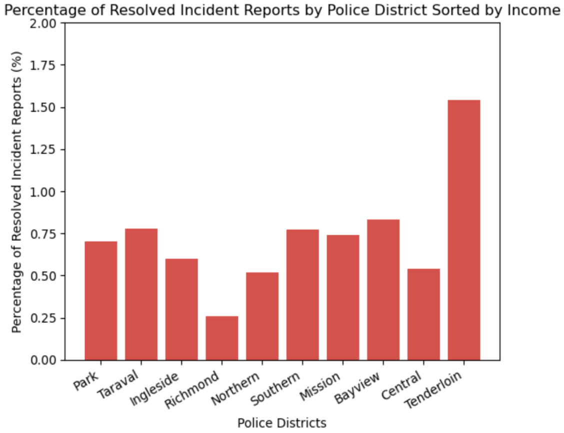


Figure 9. Percentage of Resolved Incident Reports by Police District sorted by Income

Conclusion and Future Data Explorations

Our study had several interesting takeaways regarding the relationship between thefts from vehicles and the COVID pandemic, as well as police district and resolution trends. More specifically, the monthly average number of thefts from vehicles drastically decreased during peak-COVID, and is once again on the rise during post-COVID, although the current average has not yet reached pre-COVID numbers. This is likely due to the stay-at-home orders

[6] *Map of all ZIP codes in San Francisco, California - updated December 2022*. Zipdatamaps.com. (n.d.). Retrieved December 3, 2022, from https://www.zipdatamaps.com/zipcodes-san-francisco-ca
[7] *Ibid.*

that were implemented at the onset of the pandemic, which prevented people from leaving their house using their cars. Moreover, when examining the number of thefts by day of the month, we noticed that the last few days of the month (when bills and rent are usually due) had the highest number of thefts during the peak-COVID period. This might be due to the financial strain that the pandemic caused, as pre- and post-COVID did not trend towards the end of the month in the same drastic way. We also noted that for all three time periods, Friday had the highest average number of thefts, which could be due to the fact that Friday is the day when most people leave their car parked for long periods of time at night.

In terms of district trends, the districts with the highest number of thefts were Northern and Central districts and the ones with the lowest numbers were Tenderloin, Ingleside, and Bayview across all three time periods. Our analysis did not provide an explanation as to why these districts had the highest or lowest thefts, but this can potentially be explored in a future study. Another consideration is that we did not normalize our study by the size of police districts, so the differences in average number of thefts by district could be simply due to the size of the district.

Lastly, our study showed that the Tenderloin district had the lowest percentage of “Open or Active” cases, while Richmond had the highest percentage. Furthermore, the relationship between household median income and the percentage of “Resolved” incident reports was difficult to define and the resolution rates for each police district were extremely close (less than a percent difference between each). The percentage of resolved cases for thefts from vehicles was also considerably lower than the percentage of overall resolved cases for all crimes, likely due to the fact that such thefts are naturally more difficult to resolve.

A limitation of our study is that our data relies heavily on incident reports that may only happen once the vehicle owner notices that a theft has occurred. This means that the timing of the incident is a rough estimate and also relies on the assumption that people notice and report car break-ins relatively quickly on average. As a result, our month of the year analysis is perhaps more accurate than the day of the week analysis.

We believe the local government can utilize studies like ours to change their crime prevention strategies, make plans for different types of surveillance, and erect signs that warn people of these thefts. Political candidates who are looking for issues that are of interest to the residents of SF can also address and further explore this issue. Regardless, more thorough research needs to be done in the future to gain a more holistic picture of these thefts. For example, we can investigate the types of cars that are being targeted as well as the features that are shared among them, such as certain makes or models, whether the windows are tinted, if there were items in the backseat, etc. Further research can also be conducted on the location of these crimes, particularly whether they were happening in residential areas (in front of homes) or commercial areas (in front of restaurants, in parking lots). Lastly, it might also be interesting to examine how much of an issue SF residents perceive the crime to be, although that might have been explored in past studies.