# Analysis of Parking Citations in San Francisco

Matthew Chen

## 1 Introduction

The San Francisco Municipal Transportation Agency (SFMTA) keeps an updated database on all parking tickets collected in the city since 2008, including information about the parking violation type, the street address where the violation occurred, the date of when the citation was issued, and more [1]. The overall goal of my project is to analyze these citations, how much revenue they generate, and how they are geospatially and temporarily distributed. I will approach this with a comprehensive pipeline including data collection, data processing, data analysis, and database upkeep.

In this project, we ask the following research questions:

1. How does revenue from parking citations vary geographically in San Francisco?
2. How does the number of parking citations change with the time of day? What are the most common violations during these times?
3. How has revenue from parking citations changed with time (i.e. over the last 5 years)? Are there seasonal trends on the monthly scale? How has the COVID-19 pandemic affected parking citation revenue?

The GitHub Repository for the results and code of this project can be found here:

https://github.com/Matt2371/SF-parking-citations

## 2 Methods

In this project, data was primarily sourced from the San Francisco Municipal Transportation Agency (SFMTA), who maintains and updates publicly available parking citation data from San Francisco. The dataset at the time of this writing includes about 20 million individual parking citations since the year 2008 and amounts to about 2.4 GB of data [1]. We are also interested in conducting geospatial analysis at the zip code level, although unfortunately this information is not included in the main dataset. Thus, zip codes were matched with street addresses using the USPS ZipCode Lookup API [3]. With this data, we create and maintain a SQL database which we analyze, focusing on how the number of citations and estimated revenue change with time and geospatially by creating interactive visualizations. Overall, the workflow of this project is described as follows:

1. Obtain a static download of SFMTA parking data and create a SQL database with its contents.
2. Update the database with new parking citations as they occur by calling an API from the SFMTA [2].
3. Subset the data where GPS coordinates exist, and conduct geospatial data analysis on the individual parking citation level.

4. Subset the data and match street addresses with zipcodes using the USPS ZipCode Lookup API, and conduct geospatial data analysis on the zipcode level.
5. Repeat steps 3-4, query the database, and subset and process the data as needed to conduct a robust data analysis.

## 2.1 Data Acquisition

### 2.1.1 Initial Data Download and Database Creation

As mentioned above, parking citation data from the SFMTA was initially downloaded as a csv file. The contents of the data are then used to create a SQLite Database, which is accomplished by the executing the script `create_database.py`. The script creates a database named `sfmta_parking_citations.db`. This only needs to be run once during the lifetime of this project.

Notably, datetime information was converted into ISO 8601 format, which is how SQLite indexes datetime objects although they are still treated as strings.

### 2.1.2 SFMTA API

The aforementioned database can be updated with new parking citations as they occur, which is primarily accomplished by running the script `update_database.py`. The script works by querying the (local) database and finds the date associated with the last entry. Then, data between the present date and the date of last entry is fetched from the SFMTA parking citations API, one day at a time. A python function named `sfmta_parking()` fetches data from the SFMTA parking citations API and is written in the script `fetch_api.py`. The function takes in as parameter input the start and end date of the request. The SFMTA parking citations API is capable of handling SoQL (Salesforce Object Query Language) Queries as a parameter to the HTML request, which we take advantage of to request data from the needed dates as follows:

```
params_dict = {'$where': 'citation_issued_datetime BETWEEN ' + '\'' + past_datetime +
'\'' + ' AND ' + '\'' + present_datetime + '\'', '$limit': limit,}
```

Here, past_datetime and present_datetime are the desired start and end dates of the request in ISO 8601 format. In our case, we use the query to fetch data one day at a time from SFMTA in order to update the local SQLite database, but the function/query is flexible to use in other dates as well.

It is important to note that use of the SFMTA parking citations API may result in throttling if authentication is not obtained. Signup for authentication can be obtained here: https://data.sfgov.org/profile/edit/developer_settings. Additionally, the full documentation for use of the API is available at the reference link [2].

### 2.1.3 USPS ZipCode Lookup API

In order to do a zipcode level geospatial data analysis, we need to match street addresses with their corresponding zip-code using the USPS ZipCode Lookup API [3]. The API both handles

inputs (street address, city, and state) and outputs (zipcode4, zipcode5) as XML, and takes up to 5 addresses at a time. In this project, we mainly stick with the zipcode5 output. An example of an input XML I constructed for the query is provided below (only one address request is shown, but the procedure can be extended to up to 5 addresses):

```
# Build XML input (takes up to 5 addresses)
xml = "<ZipCodeLookupRequest USERID=\"" + app_token + "\">" + """
<Address ID="1">
<Address1></Address1>
<Address2>""" + str(street_addresses[0]) + """</Address2>
<City>San Francisco</City>
<State>CA</State>
</Address>
```

Additionally, there are edge cases where a preceding 0 character breaks the address search (21 01$^{ST}$ Street instead of 21 1$^{ST}$ Street). These cases are pre-processed using regular expressions. Overall, this API call is summarized in the function `usps_zipcodes()` in the `fetch_api.py` script, which will take as parameter input a list of up to 5 street addresses and return their corresponding zipcodes. A challenge with this is that the USPS ZipCode only processes 5 addresses in about 0.5 seconds, which means it is impossible to match every address in the database with zipcodes. Thus, for the zip code level geospatial analysis (and many other analysis in this project) we take a random subset of the data. The `zip_codes()` function in the `get_data.py` script will take parking citation data from a given year, take a random sample as a subset, and match each address with up to their corresponding zipcodes, up to 5 at a time.

As before, authentication is needed to use the API to avoid throttling. One can signup for a token at the following site: https://registration.shippingapis.com/

### 2.2 Revenue and Counts Estimation

In a lot of cases in this project, counts of parking citations and total revenue need to be estimated from a subset of the data, due to the size of the total dataset and limitations on computational power. The default subset fraction, p, used here is 0.001 but the code and methodologies presented are flexible to larger subsets as well.

To estimate total annual revenue, we assume that the random subset (from a given year) is a sufficiently accurate representative of the population dataset, and estimate total parking citation revenue as follows:

$$\widehat{R_i} = \frac{\sum_{k=1}^{n_i} F_{ik}}{p} \qquad (Eq.\ 1)$$

where $R_i$ is the estimated total revenue of the i$^{th}$ zip-code for that given year, $F_{ik}$ is the k$^{th}$ fine amount in the i$^{th}$ zip code in the subset, and p is the subset proportion as aforementioned.

Similarly, we can estimate the monthly total number of citations based on the subset using similar assumptions as above:

$$\widehat{N}_i = N_i/p \qquad (Eq.\,2)$$

Where $N_i$ is the number of citations for the $i^{th}$ month in the subset, and p is the fraction of the subset.

# 3 Data Analysis

## 3.1 Time of Day Analysis

### 3.1.1 Individual Parking Citations

Here, we analyze the geospatial distribution of individual parking citations. In a previous analysis, we find that the City of San Francisco has depreciated the tracking of GPS coordinates after the year 2019. Thus, we take 2019 data as the latest year where geometries exist and take a random subset (for this example, we take a subset fraction of p=0.005). Using the python package Folium, we create an interactive map of individual parking citations color-coded by the time of day when the citation was issued. The map also has a hover/tooltip which was added using custom HTML code, and reveals additional information about the individual parking citation clicked.

In the repository, the interactive output can be found in `html/citations_over_time_2019_map.html` and a screenshot is also included below (Figure 1).

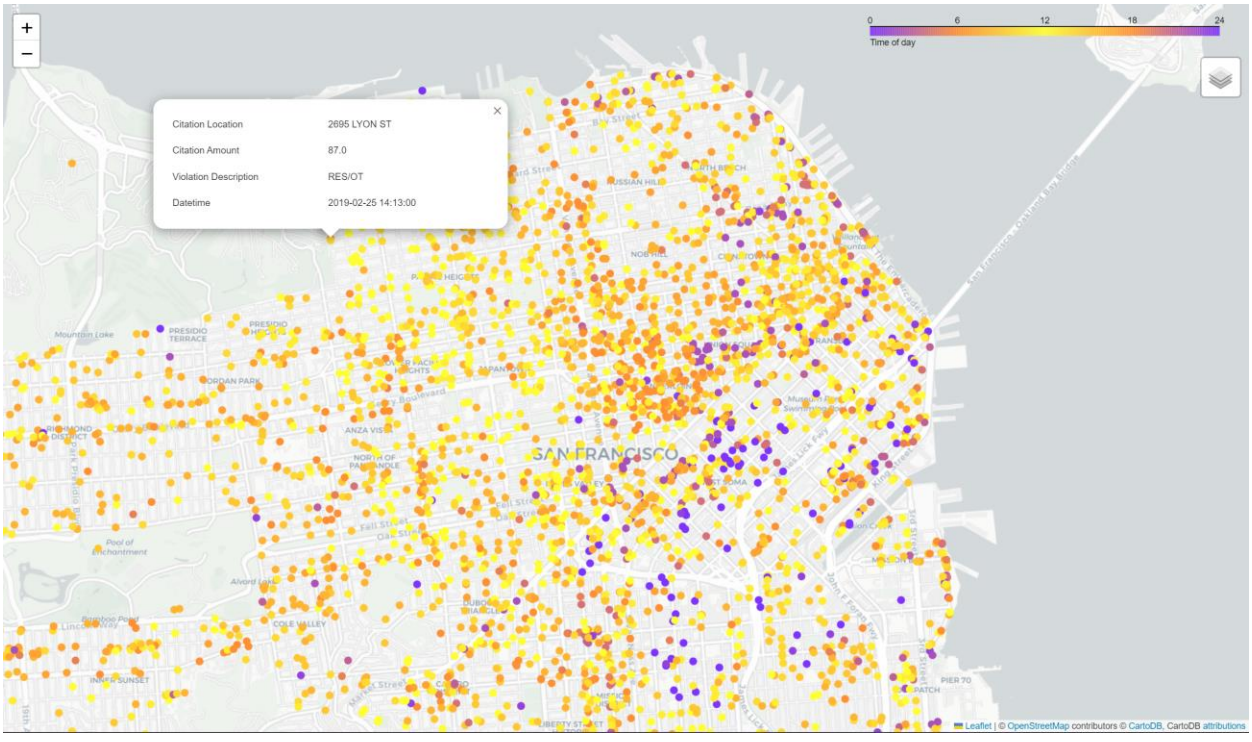


*Figure 1. Map of individual parking citations. Interactive version: html/citations_over_time_2019_map.html*

Here, we find that the majority of parking citations occur between 6am and 6pm, although citations that occur past these hours do occur. We also find that the downtown area and areas nearby have a greater density of parking citations compared to the more residential areas. Finally, we observe that very few parking citations were issued in the Presidio, Golden Gate Park, and Treasure Island. This is unsurprising, given that these areas have less car traffic and available street parking compared to more densely populated regions.

### 3.1.2 Fraction of Parking Citations

We also explore the fraction of total parking citations that occur vs. the time of day as a bar graph (Figure 2). To be consistent with Section 3.1.1, we again use the same subset of data from 2019. An interactive version of Figure 2 is available in the repository at `html/citations_over_time_2019.html`.

Here, we find that the most parking citations were issued at 12pm, and become much less common past 6pm and before 6am. This is consistent with our findings in Figure 1.

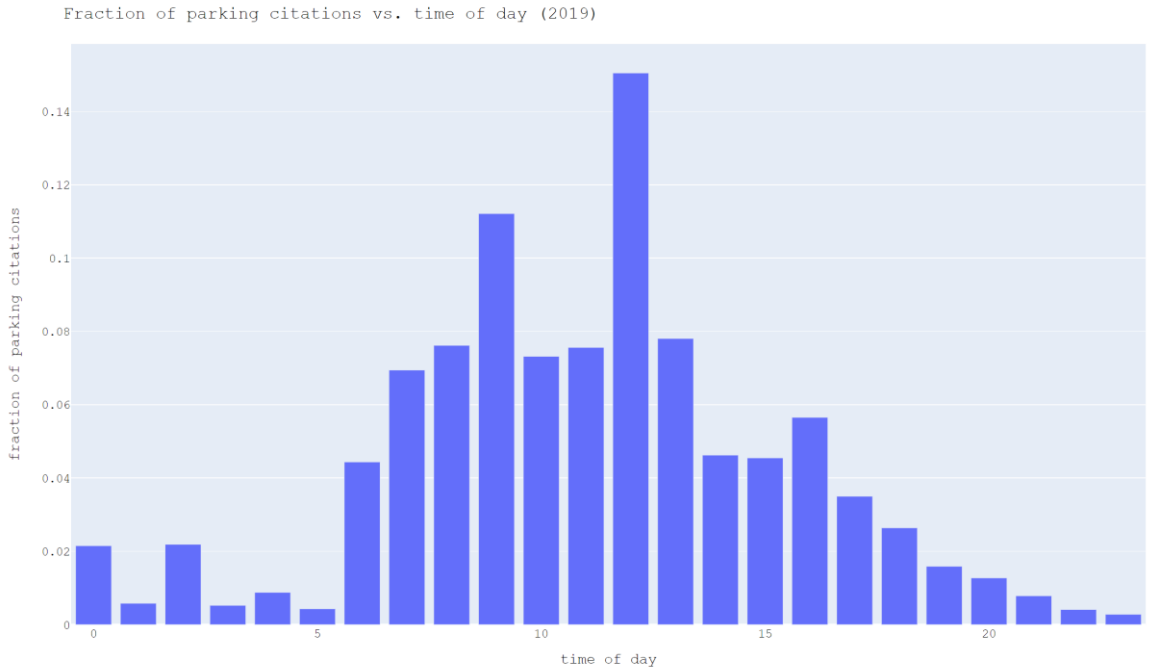


*Figure 2. Fraction of parking citations vs. time of day in 2019. Interactive version: html/citations_over_time_2019.html*

We also explore the fractions (of total violations) of different types of violations as they evolve with the time of day (Figure 3). An interactive version of Figure 3 can be found in the repository at `html/violation_types_over_time_2019.html`.

Here we find that across all times, street cleaning is the most common violation type, followed by parking meter citations outside of downtown (MTR OUT DT). Unfortunately, the SFMTA data documentation does not include much detail about the meaning of the different violation

codes. Additionally, we find that before 6am, the predominant violation type is street cleaning (Figure 3b).

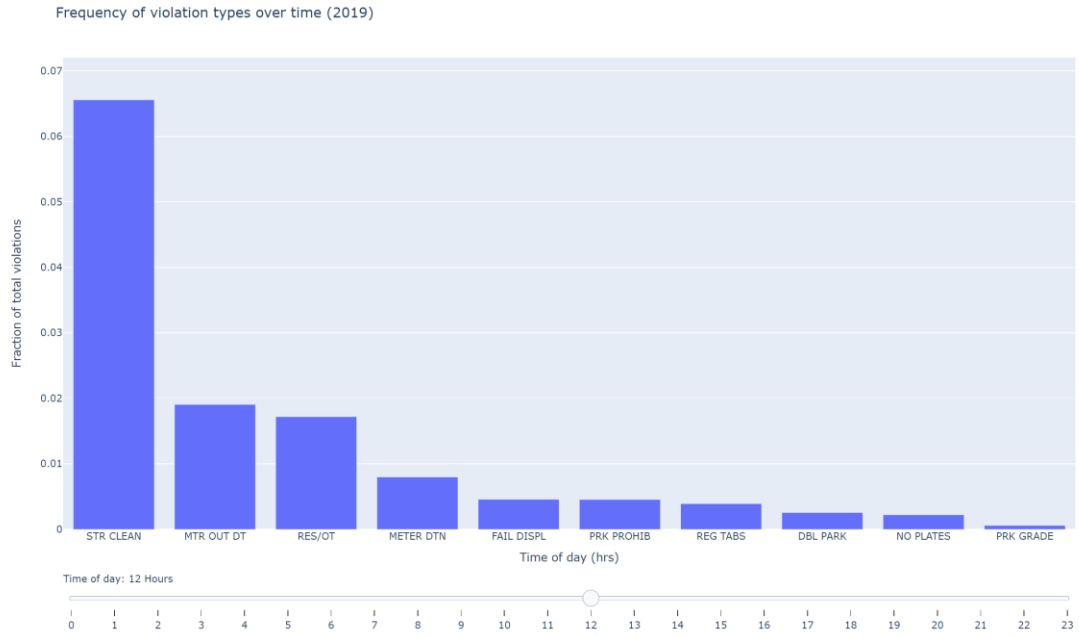


*Figure 3a. Frequency of violation types of time as the fraction of total violations: 12pm is shown above. Interactive version: html/ violation_types_over_time_2019.html*

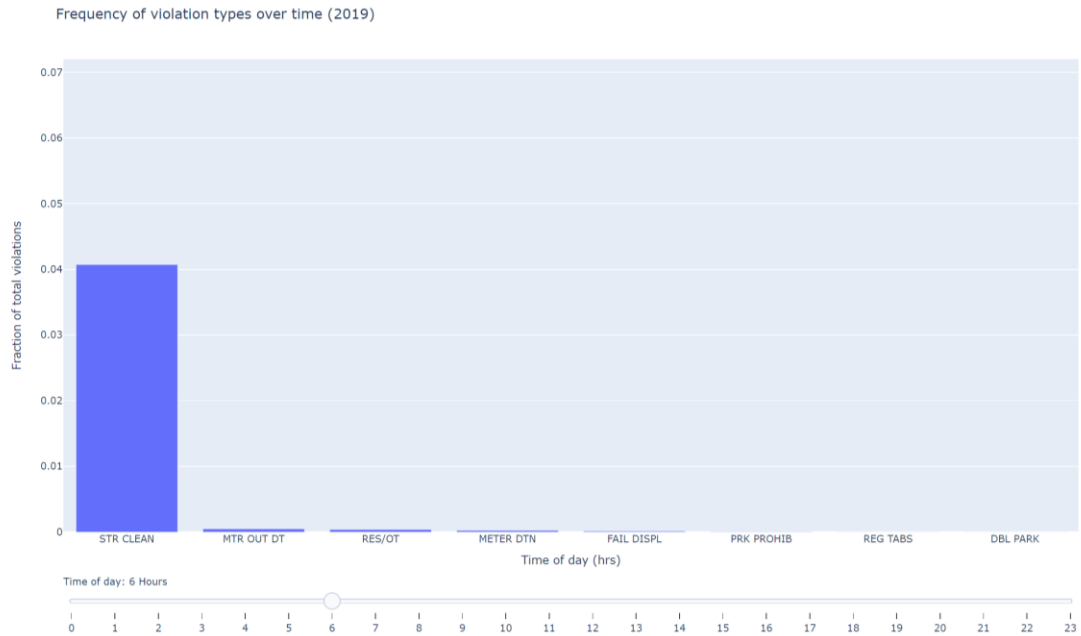


*Figure 3b. Frequency of violation types of time as the fraction of total violations: 6am is shown above. Interactive version: html/ violation_types_over_time_2019.html*

## 3.2 Estimated Revenue

### 3.2.1 Estimated Revenue by Zip Code

Here, we use a subset of the 2022 data (p=0.001) and match street addresses with zip codes as described in the Methods section. Descriptive statistics of estimated revenue at the zip code level in 2022 is shown below in Table 1. We find that estimated annual revenue ranges between about $174,000 to $9.5 million at the zip code level.

*Table 1. Descriptive statistics of estimated revenue by zip code in 2022*

| Estimated Revenue by Zip Code (2022) | |
|---|---|
| zipcodes | 25 |
| mean | 3.88E+06 |
| standard deviation | 2.53E+06 |
| min | 1.74E+05 |
| 25% percentile | 1.70E+06 |
| median | 3.60E+06 |
| 75% percentile | 5.75E+06 |
| max | 9.47E+06 |

Figure 4 below shows a choropleth of estimated annual revenue in 2022. An interaction version of the map can be found in the repository at `html/estimated_revenue_by_zip_2022_map.html`. The zipcode geometries were obtained from [4] and also includes population metrics from 2010 and the area of the zipcodes. Here, we merge the estimated revenue values with the geojson data and create the choropleth in Folium.

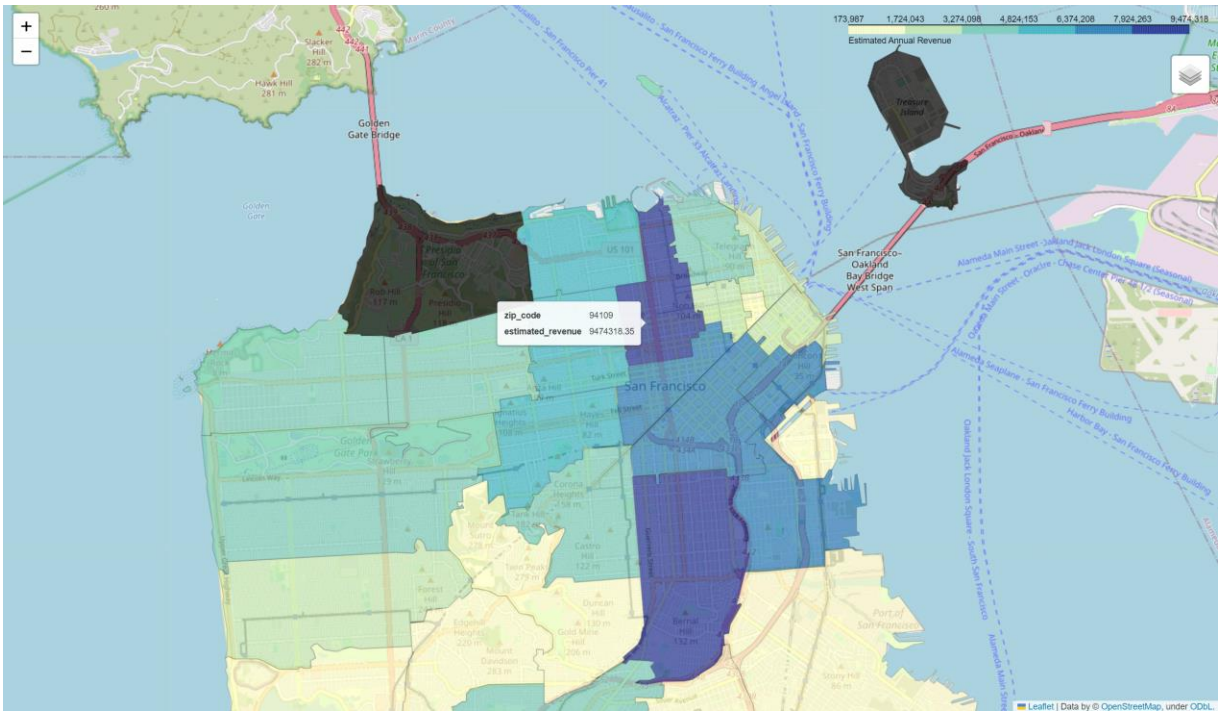


*Figure 4. Choropleth of estimated annual revenue by zip code in 2022. Interactive version: html/estimated_revenue_by_zip_2022_map.html.*

We find that the highest earning zip codes are in the downtown and surrounding areas, which are in more high-traffic areas as expected. The lowest revenue zip codes are also unsurprisinglymore in the south of the city, where the region is more mountainous and residential with far less car traffic. We also find no citations in the Presidio or Treasure Island, which is consistent with earlier results.

Table 2 shows the correlation matrix between estimated revenue, the area of the zip code, and the population. We find that there is a weak correlation between population and estimated revenue. This makes sense, since car traffic may be more correlated with parking revenue than population. For example, there may not be a lot of residential population in downtown areas, but the car traffic and estimated revenues are high. Interestingly, there is a weak negative correlation between revenue and zip code area. This may because large zip codes are more residential and therefore have less traffic, as seen in the larger positive correlation between population and area.

*Table 2. Correlation matrix on zip code level metrics*

| | estimated_revenue | area_sqmi | population_2010 |
|---|---|---|---|
| **estimated_revenue** | 1.00 | -0.20 | 0.36 |
| **area_sqmi** | -0.20 | 1.00 | 0.52 |
| **population_2010** | 0.36 | 0.52 | 1.00 |

Finally, we study how total estimated annual revenue has changed over time, specifically from the years 2018-2022. This was done by estimating the total revenue from these years and summarizing them (Figure 5). We find that total revenue was relatively consistent between the years 2018 and 2019, but dropped significantly in 2020. The COVID-19 pandemic is likely to be the cause of this, as it is likely to significantly reduce car traffic in San Francisco, especially in high revenue areas such as downtown where there might be a lot of office space. Revenue levels increased significantly in the following years but is still lower than pre-pandemic levels.

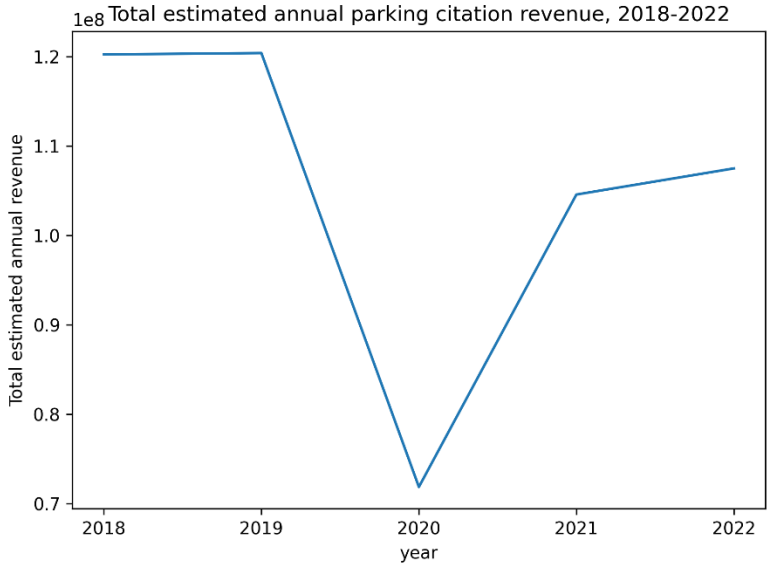


*Figure 5. Total estimated annual parking citation revenue between the years 2018-2022.*

### 3.3 Seasonality of Parking Citations

Figure 6 below shows the monthly estimated total counts of parking citations, based on the same subsets of data from 2018-2022 used in Figure 5. Except for 2022, which was impacted by the COVID-19 pandemic, we see no clear seasonality in the data. The (estimated) monthly counts of citations appear to be mostly random. In 2022 however, we see a clear downward dip when the pandemic started.

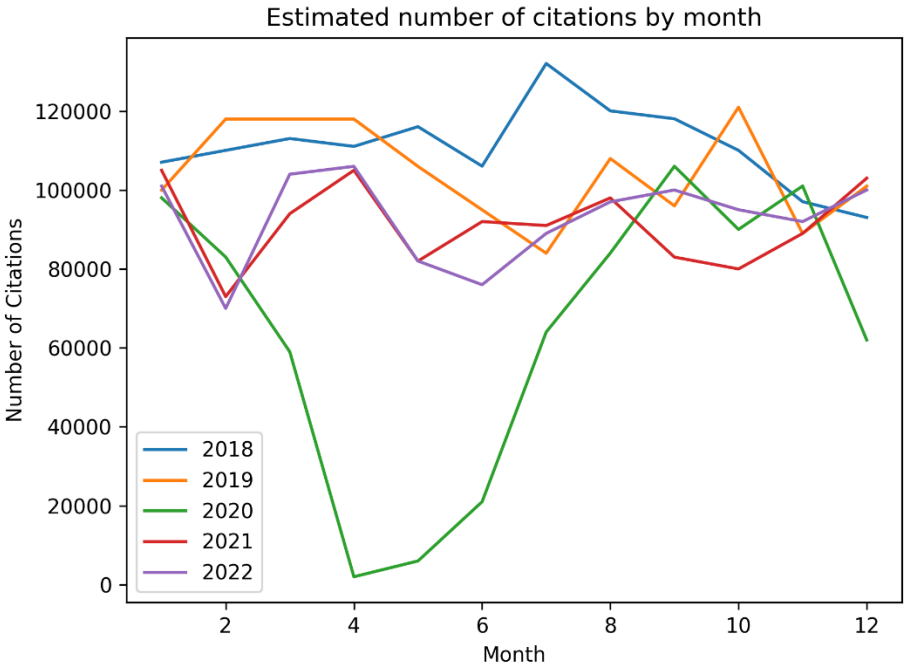


*Figure 6. Estimated monthly counts of parking citations for the years 2018-2022*

## 4 Conclusions

In this project, we source parking data from the SFMTA, create and maintain a database with the data, and conduct a data analysis focusing geospatial trends on both the coordinate individual and zip code level, as well as temporal trends on the annual, monthly and, hourly time scale. We focused primarily on estimated revenue, counts of citations, and frequencies of different types of violations.

Our key findings include:

1. Downtown and areas near downtown are associated with the greatest number of citations and the highest revenue. However, this is not necessarily strongly correlated with population. We suspect that this would be more correlated with high traffic since office and commercial spaces may not correspond to high population, but further research is needed to confirm this.

2. We find that the majority of parking citations occur between working hours (6am and 6pm) and that street cleaning violations are the most common violation type. Citations that occur outside of working hours are primarily street cleaning violations.
3. The COVID-19 pandemic may have caused a significant decrease in parking revenue in 2022, and levels have still not returned to pre-pandemic levels (and may never return due to new work-from-home trends). With the exception of the pandemic, there is no clear reason to suspect that parking revenue changes much annually.
4. There are no clear monthly trends in the number of parking citations issued, with the exception of the timeframe during the onset of the COVID-19 pandemic in 2020.

Future work may include the prediction and forecasting of parking revenue, given more time and resources. For example, one could attempt to predict potential revenue from a given location given observed revenues using time-series forecasting or using a recurrent neural network. This would be especially beneficial with greater computational resources so one can take full advantage of the large dataset.

# 5 References


1. SFMTA Parking Citations: https://data.sfgov.org/widgets/ab4h-6ztd
2. SFMTA API: https://dev.socrata.com/foundry/data.sfgov.org/ab4h-6ztd
3. USPS ZipCode Lookup: https://www.usps.com/business/web-tools-apis/address-information-api.htm
4. SF Zip Codes and Population GeoJSON: https://data.sfgov.org/Geographic-Locations-and-Boundaries/San-Francisco-ZIP-Codes/srq6-hmpi


**Github Repository:**

https://github.com/Matt2371/SF-parking-citations