Estimating the Probability of Receiving a Parking Meter Violation in San Francisco

Tim Tantivilaisin
University of California, Berkeley
timt@berkeley.edu

Abstract

This paper presents a comprehensive review of the steps our group has taken over the semester to rigorously calculate the probability of receiving a parking ticket in San Francisco. We go through the original preliminary approach and questions, necessary pivots, data set, final model, applications, and limitations.

1 Introduction

The initial goal of our project was rigorously answer the following two questions about San Francisco (SF) parking citations. The first being: what is the probability of receiving a parking ticket given a time and place in San Francisco given that the individual is committing an infraction? The second being: are these citations being given out fairly? Or to put it into other terms, would one be less likely to receive a parking ticket if they are parked in a wealthier neighborhood? This paper will go over how these questions and our approach to answering them changed over time.

Before we proceed further, we will go over a very brief overview of the subject matter. Parking violations in SF are given out by the San Francisco Municipal Transportation Agency (SFMTA) under the division of parking enforcement. Some infractions that are enforced are: not paying for meters, color curbs, double parking, etc. The enforcement of these rules come down to Parking Enforcement Officers (PEO). These PEOs are in small vehicles that patrol the streets of SF looking for cars that violate the citys laws. While this may seem like just a small operation, just in 2022 alone there was approximately $100,000,000 in revenue generated through these tickets alone.

Returning to the problem of prediction, if answered, can help both the citizens of SF in addition to the SFMTA. When citizens are well-informed about the probability of receiving a traffic ticket, they can make better decisions regarding their driving behavior and compliance with traffic laws. This knowledge can lead to several benefits for both citizens and enforcement agencies. A better understanding of traffic ticket probabilities can help drivers avoid violations and subsequent penalties. This can lead to a reduction in legal costs for citizens and a decreased administrative burden for enforcement agencies, as there will be fewer cases to process. Knowledge of traffic ticket probabilities can provide valuable insights for policymakers, enabling them to develop targeted interventions and initiatives aimed at improving road safety and compliance. This can lead to more effective and evidence-based traffic policies.

The structure of the rest of the paper will be the following: data description, why our original questions and modeling failed, our final modeling steps, application, and potential avenues for improvement.

2 The SFMTA Parking Dataset

The data are directly from data.sfgov.org where every citations information is uploaded in tabular form daily starting from 2008. Giving us the time, location (longitudinal), address, violation type, etc. We downloaded it directly as a .csv before importing it onto Python for our data wrangling. We first had to wrangle our data into data types that we could work into our geospatial and temporal analysis. To make the analysis more manageable, we filtered the data to only January 2022 to February 2023 and just meter violations.

The second dataset that we incorporated into our analysis is a street cleaning dataset. This dataset is also in a tabular csv, giving us the schedule of street cleaning corresponding to each street. Most importantly, it included the endpoints of each street segment. Where street segment is defined as one side of a block, that is intersected at each end by two or more cross streets. This dataset gives us the added granularity of not just looking at entire streets, which is an issue as there is heterogeneity in the length of streets. I.e., the number of tickets given on a street that is 5 miles long is not a direct comparison to one that is 200 feet long. Furthermore, since we are given two endpoints, we can directly calculate the distance of each street segment, something we were unable to do before with the original data.

The last two datasets we incorporated into our analysis parking meter locations and meter transactions. Where the meter locations dataset contained every meter in San Francisco, which we were able to correspond to street segments. Finally, the transaction dataset contained every payment corresponding to a meter in San Francisco.

3 Why Our Original Approach Failed

When constructing our preliminary statistical model (before the midterm presentation), we chose the Poisson regression. We concluded that it was the most appropriate as we were trying to approximate the rate of tickets at each location, given a time parameter, over a duration of time. However, the downside of this model is that we still could not solve the problem of not having the denominator in the following equation: Before fitting the model, we defined a training and testing split. Training the model on January 2022 and testing it on February data. The specific model that we ultimately decided to move forward with was from CatBoostRegressor package with the Poisson objective. So given the features of longitudinal coordinates of a street section, citation type, and lag variables (of two weeks), it would predict the number of citations that would occur by street section on each day in our test set. Ultimately giving us the result of: R2 = 0.237 and the RMSE of 0.28.

$$\mathbb{P}(\text{Ticket}|\text{Parked Illegally}) = \frac{\mathbb{P}(\text{Ticket } \cap \text{ Parked Illegally})}{\mathbb{P}(\text{Parked Illegally})}. \tag{1}$$

It became evident that the Poisson Regression model reached a dead end due to the limited information available. Even if we could accurately predict the count, it would be impossible to calculate the denominator. Thus, we were unable to calculate probabilities with this approach.

This realization prompted us to reconsider the probabilities we were trying to calculate and instead define a proxy probability that could be calculated. Before introducing this proxy, we made certain modeling assumptions. We assumed that if a car is committing an infraction and a parking enforcement officer is on the street, then the car is guaranteed to receive a ticket. Additionally, we assumed that all parking spots are consistently occupied. Under these assumptions, our focus shifted from the number of cars committing infractions to the mere presence of an infraction. Therefore, our goal became predicting the probability of enforcement officers being present on a given street section.

The next section will outline how we notated our probabilities, prepared data, and calculated our probabilities.

4 Final Model

4.1 Notation

Before delving into data sources and determining the necessary information for estimating probabilities, let us establish some notation and frame the problem at hand. We have defined the following variables:

$$W \in \{\text{Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday}\}$$

$$S \in \{1, 2, \ldots, N\}$$

$$T \in \{\text{9:00am}, \text{9:15am}, \text{9:30am}, \ldots, \text{5:45pm}\}$$

$$E = \text{event that enforcement happens}$$

$$I = \text{event that illegal parking happens.}$$

Where W is the set containing the days of the week. S is the set containing the segment ID of all unique street segments in San Francisco. T is the start time bin incremented by 15 minute intervals containing all times when parking meters are enforced. E is the event that enforcement happens and I is the event that someone is parking illegally.

Suppose we are at time t, on street segment s, on weekday w. Then we want to know the probability that we get a ticket given that we parked illegally. We denote this as the following:

$$\begin{aligned} p_{t,s,w} &= P(E \mid I, T = t, S = s, W = w) \\ &= \frac{P(E \cap I \mid T = t, S = s, W = w)}{P(I \mid T = t, S = s, W = w)} \end{aligned} \tag{2}$$

Then applying the formula conditional probability, we get line two. This is ultimately the form we are trying to calculate. We can now construct estimates of the numerator and denominator from the data.

4.2 Preparing Data for Analysis

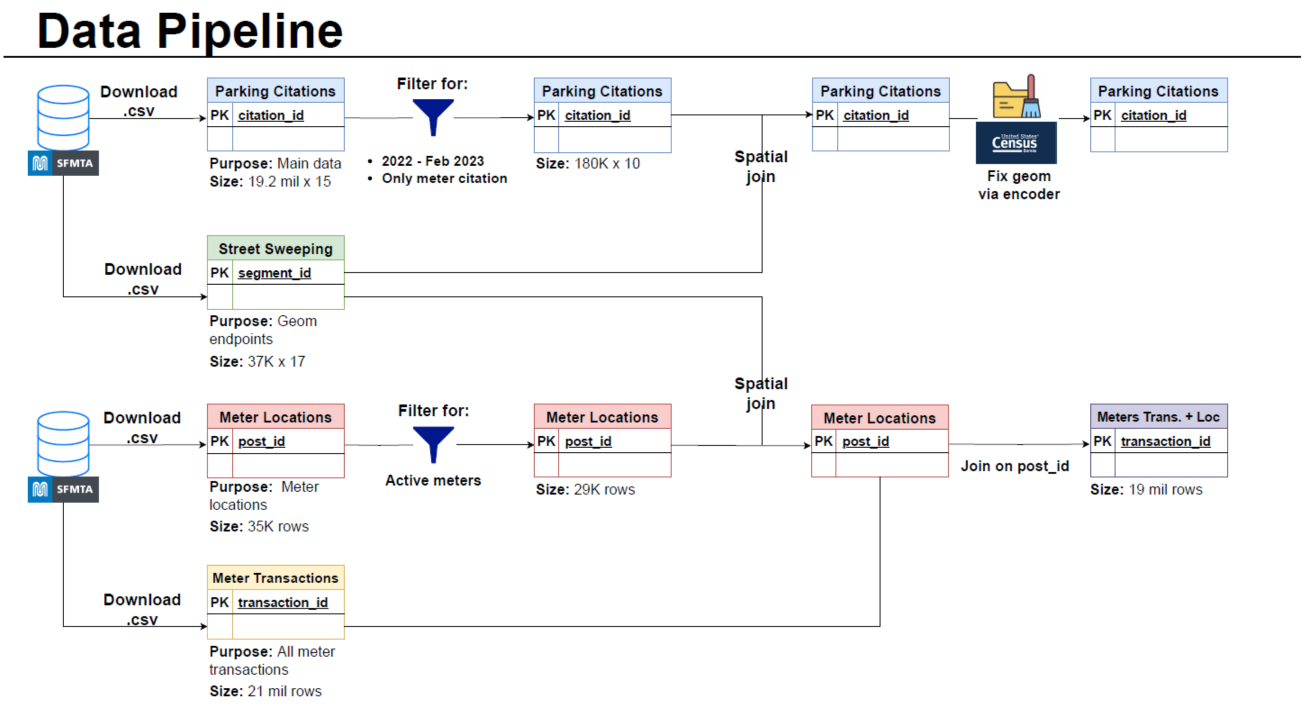


Figure 1: Data pipeline for the final model.

In our analysis, we are working with four distinct tables, each color-coded in Figure 1. The top sequence of tables corresponds to the data that provides us with the numerator for our calculations, while the bottom sequence represents the tables that contribute to the denominator. The primary dataset we are utilizing is the SFMTA parking citations dataset, represented by the blue tables. This dataset comprises 19 million rows, with each row corresponding to a unique citation incident. We applied filters to consider only violations that occurred between the years 2022 and 2023 and meter violations.

In order to obtain the street ID associated with each citation incident, as the blue dataset only provided latitude and longitude information, we performed a spatial join with the street sweeping dataset, represented by the green table. This dataset contains the geometric endpoints of each street segment. To ensure data quality and correct for any inaccuracies in the geometric encodings, we utilized the US Census geoencoder. This step was crucial in achieving the final dataset required for estimating the numerator of our probability calculation.

To estimate the denominator, denoted as $P(I)$, we need to identify all instances of illegal parking. In order to achieve this, we focus on meter violations since there exists comprehensive transactional data for each meter in San Francisco. This information is available in the yellow dataset. By considering the transaction data and the corresponding meter locations, we can infer the times when parking meters were unpaid. Under the modeling assumption that highly trafficked spots are consistently occupied, these unpaid instances represent cases of illegal parking. By utilizing this dataset, we can estimate the denominator required for our probability calculation.

We acquired the meter location dataset from the SFMTA, represented by the red tables. We applied filters to consider only active meters. Next, we performed a spatial join with the street sweeping dataset, allowing us to associate unique street IDs with the meter locations. This step, represented by the green tables, helped establish the relationship between meter locations and specific street segments.

Finally, we joined the resulting dataset with the meter transaction dataset, represented by the yellow tables. This combination enabled us to obtain all transactional data associated with each meter located on each street segment. By completing these steps, we obtained the final tables necessary to estimate the denominator for our probability calculation.

4.3 Final Modeling Steps

We begin by outlining to steps to calculating the following probability illustrated in Figure 2:

$$P(E \cap \widehat{I \mid T = t, S = s, W} = w) = \frac{\sum_{\text{weekday(date)}=w} \mathbb{1}[\#\text{tickets}(s, t, \text{date}) > 0]}{\#\text{weekday(date)} = w} \tag{3}$$

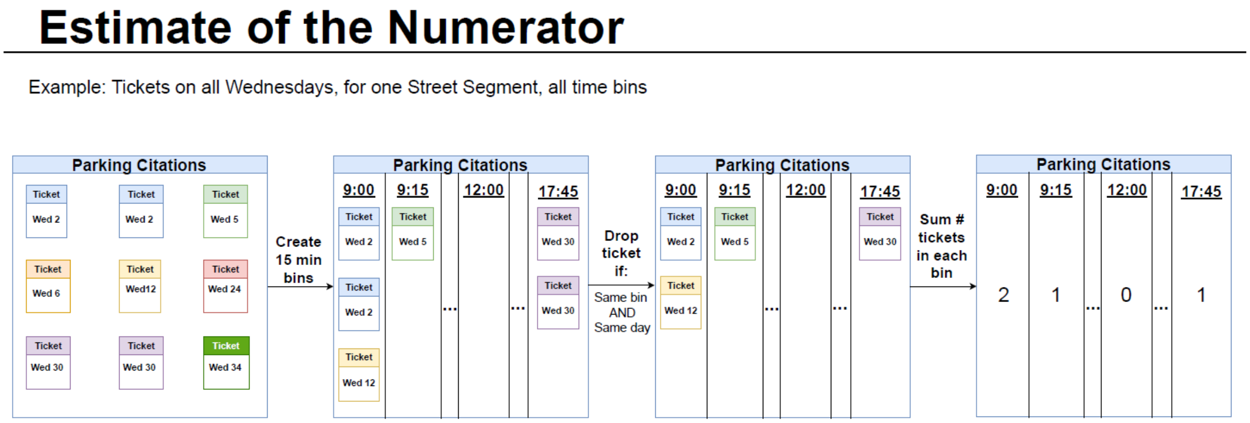


Figure 2: Diagram detailing how to calculate the numerator.

Without loss of generality, we focus on an example where we are on a specific street segment, denoted as s, and a particular day of the week, such as Wednesday. We begin by taking each citation incident that occurred on a Wednesday within street segment s and assigning it to a fifteen-minute time bin. For example, a ticket issued at 9:02 would be allocated to the 9:00 bin, while a ticket at 9:23 would be assigned to the 9:15 bin, and so on.

To obtain a single representative ticket for each time bin on Wednesdays, we record only one ticket if there were multiple citations during the same Wednesday period. This step ensures that we have an indicator of whether or not there was at least one instance of enforcement and illegal parking on a specific day. Finally, we calculate the sum of the number of tickets within each time bin and divide it by the total number of Wednesdays present in the dataset. This computation provides us with an estimate of the probability of $P(E \cap I|W = \text{Wednesday}, \text{Street} = s)$ for a specific time bin.

The following figure illustrates how we calculated the denominator:

$$\widehat{P(I \mid T = t, S = s, W = w)} = \frac{\sum_{\text{weekday(date)}=w} \mathbb{1}[\#\text{unpaid meters}(s, t, \text{date}) > 0]}{\#\text{weekday(date)} = w} \tag{4}$$

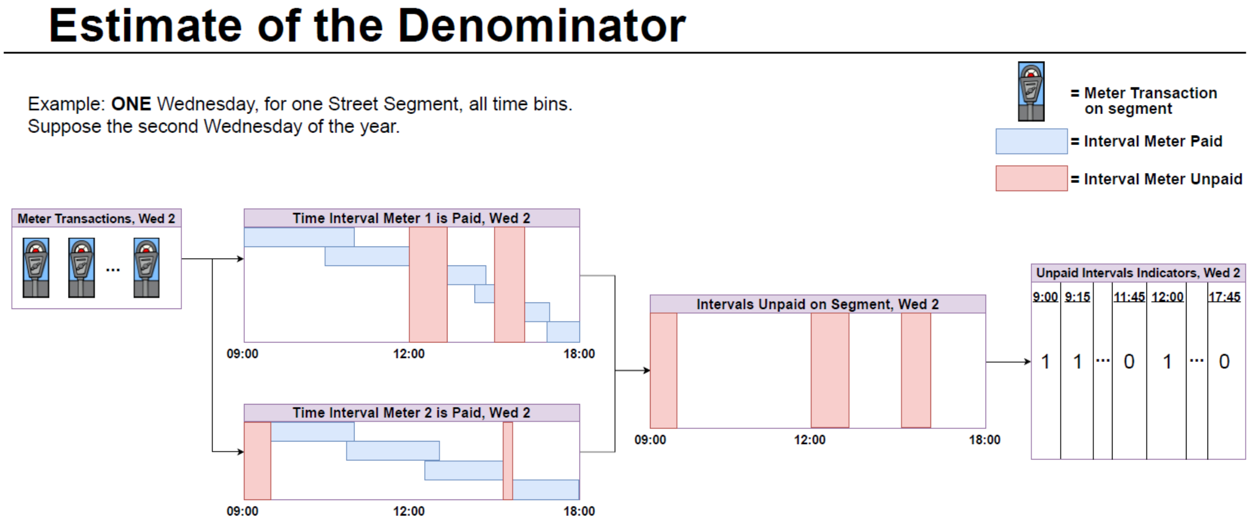

Estimate of the Denominator
Example: ONE Wednesday, for one Street Segment, all time bins.
Suppose the second Wednesday of the year.
= Meter Transaction on segment
= Interval Meter Paid
= Interval Meter Unpaid
Meter Transactions, Wed 2
...
Time Interval Meter 1 is Paid, Wed 2
09:00
12:00
18:00
Time Interval Meter 2 is Paid, Wed 2
09:00
12:00
18:00
Intervals Unpaid on Segment, Wed 2
09:00
12:00
18:00
Unpaid Intervals Indicators, Wed 2
9:00
9:15
11:45
12:00
17:45
1
1
...
0
1
...
0

Figure 3: Diagram detailing how to calculate the denominator.

Once again without loss of generality, we consider a single specific Wednesday, observing all time slots, for a distinct section of street s that contains multiple parking meters. For each meter we bin the transactions into 15-minute intervals. We also incorporate a 3-minute grace period to allow for the transition between one car leaving and the subsequent vehicle parking and completing payment.

In instances where no transactions span a 15-minute slot, or if there are transaction gaps exceeding our 3-minute buffer, we declare the meter as unpaid during that period. We then accumulate all the intervals in which at least one meter remained unpaid on that street segment. This data is used to generate an indicator for the corresponding time slot.

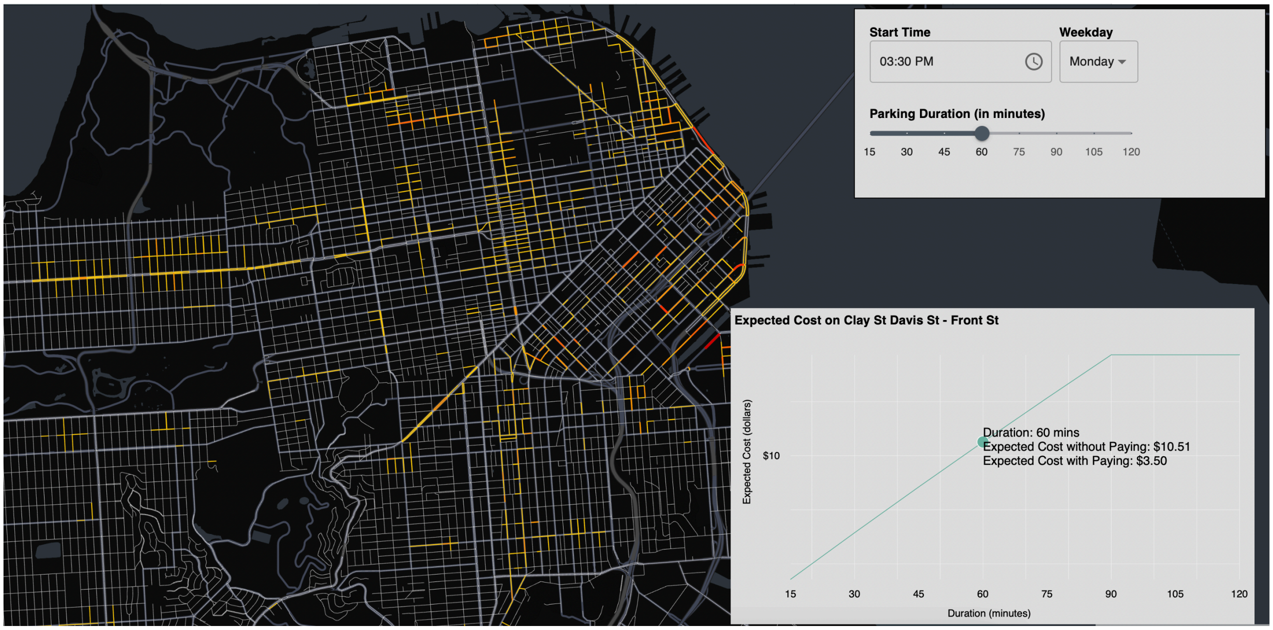
While not explicitly illustrated in the figure, we sum up these indicators across all Wednesdays and divide by the total count of Wednesdays within our dataset with respect to each time bin. This provides an estimate of $P(I|W = \text{Wednesday}, \text{Street} = s)$. To get the final probability, we simply perform the division with respect to T, S, W.

5 Application of our Analysis

5.1 The Application

To make our probabilities more concrete and applicable for San Francisco's residents, we designed an interactive web application illustrated in Figure 4. This tool allows users to select a specific time, day of the week, and desired parking duration, ranging from 15 minutes to 2 hours. It then displays the probability of receiving a parking ticket for each street segment in real-time. A color gradient is used for visual clarity, with more intense red indicating a higher ticket probability.

By hovering over a street, users can view the exact probability of receiving a ticket within a fifteen-minute interval. Additionally, clicking on a street segment triggers a displays a plot with time on the X-axis and the expected cost in dollars on the Y-axis. This feature allows users to visualize the expected cost of illegal parking on a specific street segment for a given duration, and compare this with the cost of paying for the meter.



Start Time
03:30 PM

Weekday
Monday

Parking Duration (in minutes)

Expected Cost on Clay St Davis St - Front St

Duration: 60 mins
Expected Cost without Paying: $10.51
Expected Cost with Paying: $3.50

Figure 4: Web app we created using our probabilities. In this particular example, if the user were to want to park for 30 minutes, they would want to pay for parking.

To calculate the probability multiple time intervals, we treat each interval as an independent Bernoulli and use the following equation:

$$\hat{p}_{d,s,w} = 1 - \prod_{i=1}^{k} (1 - \hat{p}_{t_i,s,w}) \tag{5}$$

Where d is the duration of time and k is the number of 15-minute segments. To calculate the expected cost without paying, we calculated: $93.50 \cdot \hat{p}_{d,s,w}$.

5.2 Limitations of Our Model

To manage the complexity of this application, we made a few modeling assumptions. First, we treated the average cost per hour for all meters, $3.50, as a constant across all meters and time intervals. In reality, this is not accurate as meter prices in San Francisco are dynamically adjusted in response to demand. Secondly, we assumed the average cost of a parking violation, $93.50, to be the standard for all tickets. However, the actual cost varies between downtown and non-downtown locations. Lastly, our model presumes that these parking spots are occupied at all times. Therefore, our model tends to be more accurate for busy areas near major points of interest, but less so for less trafficked locations. It's important to keep these assumptions in mind when interpreting the outputs of our web application.

6 Further Work

As previously noted, our assumption that metered spots are consistently occupied is only accurate in certain areas. This means that in many cases, our model may underestimate the real probability of receiving a parking ticket. Acquiring data that could help refine this assumption would require on-the-ground presence in San Francisco, which was not feasible for us during the semester. However, future improvements could be made through methods like conducting surveys to gather actual infraction rates, or installing parking sensors at a sample of metered spots.

Returning to the question of equity, our estimated probabilities could serve as a tool for assessing whether ticketing is conducted fairly across the city. To further probe this issue, we could conduct an ANOVA across different neighborhoods. This would enable us to determine if certain parts of the city are disproportionately ticketed relative to factors like the size of the area, number of meters, and other relevant parameters. Such an analysis could provide valuable insights into the equitable distribution of parking enforcement, potentially highlighting areas for improvement in the city's ticketing policy.

References

[1] Ning Jia (2022): What are the odds of getting a parking ticket in Toronto?, https://towardsdatascience.com/what-are-the-odds-of-getting-a-parking-ticket-in-toronto-1f090d d0c608

[2] Song Gao, Mingxiao Li, Yunlei Liang, Joseph Marks, Yuhao Kang & Moying Li (2019): Predicting the spatiotemporal legality of on-street parking using open data and machine learning, Annals of GIS, DOI: 10.1080/19475683.2019.1679882

[3] **Parking Citation Data:** https://data.sfgov.org/Transportation/SFMTA-Parking-Citations/ab4h-6ztd

[4] **Street Sweeping Data:** https://data.sfgov.org/City-Infrastructure/Street-Sweeping-Schedule/yhqp-riqs

[5] **Census geocoder:** https://geocoding.geo.census.gov/geocoder/

[6] **Meter Transactions:** https://data.sfgov.org/Transportation/SFMTA-Parking-Meter-Detailed-Revenue- Transactions/imvp-dq3v/data

[7] **Meter Locations:** https://data.sfgov.org/Transportation/Map-of-Parking-Meters/fqfu-vcqd

[8] **Census geocoder:** https://geocoding.geo.census.gov/geocoder/