**An Analysis of Parking Citations in the City of Baltimore**

*Tamara Monge*

January 1, 2018

1. **The Problem**

Parking violations are an increasingly common concern in metropolitan areas. Not only are they a crime, but they can reasonably be understood as a measure of the populace’s frustration with the availability of legal parking. When parking violations occur, a citation is issued and the local municipality collects the fine as revenue. In this project I analyzed patterns in parking citations issued by the City of Baltimore over the past 26 months.

1. **The Client**

The client in this study would be the City of Baltimore. Since parking fines are a significant source of revenue, the city would be interested in being able to predict whether or not a given fine would be paid based on features of the citation itself, such as: it’s location, the month, day, and hour of occurrence, the type of violation, amount of the fine, the make of the vehicle and whether or not the vehicle is from Maryland or elsewhere.

1. **The Dataset & Wrangling Steps**

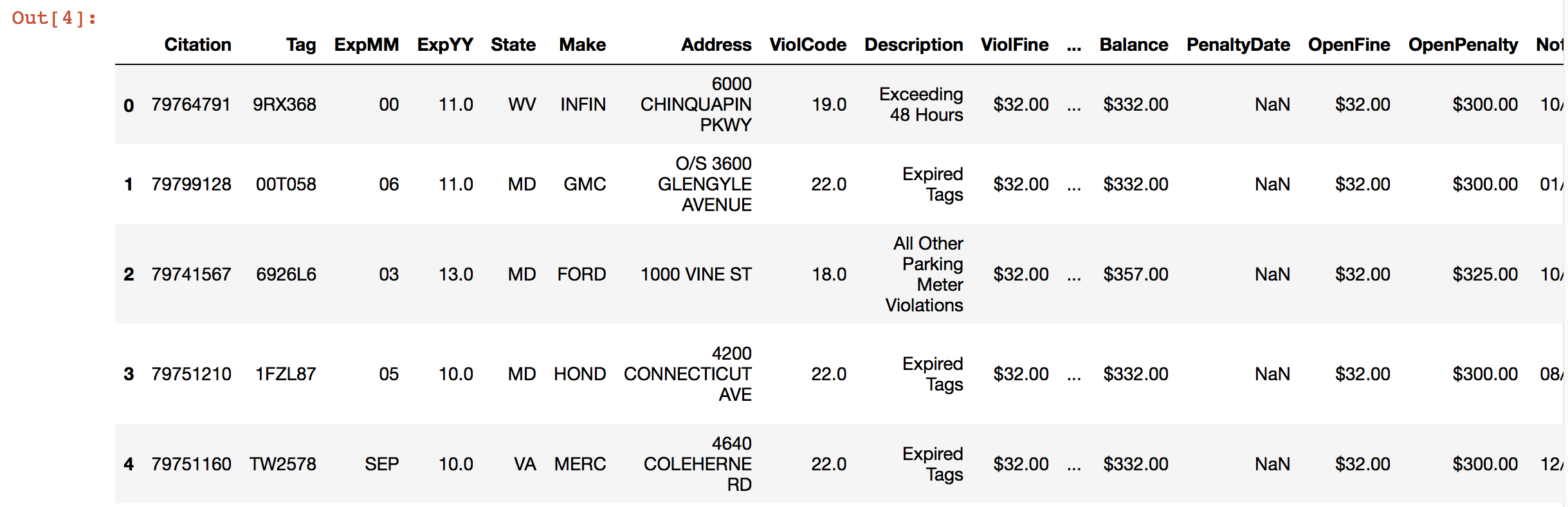
The dataset I used for this project was from the City of Baltimore website1. It is available for free download. The dataset contained two distinct cohorts of records: (1) a rolling record of all citations that were issued over the past two years, updated daily, and (2) a record of all citations that were issued more than two years ago but still had an outstanding balance. Each record contained the following information: date, time, and address of incident, violation description and code, citation number, license plate number, license plate state, fine amount, and account balance. Some records contained additional information, such as: latitude, longitude, neighborhood, police district, and council district of the incident, license plate expiration, vehicle make, and penalty amount (if any).

**3.1 Downloading and importing data**

In order to extend the length of the study beyond 24 months, I performed two downloads: one near the start of the project (September 23, 2017) and one a few months later (November 30, 2017).

Each downloaded dataset was provided as a raw CSV file. I imported them as Pandas dataframes, merged them, and dropped duplicate rows. The merged dataframe contained 1,503,910 records. The first five rows are shown in Table 1.

**Table 1. First five rows of merged dataset.**



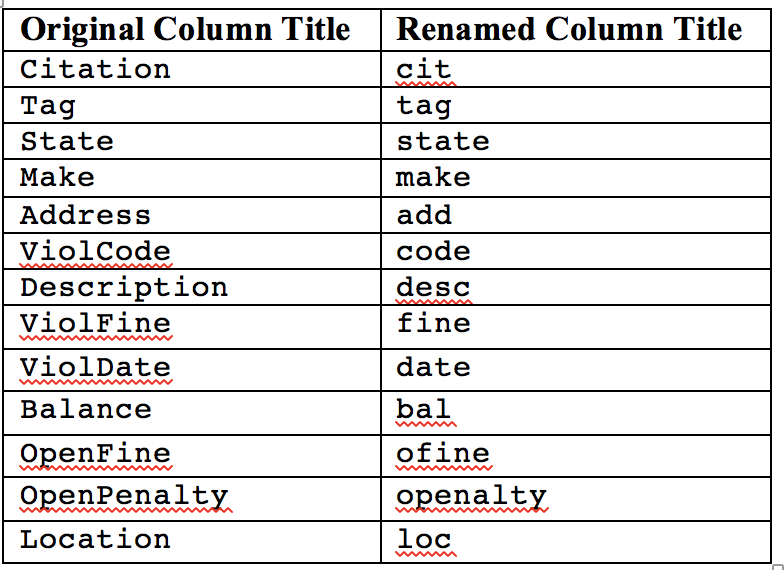
**3.2 Dropping irrelevant columns**

The dataset was large (~ 650 MB) and contained several columns which would not be needed in this study. In order to conserve space and computing time, I dropped the irrelevant columns: ExpMM, ExpYY, PenaltyDate, NoticeDate, ImportDate, Neighborhood, PoliceDistrict, and CouncilDistrict.

**3.3 Renaming columns**

In order to make data operations more convenient, I converted the column titles from title case to lowercase and shortened some of them. The original and renamed column titles are shown in Table 2.

**Table 2. Original and renamed column titles.**



**3.4 Extracting relevant (complete) cohort and assigning datetime index**

I realized that analyzing both cohorts of the dataset together would produce skewed results because one cohort was complete (it contained *all* records in the two-year history) and the other was incomplete (it contained only *older records with outstanding balances*). In order to avoid biasing my analysis toward the older, outstanding accounts, I chose to analyze only the first cohort of data. I converted the date column from a string object to a datetime object and used it to extract the first cohort (i.e., all records after September 23, 2015) to a new dataframe. The new dataframe contained 912,308 records. To assist with analysis and plotting, I then assigned date to be the datetime index of the dataframe.

**3.5 Dealing with missing values**

Approximately 29.7% of the records were missing latitude and longitude information. We attempted to remedy this by creating a dictionary of address:lonlat and using the dictionary to populate the missing values. Unfortunately, this resulted in only a very small improvement: a 0.3% increase in populated observations. However, since the dataset was so large, I found it reasonable to drop the records with missing geospatial information and conduct my study on the observations that contained spatial information. The resulting dataframe contained 641,072 records. It is important to remember this choice when interpreting the results of this study. It is possible there was an unknown pattern (period of time, geographic area, or particular officers) to the records that did not contain spatial information, such that the exclusion of those records may have introduced bias into the results.

**3.6 Removing records from outside Baltimore**

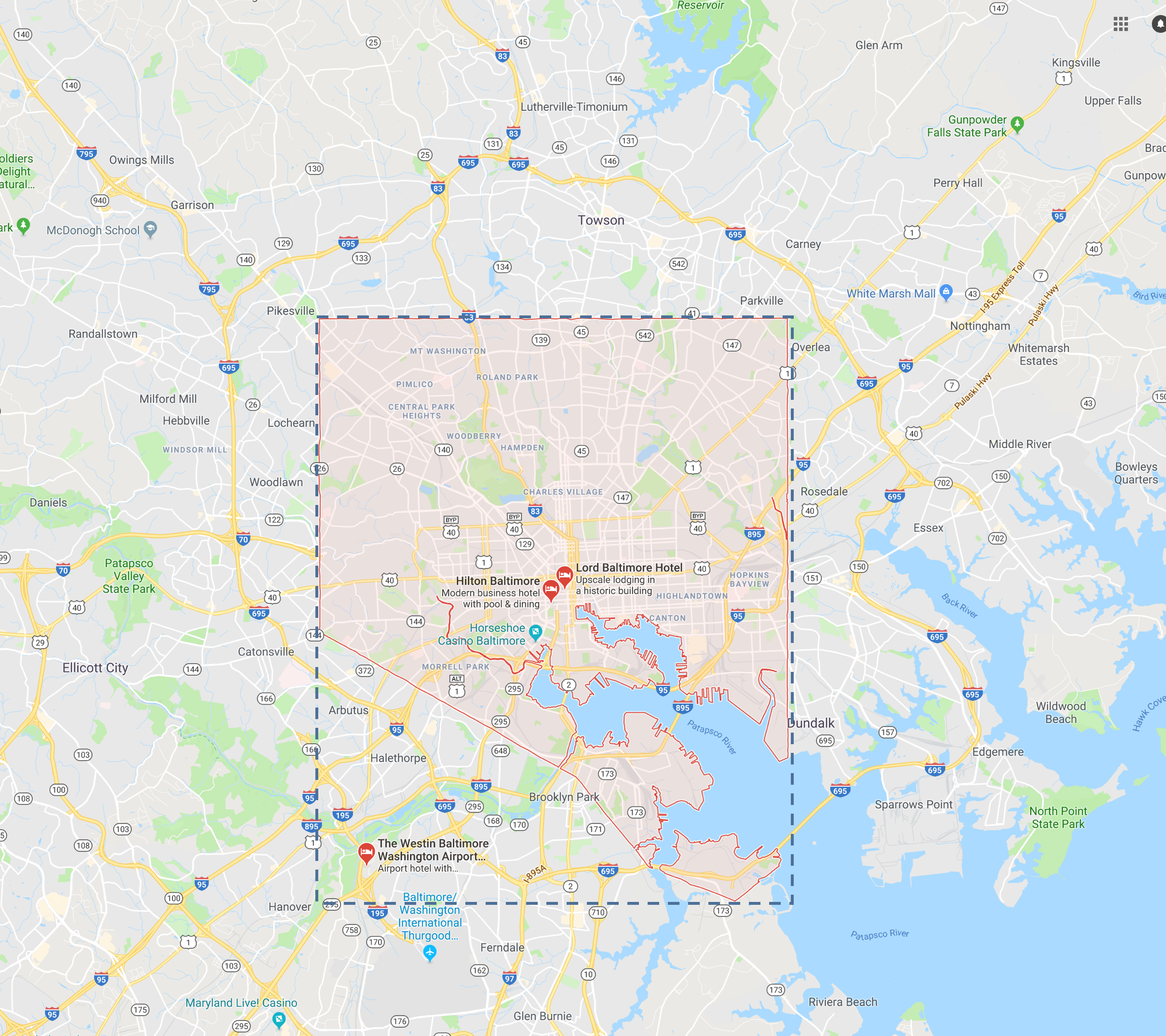
While examining the data, I noticed a small number (2.7%) of citations were issued outside of the city limits of Baltimore. This could be the result of xxx or xxx. Regardless of the reason for these extraneous records, my focus for this project was to examine parking citations in the city of Baltimore. Therefore, I removed the citations issued outside the city. The eastern, northern, and western borders of the city all fall approximately along latitude and longitude lines, while the southern side of the city has a more complex border. For simplicity, I drew a rectangle along the farthest edges of the city, as shown in Fig.1, and used that as my bounding box. This reduced our dataset to 623,639 (97.3% of records).

**3.7 Adding and cleaning columns**

The variables date and loc each contained several pieces of information that would be useful to examine independently. For ease of operations, I extracted these bits of information into separate columns named: yr, mo, day, hr, lat, lon, and lonlat.

The financial variables (fine, bal, ofine, and openalty) contained numerical information but were stored as string objects. In order for them to be treated as numerical variables, I removed the `$` character and converted the strings to floats.

Some of the string variables (add, state, make, and desc) contained more than one representation of the same value. For example, one record would indicate a vehicle make of `honda`, while another would indicate the same make as `HON`. To clean these columns, I set them to have consistent cases and consistent number of characters.



**Figure. 1. Map of Baltimore region2. Red shading indicates Baltimore city limits. Dashed rectangle indicates**

**bounding box used for selecting data inside Baltimore**

**3.8 Simplifying variables**

Three columns needed to be simplified into binary or categorical variables for use in the machine learning algorithms. With each of these columns, I retained the original data and added a simplified column.

I simplified the variable bal to a binary variable paid that indicated whether or not the balance had been paid down.

Next, I simplified the variable state to a binary variable instate that indicated whether or not the vehicle was from Maryland.

Lastly, I simplified the variable lonlat to a categorical variable quad that indicated in which quadrant of the city the incident occurred.

1. **Exploratory Data Analysis**

I have grouped the preliminary questions asked in this investigation according to what they reveal. For instance, regarding the violations themselves, what types of offenses were committed? Which were the most common offenses? With regard to the incidents, when did the most incidents take place? Which hour of the day was most common? With regard to the offenders, which vehicle make committed the most offenses? From which state did the majority of offenders come? Concerning the status of the accounts, what were the most common fine amounts? How many accounts had been paid down? Were account balances ever higher than the initial fines? Next, with regard to the volume of citations, on average how many citations were issued per day, per month, per year? Finally, with regard to city revenue, on average how much money did the city make from parking citations, per day, per month, per year?

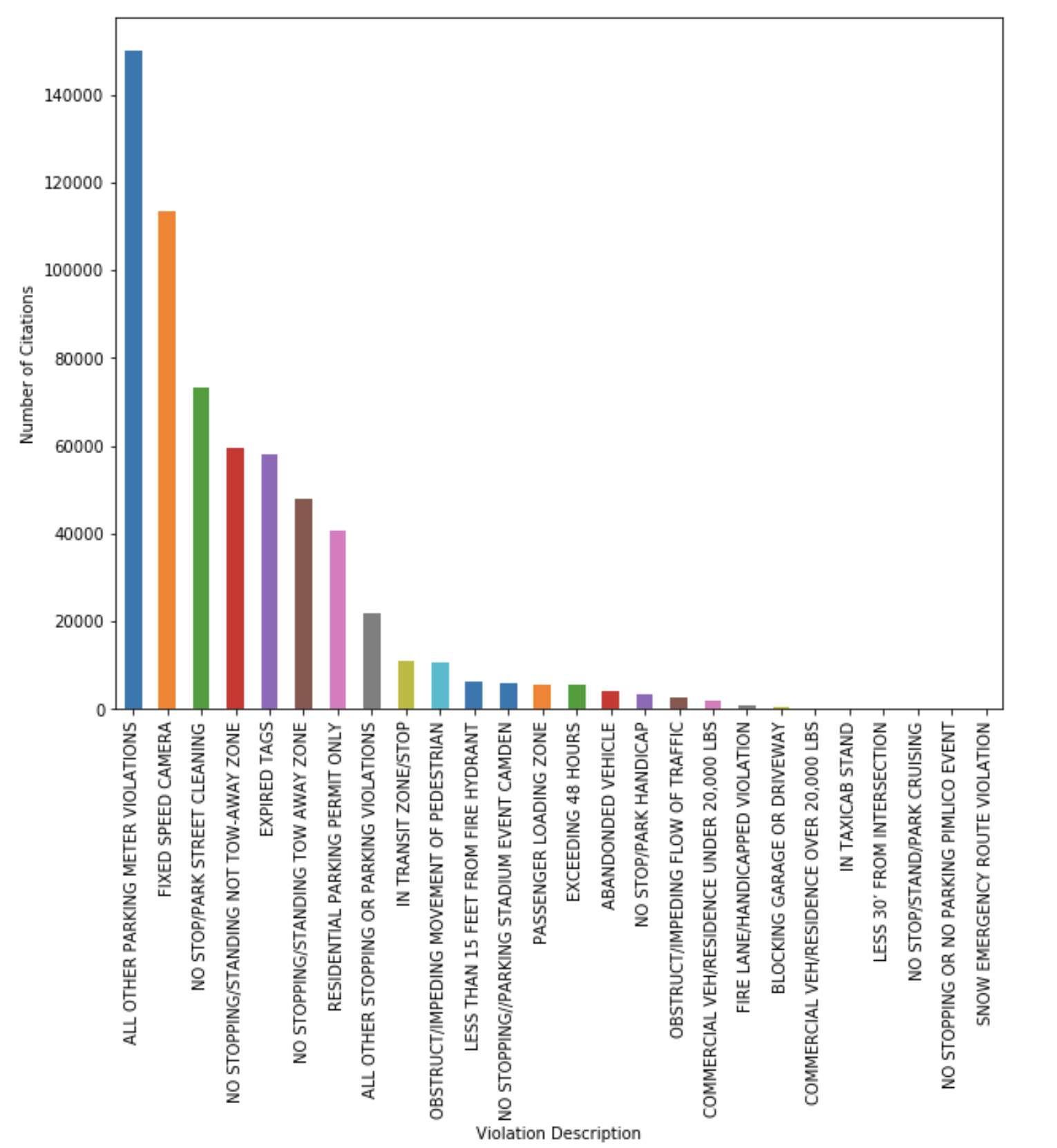
* 1. **The Violations**

Citations were issued for 26 unique violations (Fig. 2). The most common violation was “All Other Parking Meter Violations” (24%), followed by “Fixed Speed Camera” (18%), and “No Stop/Park Street Cleaning” (12%).

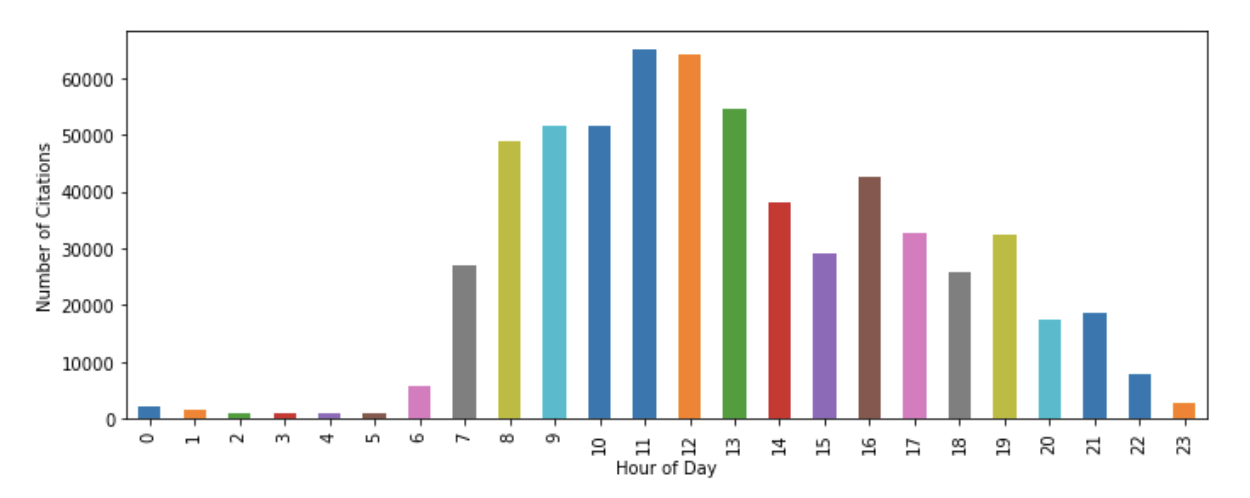
* 1. **The Incidents**

The hour of day that that saw the most citations was 11:00-12:00, followed closely by 12:00-13:00 (Fig. 3).

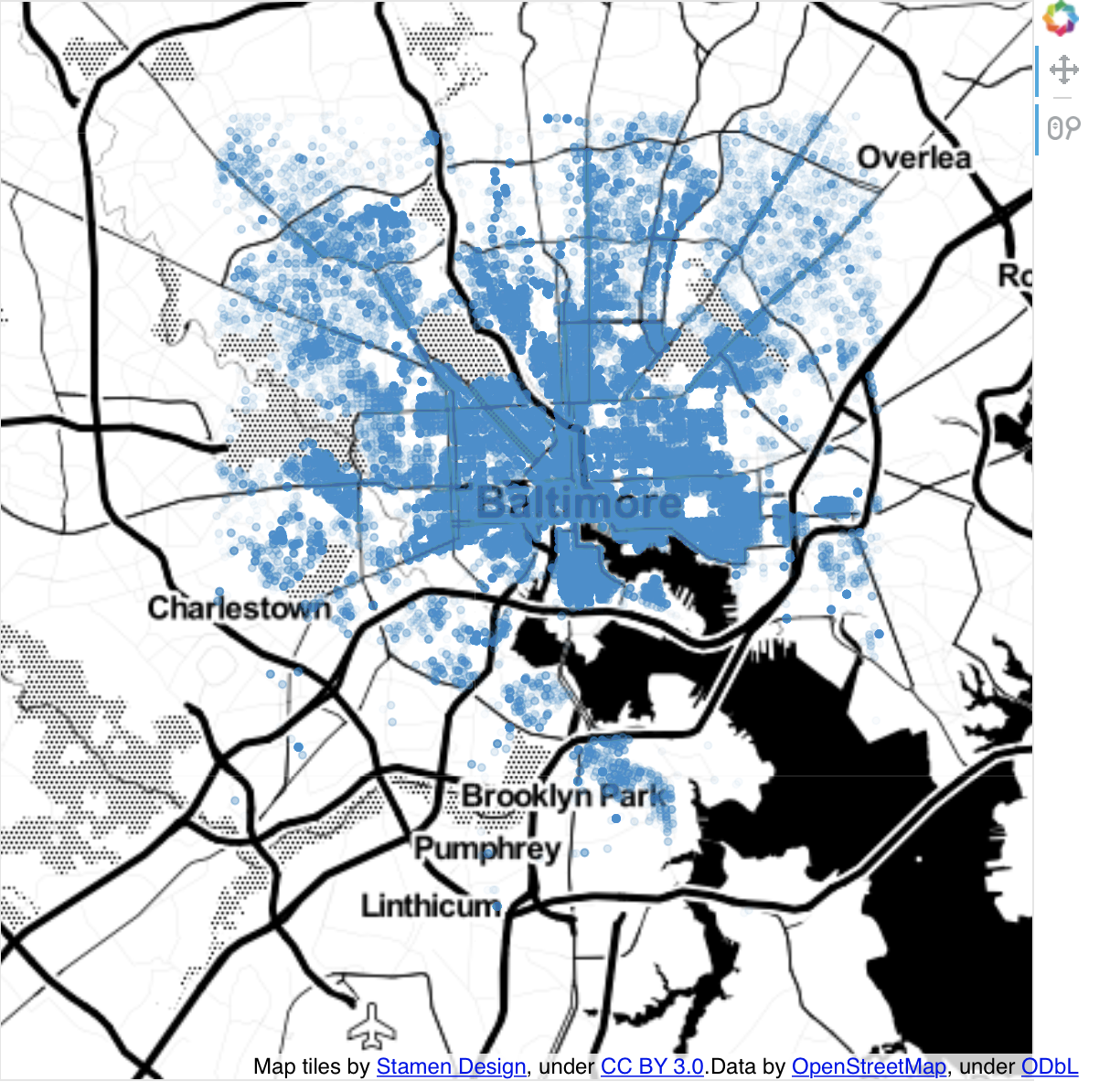
The highest concentration of citations occurred in central Baltimore (Fig. 4).

****

**Fig. 2 Volume of Citations by Description**



**Fig. 3 Volume of Citations by Hour of Day**

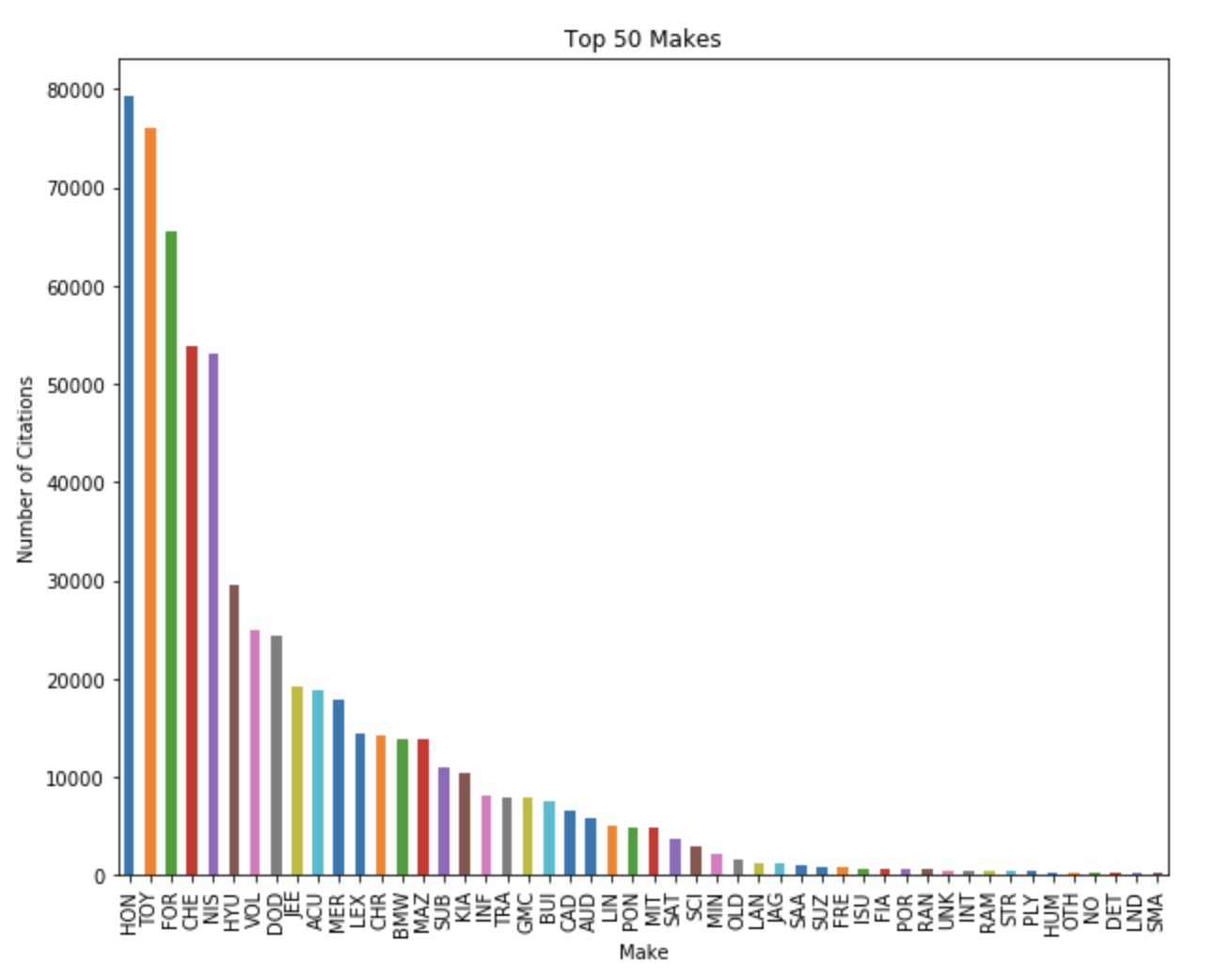


**Fig. 4 Citation Locations.**

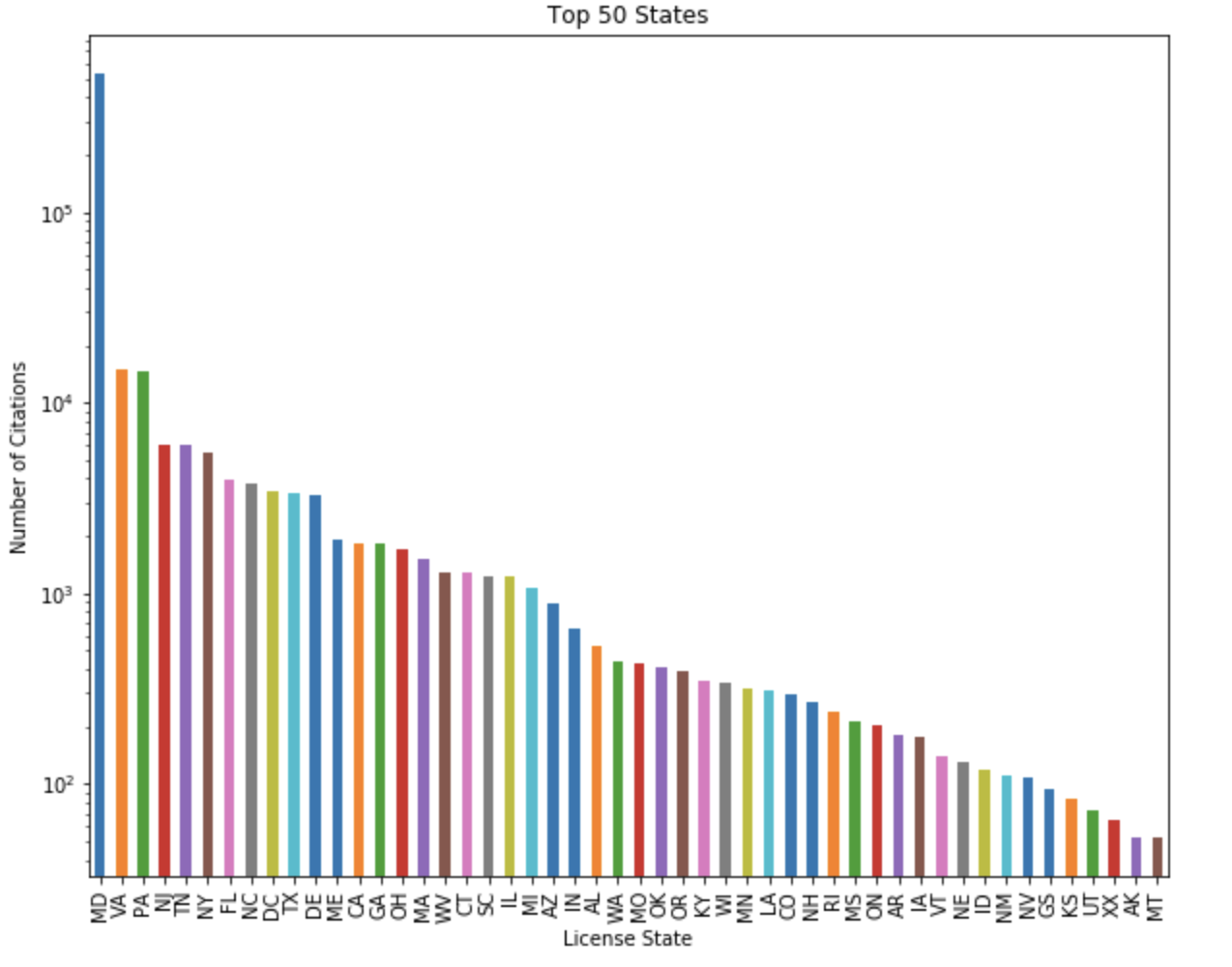
* 1. **The Offenders**

Citations were issued to 317 different makes of vehicle (Fig. 5). The makes to receive the most citations were: Honda (13%), Toyota (12%), and Ford (11%).

The most common state of offending vehicles was, not surprisingly, Maryland (86%), followed by Virginia (2.4%) and Pennsylvania (2.3%) (Fig. 6).



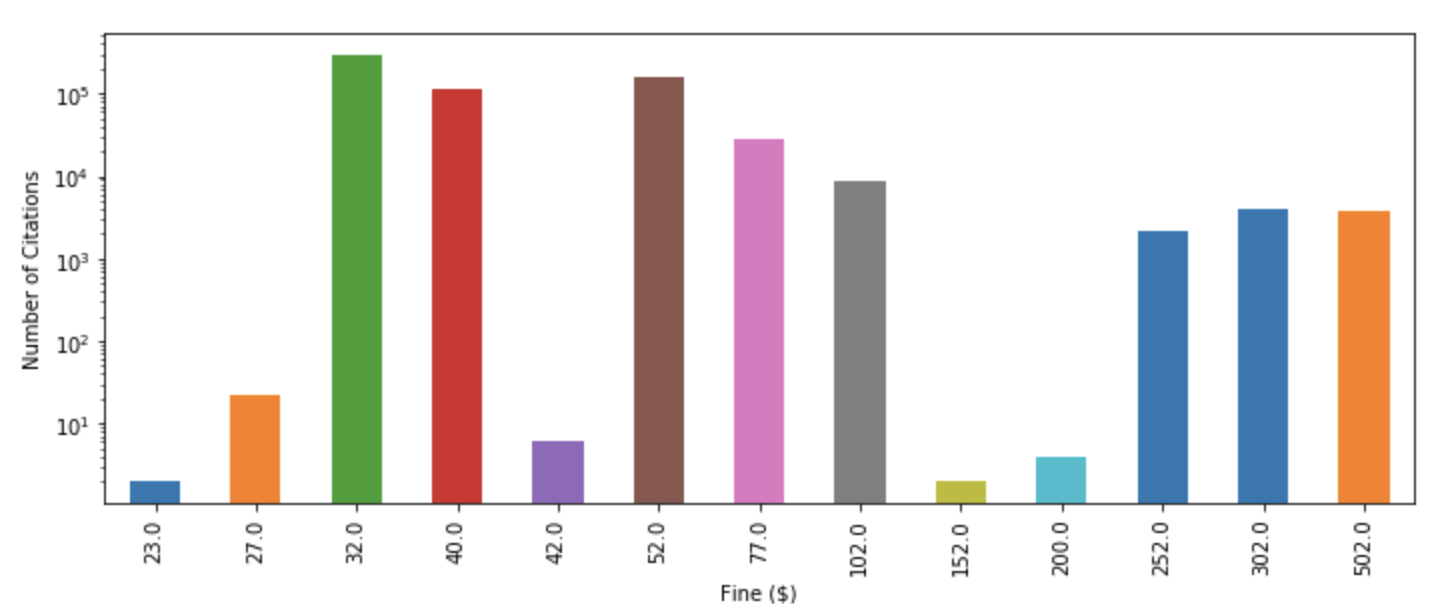
**Fig. 5 Volume of citations by vehicle make.**



**Fig. 6 Volume of citations by vehicle’s state. Note the logarithmic scale.**

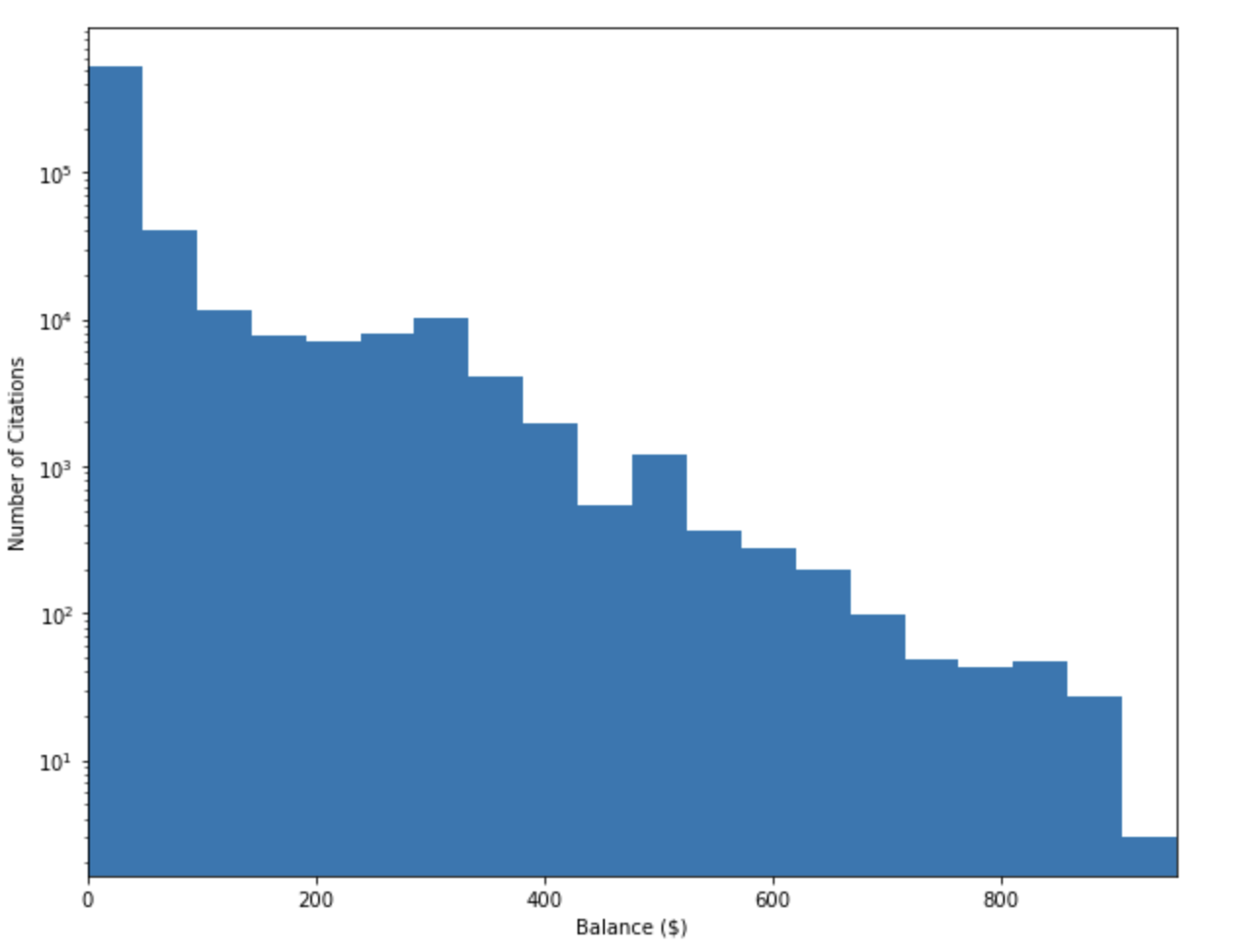
* 1. **The Status of Accounts**

The fines for parking violations ranged from $23 to $502 (Fig. 7). 48% of fines were exactly $32, 67% of fines were less than $50, and 97% of fines were less than $100.



**Fig. 7 Volume of citations by fine amount. Note the logarithmic scale.**

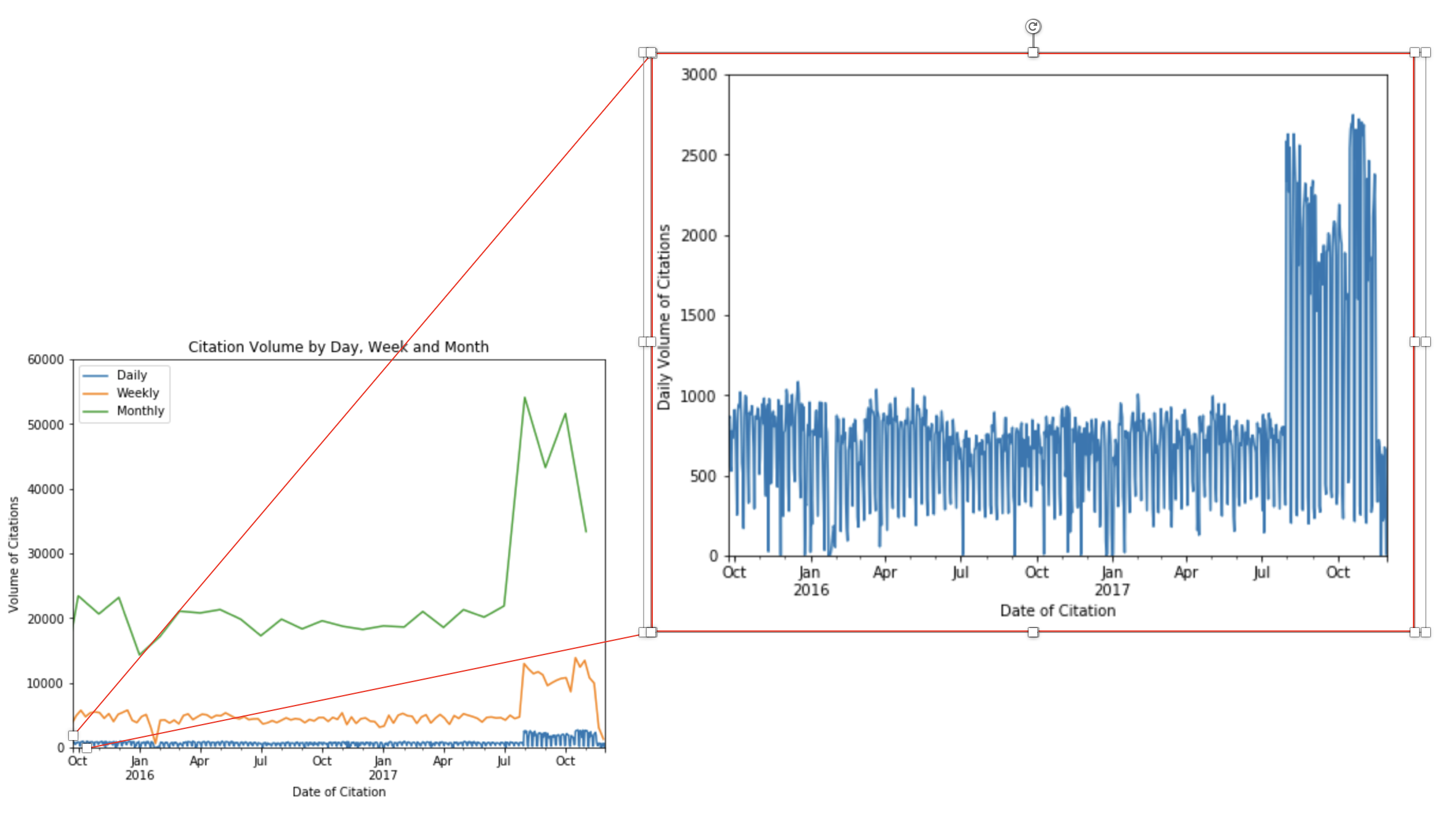
The balance due on accounts ranged from $0 to $954 (Fig.8). Some accounts had balances greater than their initial fine because they were issued a penalty for delinquent payment. 67% of accounts carried a balance of $0, meaning they had been paid down.

****

**Fig. 8 Volume of citations by balance due on account. Note the logarithmic scale.**

* 1. **Citations and Revenue Over Time**

On average, the city issued 23,082 citations per month, and 780 citations per day (Fig. 9). If we assume all fines will eventually be paid, this translates to an average revenue from parking citations of, $1.08 million per month or $36,700 per day.

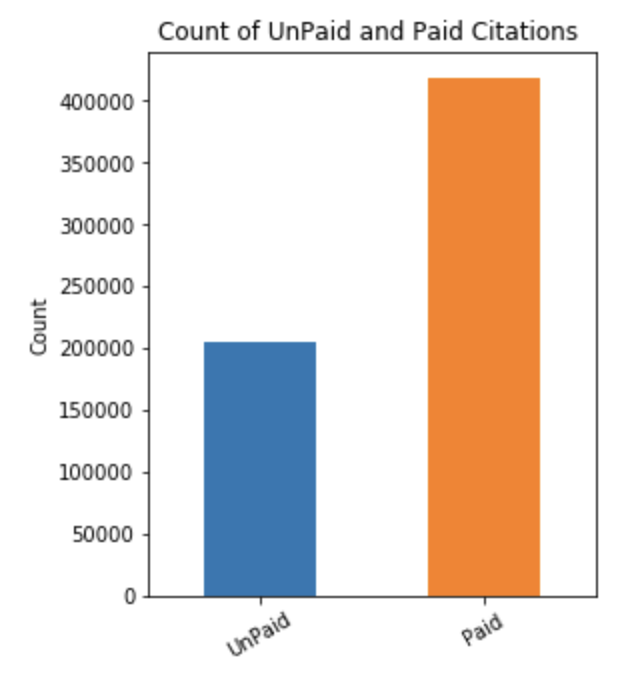
****

**Fig. 9 Citation volume over time.**

1. **Inferential Statistics**

67% of citations have been paid down (Fig. 10). Later in the project I will be building machine learning models to predict whether or not a citation will be paid (***y*** = paid) based on the other variables in the dataset (***X*** = [fine, desc, instate, make, quad, yr, mo, day, and hr]). With this in mind, I sought to determine if there existed dependencies between the predictive and predicted variables I will be using in said models.

To test for independence between paid and each of the predictive variables, I performed chi-square tests of independence. All nine predictive variables showed dependency with paid (Table 3).

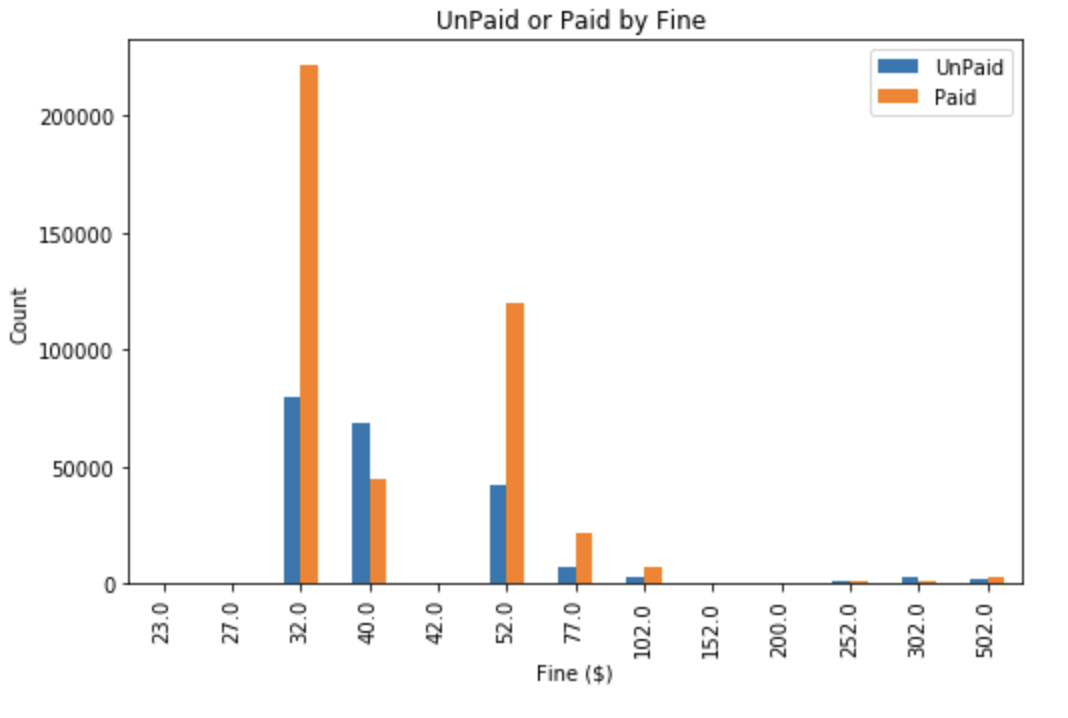


**Fig. 10 Countplot of paid and unpaid citations.**

**Table 3. Chi-Square Test Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | DOF | Critical X2 | Computed X2 | Dependency |
| fine | 12 | 26.2 | 52988 | yes |
| desc | 25 | 44.3 | 70747 | yes |
| instate | 1 | 6.6 | 2259 | yes |
| make | 316 | 377.4 | 11441 | yes |
| quad | 3 | 11.3 | 12425 | yes |
| yr | 2 | 9.2 | 67422 | yes |
| mo | 11 | 24.7 | 37408 | yes |
| day | 30 | 50.9 | 1873 | yes |
| hr | 23 | 41.6 | 4780 | yes |

Fine appears to be a good predictive variable (Fig. 11). Most citations with a fine of either $40 or $302 were unpaid, while most citations of any other fine amount were paid down.

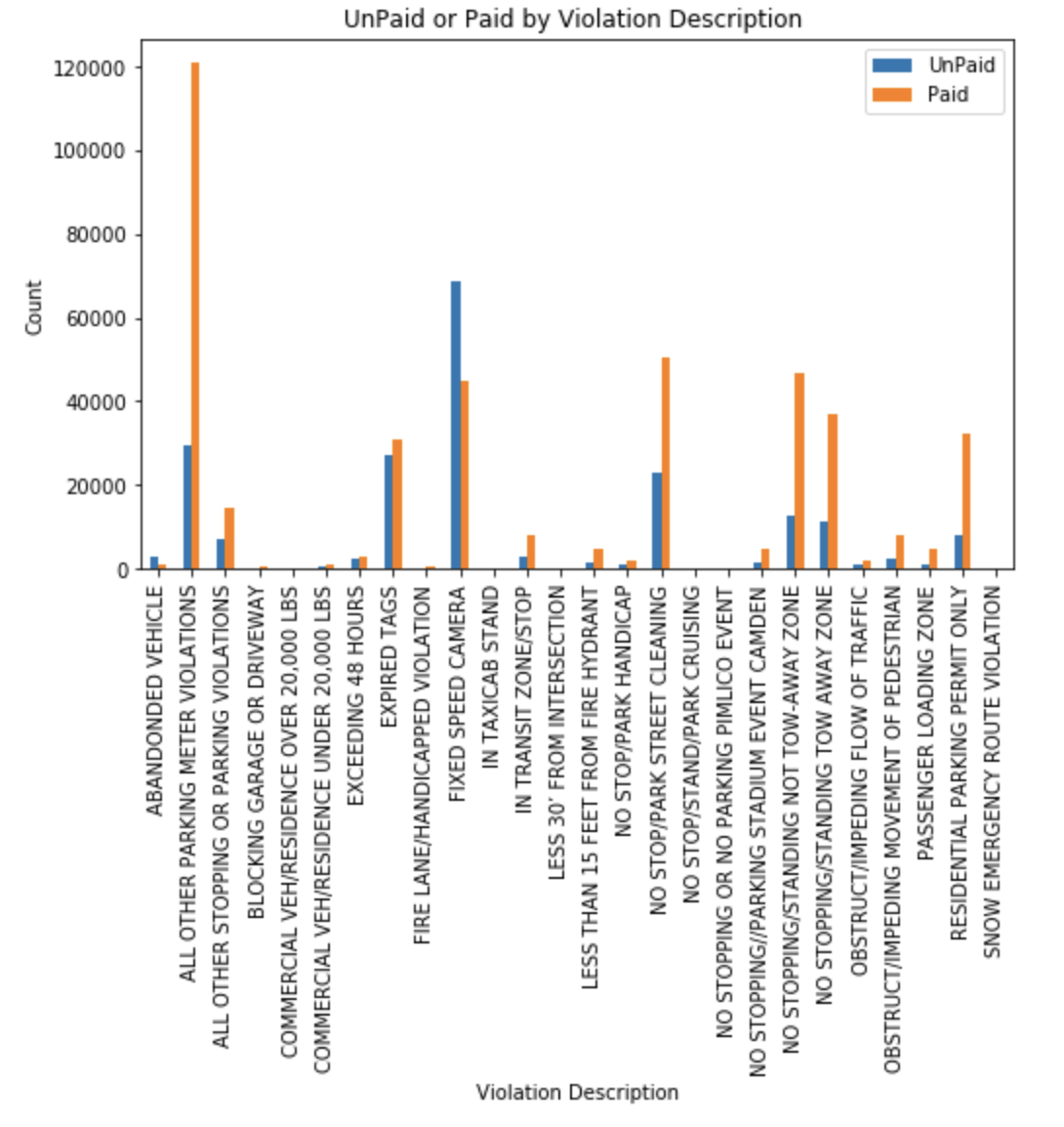


**Fig. 11. Countplot by fine amount.**

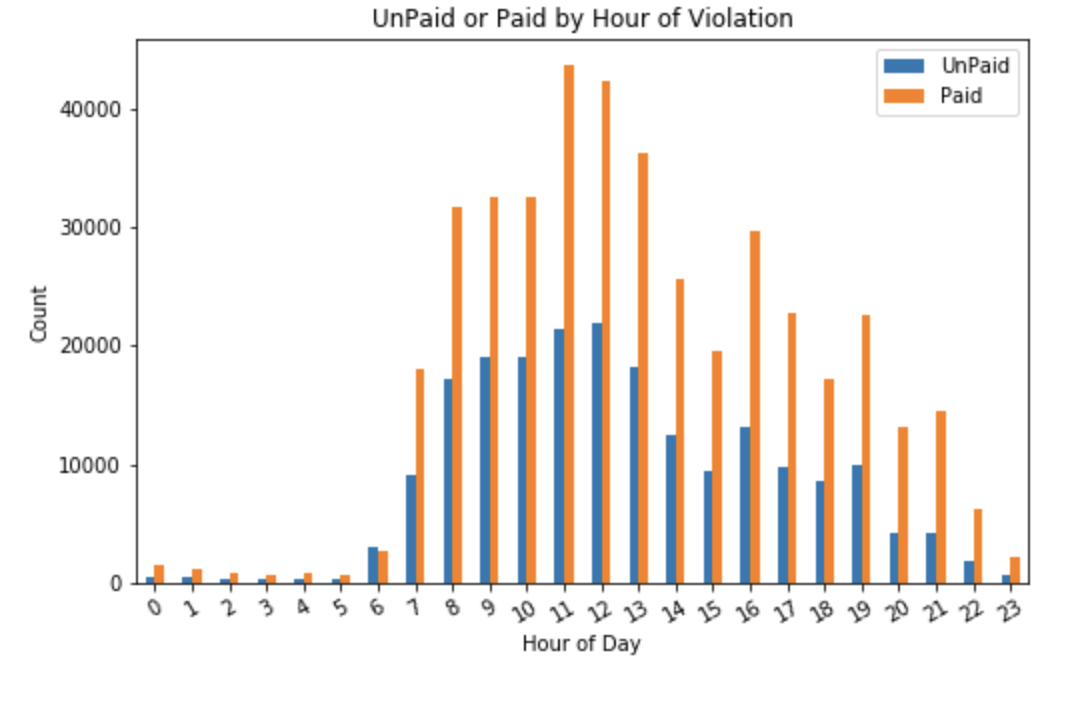
Desc also appears to be a good predictive variable (Fig. 12). Most citations with the description “Fixed Speed Camera” or “Abandoned Vehicle” were unpaid while most citations with any other description were paid down.

Instate and quad (not shown) appear to be less powerful predictors in that most citations in each class of these variables were paid.

Hour appears to be the only good predictor of a temporal nature (Fig. 13). The majority of citations issued in the 6 o’clock hour were unpaid while most citations issued at any other time of day were paid down.



**Fig. 12 Countplot by violation description.**



**Fig. 13 Countplot by hour of day.**

1. **Next Steps**

The next step for this project is to perform feature selection. After that, I will build two machine learning models: a logistic regression model and a support vector classification model. Once the models are operational, the next step will be to evaluate model performance and create visualizations. Lastly, I will make recommendations to the client and offer ideas on how to extend the study.