New York City Parking Violations

Fiscal Year 2022

Sebastion Ojeda, Dylan Power, Derek Watson, Henry Wo

Abstract

This paper seeks to examine parking violations in New York City's five boroughs for the fiscal year 2022, and explore the various factors which may influence or contribute to those violations. Using data provided through the Open Data project of New York City's Department of Finance, the authors seek to answer such questions as *When during the year are most tickets likely to be issued?* and *Is there a correlation between out-of-state license plates and violations on certain streets?* The answers to these questions could hold significance not only for the wary automobilist seeking to avoid hefty fines, but also for the policymakers who must make sound decisions on public safety transportation strategies. For example, by far the most likely street for out-of-state license plate parking violations is West 14th Street, so it could be reasonably concluded that this street is a hot spot for tourists in the city. It would be useful, then, for travel guides to mention that tourists should be wary of parking along this street while visiting the city. In this way, the questions analyzed in this paper could be used to improve the overall state of transportation and traffic in the city.

Introduction

The authors have procured a dataset composed of 15.4 million parking violations in New York City over the course of fiscal year 2022. The data is furnished by New York City's Department of Finance, and includes pertinent information on the type of infraction, location of issuance, and type of vehicle. Originally meant to guide public policy around the issuance of parking tickets, the dataset also allows the inquiring data analyst to extract useful information pertaining to where and why most parking violations are issued.

The authors have proposed a set of questions which are relevant to both the issuers of parking violations as well as the unsuspecting civilian likely to receive those violations. First and foremost, *Which streets or locations have the most number of parking tickets?* In the case that parking in those streets and locations is unnecessary, this question could help guide the operators of motor vehicles away from potential financial hazards. In the other case, this question could advise law enforcement where to deploy the most resources. Next, *What are the most common violation codes issued? Is there a correlation for certain types of vehicles?* If certain types of vehicles, say taxis, are overrepresented in the pool of vehicles receiving infractions, it may be prudent for the city to issue stricter training for the operators of those vehicles. Third, *When during the year are most tickets likely to be issued?* Like the first question, this would be useful for either advising civilians when not to park, or advising law enforcement when to patrol. Fourth, *Which make of passenger vehicle is most likely to be ticketed?* This question could be used to expose potential bias in ticketing, or perhaps to expose the buying habits of irresponsible drivers. Fifth, *Which counties have the highest number of parking violations?* Correcting for population, this question could also be used to expose either

bias when issuing tickets, or irresponsible parking habits on the county level. Finally, *What are the most common parking violations for each plate type?* This question could illuminate whether, for example, passenger vehicles are more likely to receive certain violation codes whereas taxis are more likely to receive others.

Related Work

Prior work done on similar datasets to this one is fairly robust. Using NYC ticketing data from 2013-2014, researcher Ben Wellington set out to find particular problem spots in the city: https://iquantny.tumblr.com/post/83770853308/update-single-fire-hydrant-nets-nyc-33000-a He found that a particular fire hydrant netted the city nearly $33,000 USD a year in parking tickets. Not only that, he hypothesizes that the reasons for the comparatively steep earnings from this spot are due to misleading markings on the pavement near a fire hydrant.

Framed as a python framework tutorial, data analyst Nick Cox uses NYC parking data from 2021 to discover various trends related to license plate type and vehicle make: https://towardsdatascience.com/learn-python-data-analytics-by-example-ny-parking-violations-e1ce1847fa2 His work points out, for example, that violation code 36 (speed limit in school zones) is most common among plates registered to other states, and that Hondas are most likely to receive tickets on Broadway Street.

As part of a data visualization project, analyst Steven Ginzberg breaks down parking violations by various metrics, including burrough, type of violation, and top dates of violation: https://nycdatascience.com/blog/student-works/data-visualizing-new-york-citys-parking-violation/ One notable finding is that by far the most common ticket issued in Manhattan is the "General no-standing zone" violation, which allows for loading and unloading, but only if the driver stays in the car.

Data Set

As mentioned previously, our dataset consists of over 15.4 million parking violations issued by New York City beginning on January 1, 2022 and ending on December 31, 2022. The data was sourced from New York City's Department of Finance, which has made the data accessible through its Open Data program, which is an online portal that makes public data generated by various New York agencies and organizations available for public use. The data file, Parking_Violations_Issued_-_Fiscal_Year_2022.csv, can be found at the following url: https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2022/7mxj-7a6y. This file contains 15,435,607 rows and 43 columns; each row represents an individual parking violation issued, and each column represents an attribute of the parking violation. The attributes provide detailed information on the type of vehicle, (such as the body type, color, make, and model), the type of parking violation, the issue date, the plate ID, information about where the ticket was issued such as the street number, zip code, and precinct, the summons number, and the date and time of the parking violation. All attributes are of three data types: Plain Text, Number, or Date & Time. Because most of the attributes are descriptions like vehicle makes, plate types, and violation codes, over half of the attributes are in the Plain Text data format. The rest are of the Number data format, and only one attribute, the Issue Date, uses a Date & Time data type.

The attributes that use the Plain Text data type are all nominal data, as they are simply identifiers that are used to describe data, and there is no indication of the direction or amount of difference between two data points in this category. For example, the Vehicle Body Type, Vehicle Make, or Issuing Agency attributes are all data points that can only be categorized and labeled; there is no order or rank between them.

The one attribute that uses a Date & Time data type is interval data, as the time information provided can be categorized, ranked, and have equal intervals between data points (such as seconds, minutes, hours). The attributes that use the Number data format are a mix of nominal, ordinal, interval, and ratio data. For example, the Violation Code attribute is an ordinal data type, as the codes can be categorized and ranked in numerical order, whereas the Summons Number attribute is an example of ratio data, as each summons violation can be categorized and ranked with equal intervals and has a true zero point.

Main Techniques Applied

Data Cleaning and Preprocessing

In order to begin our analysis, we started the project with a lengthy period of data cleaning and data preprocessing. We have created two data cleaning and preprocessing files, CleanData.py and CleanData_1.py which we applied to our large dataset to improve the quality of the data and the efficiency and ease of the mining process.

CleanData.py is a large python script with two functions, one to clean the 'Street Name' column, and another to clean the 'Violation County' column. The 'Violation County' column was in much better state than the column with street names, and so was much simpler to clean. New York City has five counties, or boroughs, and the counties in the data were represented by only a few different strings, typically abbreviations of the county or borough names. Thus, cleaning the county names involved creating a python dictionary in order to map the several different shorthands to one name. For example, the values 'NY' and 'MN' could be mapped to 'Manhattan', as both are abbreviations for the county name and borough name, respectively. Additionally, the borough name was chosen over the county name for the sake of familiarity.

The 'Street Name' column had much more variation than the county column, and so was significantly more difficult to clean. Each individual street could have multiple aliases, and each alias could be written with a different kind of shorthand, or even have typos. Most likely, these idiosyncrasies would be easy to parse for the police officer or county worker tasked with processing the violations, but for a data miner it could represent a huge problem. The first step was to segment the street names into different sections, which were direction, name, number, and type. The directions include the compass points, such as 'North' and 'South', the names are the unique identifiers for certain streets, such as 'Broadway' and the type can be thought of as the suffix for the street, such as 'Rd' and 'Way'. Once the sections were laid out, the next task was to parse each name to find out which sections it contained, and what the section actually was. This proved to be too difficult for the strategy used for the county column, since the dictionary would quickly grow too cumbersome to handle all the variation in the street names. Instead, regex was employed to check for some simple patterns that were common with the different sections. For example, the letter 'n', lowercase or uppercase, with a space before and after, would indicate that the street has 'North' in the name. Similarly, any sequence of letters longer than 2 or 3 characters, and which did not fit into the other categories, would have to be the name of the street. In this manner the many different street names were homogenized into one value.

Data preprocessing file CleanData_1.py uses data reduction techniques to obtain a reduced representation of our original dataset that is much smaller, efficient, easier, and faster to work with. Data reduction techniques helped reduce our dataset row volume by about 19%, from about 15.4 million rows to about 12.5 million rows, and also cut dataframe loading

times in Pandas by about half. The reduced version of our dataset still closely maintains the integrity of the original dataset, so data mining on our new data set gives the advantages of working more quickly and efficiently while also producing the same or very similar analysis and evaluation results.

As an example, CleanData_1.py keeps 17 columns of the original 43 columns provided in the original dataset. We decided to remove columns that had a lot of non-useful data such as Meter Number, Summons Number, Feet From Curb, Days Parking In Effect, and Street Codes, as we felt that these columns would provide little value in our final analysis and proposed questions and work. We also dropped columns that had a majority of garbage or null values, such as Hydrant Violation, Double Parking Violation, and No Standing or Stopping Violations. In addition to dropping these columns, the file also filters out invalid rows in the reduced dataset that are missing values for important columns that are necessary for our proposed analysis and evaluations, such as dropping rows that have invalid Registration State, Plate Type, Violation Code, Vehicle Make, Violation Time, and Vehicle Year.

We performed a final cleaning step using Tableau Prep Builder, which has a GUI that makes it easy to handle some of the simpler cleaning and preprocessing tasks. We are left with a reduced and fairly clean dataset that we can work with in the final stages of the project to help us answer our questions.

Some of our questions required context outside of our original dataset as well as preprocessing beyond our initial cleaning. The specific questions where this was necessary were: *Which makes are most likely to be ticketed?* and *Which counties have the highest numbers of violations?* In both of these cases a simple query cannot answer the question, because the most populous areas and most popular makes will have the most violations. In the notebook EDA.ipynb in our repository, you can see the process of bringing in supporting data, binning categories, filtering values and deriving new attributes in order to get to the heart of these questions. More detail on the process is given in the parts of this paper covering those specific questions.

Bringing in Supporting Data

In addition to cleaning our original dataset, we also needed to locate and integrate a number of supporting datasets to answer our questions. In total there were six, each outlined below:

The first two were included as attachments to our original data: one CSV describing the columns and one CSV defining the parking violation codes and providing fine amounts for each one. The former was useful in our initial exploratory data analysis in determining which columns would be useful to keep, the latter was useful in defining the Violation Code“foreign key” in our original dataset. If we denormalized the data we could see what type of violation each code represented in our queries.

The second two supporting datasets were GeoJSON files defining the borders of New York’s five boroughs and many police precincts. These borders are defined as a series of latitude-longitude points and can be interpreted by software like Tableau to create mapped visualizations of our data.

We also needed some data to provide context to the violation numbers. For example, what good is it knowing that the most violations occurred in a given place if we don’t know the population of that place relative to the rest. Without such context, it’s hard to draw conclusions about the data. For this reason, we had to find data about NYC population by borough and data about vehicle registrations in the state of NY.

Here are links to each dataset outlined above. The city of New York does an excellent job at publicizing data:

Violation Codes, Column Definitions:
https://rb.gy/dwicm

GeoJSON Boroughs, Precincts:
https://rb.gy/0rje2
https://rb.gy/4uufq

NY Vehicle Registration by Make:
https://rb.gy/8kull

Borough populations:
https://rb.gy/3zk56

Evaluation Methods

Because we have not proposed prediction or inference questions as part of this project, we will not be building predictive or classification models. In fact, the most obvious prediction questions we could ask (*Under what circumstances are you most likely to receive a ticket? Are some areas enforced more or less than others?)* are impossible to answer with our dataset, because every single record dataset represents a vehicle that *did* receive a ticket. To answer these questions we would need another large dataset of vehicles that parked without receiving a ticket, and no such dataset exists. In any case, if someone wants to determine if they're likely to get a ticket, it's probably best to just reference the rules in that spot!

Our questions instead are looking for patterns and insights in the data that are unclear or unanswerable without deep analysis. These questions require the data to be cleaned, preprocessed and outside data to be linked to get an answer. The bulk of our work was done in the pre-analysis phase of the project, making the analysis itself rather straightforward.

Our answers and insights will be evaluated using the following methods: data visualization, statistical measures and derived attribute analysis. Data visualization will include charts, graphs and maps. Statistical measures will include summary statistics like mean, range, and correlation coefficients as well as hypothesis testing with T-tests. Finally derived attribute analysis will involve calculating derived attributes by aggregating information from data from multiple sources and using those calculated values to draw conclusions that bring new insights.

Analysis

In terms of analysis, much of the work involved using Tableau, Pandas, matplotlib, and finding related datasets to supplement our data.

In conjunction with a GEOJSON file of NYC, using Tableau allowed us to create 'heat maps' of common violation spots, furthering our understanding of how geography plays a role in parking violations. Additionally Tableau allows for the quick creation of various types of charts, such as stacked bar charts, to help our visual understanding of the data.

Similar to Tableau, python has optional libraries which allow for the manipulation of data and the creation of various plots. Using Pandas, we were able to read in the data objects from a .csv file and manipulate them into various groups. This allowed for us to easily calculate various statistics values, such as min, max, and count. Used in conjunction with Pandas, matplotlib can easily create visual representations of the data, such as scatter plots and bar charts. As an example, for the plots involving the busiest streets per borough, the data was read in from the csv using a Pandas method. Further Pandas methods were used to select only those violations occurring in a given borough. After that, the data was grouped by the street name and the count attribute was selected. The data was filtered to select only those rows where the counts were higher than a certain threshold, and the remaining data was plotted in a horizontal bar chart using matplotlib. The entire process only requires only a dozen or so lines of code, and was incredibly useful for visualizing data.

Key Results

As described above, one key finding from the data was a list of the busiest streets in each borough, and their respective violation count over the course of the year.

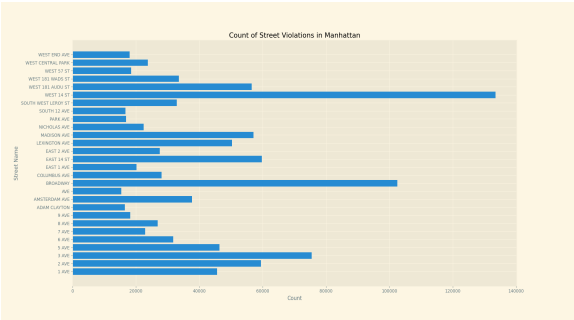


Fig. 1 - Violation count of busiest streets in Manhattan

From the bar graph, it is clear that West 14th Street has the highest violation count in Manhattan, though it is uncertain why this is the case. At a glance the street doesn't appear to contain more attractions or tourist spots than other streets, so it wouldn't appear that the street receives more traffic than others. After some digging, it appears that a portion of 14th street was turned into a bus route, but only during rush hour. It also appears that this change took effect in 2019, just three years before the collection of the data in this report.

https://pix11.com/news/l-train-shutdown-to-close-portion-of-14th-street-during-rush-hour-mta/

It is very likely, then, that 14th street has a proportionally large share of violations because the city restricts access to cars at certain times of day. A driver unfamiliar with the bus system in New York might not realize that their car will be parked illegally at certain times of the day, and thus West 14th Street incurs more violations.

Further analysis into parking violations based on street names seems to support this hypothesis:

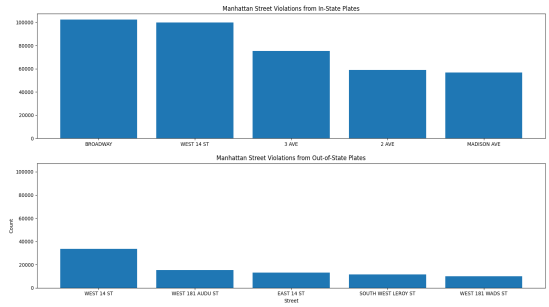


Fig. 2 - Violations on busy streets based on plate state

As expected, out-of-state license plates are responsible for far fewer violations than New York plates. However, it is worth noting that for New York plates, Broadway is the most common street for violations, and is almost tied with second place, West 14th Street. On the other hand, Broadway street doesn't even show up in the top five streets for out-of-state plates, for which West 14th Street is the most common. This seems to support the hypothesis that the comparatively large amount of violations on West 14th Street is due to the tricky nature of the bus timing, which locals might be more used to.

Additional information can be gleaned by breaking down the violations by not just street name, but also by the hour:

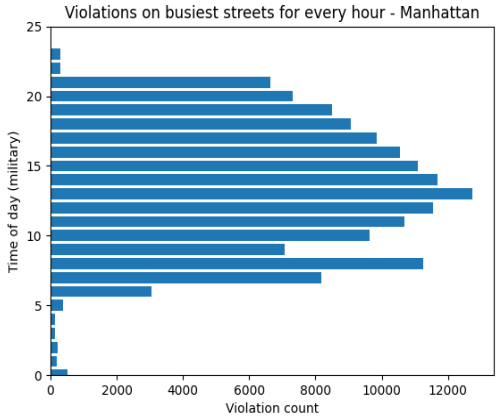


Fig. 3 - Violation count of busiest street by the hour

Figure 3 above shows the violation count for the busiest streets of every hour. Interestingly, violations are almost non-existent at night when

compared to daytime, suggesting either that police are less active at night, or that people park their cars illegally overnight less often. Somewhat unsurprisingly, there is a large peak at noon, during the lunchtime rush-hour, and there is another peak around 8am, corresponding to the morning rush-hour. It's hard to say why people would receive so many tickets during the time when they should be driving to work, but it's possible that the spike could be due to parking for breakfast.

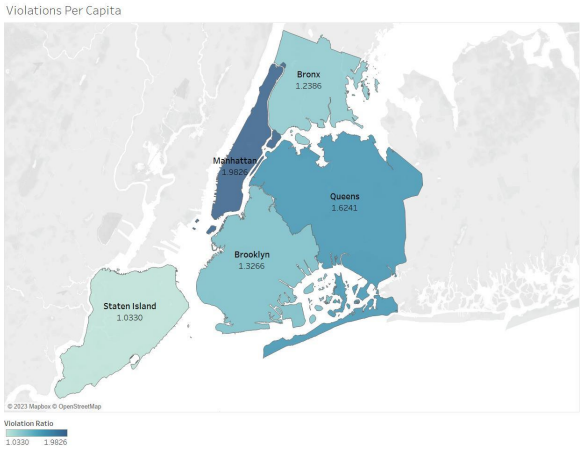


Fig. 4 - Per Capita Tickets by Borough

County	Pop.	Violations	Per Cap.
Manhattan	1.6M	3.16M	1.98
Queens	2.28M	3.7M	1.62
Brooklyn	2.59M	3.44M	1.33
Bronx	1.38M	1.71M	1.24
Staten Is.	0.49M	0.51M	1.03

Table 1

Figure 4 and Table 1 attempt to answer the question *"Which counties (boroughs) have the highest number of parking violations"*. They do so by dividing the population for each county by the number of violations for each county to calculate a per capita violation number. This number has then been used with the GeoJSON county boundaries to create a heatmap that visually shows the areas of the city that have the most parking violations per capita.

Manhattan, as one might expect, has the most by a wide margin. This is likely a combination of a large number of tourists and a high population density creating a premium on space, resulting in it being easier to find violations. In contrast, Staten Island has about half the violations per capita as Manhattan. Being suburban to semi-rural, Staten Island residents are more likely to have personal parking in their homes or access to street parking with less strict rules.

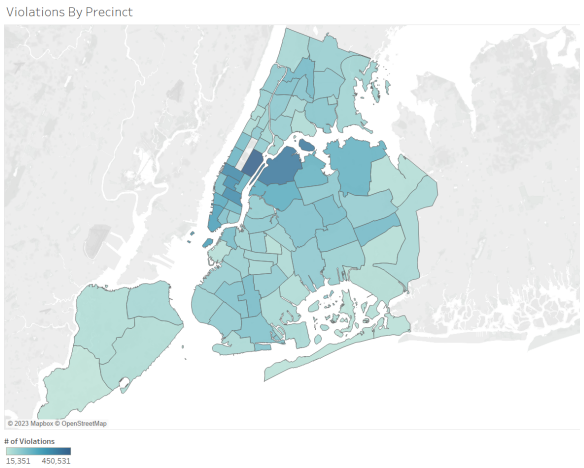


Fig. 5 - Tickets by Police Precinct

Figure 5 is a more granular version of Figure 4, but is not normalized by population. No data was available for population by police precinct, so the violation numbers here are raw values. This map looks at the city on the police precinct level, showing the number of tickets in each precinct ranging from about 15,000 in southwest Staten Island to 450,000 in the Upper East Side of Manhattan. One can see here that Manhattan south of Central Park and Queens near the Queensboro and Whitestone Bridges have some of the largest raw numbers of violations. These are likely the areas contributing most to Manhattan and Queens having the greatest number of per capita violations in the city.

Next, we will be answering the questions *"What are the most common violation codes issued?"* and *"Is there a correlation for certain types of vehicles?"* First, we can see from Fig. 6 that the most common violation codes are 36 (Exceeding the posted speed limit in or near a designated school zone) and 21 (Street Cleaning). The high number of violations for speeding is likely due to the fact that this is an automatic violation while speeding through a school zone as opposed to other violations that require issuing by a parking enforcement officer. The second most common violation indicates that vehicle owners likely forget to move their vehicles on the mornings that street cleaning is scheduled.

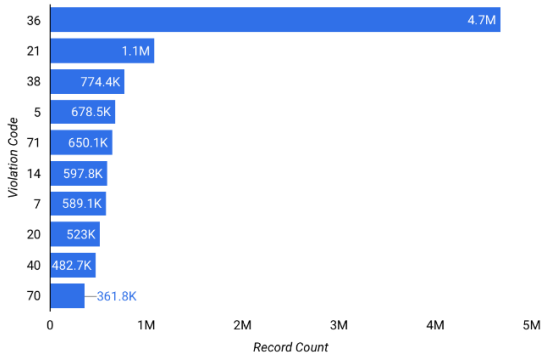


Fig. 6 - Tickets by Violation Code

As we look more closely at the kinds of vehicles and violation codes, we can clearly see that suburbans are the most commonly ticketed vehicle type.

	Vehicle Body Type / Record Count			
Violation Code	SUBN	4DSD	VAN	PICK
36	2.1M	1.2M	86.4K	125K
21	580.3K	380.1K	43.1K	23.9K
38	397.2K	221.7K	87.8K	23.8K
5	267.6K	178.9K	32.3K	11.5K
71	307.8K	244.2K	34K	21K
14	211K	135.6K	140.2K	15.3K
7	241.2K	165.4K	15.8K	12.9K
20	207.8K	124.5K	119.9K	19.6K
40	221.4K	162.7K	36.8K	12.7K
70	182.8K	128.1K	17.3K	10.8K

Fig. 7 - Tickets by Vehicle Body Type and Violation Code

Furthermore, we were able to delve more deeply into the violations data by examining the violations codes and their respective counts for both vehicles that are registered within the state of New York and for vehicles that are registered outside the state. As expected, violations for vehicles that are registered within New York exceed that of those for vehicles registered outside the state, with 9,164,309 violations and 1,281,502 violations, respectively. Out of the parking violations issued to vehicles registered in New York, the top three most common violation codes were:

Code 36 (Exceeding the posted speed limit in or near a designated school zone.): 3,727,641 violations

Code 21 (Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.): 1,080,825 violations

Code 38 (Failing to show a receipt or tag in the windshield.): 774,299 violations

Together, the top three most common violations make up nearly 61% of all violations issued to vehicles registered in New York. This is very similar to the overall violations data presented previously.

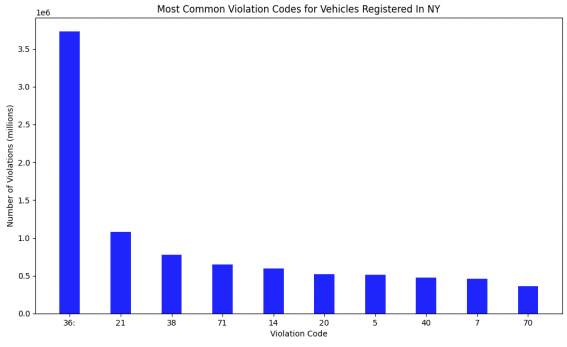


Fig. 8 - Most common violations for vehicles registered in New York

We did a similar analysis for violations issued to vehicles registered outside of New York. For these vehicles, the most top three most common violation codes were:

Code 36 (Exceeding the posted speed limit in or near a designated school zone.): 955,094 violations

Code 5 (Failure to make a right turn from a bus lane): 161,4125 violations

Code 7 (Vehicles photographed going through a red light at an intersection): 12,6441 violations

Together, the top three most common violations make up almost 97% of all violations issued to vehicles registered outside of New York.

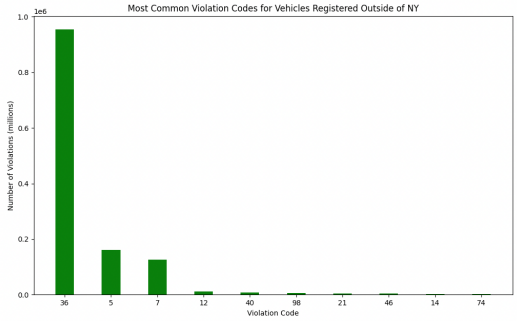


Fig 9 - Most common violations for vehicles registered outside of New York

Next, we will answer the question "*When during the year are most tickets likely to be issued?*" In order to do so, we want to look at the number of tickets issued on a monthly basis. As you can see in Fig. 10, there is a mostly consistent number of tickets issued throughout the year.

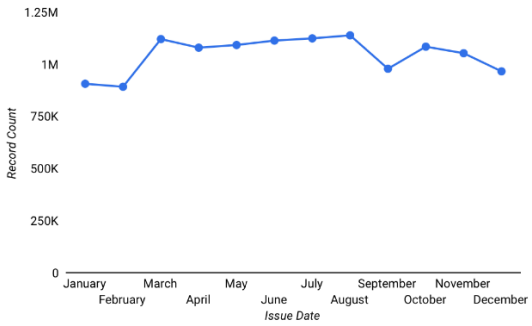


Fig. 10 - Tickets Issued by Month

When we take a look at the daily view, we can see again that there is a steady number of tickets issued throughout the year, with some predictable drops on days of the week that parking is not enforced (e.g. Sundays and Holidays).

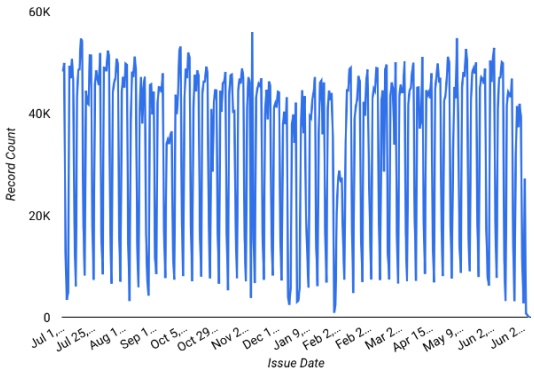


Fig. 11 - Tickets Issued by Date

The final question we addressed in our project was *Which make of passenger vehicle was most likely to be ticketed?*. This question required bringing in an outside data source to first determine which vehicle makes were most commonly registered in New York. The NY DMV had such data, and it was filtered to include only passenger vehicle body types from the most common makes. This data was then aggregated with our original dataset, which had to be further filtered to only include vehicles from NY with passenger plates and common makes.

The two datasets were joined on Vehicle Make, so that a common data frame was produced containing Percent of Total Registrations and Percent of Total Tickets for each make of car. A disproportionality measure was then calculated which was the ratio of tickets to registrations for a make, divided by the ratio of tickets to registrations for all other makes.

The following are the results, showing which makes are ticketed disproportionately more than others:

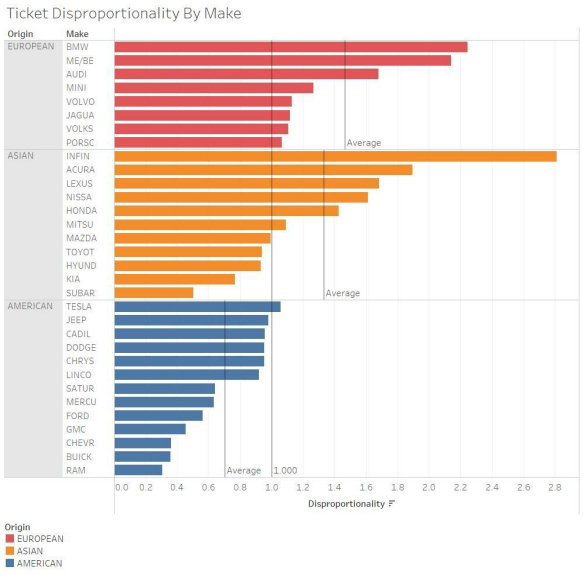


Fig. 12 - Rate of Ticketing by Make

One clear implication of this chart is that foreign cars are ticketed at a significantly higher rate than domestic cars are. Almost all American made cars are ticketed at a lower rate than their registration numbers would suggest, while many foreign cars are ticketed at a disproportionately higher rate than they are registered at.

To see if the difference could be explained by random variation in the data, a t-test was performed, checking if the difference in rate of ticketing between foreign and domestic made cars was significant. The T-statistic was 3.98, and the resulting P-value was 0.0004, meaning it is extremely unlikely that the difference is because of chance.

Further research would be needed to determine if this result is because of a bias in traffic enforcement, an incomplete dataset or an error in data processing.

Applications

A few interesting findings from our analysis of the data could be used for applications related to both touristry and city governance. For starters, it's clear that West 14th Street has the highest rate of parking violations, and is more commonly violated by out-of-state plates than New York plates. Based on our hypothesis that the high violation count is due to the intermittent bus schedule, the city might be encouraged to post more robust signage at that location. Any effort made to increase awareness about the restrictions of passenger vehicles along stretches of West 14th Street might go a long way towards preventing negligent parking.

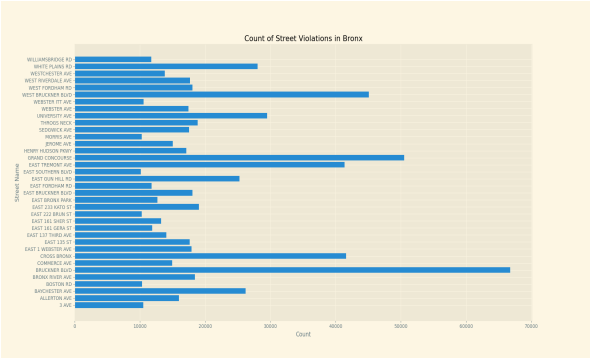
Furthermore, data about the busiest streets for violations every hour could be used to help apportion traffic police during rush-hour. It's possible that parking enforcers are overburdened along those streets, which might make it difficult to catch every violation. However, it's hard to say if this would be appropriate, since those streets might have the most violations *because* they have more police patrols already. In that case, survivorship bias might be skewing the data.

Another potential application of our gathered data could be aimed at reducing the violations for parking code 36 (exceeding the posted speed limit in or near a designated school zone), which is the most common violation for both vehicles registered within and out of New York. Since this violation occurs at a significantly higher rate than any other violation, focusing on this specific type of violation could have the highest potential impact in reducing overall parking violations. One possible solution to increase awareness of designated school zones may be through increased construction of high visibility

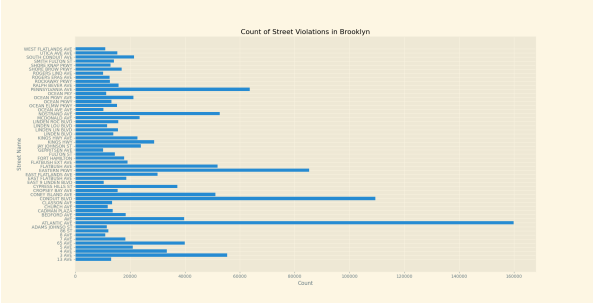
signage around these zones and increased public awareness through public outreach campaigns.

Visualization

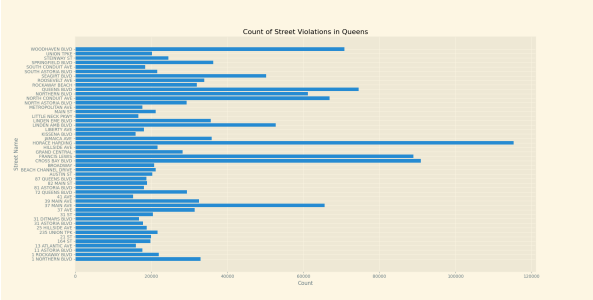
Supplemental Graphs Based on Street Names



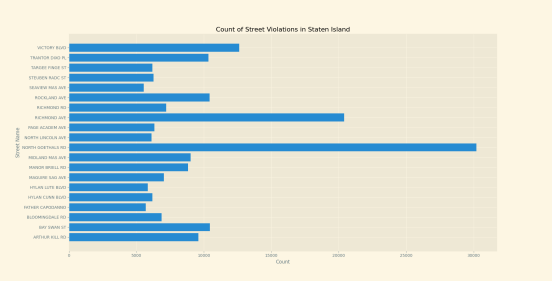
Violation count of busiest streets in the Bronx



Violation count of busiest streets in Brooklyn

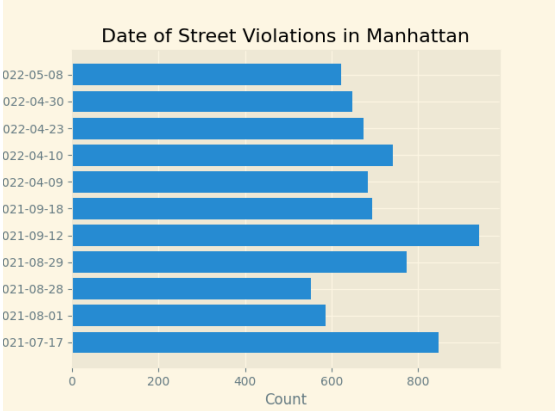


Violation count of busiest streets in Queens

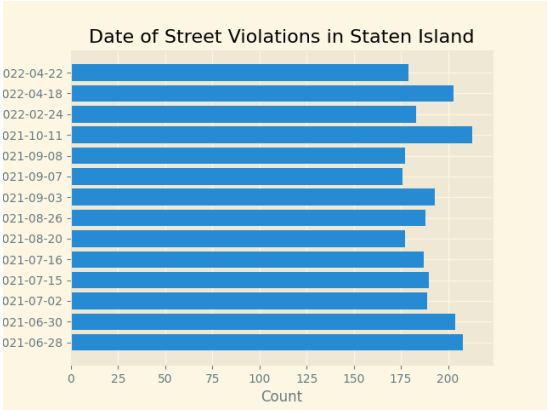


Violation count of busiest streets in Staten Island

Supplemental Graphs with dates of violation



Busiest dates for busiest street in Manhattan



Busiest dates for busiest street in Staten Island