

The Worst Streets in New York

Timur Boskailo[†]

College of Engineering and
Applied Science
University of Colorado
Boulder
Boulder, CO, USA
tibo7177@colorado.edu

Tyler Cranmer

College of Engineering and
Applied Science
University of Colorado
Boulder
Boulder, CO, USA
gecr2427@colorado.edu

Hitomi Imai

College of Engineering and
Applied Science
University of Colorado
Boulder
Boulder, CO, USA
hiim9551@colorado.edu

ABSTRACT

For a majority of Americans, cars are a ubiquitous part of modern life, with only about 9% of Americans reporting no access to a vehicle [8]. In New York City, a city famous for its public transportation, about 45% of households own a car. This number fluctuates across the five boroughs, with Manhattan having the lowest percentage of households with a car, 22%, and Staten Island having the highest percentage, 83% [7]. Along with car ownership comes another aspect of modern life, traffic citations, with about 10 million traffic citations given out yearly in New York City. These tickets created over half a billion dollars in revenue for New York City in 2014 alone [9]. Data mining of traffic citations can provide actionable insights for police departments, car insurance companies, as well as regional and state governments, and the federal government as well, which has agencies like the NHTSA in charge of related policies. While some studies of this variety use insurance data, and others use self-reporting questionnaires, ours will use a large data set of traffic parking citations in New York City taken from 2016-2017. Data mining of this data set can provide correlations between the many attributes provided by the data

set, such as citation type (illegal parking, expired meter, etc), vehicle brand, location of incident, and the issue date. We propose that finding patterns in the data, and creating classifiers, as well as data cleaning and organization, can provide insights that can be useful to numerous organizations.

CCS CONCEPTS

- Information Systems - Data Mining

ACM Reference format:

Timur Boskailo, Tyler Cranmer and Hitomi Imai. 2021. The Worst Streets in New York. In *CSPB 4502 – Data Mining. Boulder, CO, USA, 3 pages.*

1 Problem Statement/Motivation

In New York City, ~10M parking tickets are issued every year. The dataset on the parking tickets issued in NYC are collected by the NYC Department of Finance to aid in ticket resolution and guide policymakers.

Our team is motivated to analyze such data sets to examine whether there are any noteworthy patterns between violations and types of cars, seasonality and locations:

- Are there any patterns between types of violations (violation codes) and features of cars (such as vehicle types, makers and colors of cars)?
- Are there any patterns in which agencies issue tickets?
- Are there any seasonality in violation? Which months show more violations?
- Are there any trends in the locations (registration states and locations of the violation)?

2 Literature Survey

There are a large number of empirical studies related to traffic violations. Many of these studies attempt to seek correlation between traffic violations and traffic accidents. Many of the studies also use self-reporting to generate data from questionnaires, which leads to some methodological issues. In general, people tend to underreport their traffic violations [1]. There are other methodological questions at play when studying traffic violations. For instance, mileage is a major factor to be considered when looking for correlations. More time spent on the road, means an increase in exposure to both violations and accidents. However, car mileage can also not be directly linked to the owner, because owners change. Insurance data can provide changes in car mileage during a specific owner's use. In other

words, annual mileage [1]. Another factor to consider is the age of the car. Older cars tend to get more vehicle-related offenses, and older vehicles are generally owned by younger drivers.

In addition to the studies of correlation between violations and accidents, other studies seek to find correlation between age and/or gender, and traffic violations. In general the studies seem to show that male drivers have more accidents than female drivers, and are more likely to engage in unsafe driving behaviors, including drinking under the influence of alcohol and speeding. While women on the other hand are more likely to get into accidents as a result of judgment errors [5]. In addition, there seems to be a correlation between traffic accidents and age.

Middle-aged drivers tend to have the lowest accident rate, while young and old drivers have an increased rate of traffic accidents [5]. Moreover, there are other factors to consider, such as time of day, weather conditions, and road conditions. For example, certain types of accidents are more common during evening or at night, while others are more common in heavy traffic [6]. Other studies [3,4] look for correlations between driver behavior and traffic violations. Again seemingly finding the obvious that drivers prone to risky behavior are more likely to be involved in accidents. "Traffic violations and Insurance Data" is a study that uses insurance data from the largest insurance company in Sweden [1], instead of self-reporting questionnaires. Among other motivations, the study seeks to correlate car type with traffic violation, arguing that "status brands" such as BMW, Porsche, and Lamborghini, are more likely to be involved with speeding violations than "family brands" such as Volkswagen, Saab, and Volvo. However, despite the existence of these studies, we were unable to find any that deal with parking violations. The main focus of

the studies is instead on speeding and accidents. We on the other hand would like to investigate parking tickets.

3 Proposed Work

Our team will address major tasks in data preprocessing, including data cleaning to handle noisy data and data reduction to remove irrelevant attributes in the dataset. Furthermore, because data files are organized by fiscal year, we will integrate different datasets to combine data from different sources and transform data for discretization. The plans are to keep the data initially separate so we can do analysis on each year. After finding trends, we will be running a similar analysis to a data frame that has the entire four year set combined. Our thoughts in doing this is that during the exploration phase, we would like to reduce the size of the data frame we are working with. This is going to reduce the computational efforts of our local machines and reduce the initial time of running various analytic procedures.

From the initial investigation, there will be a significant amount of effort with cleaning the data. Out of the 53 attributes, roughly 18 of them comprise of 50% or more null values. After further investigation into those categories of attributes, we realized there are a few columns in which the presence of a null value represents a boolean value of false. i.e. The attribute of unregistered vehicles has a 0 or null value. Where the presence of a 0 represents a vehicle that is unregistered, and a null value represents a registered vehicle. So, with that in mind, the team will have to comb through each attribute to make sense of how the values are being represented. The team will be reducing the data by dropping certain attributes that have over 50% of true null values. We believe the attributes that have more than 50% missing data would skew the overall analysis and correlation between attributes. The data set will also have to go through some data transformation. We noticed that there are a

handful of attributes where the data type doesn't match the data that is being represented. The team will have to scrutinize each category in detail to correctly define its correct data type.

After completing our initial cleaning phase, we will start our analysis of the data sets. Our data set looks to be a better fit for the style of correlation analysis. We believe there will be strong correlations between the vehicle attributes, location, and the types of tickets they received. Which is similar to the findings of our previous literature survey. We hope to find some novel findings that haven't been discovered yet.

4 Data set

NYC Parking Tickets (Source: The NYC Department of Finance):

<https://www.kaggle.com/new-york-city/nyc-parking-tickets>

The data set is about details on NYC parking tickets, including the types of violating cars, dates and locations of the violations, and the violating codes. The data set was collected by the NYC Department of Finance for the periods from August 2013 to June 2017, and there are 51 attributes (for fiscal year 2017 file, 43 attributes only) which are organized in four files classified by fiscal years.

Major attributes of the data set we plan to consider are:

- Plate Type (nominal / string): is the vehicle a motorcycle, private car, commercial truck, or other
- Plate ID (nominal / string): unique identifier for violating vehicles
- Registration State (nominal / string): states of registration

- Issue Date (integer / string): dates of ticket issuance
- Violation Code (numeric / integer): violation codes for parking tickets - refer to this website <https://www1.nyc.gov/site/finance/vehicle/s/services-violation-codes.page>
- Vehicle Body Type (nominal / string): body types of violating vehicles. Reference is: <https://data.ny.gov/api/assets/83055271-29A6-4ED4-9374-E159F30DB5AE>
- Vehicle Make (nominal / string): makers of violating vehicles
- Issuing Agency (nominal / string): issuing agencies for parking tickets
- Violation Precinct (nominal /string): in which police precinct did the violation occur
- Violation County (nominal / string): county of violating violation
- Meter Number (integer): unique identifier for parking meter
- Street Name (nominal / string): street names for parking violations
- Vehicle Color (nominal / string): colors of violating vehicles
- Vehicle Year (interval / integer): years for violating vehicles

5 Evaluation Methods

We intend to use the following data mining techniques to evaluate our data:

- Correlation analysis, Interestingness, Sequential time series pattern.
- Hold-out by splitting our dataset into training and test sets.
- Linear regression to determine the relationships between certain variables.
- Clustering

6 Tools

- Python
- Pandas
- NumPy
- Matplotlib

7 Milestones

Our team completed addressing the process of cleaning up the data. Out of the fifty-two attributes, ten were dropped due to the fact they had over 50% null or missing values and 9 attributes were dropped because we found that the data recorded to be irrelevant or duplicates data found in other attributes. We had to perform data transformation

Data cleaning:

We examined the 44 attributes and removed the ones which were incomplete, duplicative or irrelevant for our analysis. The following is a list of removed attributes. See Fig 1, for a breakdown of starting attributes.

- Summons Number
- Time First Observed
- Violation In Front Of Or Opposite
- Intersecting Street
- Date First Observed
- Law Section
- Sub Division
- Violation Legal Code
- Days Parking In Effect
- From Hours In Effect
- Unregistered Vehicle?
- Violation Post Code
- No Standing Or Stopping Violation
- Hydrant Violation

- Double Parking Violation
- Latitude
- Longitude
- Community Board
- Community Council
- Census Tract
- BIN
- BBL
- NTA

The attributes that we will be using for our analysis are as follows. Plate ID, Registration State, Plate Type, Issue Date, Violation Code, Vehicle Body Type, Vehicle Make, Issuing Agency, Vehicle Expiration Date, Violation Precinct, Issuer Precinct, Issuer Code, Issuer Squad, Violation Time, Violation County, Street Name, Vehicle Color, Vehicle Year, and Meter Number.

In addition to removing attributes, the ‘Issue Date’ attribute was converted into a pandas datetime format to prime it for analysis. We also discovered that the ‘Vehicle Make’ attribute has many misspellings and many abbreviations that we are unable to decipher. For this reason, we took out any ‘Vehicle Make’ values that had less than 500 tickets. So far, we have performed the following analyses. Figure 2 shows the top 20 violations and the number of tickets per violation. The most cited violation is speeding in a school zone, followed by parking illegally on a street during street cleaning. Figure 3 shows the number of tickets handed out per time of day. Most tickets are handed out between 7AM and around 2 PM. Figure 4 shows the percentage of tickets handed out by street name. The streets

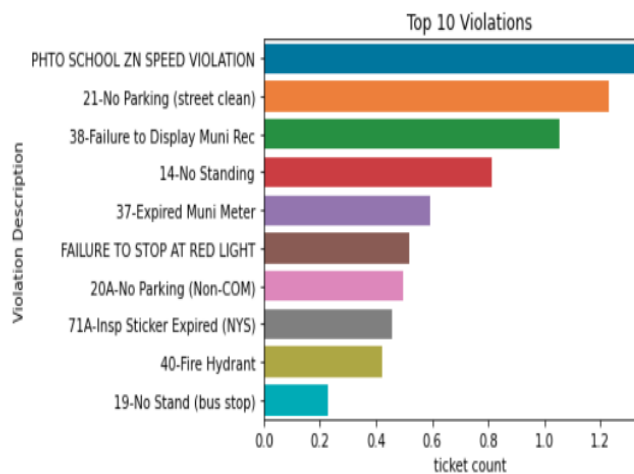
where the most tickets are handed out are Broadway, 3rd Ave, and Madison Street. Figure 5 shows the number of violations given per month. The month with the highest number of tickets is June, while July has the least amount of tickets given.

In addition to these figures we also created lists that show the type of violation per car brand, and the number of tickets each car brand has received, organized by the type of violation.

8 Results

Top 10 Violations

Our team found that from the 107 different violations that were given out, over 66% of tickets were from 10 different violations. Where speeding in a school zone was the top contender with 13% of total tickets and street cleaning consisted of 11% tickets.



Registration States of Violating Cars

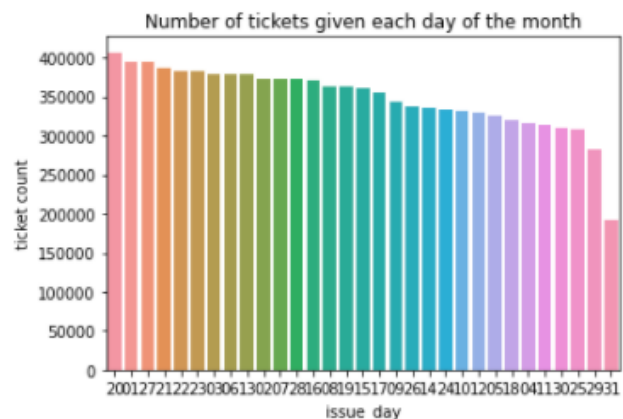
We examined if there is any tendency between violating codes and registration states. For the sake of analysis, we broke down by home areas, specifically ‘cars registered in NY States’ or ‘cars registered outside NY States’. Our team focused on top 3 violations codes (i.e, 21 - Street Cleaning - No parking where parking is not allowed by sign, street marking or traffic control device; 36 - Exceeding the posted speed limit in or near a designated school zone; 38 - Failing to show a receipt or tag in the windshield) and found that while approximately 78% of the

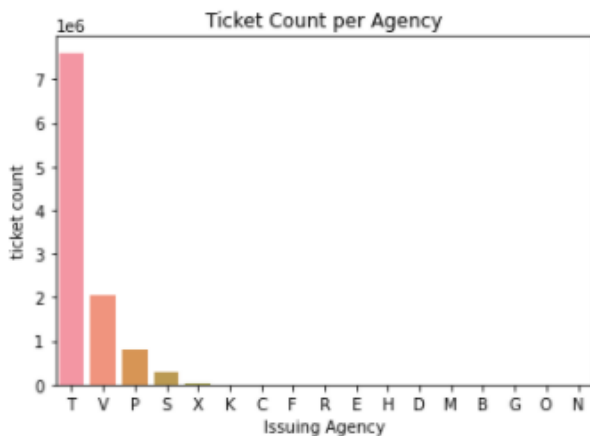
violating cars are registered in NY, there are certain tendency that some codes are more likely violated by local cars registered in NY States. For example, Code 36 is relatively more violated by cars registered in NY States, while Code 21 is relatively less likely to be violated by cars registered in NY States. The result is shown in the table below:

| | Total Violations | Violation Code 21 | 36 | 38 |
|--------|------------------|-------------------|-------|-------|
| NY | 0.784 | 0.748 | 0.862 | 0.793 |
| Not NY | 0.216 | 0.252 | 0.138 | 0.207 |

Ticket Numbers During the Month

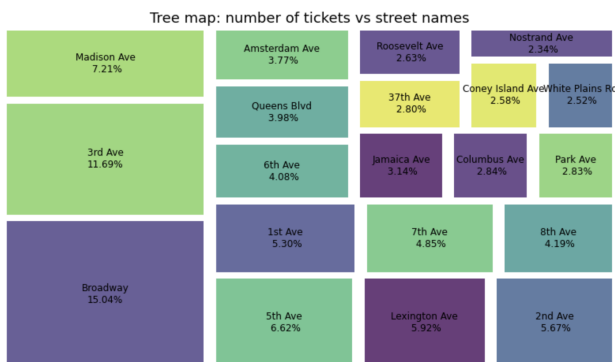
Our team found that police agency T had by far the most tickets issued throughout the course of the year. What's interesting to see is that the ticket issuances on a monthly basis slowly declined through time and had a swift drop off at the end of each month. One can think that the infamous theory of police ticket quotas might exist.





Ticket issuances throughout the day favored the standard 9 to 5 work day with the peak number of tickets given out between the hours of 9am and 1pm. As one would expect, the worst times to be parking around the city is during the middle of the day.

Dispersion of Tickets Among City Streets

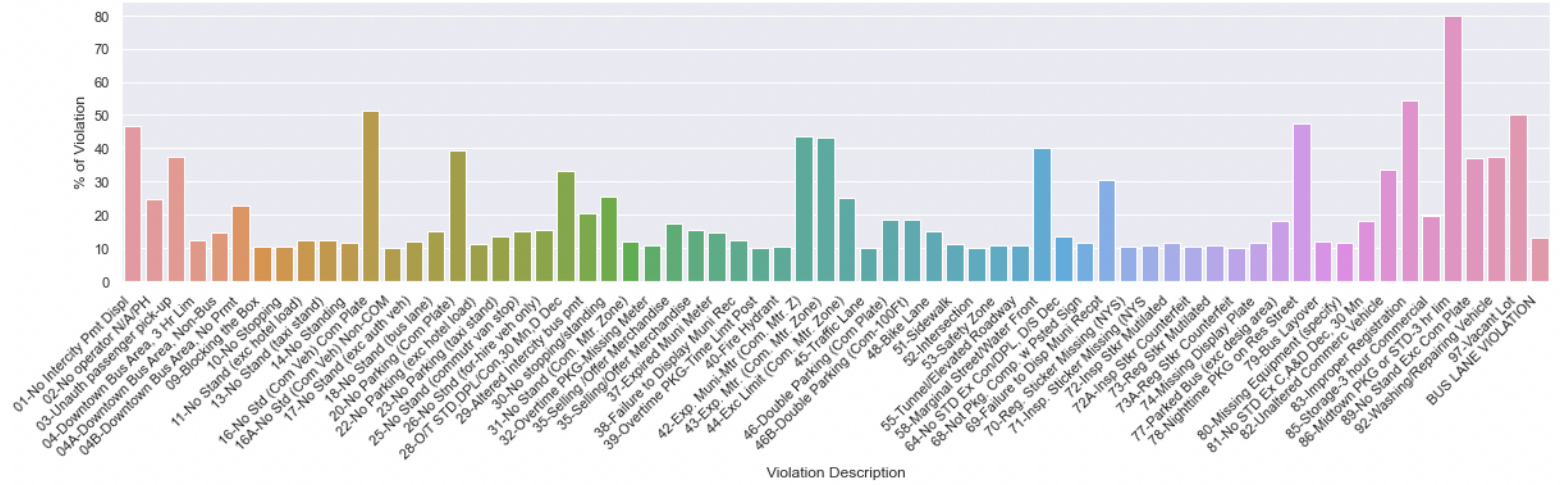


As clearly visible in the graphic above, Broadway, 3rd Avenue, and Madison avenue, had the highest percentage of tickets, with 15.04%, 11.69%, and 7.21% of all tickets handed out respectively. Combined, these 3 streets make up 33.94% of all tickets handed out.

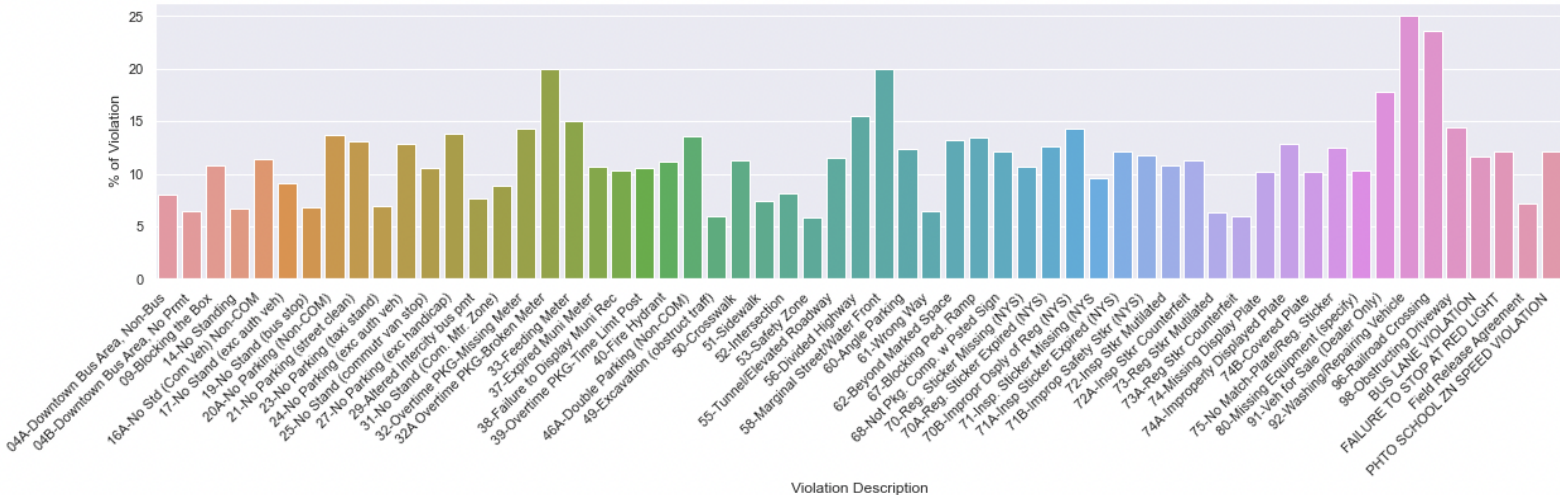
Violations for Top 3 Makes

The top three makes, in terms of total volume of tickets are Ford, Toyota, and Nissan, which make up about 33% of all tickets. The graphs below show each of their ticket breakdowns, in terms of the percentage of each violation that the respective make represents. Not all violations are shown, rather, there is a cutoff for each make, usually 5% or 10% of the violation. We see that there is a fairly consistent spread across the various violations, but each make has specific violations where they take up a larger fraction of the violation tickets. For example, Ford accounts for just over 50% of all tickets given for violation 16, No standing, for a commercial vehicle. Ford also makes up about 80% of all tickets given for violation 86, Midtown Parking or Standing, or 3 hour standing limit, and also about 50% of all tickets for violation 97, illegal parking in a vacant lot. Honda on the other hand has these ticket peaks at violation 92, illegal location for washing or repairing a vehicle, taking up about 25% of all tickets, violation 96, illegal railroad crossing, also about 25%, and violation 32A, overtime parking at a broken meter, 20%. Toyota sees these spikes for about 6 violations, with the two top ones being violation 64, no standing except for consulate or diplomat plates, and violation 52, standing or stopping in an intersection, with about 50% and 40% of all violations respectively. It's interesting that these top 3 makes, account for a lot of the tickets given out, but that they each have specific violations in which they exceed the others.

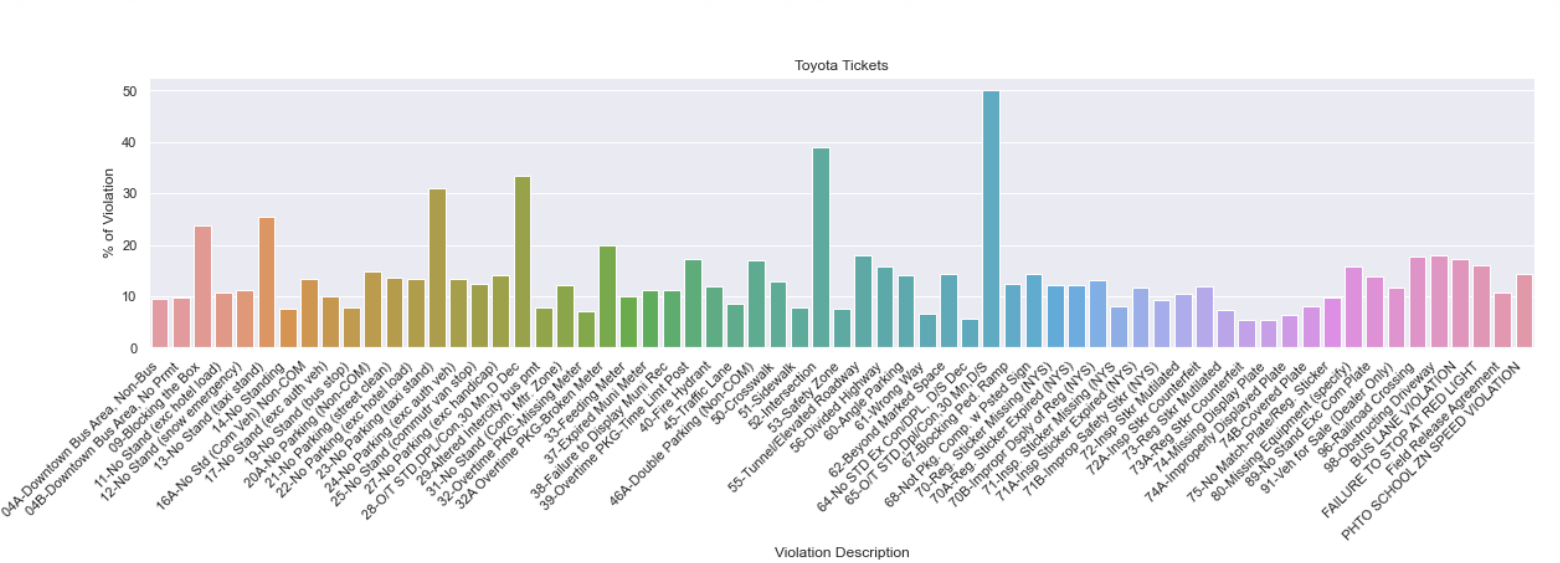
Ford Tickets



Honda Tickets



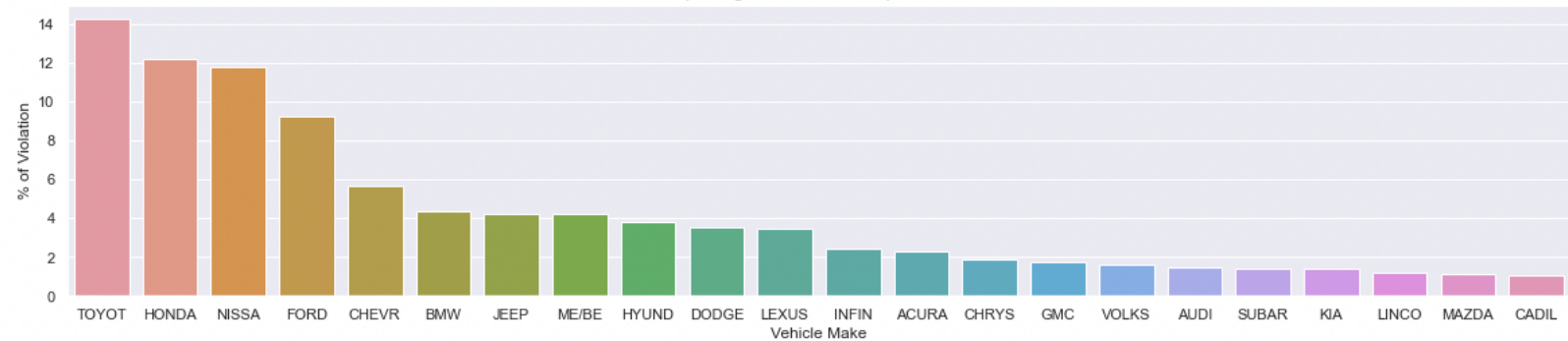
Toyota Tickets



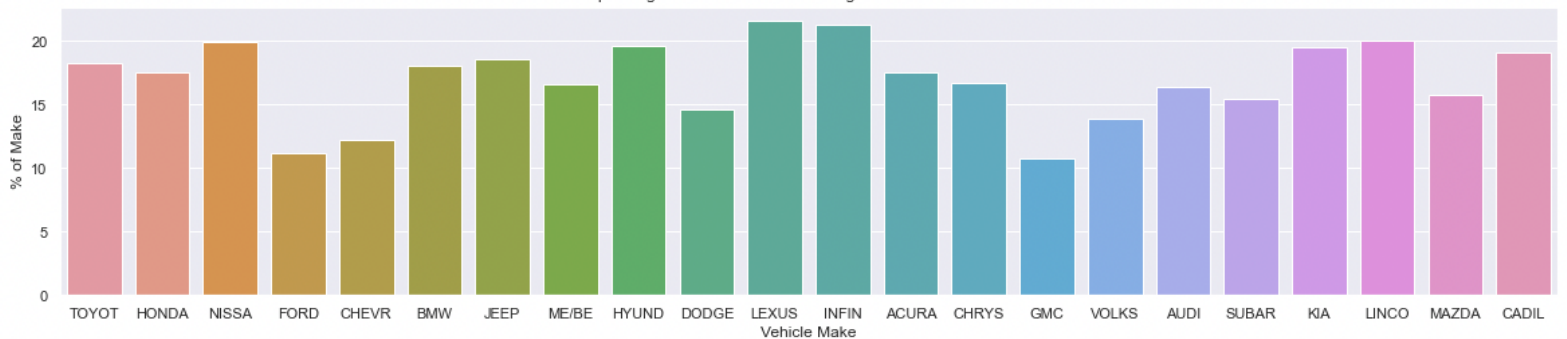
Speeding in School Zone

The violation with the greatest number of tickets is Speeding in a School Zone, taken with a speeding camera, with about 1,400,614 tickets. The following graphs show a breakdown of car makes for this violation. The first graph shows the top violators in terms of what percentage of the violation they represent. The second graph shows these top violators, and what percentage of their respective violations, speeding in a school zone accounts for. We see for example that Lexus accounts for just under 4% of all speeding in a school zone violations, and yet this violation accounts for just over 20% of all violations for Lexus.

Speeding in School Zone - Top Violators



Speeding in School Zone - Percentage of Make's Total Tickets



Results Conclusion

Overall, we were able to extract a lot of interesting information from the data. We saw which streets account for the most tickets, which vehicle makes are most prevalent, and which tickets they tend to get. We saw breakdowns of the top violation and which vehicle makes are most responsible, and we also saw the top three vehicle makes and the types of violations they tend to get. In addition we looked at the time of day, as well as the month, to see if we can spot any patterns in violation activity. However, given the immense nature of the data set, there is much more to be learned. Further treatment would require finding external data on the total number of registered vehicles for each make in the state of New York and using this information to normalize the data to account for the differences in vehicle make prevalence. In addition, given the large number of attributes in the data set, many more connections can be explored, such as which tickets are most prevalent on which streets and among which vehicle makes.

Applications

There are two general areas of application for investigating this data set; one is from the perspective of policing and public office, and the other is from the perspective of the general public or commercial enterprises that use vehicles. From the perspective of NYC government, we see that our data can be used to determine allocation of resources for police officers. It is clear which streets get the most tickets, and which violations are most prevalent. It is then up to public officials to investigate whether more resources need to be sent to high ticket areas to help police these infractions, or if more resources need to be sent to areas with a low number of tickets, if the number is low simply because of lack of policing. This applies to police departments as well.

Furthermore, we see evidence that some meters are consistently broken, and this could be investigated as well. However, it is also important to look at this data from a different perspective. On the one hand, government officials can step up police presence and issue more tickets, or they can take the perspective of the people who are getting the tickets. In general people are not willing criminals. breaking the law for fun. Rather they are people going about their daily work and tasks as best as they can. It is possible that certain tickets are constantly being handed out simply because the city design is inconvenient for its residents. If a lot of parking tickets are being given out in a certain area, it could be because there is simply not enough adequate parking. If commercial vehicles keep getting tickets for standing in places where they are not allowed to, then what does this say about the design of those areas? Are there ways to help the citizens of the city instead of punishing them? This general idea leads us to the second general perspective, which is that of the average citizen of New York City. It is clear which violations one should be careful of receiving in which areas. For example, we saw that a lot of tickets are given out for parking in a location scheduled for street cleaning. This is a good reminder to always look out for signs indicating future street cleaning whenever parking. The data could also be a way for citizens to organize and present the city with grievances. If people are getting ticketed for just trying to go about their lives, then maybe it is time for the city to look into restructuring parking and streets to better accommodate its people, instead of ticketing them in such excess.

REFERENCES

- [1] Arvidsson S. Traffic violations and insurance data : a note on the role of age, gender, annual mileage and vehicle brand [Internet]. Stockholm; 2011. (Working papers in transport economics). Available from: <http://urn.kb.se/resolve?urn=urn:nbn:se:vti:diva-637>
- [2] Josep Castellà, Jorge Pérez, Sensitivity to punishment and sensitivity to reward and traffic violations, Accident Analysis & Prevention, Volume 36, Issue 6, 2004, Pages 947-952, ISSN 0001-4575,
- [3] Walter Renner, Franz-Georg Anderle, Venturesomeness and extraversion as correlates of juvenile drivers' traffic violations, Accident Analysis & Prevention, Volume 32, Issue 5, 2000, Pages 673-678,
- [4] Sophia Vardaki, George Yannis, Investigating the self-reported behavior of drivers and their attitudes to traffic violations, Journal of Safety Research, Volume 46, 2013, Pages 1-11,ISSN 0022-4375,
- [5] Dana Yagil, Gender and age-related differences in attitudes toward traffic laws and traffic violations, Transportation Research Part F: Traffic Psychology and Behaviour, Volume 1, Issue 2, 1998, Pages 123-135, ISSN 1369-8478
- [6] Guangnan Zhang, Kelvin K.W. Yau, Guanghan Chen, Risk factors associated with traffic violations and accident severity in China, Accident Analysis & Prevention, Volume 59, 2013, Pages 18-2, ISSN 0001-4575
- [7] NYCEDC, 2021. New Yorkers and Their Cars | NYCEDC. [online] Edc.nyc. Available at: <<https://edc.nyc/article/new-yorkers-and-their-car#:~:text=According%20to%20recent%20census%20estimates%2C%20%5B1%5D%20almost%201.4,almost%203%20percent%20that%20own%20three%20or%20more%21%29.>> [Accessed 25 October 2021].
- [8] Peterson, B., 2021. Car Ownership Statistics (2021 Report). [online] ValuePenguin. Available at: <<https://www.valuepenguin.com/auto-insurance/car-ownership-statistics>> [Accessed 25 October 2021]
- [9] The number of parking tickets issued New York City police officers in the last week has plummeted 93% compared to 2014. (n.d.). How much money NYC makes from parking tickets. CNNMoney. Retrieved November 28, 2021, from <https://money.cnn.com/2015/01/06/news/economy/nypd-tickets/index.html>.