# 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. 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 citations in New York City taken from 2014-2017. Data mining of this data set can provide correlations between the many attributes provided by the data set, such as citation type ( illegal parking, speeding, 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 Application---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 milage 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.

## 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:

- Summons Number (nominal / integer): unique identifier for particular summons
- Plate ID (nominal / string): unique identifier for violating vehicles
- Registration State (nominal / string): states of registration
- Issue Date (interval / string): dates of ticket issuance
- Violation Code (numeric / integer): violation codes for parking tickets
- Vehicle Body Type (nominal / string): body types of violating vehicles
- Vehicle Make (nominal / string): makers of violating vehicles
- Issuing Agency (nominal / string): issuing agencies for parking tickets
- Vehicle Expiration Date (interval / integer): expiration date for violating vehicles
- Violation County (nominal / string): county of violating violation
- Violation In Front of or Opposite (nominal / string): the places where tickets were issued ("front of" or "opposite")
- 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 out dataset into train and test sets.
- Linear regression to determine the relationships between certain variables.
- Clustering

## 6 Tools

- Python
- Pandas
- NumPy
- Matplotlib

## ACKNOWLEDGMENTS

To Digby, my rock and the light of my life.

## 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-cars#:~: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]