# An Exploration of the Traffic Violations in NYC

Each year in NYC, there are over 2.5 million violations committed. Many of these violations are dangerous actions that can result in death. Finding out if there's any correlation between the violations could help determine what causes these incidents and if there's any way to prevent them, and make the city of New York more safe. Examples of possible correlations we're looking for is an increase in violations in certain regions, or violations committed by certain brands of cars, or certain types of cars. Finding correlations between traffic violations and one of those areas of possible correlation will be useful in finding out what changes need to be implemented in order to best reduce traffic violations.

## Data Sources

Database Link: https://www.kaggle.com/datasets/new-york-city/nyc-parking-tickets
The database contains data pertaining to all traffic violations that took place in New York City. We are specifically looking at the data from the year 2017, and taking a small sample of 3000 from the initial data.
Violations Key: https://www.nyc.gov/site/finance/vehicles/services-violation-codes.page
Vehicle Make Type Key:
https://data.ny.gov/api/assets/83055271-29A6-4ED4-9374-E159F30DB5AE

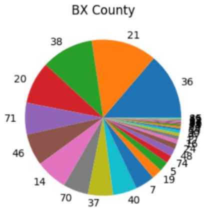
## Data Cleaning/Processing

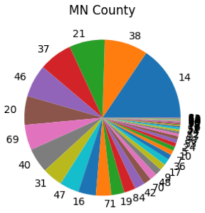
1. **Violation Time** and **Issue Date** split up the exact times of the violations. They can be combined in order to create a new datetime column. Before this step, the **Violation Time** needs to be cleaned so that all the times are in accurate format to be converted into datetime format, and remove all errors in inputting time, which can be done through string manipulation. Then we use pandas to convert the new combined column into python's datetime data type.
2. There are typos in the **Street Name** column in the csv, where sometimes the **Intersecting Street** name will sometimes appear in the name of the **Street Name** column. We can resolve this error by simply stripping all the letters after the "@" symbol.
3. **Violation Description** is inconsistent in describing the type of violation committed. We need to standardize the way that violations are described so that it will always list the
4. Make it so that borough names are listed in a standardized format. This data populates the **Violation County** column. It uses a dictionary of all the original values in the csv, and converts them into a standardized format that represents the borough names. For instance, Brooklyn, which is represented by both "King's County" and "Brooklyn" is now consistently represented as "BK".
5. Remove all features/columns that are not being used, and get rid of all null values in each of the relevant columns.

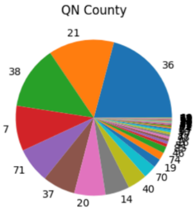
# Exploratory Data Analysis

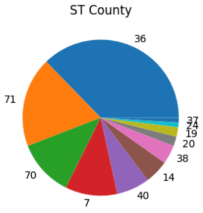
To understand the data and make any initial assumptions, we perform EDA steps to start forming the base for any model to come. After cleaning, we made multiple graphs from which numerous hypotheses can be made.

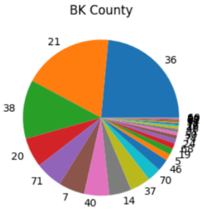
1. Bar Graph - Number of particular violations reported. We made a bar graph to represent all the different violations in the dataset. This is to provide a whole picture of numerous violations all over NYC across the boroughs. Moreover, this helps us identify the most common violations. From the graphs, we can see the Violation code 36, 21, and 38 are the most common occurring violations.
2. Histogram - Identify the violations based on each borough. Using histogram, we were able to make bins for each borough and add all the violations reported in the area. This helps us identify the area with the most violations. Based on the graphs, Manhattan was reported with the most number of violations, followed by Brooklyn and Queens. This could be either the traffic coming into the city or residents of those areas.
3. Pie Chart - Identify the most common violations in each borough. Using the dataset's wide range of columns, we were able to to find violations code and their description for each violation reported. With pie charts, we are able to map out the all the violations reported in the area to find the most commons ones. This would help us understand the violations and form any possible hypothesis as to why there are violations. We can see that the most common violation code is 36 and the common one across all borough except Manhattan. For Manhattan, we can see that the top 5 violations are space related. Code 14 being the standing in a no standing zone. Code 38 and 37 being Parking Metre violations, staying there for more time or not displaying the parking ticket. Code 21 - Parked during street cleaning and Code 46 - Standing or parking on the roadway side of a vehicle stopped.





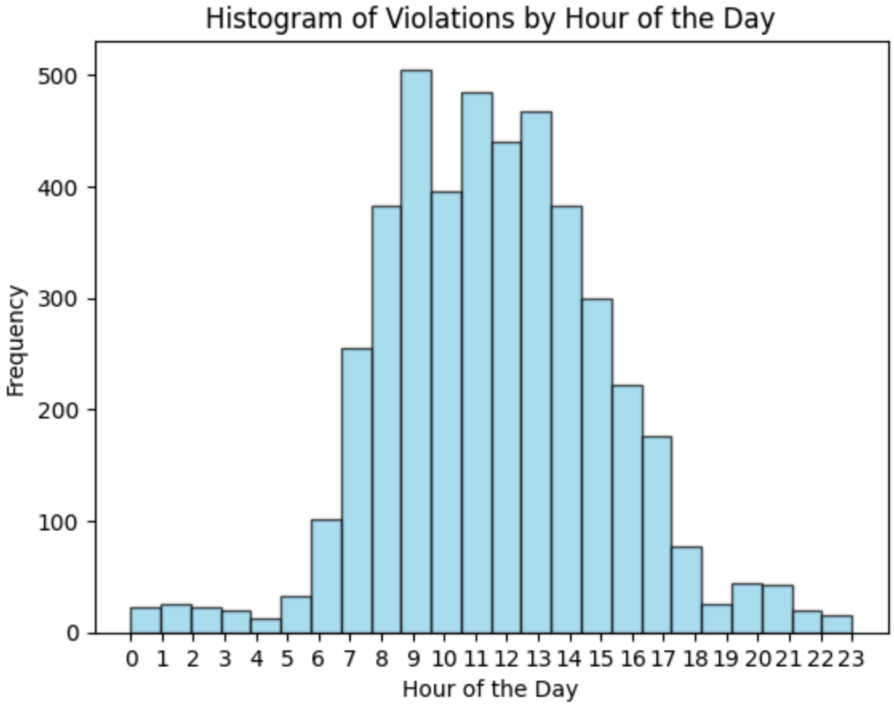






4. Pie Chart - Identify the model type for violations. Using the vehicle make type could be important to form any hypothesis in understanding along with the violations. Based on the area and violation type, we could come up with a hypothesis or reason as to why there would be a violation reported. Moreover, we could also see the size of the reported Vehicle Make, which would help us understand violation trends. There are also a few outlining models like bus, boat, dump truck and more. It would help in distinguishing residential and commercial vehicles.
5. Histogram - identity the violation frequency throughout the day by the hour. This histogram shows the data arranged based on the violation by the hour of the day. We can see a bell curve in the data distribution where the peek time for any violation ins right around noon. And there is a slight skew of data during the middle of the night and a dip early morning. This could be for the time during when people are asleep. The highest

peak indicates that there is a lot of traffic around all the boroughs where multiple violations are happening.


Histogram of Violations by Hour of the Day
Frequency
Hour of the Day