**Analysis of New York City’s 2017 Parking Ticket**

**using Hadoop, Apache Pig, and Power BI**

Javier Nava

Oscar Munoz

Jesse Fabian

Ky Hoang

Department of Information System, California State University

Los Angeles

**Abstract**: Through this project, we analyzed the New York City’s 2017 parking ticket violation dataset and the relationship between different factors and their possible contribution to receiving parking violations. Could the color or the model of one’s car influence New York City’s officers to give parking citation? Are there any indicators that may show why innocent people are receiving ticket violations? [2]Using the dataset of New York City’s 2017 parking violation from kaggle.com, we were able to compare different attributes that were linked to each violation. Since the size of our dataset is 2 Gigabytes, using Hadoop Distributed File System made it easier for storing and cleaning our data. Through the use of Hadoop, Apache Pig, and Power BI, we were able to store, clean, filter and analyze the data, which allowed us to create insightful findings. By cleaning and filtering our dataset through Pig, we were able to export a dataset with the necessary fields that would help us analyze and visualize different findings and information.

**1. Introduction**

For many of us who have received one or more parking violations, we often come to the conclusion that we unknowingly parked at the wrong spot at the wrong time. However, there are also times when parking violations are falsely given out with biased tickets given to one vehicle but not the other. [1]Upon research, were not surprised to see that NYC placed first among the top ten cities with the highest revenues from parking violations. Fortunately, the dataset of New York City’s parking tickets in the fiscal year of 2017 could help us find some insights on what different factors may influence one to receive a ticket violation. From the use of this dataset, we will be analyzing different factors listed within the ticket that may have contributed in receiving the violation, e.g. vehicle color, make, type, state registered, the location of parking, etc...

**2. General Tools & Methods**

In order to answer the burning questions in regard to traffic tickets in the city of New York, we required a Hadoop cluster that was able to read and filter large amounts of data. Since the dataset was around two gigabytes in size, the cluster had to be powerful enough to manage the data quickly and elegantly. The cluster we used for this project was composed of five nodes, which had a cumulative total of one hundred and fifty gigabytes of memory, and had a speed of around 2,195 MHz. Apart from the physical hardware we also used Amazon web services to host our dataset, Apache Pig to check and filter our data and Power BI to help visualize our findings.

The original traffic ticket data set found on Kaggle contained multiple years of data, and in order to help limit the scope of the project we chose to focus on the traffic tickets given out in 2017. This meant that now we needed to find a place where we could host the now reduced dataset in order to make it available for each member of the group to download it freely. Once the dataset was available to every member we each began to study the content of the dataset to see what information was provided. Upon seeing the content of the dataset we realized that the data consisted of twenty-three different columns. Most of the columns provided were not really necessary or important to what we were trying to answer so naturally we decided to divide the dataset into smaller more relevant portions. In order to begin the partitioning of the dataset we first had to define a schema for the data and its twenty-three different columns. Once that was done we checked to make sure that the data was properly partitioned by using the DESCRIBE Pig command.

With the dataset now properly defined we began to generate a whole new dataset that only contained the appropriate columns for our research purposes. Once that was done we stored the results of the filtering and divisions onto our Hadoop filesystem. Once we made sure that the results were saved, we then converted all the different result files into one single CSV file to be used for visualization.

**3. Findings**

There are many ways that we can present our findings. We decided to use Microsoft Power BI. We had to upload our CSV file to Microsoft Power BI. By using Power BI we were really able to visualize our findings, making it easier for the average person to look at our findings.

When it came to our findings we were very curious and excited about what we were going to find. We decided to limit our objectives to the following questions. What are the top 5 car brands that had the highest amount of violations? What car did colors attract the most tickets? What states had the highest number of cars registered that got the highest amount of tickets? What were the times that had the highest amount of tickets? What were the locations that had the most violations?

When it came to the first question we decided to use a bar graph to show our findings. Figure 1 displays the answer to our first objective. We found that the top 5 car brands that had the highest amount of tickets were Ford, Toyota, Honda, Nissan, and Chevrolet.

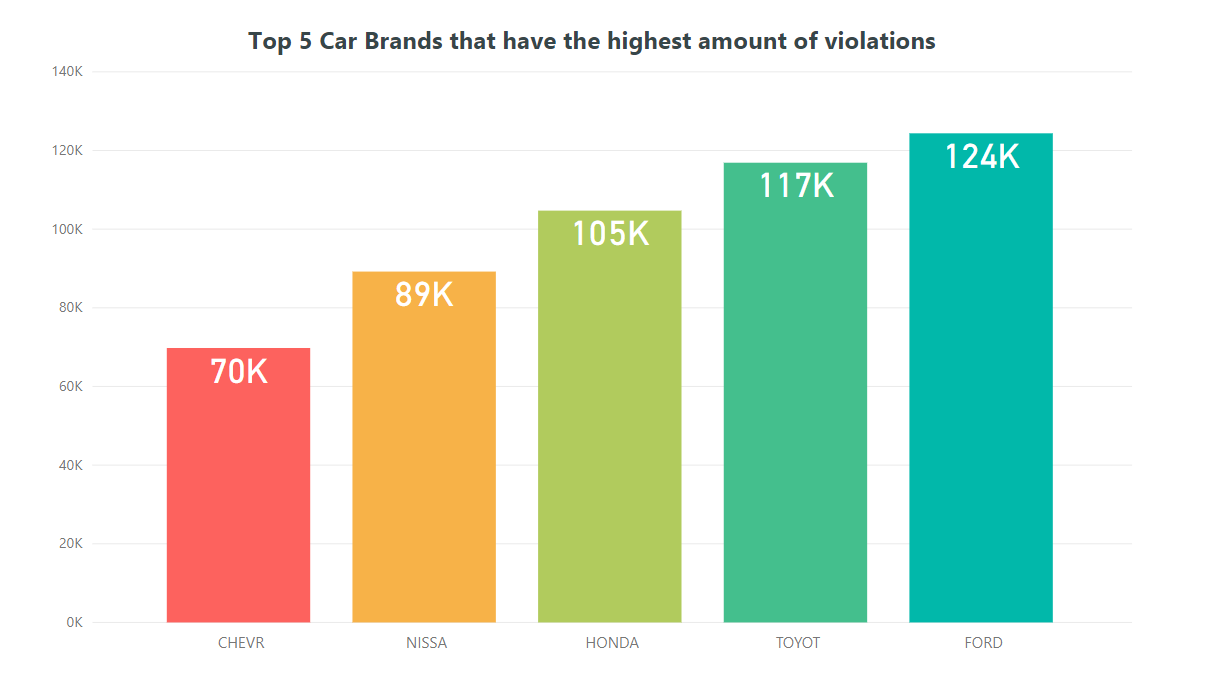
****

Figure 1 Car brands with the highest violations.

Moving forward to our second finding, we noticed that we ran into an issue. We noticed that the all of the agents that inputted the data had different codes for colors. This is bad because it throws off our data. However, we were still able to list the top 5 color codes. Figure 2 displays that the codes were GY (Grey), WH (White), BK (Black), WHITE, and BLACK. With this given data we can see that black, white, and grey cars got the highest amount of tickets.

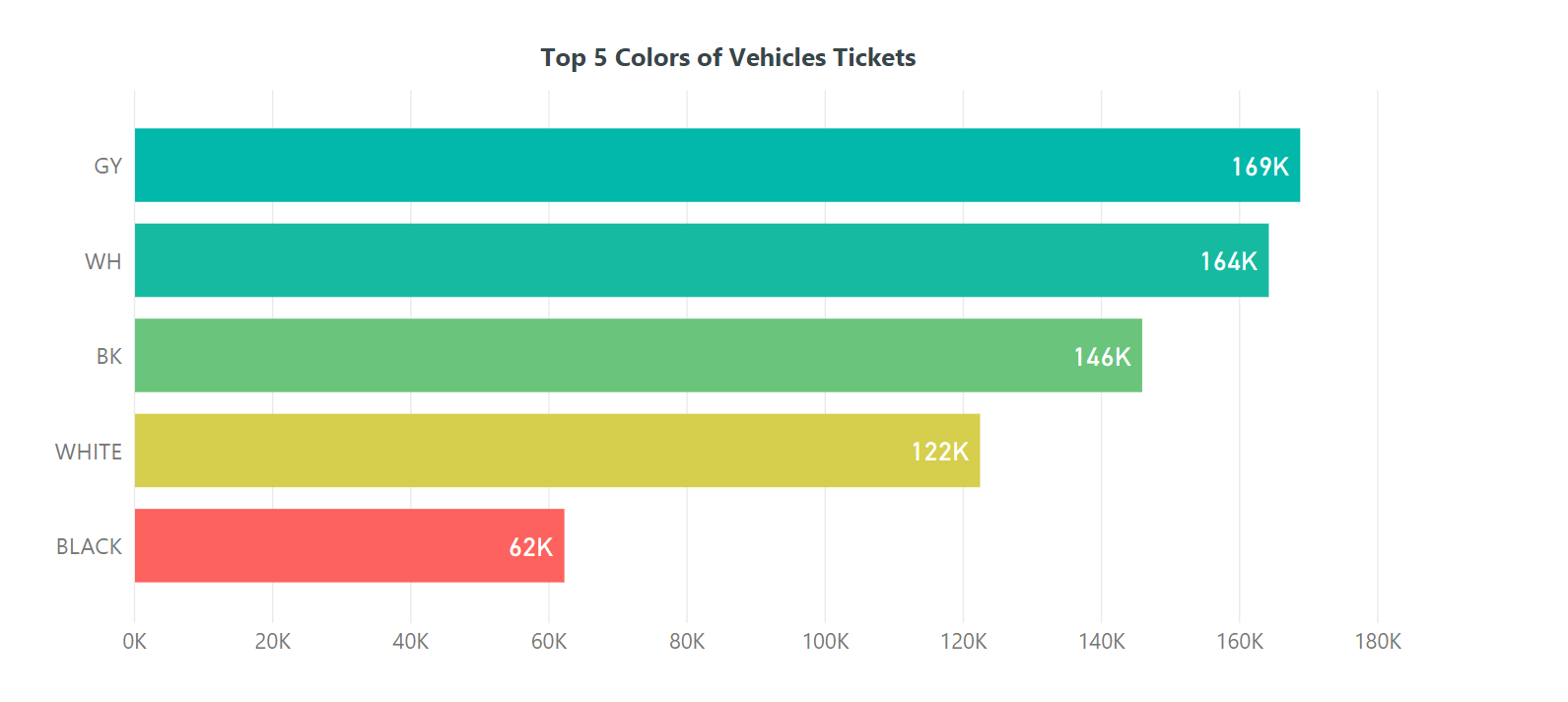
****

Figure 2 The color of cars that attracted the most violations.

Moving forward to our third finding, we decided to use ArcGIS map (Power Bi feature). Figure 3 displays the answer to our third objective. This map works as a heat map that shows what states had the highest number of cars registered that got the highest amount of tickets. Apart from New York, we found that a large number of cars that were registered in Miami had violations. Another state that had a large number of violations was Texas.

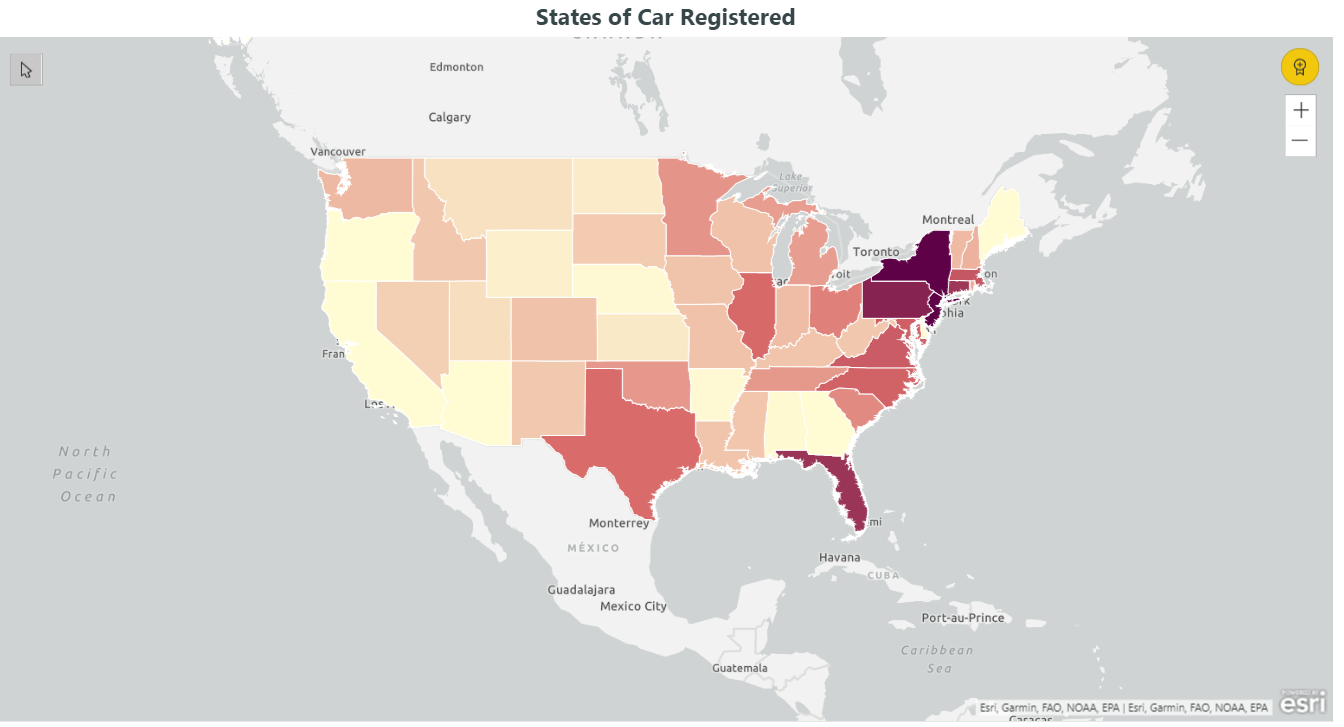
****

Figure 3 A heat map that shows what states had the highest number of cars registered that got the highest amount of tickets?

Moving on to our fourth objective, Figure 4 displays the so-called “rush hour” times. In other words, these were the exact times that cars had the highest amount of tickets.

The times that cars got the most violations were 8:36 am with 2,664 violations , 11:36 am with 2,633 violations , 11:40 am with 2,626 violations , 9:06 am with 2,486 violations and 9:36 am with 2,453 violations. A good logical reason for most violations being during these hours is the possibility that those are the times that the city street sweeping is completed.

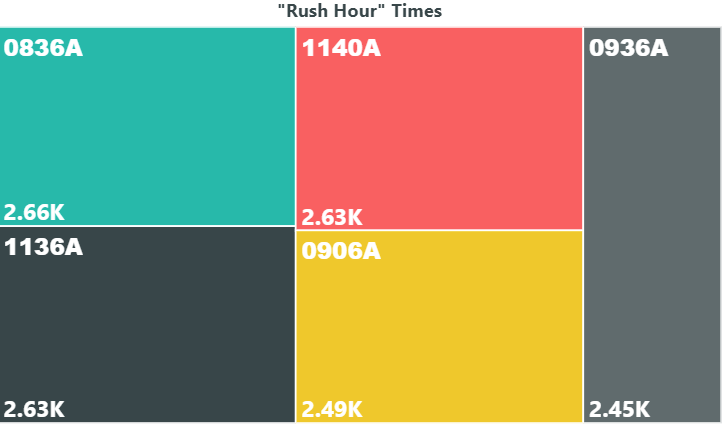


Figure 4 The times that had the highest amount of tickets.

The last objective that we were hoping to complete was which locations in New York had the highest amount of violations. To do this we used a regular map on Power BI and plotted our data on the map. An issue that we ran into was the fact that different interactions came out in different states all over the United States. We had to compromise and zoom into only New York to see the accurate locations. We found that the following locations had the most amount of tickets: Manhattan, Brooklyn, Queens, and the Bronx. To run a deeper analysis we decided to zoom into the map to find which locations clustered. Figure 5 shows an overview map of New York that shows the locations that cars got violations.

When looking at Figure 6 we saw that there was a cluster of locations. With Figure 7 we were able to see the exact street that gave the most amount of violations.

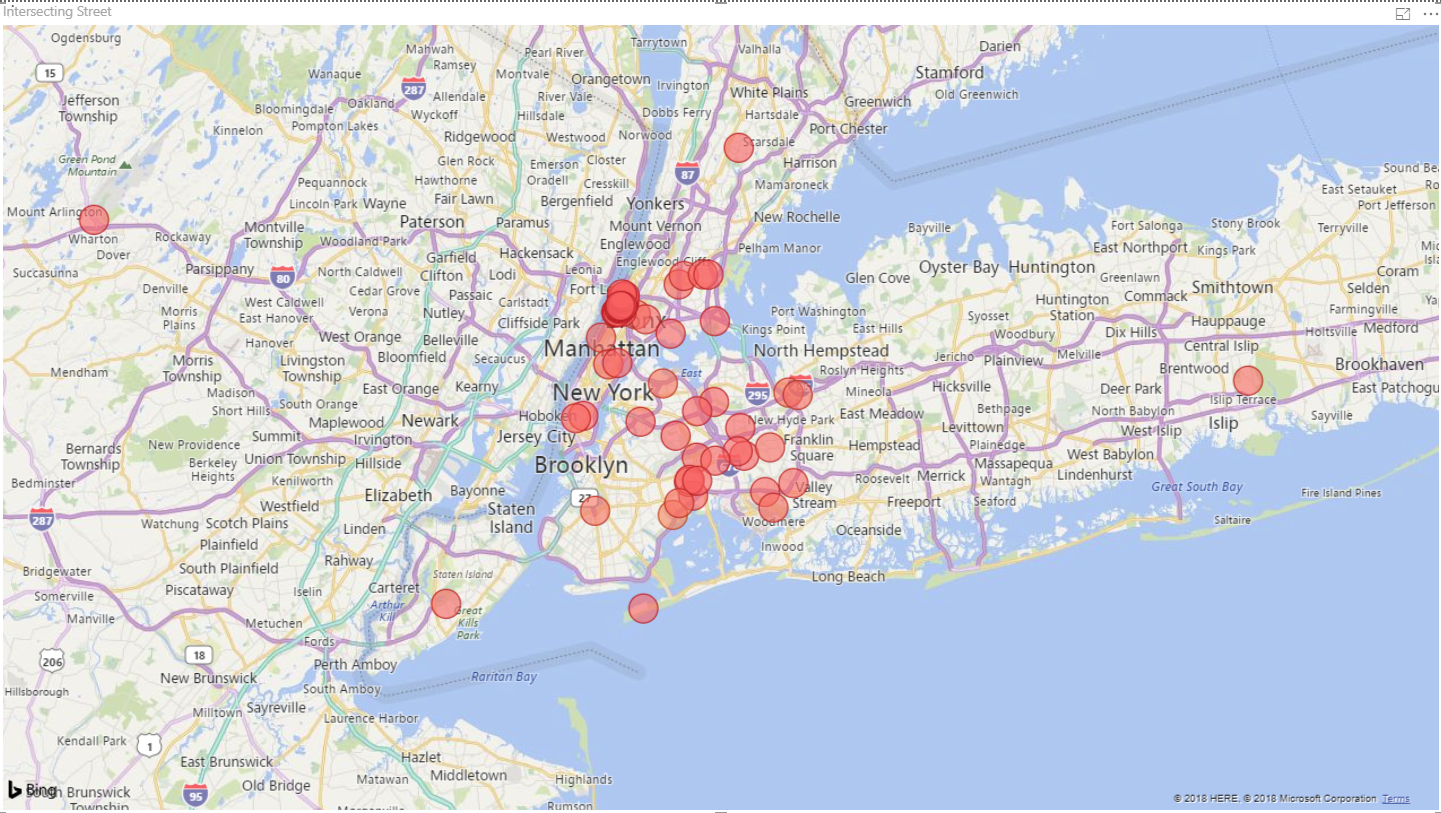


Figure 5 An overview map of New York that shows the locations that cars got violations.

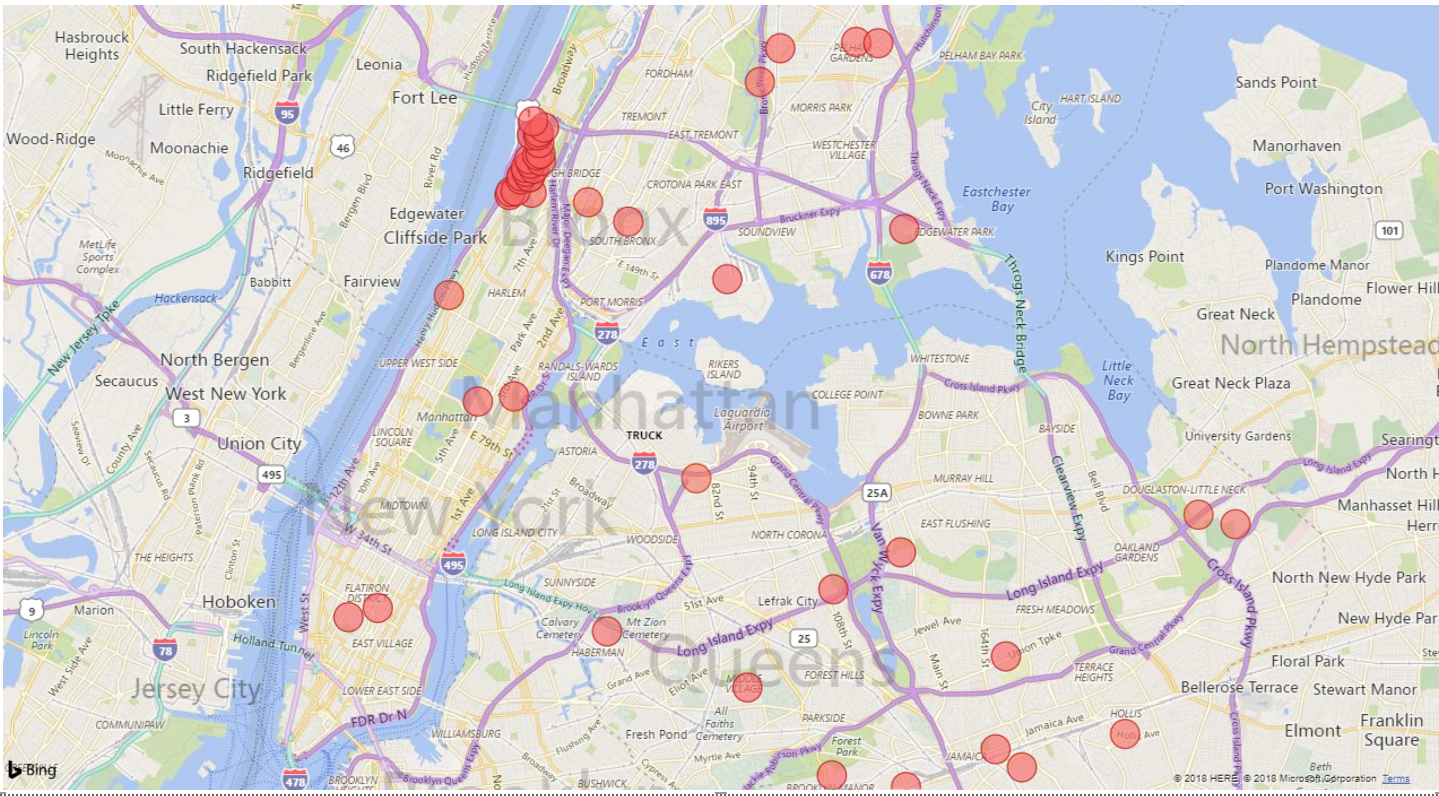


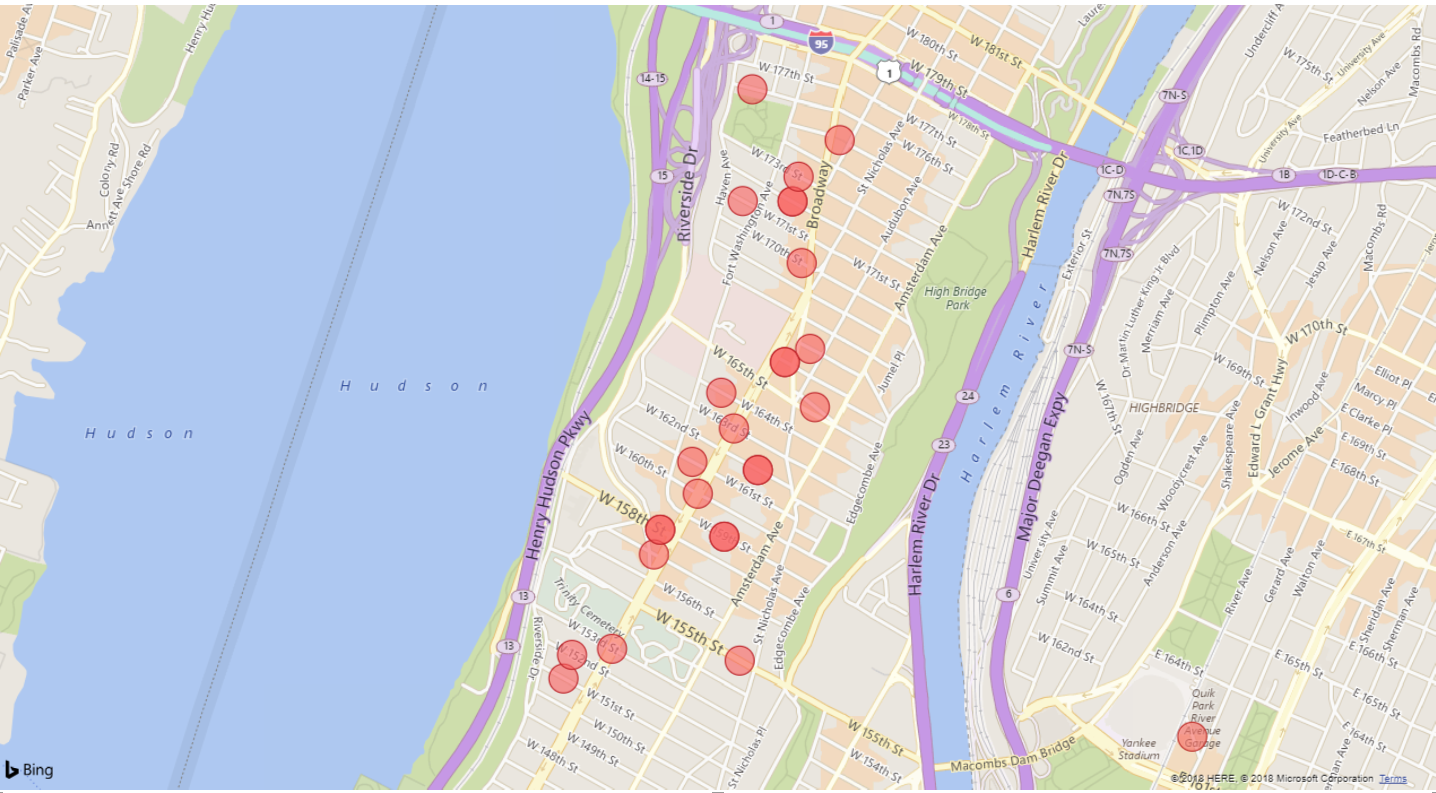
Figure 6 A zoomed in map of New York that shows us the distribution of locations by city. 

Figure 7 A map of New York zoomed into the locations that were clustered the most with violations.

**4. Conclusion**

Through our research and analysis, we discovered many interesting variables pertaining to the highest amount of parking violations in New York. First we started with the brands of cars, in which Ford was the highest. Then we filtered for color, where we discovered white cars had the highest amount of parking violations. For a more in practical analysis, we checked which times had the highest amount of parking tickets, and it turns out to be during the morning, specifically 8:36, 11:36, and 11:40 am. Along with the times we checked for the locations, and it turns out the highest concentration on the street Broadway.

**References**

[1]<https://www.kaggle.com/new-york-city/nyc-parking-tickets#Parking_Violations_Issued_-_Fiscal_Year_2017.csv>

[2]<https://www.parkingpanda.com/blog/post/top-10-cities-with-highest-revenues-from-parking-violations>

[3]<https://github.com/oscarmnz8/Big-Data-NYC-Parking-Tickets>