# Open Parking and Camera Violations Analysis

By Diana Konsevych, Dinara Ibotova, Beelen Guzman, Ray Kim

CIS3120 Bilal Abdulrahman

December 11, 2024

**Introduction:**

Statement of the problem: Living in cities such as NYC can create challenges like traffic or parking violations.

Objective: Our objective is to analyze the parking violation data from NYC Department of Finance (DOF) in the year 2024 to identify pattern, trends or insights of car violations. Our goal in forming questions is to address and analyze parking violations and the issuing agencies while providing recommendations to improve on passengers to drive safety.

Importance of the project: The analysis of parking and traffic violations data is the first step in helping reduce the number of violations. It can be used to create targeted marketing campaigns that would encourage each group of drivers to drive safer. Alternatively, it can give government agencies insights on whether the current fine amounts are effective.

**Data Overview:**

Source: The dataset was provided by the Department of Finance (DOF) which is available at NYC OpenData, which now ones this dataset. It provides records of parking violation issues of Year 2024.

Type: The dataset has been organized into a table format, with related attributes. It included numerical values, categorical and time series.

Size: Originally, the dataset had over 122 million rows of data. However, upon numerous attempts to download the data, all our computers would crash. Thus, we decided to use the API point with a limit of 1 million rows, as it was the optimal amount that would not crash our computers.

**Methodology:**

Key steps in analysis: First, we used API to load our data. Then, because we could not use all 122 million rows, we saved our API output into a csv file, which we re-uploaded and used for all further analysis. Afterwards, we removed the rows with null values and renamed ambiguously named columns. Then, for every question on our agenda, we grouped and filtered the data accordingly (e.g., find the top 5 violations and top 5 licenses with most violations). We used .dropna() .rename(), as well as groupby().size() and .nlargest()

Tools and Software used: Python and python libraries such as Pandas, Requests, and Seaborn. We used Deepnote to collaborate in real time. Then, we downloaded the jpynb to be submitted.

Challenges:

One of the biggest challenges was the amount of data as briefly mentioned in the data overview. Due to the way we used the API endpoint, every time we would run the notebook, the data would change and present new errors, the graphs would look different. Thus, we decided to download the result of the

API interaction into a csv file and load it back into our workspace for consistency of results and being able to actually address errors in code.

Another challenge was violation time, we had a hard time manipulating the column every time we tried to change the format all the values in violation time became NaN values. Due to having continuous trouble with it we have decided as a team not to work on the violation time and concentrate on other parts of the data.

One challenge we were able to fix was an issue we had with counties being named differently. As we were working on the question which county received the most tickets/violations we noticed that some of the counties were being counted separately because they were inputted into the database with a different title. For example, Queens County would have violations under county names. Q, QNS, Qns, and QN. To fix this we created a function bucketing all of the county titles into either QNS, NY, Kings, Bronx, and Rich counties.

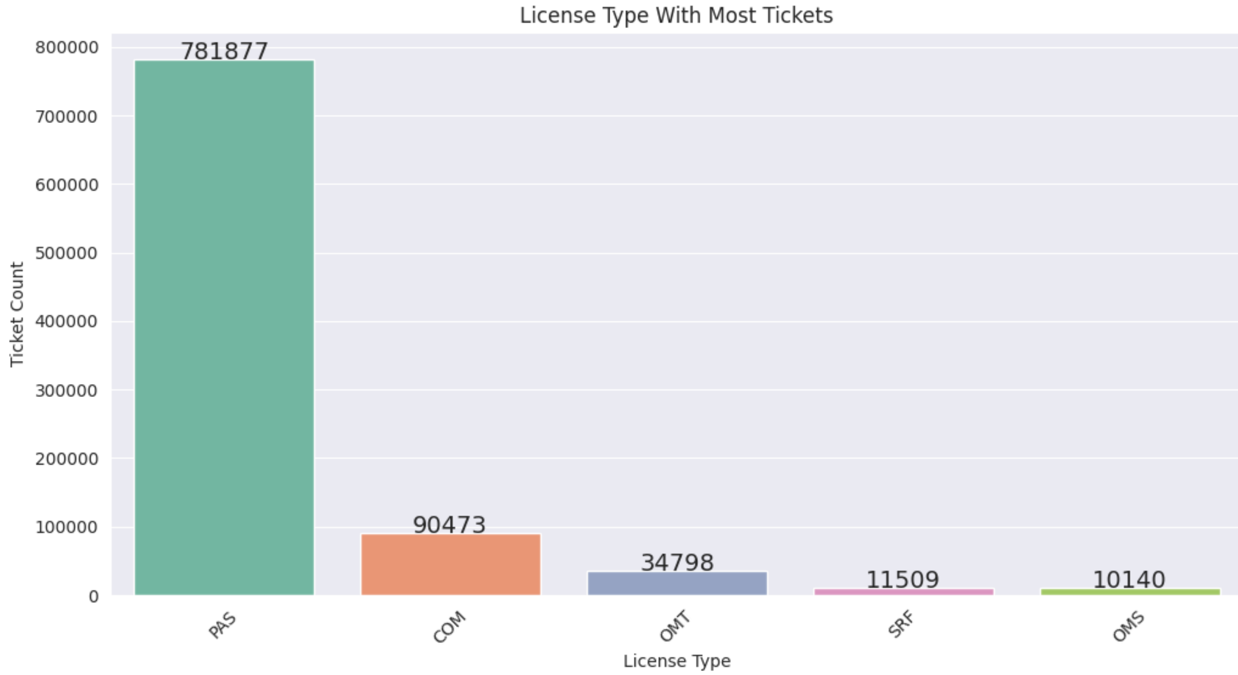
**Key Analysis and Findings (more detailed in jpynb) and Discussion:**

<u>Which license type has the most violations?</u> Analysis included grouping by license_type and then filtering by the top 5 license types with most violations (if we didn't, the graph would look incredibly convoluted, and you would not be able to draw any insights from it).

Key insights: the top 5 licenses by most violations are as follows (in descending order): PAS (passenger vehicle (non-commercial)), COM (commercial vehicle (e.g., trucks, delivery vans, taxis)), OMT (Standby Vehicle (e.g., taxi & limousine commissions)), SRF (Personalized Plates (e.g., cars, mini-vans, SUVs etc.)), OMS (Special Omnibus Rentals). The personal passenger vehicles (PAS) have a drastically higher ticket count than other license types with 781,877 violations, followed by commercial vehicles with 90,473 violations, followed by taxis and limousines with 34,798 violations, and vehicles with personalized plates and special omnibus rentals with 11,509 and 10,140 violations respectively. Does that mean that passenger vehicles have more reckless drivers? Not necessarily. Perhaps, there are just way more PAS vehicles on the road than commercial, public transport, and rental type vehicles. But unfortunately, this data was not available to us (limitation of the analysis). The overall findings match our hypothesis.

How can this be used? This data can be used in a targeted marketing campaign to decrease the number of violations per specific group of drivers.

Additional limitations of the analysis: as for our whole project, we are only using a million rows of data out of over 122 million because that is all our computers can handle. So, the results may not be a 100% representative of reality.




**<u>What are the top 5 violations and their corresponding fine amounts?</u>**

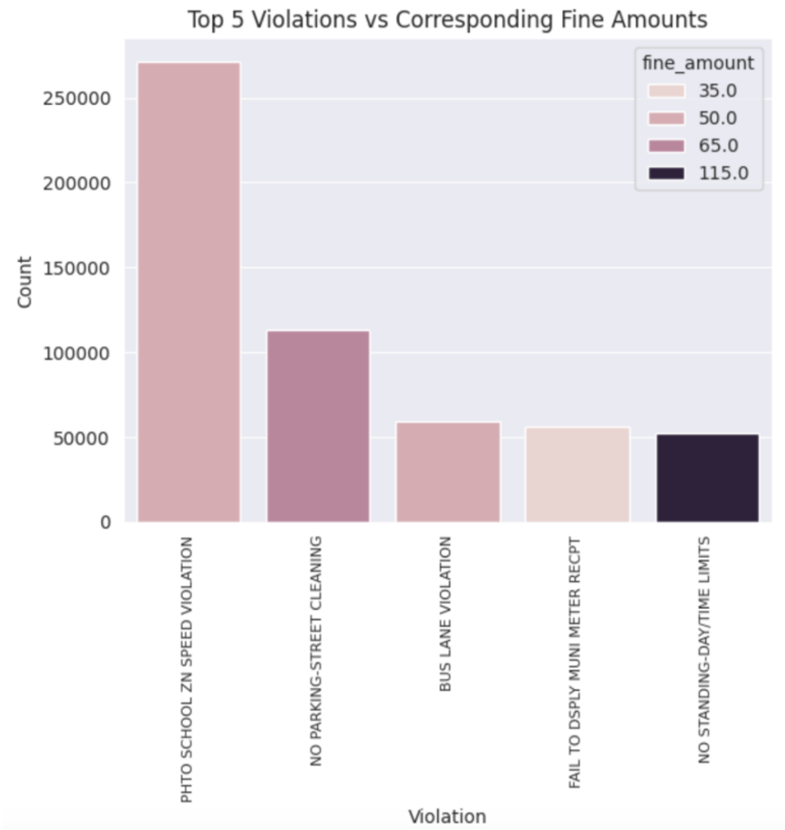
The analysis included grouping by violation and fine amount, as well as using nlargest() to find the top 5.

Insights and Discussion: The top 5 violations are (in descending order): Photo school zone speed violations with the fine amount of $50, No parking - street cleaning with the fine amount of $65, Bus lane violations $50, Fail to display muni meter receipt (muni meter receipt refers to the parking meter receipt) with the fine amount of $35, No standing - day time limits with the fine amount of $115. So, as you can see, the most common violations are not necessarily the most expensive. In fact, the most expensive among the top 5 most common violations (no standing - day time limits) is the least common. However, the fee amounts also do not decrease with the amount of violations per category. What does that mean? It is reasonable to infer that there is no apparent correlation between the fine amount and reduction in violations. In other words, a rudimentary increase in the fees will not necessarily reduce the prevalence of a given violation. In our opinion, the price increase might have to be drastic to be effective.

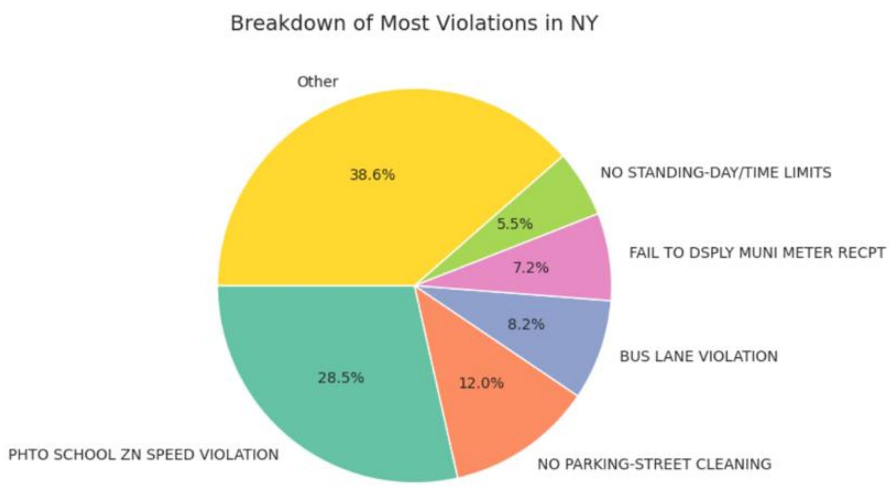
How can this data be used? It can be used to lobby for significantly higher fine amounts under the premise that small fines do not seem to do much to decrease the number of violations.

Comparison with initial hypothesis: I expected the number of violations to decrease with fine amounts. The small increase/decrease in fine amounts does not seem to influence the violation count per category.

Limitations of the analysis: as for our whole project, we are only using a million rows of data out of over 122 million because that is all our computers can handle. So, the results may not be a 100% representative of reality.




| violation | fine_amount | count |
|---|---|---|
| PHTO SCHOOL ZN SPEED VIOLATION | 50.0 | 271523 |
| NO PARKING-STREET CLEANING | 65.0 | 113386 |
| BUS LANE VIOLATION | 50.0 | 59014 |
| FAIL TO DSPLY MUNI METER RECPT | 35.0 | 55839 |
| NO STANDING-DAY/TIME LIMITS | 115.0 | 52215 |




**<u>How are the top 5 violations distributed per top 5 license:</u>**

The first step creating a line chart was to group and count data by the license and violation. Followed by identified top five violations and license types using nlargest(). After we filtered the data for violations and license type columns. Once all grouping, filtering and sorting is done we create a pivot table for plotting. After pivot table we were able to create line plot.

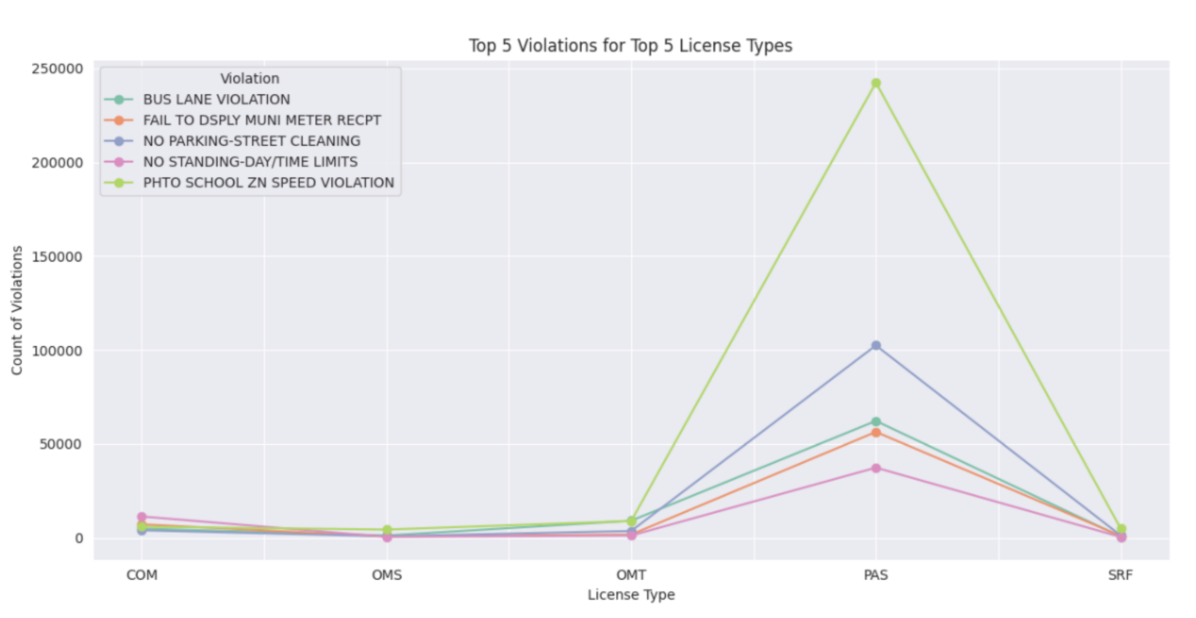
As you can see, although the top 5 violations are in the order shown in the previous bar chart, they are distributed differently among the top 5 license types.

It is interesting to see the difference in prevailing violations per license type. For instance, passenger vehicles (PAS) get a lot of speed violations in school zones (perhaps rushing to drop their kids off at school), while taxis and limousines (OMT) get mostly bus lane violation tickets (perhaps picking up or dropping off their customers).

How is this useful? Hypothetically, this data could be used to make targeted marketing campaign to encourage each group of drivers to drive safer.

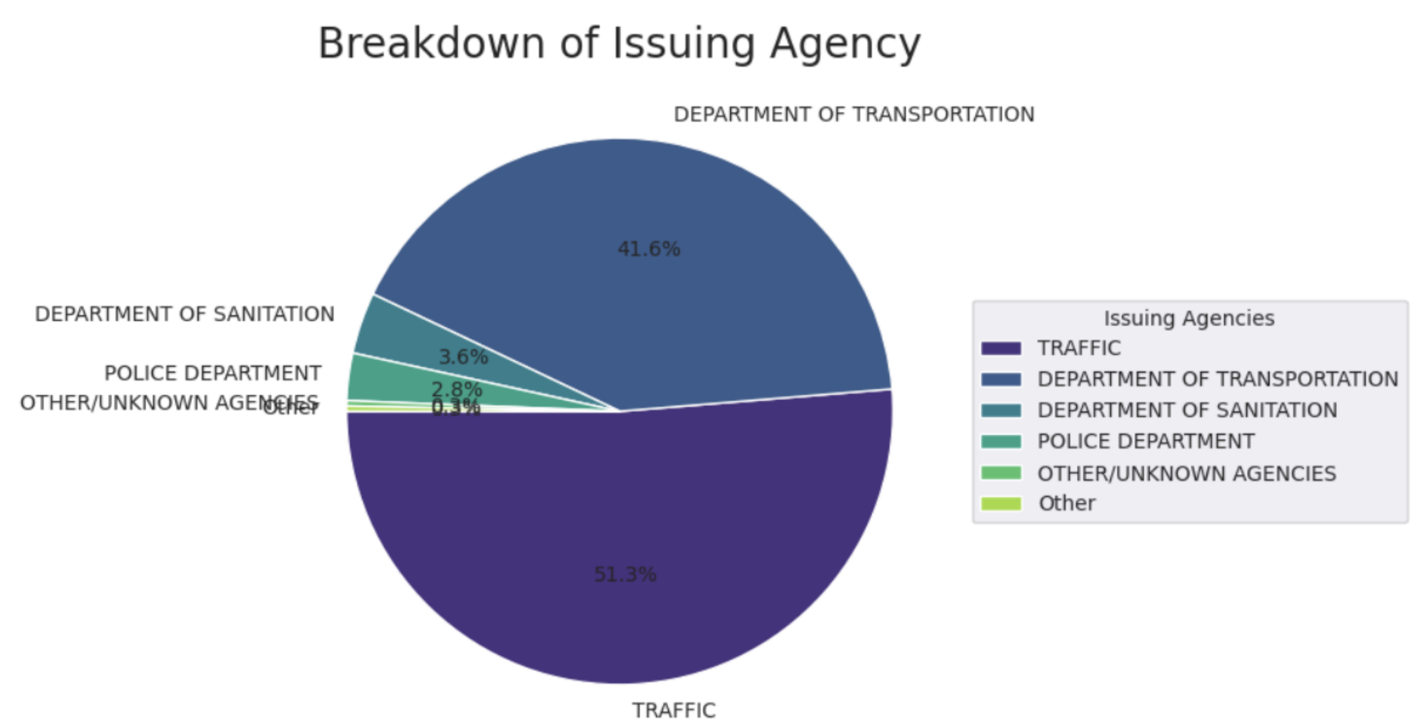
Limitations: same as for the rest of the project, as mentioned in previous questions (only able to use a limited portion of the dataset for analysis due to the size of the overall dataset.

| license_type | violation | count |
|---|---|---|
| COM | NO STANDING-DAY/TIME LIMITS | 11278 |
| | FAIL TO DSPLY MUNI METER RECPT | 7094 |
| | PHTO SCHOOL ZN SPEED VIOLATION | 5824 |
| | BUS LANE VIOLATION | 4782 |
| | NO PARKING-STREET CLEANING | 3905 |
| OMS | PHTO SCHOOL ZN SPEED VIOLATION | 4301 |
| | BUS LANE VIOLATION | 1134 |
| | NO PARKING-STREET CLEANING | 717 |
| | FAIL TO DSPLY MUNI METER RECPT | 613 |
| | NO STANDING-DAY/TIME LIMITS | 424 |
| OMT | BUS LANE VIOLATION | 9024 |
| | PHTO SCHOOL ZN SPEED VIOLATION | 8860 |
| | NO PARKING-STREET CLEANING | 3550 |
| | FAIL TO DSPLY MUNI METER RECPT | 1658 |
| | NO STANDING-DAY/TIME LIMITS | 1143 |
| PAS | PHTO SCHOOL ZN SPEED VIOLATION | 242380 |
| | NO PARKING-STREET CLEANING | 102405 |
| | BUS LANE VIOLATION | 62182 |
| | FAIL TO DSPLY MUNI METER RECPT | 56326 |
| | NO STANDING-DAY/TIME LIMITS | 37293 |
| SRF | PHTO SCHOOL ZN SPEED VIOLATION | 4847 |
| | NO PARKING-STREET CLEANING | 1151 |
| | FAIL TO DSPLY MUNI METER RECPT | 957 |
| | BUS LANE VIOLATION | 540 |
| | NO STANDING-DAY/TIME LIMITS | 387 |



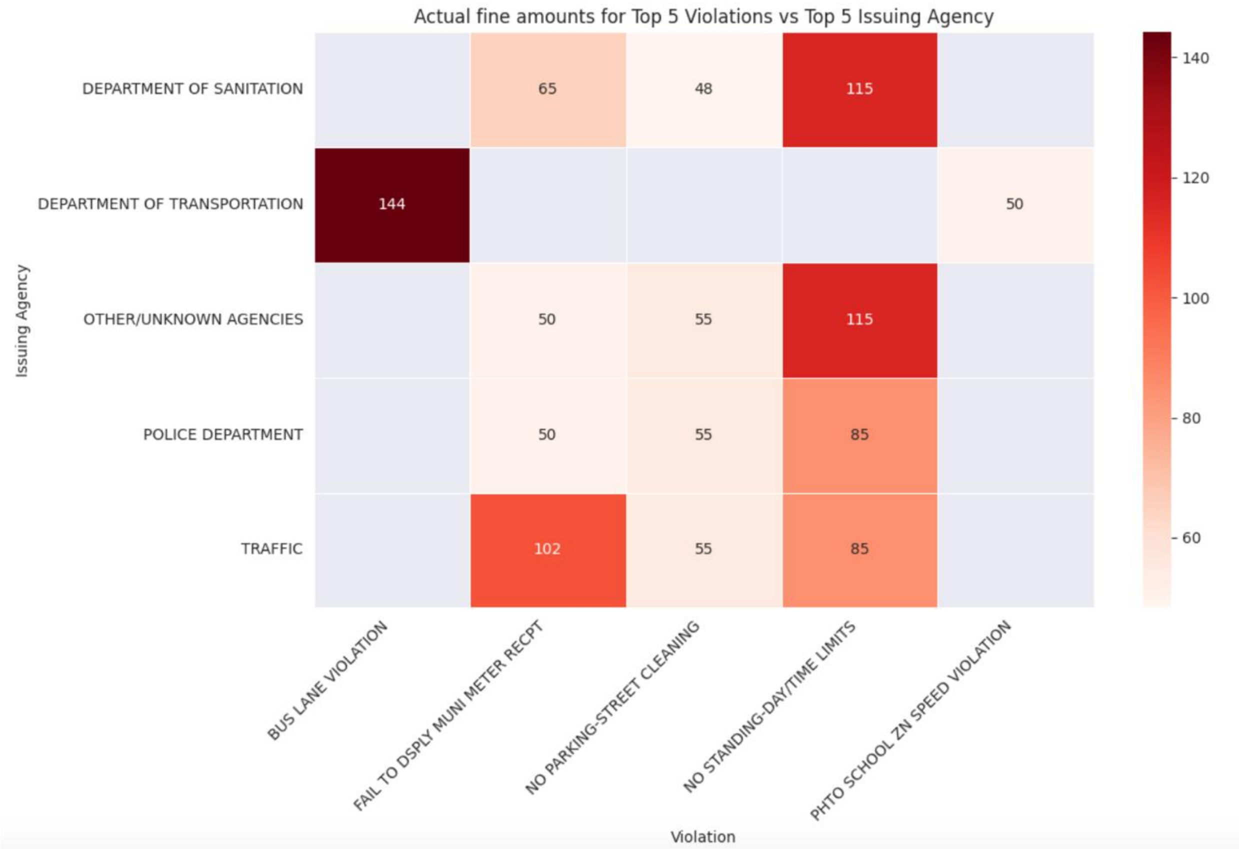

**<u>Which issuing agency gives the most violation tickets? :</u>** To determine which agency has contributed the most to violation tickets, the analysis involved keys steps when using the dataset. The first step in doing so was to calculate the total number of violations that were issued. In doing so I used the counts function to provide a clear breakdown of # of tickets coming from each agency. Since there were a lot of agencies with small number of counts, we decided it would be best to illustrate the top 5 agencies since there are 2 agencies that take majority of the violation. I then proceed to group smaller agencies into the same category under "Other". This will allow for the visual to be more easily appealing.

Based on the data provided for the issue agency, it reveals that most of the violations are issued by Traffic,51.3 %. Following, the Department of Transportation being nearly half of the contribution, 41.6%. With both agencies combined, it contributes 92.9% of the violations, indicating the most of enforcement would contribute to transportation issues. Aside from that there are smaller contributions from the following agencies, contributing 6.9%. These smaller agencies are the Department of Sanitation (3.6%), Police Department (2.8%) , Other/Unkown Agencies (0.2%) and Other (0.3%). Although these are less contributions being made into violation, their minimal involvement can reflect areas for improvement in enforcement. Therefore, it is essential to monitor if any changes are made to help improve in areas in need.



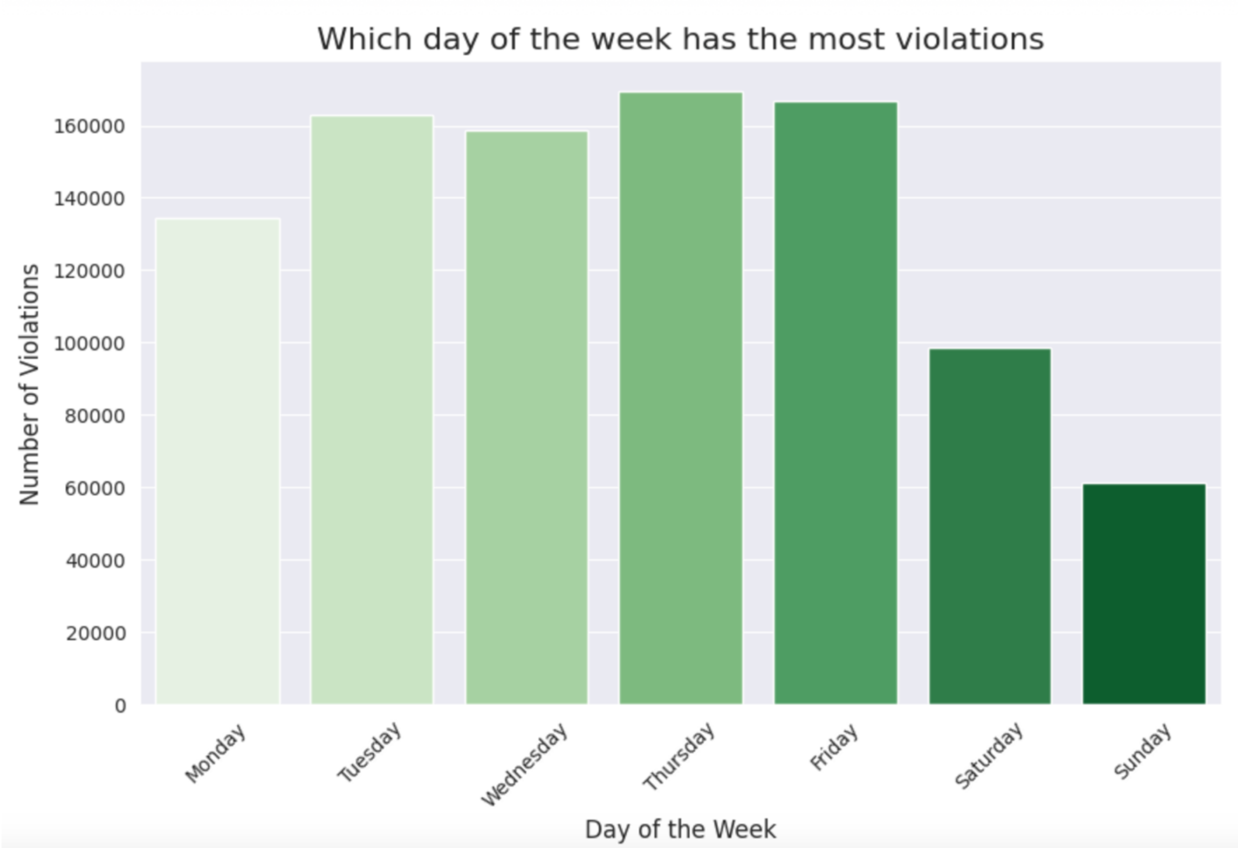

**<u>How do fine amounts vary across issuing agencies and violations?</u>** The first step of creating heatmap between the issuing agency, violations and fine amount involved converting fine amount column into a numeric value, as a heatmap would not work. Following the conversion of fine amount column into numeric value, I grouped and filtered the data by issuing agency, violation, and fine amount. Specifically, I wanted to visualize the top five violations as well as the top five issuing agencies. Before the final step of plotted heatmap had to create a pivot table.

As we can see from the heatmap, sometimes the amount of violation fee might change depending on the issuing agency. For example, as we can see the "No standing -day/time limits violation fee amount varies significantly by the department of Sanitation charging 115$ meanwhile the police department and traffic agency provide a lesser fee of 85$. There's another violation which is "Fail to display muni meter receipt" where the traffic issuing agency charges more than other agencies. So, in this heatmap visualization, we can see that sometimes depending on the issuing agency the fee amount changes.



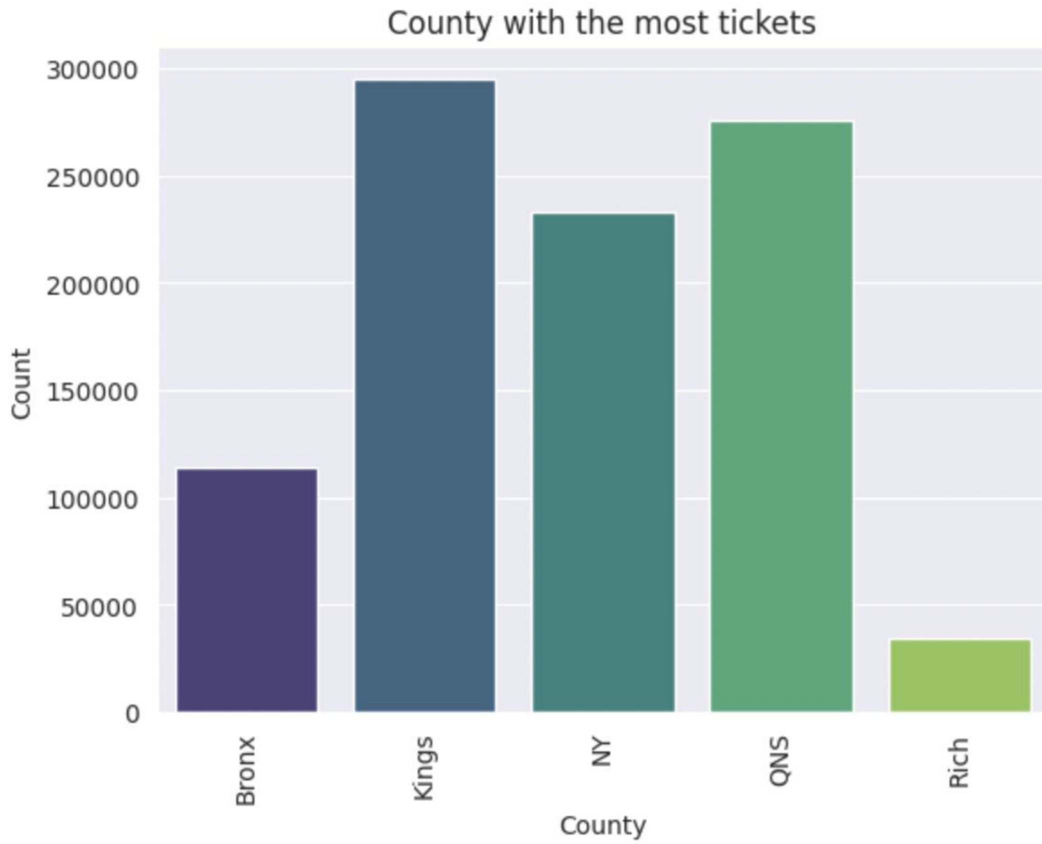

**<u>What day of the week have the most violations occurred?</u>** As the first step in formulating the analysis, the issue date was first transformed into the datetime format. The next step was to obtain the day of the week of the issue date. In the analysis I undertook, I converted some columns into categorical as well as numerical data before proceeding to group the various columns so as to be able to run the create pivot as well as group by without any hitches. Then counted the number of the days of the week and violation columns and followed them with grouping. The next code handles all the visuals in a way that the isolated day of the week can be shown. I arranged them in the format of the day of the week, because otherwise the week of the days was not in increasing order. The violations were then counted, and the highest number of violations by day was also recorded. At long last the bar plot was created.

TAs seen from the breakdown, most violations occur during weekdays, with Thursday contributing the most at 169,481 violations, followed closely by Friday at 166,554. Tuesday and Wednesday also comprise a significant share, with 162,918 and 158,467 violations, respectively. Together, these four weekdays account for the majority of violations. This indicates that enforcement or traffic activity is significantly higher during the workweek. As for days with fewer violations, Monday, Saturday, and Sunday collectively account for a smaller share, with Monday: 134,512 violations Saturday: 98,602 violations Sunday: 61,226 violations. Sunday's violations are approximately 37% lower than Saturday's, which could be because usually Sunday's parking is free in NYC.



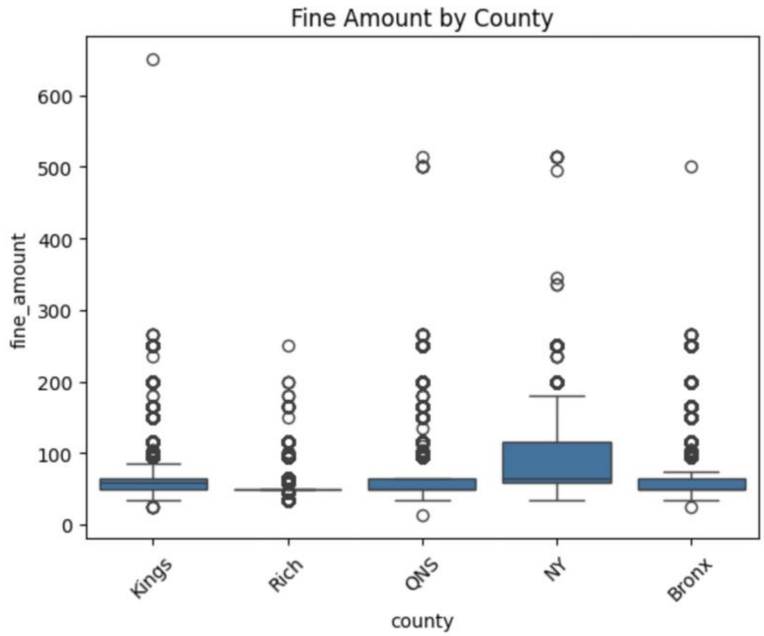

**Which county received the most tickets/violations?**

One question that arose during our conversations as a group was which county received the most violations. With the number of tickets we were working with in our database we were curious to see just how many violations/tickets each county received. The first step to our analysis was organizing the counties to make sure violations were getting bucketed to the correct location. After this, we plotted a bar chart based off violation counts per county. From our bar chart we can clearly see that King's County had the highest number of traffic violations in NYC at 29,5016 traffic violations and Queens came in second with 27,5540 traffic violations. At first assuming the population and the reputation of Manhattan traffic, I assumed NY county would come first, but NY county came in third at 23,2931 violations. In the case of Kings County there could be a multitude of reasons this specific county received more violations than the rest. Kings County could have more vehicle owners, more cameras / officers monitoring motorists, more drivers being careless with their driving, etc. The reason NY came in third can derive from the fact that more people tend to take public transportation rather than their cars, people are more cautious with their driving, or violations are not being carefully monitored.




With this we thought it would be a good idea to show the distribution of fine costs between each county and to complement our analysis, we decided to use a boxplot to visualize the distribution of violation costs. In the boxplot, we see that the fines are consistent between the counties.  The county with the most variation between fine amounts within the inner quartile is NY county, and it can be shown from the box's size. The county with the most Outliers is Kings County. They also have the highest fine given at 650 dollars indicating a more serious violation. Most of Kings County's fines amounts fall below the 100 range and it shows that most violations are not too severe.

With this information we can take steps to combat open parking / camera violations. For starters, in counties with higher violation counts we can set up more cameras/traffic officers to deter people from committing any violations. We can compare the counties with lower violation counts such as Bronx and Richmond to see what is being implemented in those counties that are not being implemented in Kings County and make changes. Finally, seeing how most violations tend to happen at a lower fine amount, we can start to raise fines to stress the severity of violations. As prices start to rise people tend to think twice about committing traffic violations.




**Conclusion:**

Summary of project objectives and how they were met: The main objective of our project was to analyze parking violation data from DOF and NYC Open Data. Individual questions were answered in order to approach the project objectives and during our individual analyses we identified patterns, such as high-violation counties, peak days, types of licenses, and violation types. We made sure that our visualization and analysis were consistent with our findings by checking them several times and were clear and easy to comprehend.

Recap of most important findings:

- Passenger license type receives a significantly larger number of tickets.
- The overall most common violation is "Photo School zone speed violation;" and, a small variation in fine amounts does not seem to influence the number of violations.
- However, the top violation varies per license type. For instance, commercial vehicles' top violation is "no standing – day/time limits."
- Thursdays are the highest ticket receiving day, and Sunday is the least number of violations.
- The agencies that were responsible for most of the violation are: "Traffic" and "Transportation Departments

<u>Recommendations for further analysis:</u> It would be useful to gain access to data about drivers' ages, number of registered vehicles per category as compared to number of vehicles per category per violation. It would also be good to see weather data per day and how that may or may not impact the number of violations and if yes, which ones specifically and for which types of vehicles (or license types).

**References:**


Department of Finance, NYC Open Data. “Open Parking and Camera Violations” NYC Open Data. Updated November 26,2024 https://data.cityofnewyork.us/City-Government/Open-Parking-and-Camera-Violations/nc67-uf89/about_data

GeeksforGeeks. “Box Plot.” GeeksforGeeks, 12 Mar. 2024, www.geeksforgeeks.org/box-plot.

Matplotlib.”Labeling a pie and a donut”, n.d, https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_and_donut_labels.html#sphx-glr-gallery-pie-and-polar-charts-pie-and-donut-labels-py


**Tools Recap:**

- Python
- Excel
- Chat GPT *(for help in solving errors and understanding concept out of this class’ scope)*
- Jupyter Notebook
- DeepNote
- Class Notes