August 2023 Traffic Violations: A Visual and Analytical Study

By Zachary Perry, Daniall Masood and Ramana Bhaskar Kosuru

This paper presents an exploratory data analysis (EDA) and visualization of the August 2023 traffic violations dataset, comprising over 100,000 entries related to moving violations issued within the month. Our study focuses on uncovering the factors that affect the fine amount and attempting to accurately predict it using machine learning models. Key aspects of the dataset include the spatial distribution of violations, temporal patterns in issuance, and the financial implications of fines and penalties. Preliminary findings suggest a complex interplay between location, time of day, and enforcement intensity. Visualization efforts highlight geographic hotspots for violations, categories of citation, and temporal trends that could guide further policy and enforcement strategies. This analysis serves as a foundation for deeper investigations into the factors influencing fine amounts and their implications on urban management and policy formulation.

# Introduction

The choice of data in this analysis stems from its suitability for visualization and its spatial components. The dataset, consisting of moving violations issued in August 2023, offers a comprehensive log that provides insights into traffic enforcement patterns across various regions and timeframes. Sourced from Open Data DC, it comprises over 100,000 entries on traffic violations, laying a robust foundation for exploratory data analysis (EDA) and subsequent visualization.

The inclusion of latitude and longitude coordinates enables spatial plotting, presenting a powerful visualization method. Moreover, the data's granularity allows for aggregation into various units, giving flexibility for analysis. With the inherent detail in the Coordinate Reference System (CRS), aggregation can be performed at multiple levels, such as wards or zip codes, to facilitate further analysis. Additional data, such as income levels, crime rates, or population demographics, could be integrated to enhance the dataset's explanatory capabilities.

The primary objective of this analysis is to unearth trends within the fine amount that can guide policy adjustments. This study particularly focuses on the geographic distribution of violations, temporal trends in their issuance, and the economic impact reflected through imposed fines and penalties. The inclusion of the agency that issues the fine and the type of citation issued also have to be accounted for when making models. It is important to account for all of these aspects but initial exploration can also be a tangential study to capture specific aspects of it.

To augment the exploratory data analysis, this study employs specific predictive modeling techniques aimed at enhancing our understanding of traffic violation fines. A Random Forest model predicts the fine amount according to the geospatial aspect of the data with the attempt of accurately capturing that component's effect on fines. Concurrently, a Logistic Regression model forecasts the fine amounts, using a host of variables to attempt a broad stroke approach where many aspects of the model are used to make an initial model that captures a large amount of variation but may contain multicollinearity or have high error.

The paper itself will first detail the cleaning and preparing the data for analysis, including addressing missing values, handling non-normal distributions, and ensuring the data's representativeness for modeling purposes. Following this preparatory phase, a detailed exploration of the data through statistical summaries and visual representations will be conducted. Finally, through the construction of a variety of models we can attempt to explain and predict the fines issued in DC. This exploration aims to reveal underlying patterns that may not be immediately apparent from the raw data alone, providing a comprehensive overview of the enforcement landscape during the period under study.

Through analysis of this dataset, we aim to contribute to the broader discourse on traffic management and law enforcement, offering insights that can refine approaches to traffic violation enforcement and enhance public safety and urban mobility. The findings are expected to assist policymakers and law enforcement agencies in devising strategies. Identifying potential over policing present in the lower income regions of DC.

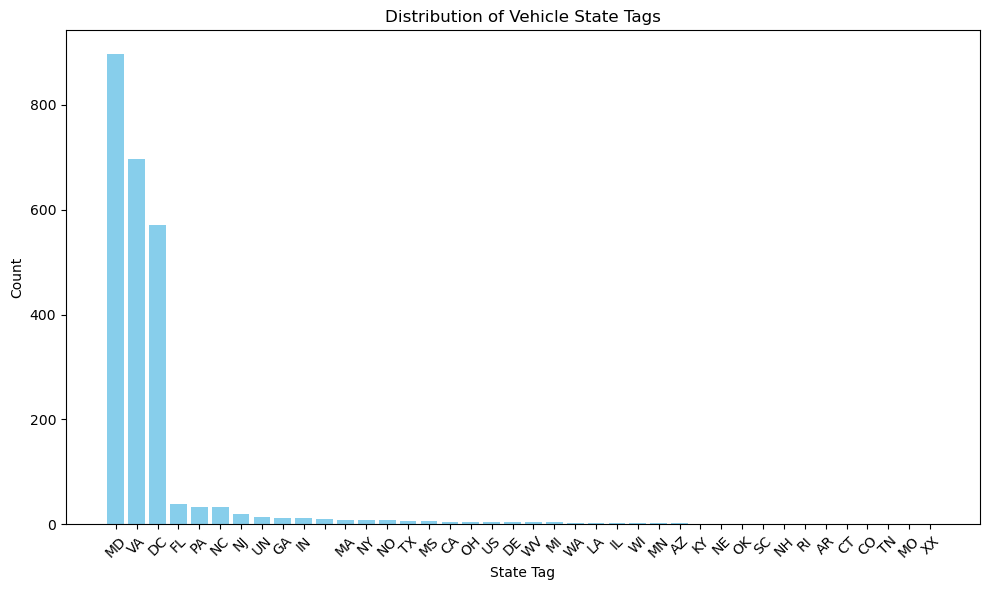
# Data

The initial dataset of moving violations contained 30 columns and 103,203 rows. The preprocessing of the data focused on removing columns that did not fit the analysis that needed to be done as well as columns that contained mostly empty or null values. Some columns had to be normalized for models to be fitted and for analysis to be run. Values in different columns had to be normalized to a certain format to ensure that when tests were run then errors could be avoided and the models would be able to run efficiently and without issue.

Two columns , “DISPOSITION\_DATE” and “BODY\_STYLE”, were dropped from the dataset as they did not contain many values that were not null or empty. These columns would not contribute to the analyses and models later on and would inevitably be removed anyways. Values in columns that contained date data were ensured to be in a consistent date-time format. Values that were deemed categorical were ensured to be in a string format and values that were numbers were converted to int64 to have stability and consistency within the data.

Initial descriptive statistics were performed to create charts and graphs to show small correlations and trends within moving violations in August 2023. These violations were issued in the District of Columbia, so some of the following charts and spatial data are going to draw conclusions based on the geography of the city as well as the socioeconomic status of different locations within the city.

State tags indicate which state the vehicle was registered in and can indicate which state the car is from, even if the resident is from DC. This could help in analyzing the driving habits of drivers from different states. However, in the data we see that many cars that were pulled over for a violation had no tags at all. From the entire dataset, 99,984 vehicles that were pulled over for a violation had no state tags. Some of the vehicles did get a violation for no tags but majority of the vehicles received other violations. From the vehicles that did have state tags and received a violation we observe that the majority of tags were from DC, Maryland, and Virginia (DMV). 969 vehicles were from the state of Maryland alone and Virginia had the next highest vehicles receiving violations (Fig 1). Now this does not indicate that drivers from these regions are worse drivers than any other regions. Keeping in mind that people drive into DC from Maryland and Virginia for jobs, school, entertainment, and various other reasons. The large skew that is present of state tags from these 3 regions merely indicates that majority of all drivers are from the DMV area and not just drivers that receive violations. Further testing and analyses must be done to check for correlation between states and the number of violations received by state.

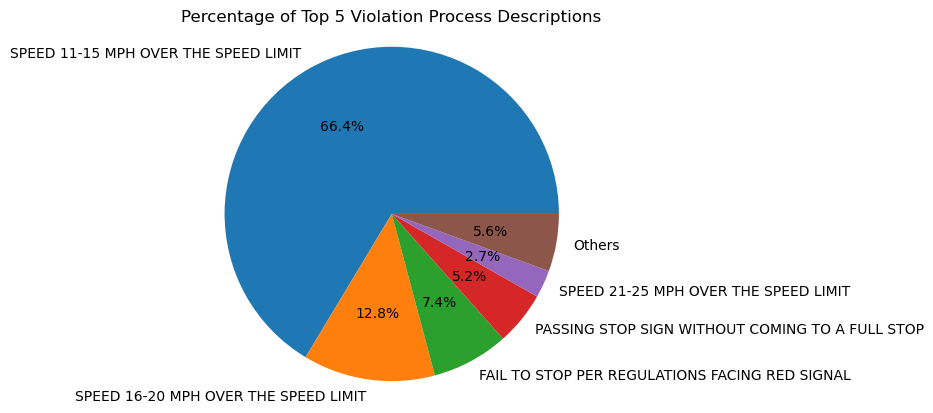


**Figure 1. Vehicle State Tag Distribution.** Majority of vehicles that received a violation are from Maryland, Virginia, and DC. This does not indicate a correlation between states and receiving a violation.

There are many different type of violations that can occur when driving. However, some violations are more likely to occur compared to others. There are violations that are more minor and violations that are more major. We observed that within the top five violations, three of the violations were related to the driver driving the vehicle over the speed limit by a certain margin. The other two violations were also minor violations that related to how the driver was operating and failing to observe small traffic laws.

From the processed data set, we observed that 66.4% of the violations in DC were due to the driver going over the posted speed limit by 11-15 MPH (Fig 2.). This indicates several things. One observation is that speeding over the limit by less than 11 MPH is either not a major offense and is not documented or that this offense is not treated as a violation. We also observe that this is the most common type of violation within this data set. This can signify how many drivers commit this violation either on purpose or due to slight negligence. Within the city going over 11 MPH is dangerous, yet is the most common violation indicating that majority of drivers commit this offense and even with so many incidents this violation is not going down. This analysis can be indicated by the fact that speeding over 16-20 MPH (12.8%) and speeding over 21-25 MPH (5.2%) over the posted limit are within the top 5 violations as well.

We can also draw the conclusion that other types of violations outside of these five are not as commonly caught or done. Major violations could include striking a pedestrian or driving while distracted. These violations are coded in the data set and have very low counts in the data. These are very major violations and are more often than not caught by officers. So to not see them counted as frequently in this data set can indicate that these violations are not committed as often and people are more wary about committing these offenses while driving.

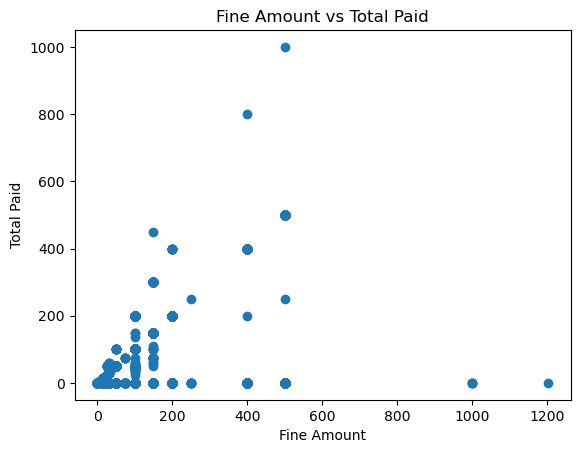


**Figure 2. Top 5 violations committed in DC in August 2023.**

The fine amount of the violations in the data set were mainly around $100 (Supp Fig 1). This indicates that the amount fined overall was very low and the average fine amount was relatively average as compared to the average driving fine amount from 2022. The average fine amount in 2022, according to Advisement, was around $150. So for most of the fine amounts being $100 was typical. This also aligns with the top five violations observed. The top five violations can be considered as minor violations so for the fine amount to be around $100 can be justified and normal. We did observe some very high fine amounts at $1000 and $1200 as well. The violation for $1200 was for overweight motor vehicles, while the violations for $1000 range from unauthorized tint to uninsured vehicles.

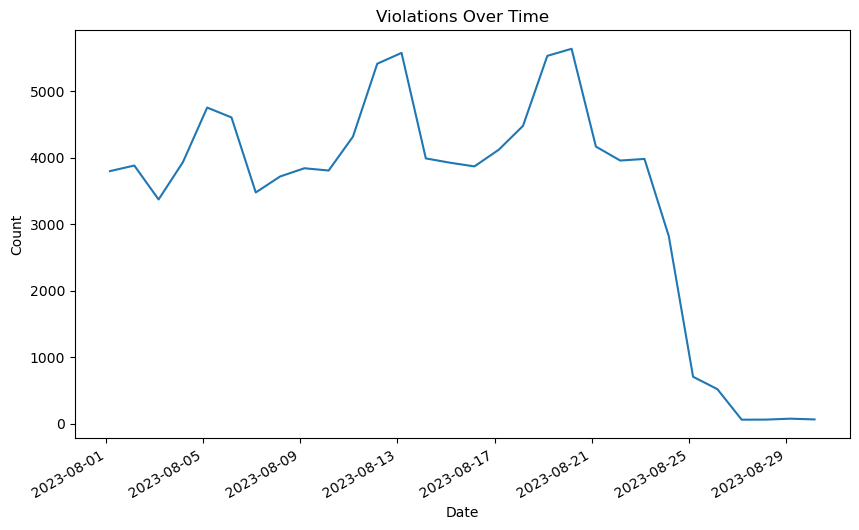
We then observed the relationship between fine amount and total amount of fine paid. More often than not we see that at times the fine amount is not paid or not paid in full (Fig 3). We see that even for the low fine amounts the fine more often than not goes unpaid. There is no clear correlation between the fine paid and the fine amount.

Some interesting points of data worth mentioning are the high amounts of fines ($1000 and $1200) went. In some cases the total fine amount paid was in fact higher than the initial fine amount. This could mean that DC puts penalties on fines that are paid after the due date. We see many instances of $100 fine amounts having higher total fine amounts paid. Some fine amounts have double the paid amount associated with them indicating that fine was paid later and more penalties were accrued. It would be interesting to observe this trend further and look at the demographic statistics of who is able to pay the fine amount and who avoided the fine amount overall.



**Figure 3. Fine amount vs Total Paid**

The inclusion of certain features within the dataset which contains information regarding date and time of issuance enables the exploration of temporal patterns embedded within the issuance of moving violations. We were interested in using these features and observing any trends, fluctuations, and potential anomalies in traffic law enforcement practices within the District of Columbia. By scrutinizing these temporal patterns, anomalies, if present, can be identified, signaling instances where enforcement may be inconsistent or irregular. Moreover, such analysis could serve as a foundational step toward enhancing the effectiveness and fairness of traffic law enforcement strategies.



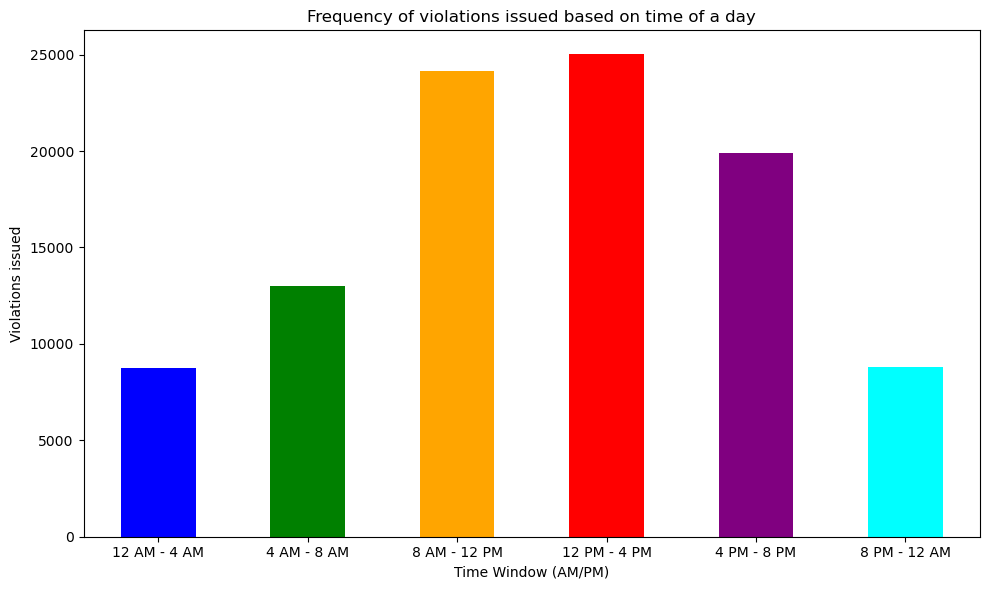
**Figure 4 : Frequency of Violations over time**

To observe these temporal trends,we utilized the features like OBJECTID,which serves as the unique identifier for a moving violation issued on a particular day at a certain time, and ISSUE\_DATE, which represents the date at which the moving violation was issued. The above line chart illustrates the variations in the frequency of violations over the month of August in 2023.

At certain points in the chart, peaks and troughs can be observed and considered usual. But after August 21, 2023 there was a sudden dip in the cumulative count of moving violations issued by all the agencies and the number almost reached close to zero. This decreasing slope over the last ten days of the month indicates that the rate at which the moving violations were issued decreased either because there is an anomaly in the traffic violation enforcement system or the number of violations perpetrated by the individuals was actually declined.

But we cannot conclusively determine whether the decrease is due to an anomaly in the system of traffic violation enforcement. To explore more on this, we were restricted by the data we possess as all the data sourced from public records is from August, 2023.

We further observed the distribution of the violations during a time of a day in August, 2023 as our data set consists of a feature ISSUE\_TIME, which represents the specific time at which the moving violation was issued by an agency. In the dataset the time was recorded in the military time format. The issued time was categorized into one of the six bins of four hour time windows that were created. The bar chart of frequency of violations was observed with time of a day as a dependent variable that was labeled into 12-hour time format. These are the combined violations issued over the month of August, 2023 by all the traffic law enforcement agencies in the District of Columbia.



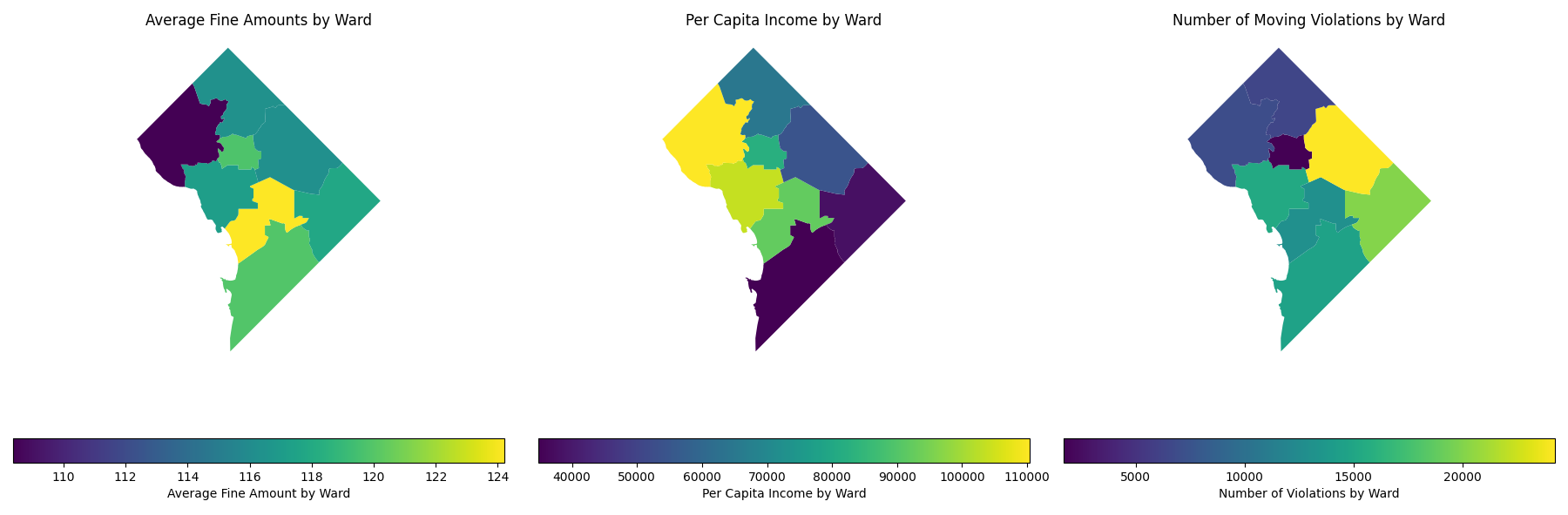
**Figure 5 : Frequency of Violations based on time of a day**

In the bar plot, a noticeable pattern resembling a normal distribution emerges. The counts of violations issued in the time windows 8AM to 12PM, 12PM to 4PM and 4PM to 8PM exhibit relatively higher frequencies of violations compared to the other time windows.

This kind of distribution was expected as there would be higher vehicular activity during the day time rather than the hours of night. But we cannot eliminate the idea that there could be unequitable traffic law enforcement in the District of Columbia as there are higher chances of occurrence of moving violations during nighttime hours.To further work on this hypothesis and draw any further conclusions the records are restricted to August, 2023.

## Spatial Component of The Dataset

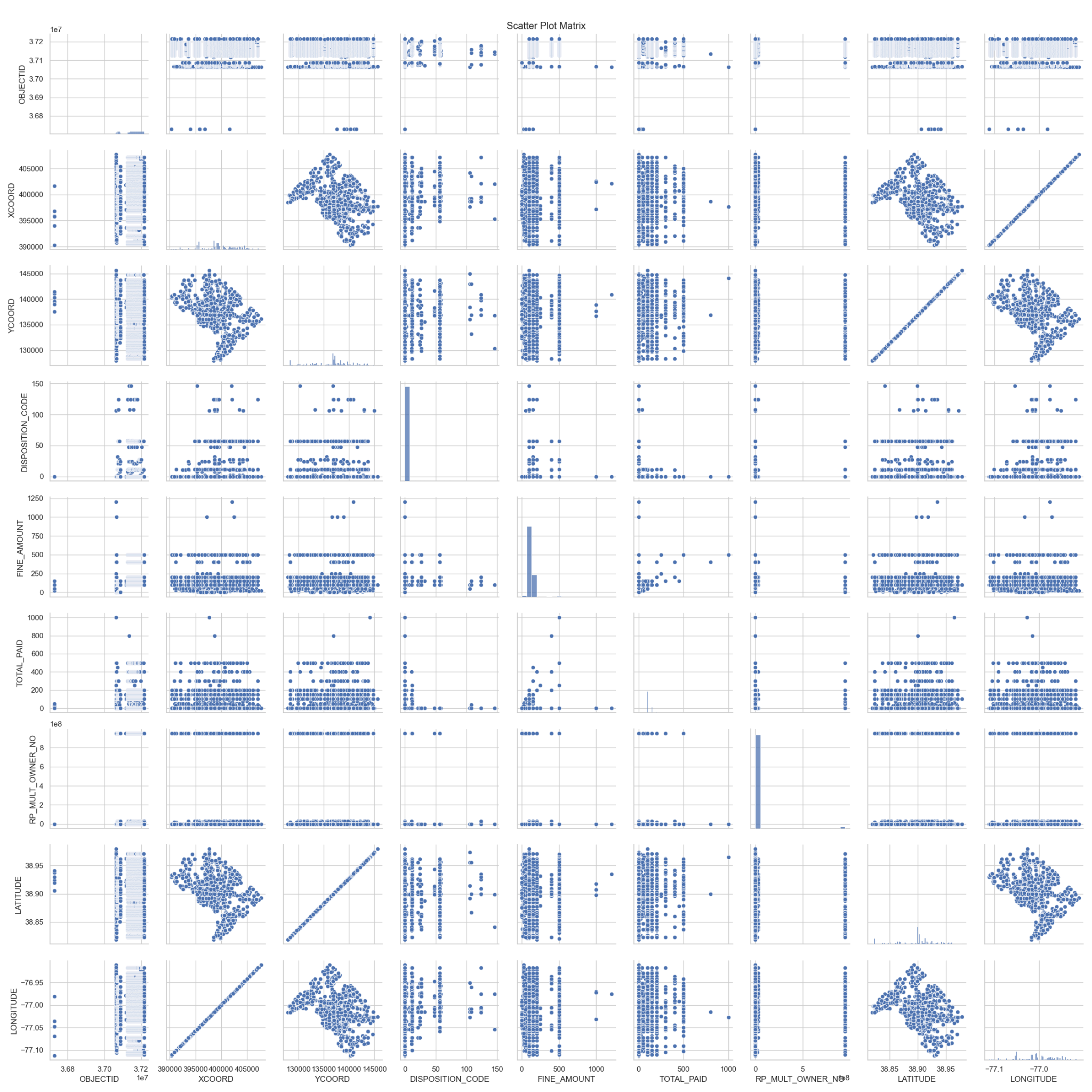
Spatial data at a pixel level is critical for visualization because it allows for a detailed view of spatial information. By plotting spatial data, analysts and viewers can observe subtle variations and patterns. This pixel specificity enables the creation of descriptive maps at any geographical level required. Granular data is essential for identifying trends for targeted interventions and policy-making. Use of spatial data at this level of detail enhances the clarity of visual representations. Easily distributing data in a means that is identifiable to any group is its greatest strength, with other plots of the same data conveying information that is much harder to understand.



**Figure 6 : Triple spatial plot of the average fine amount, per capita income, and number of violations by ward**

The triple plot image displays the spatial representation of the data across different wards, showing the average fine amounts, per capita income, and number of moving violations. This spatial visualization is useful for a few reasons:

1. The spatial distribution of average fine amounts by ward shows variance wards, indicating a pattern in enforcement or compliance that warrants further investigation.
2. The per capita income by ward, when also displayed with the average fine amounts suggests this regressive relationship in policing practices.
3. The number of moving violations by ward is also necessary in identifying over policing as it shows that the amount of citations issued is much greater in the south east as opposed to the north west. Without a plot of the population it is unknown if this is due to the population of another plot showing evidence of over policing.



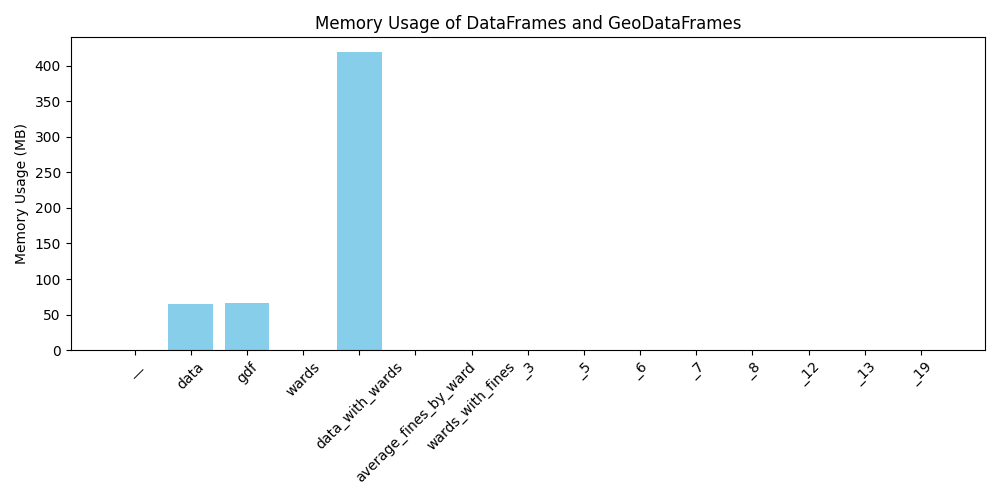
**Figure 7 : Large pairplot showing the scatter plots of all the numerical variables in the dataset**

The pair plot visualizes relationships between multiple variables, in this case, including geographic coordinates, fine amounts, and paid amounts. Here's analysis based on the provided pair plot for those specific columns:

1. Geographic Coordinates (Latitude and Longitude):The plots of the geographical components are perfect, with it simply being mirrored across the axis but each plot makes the shape of DC.
2. Fine Amount vs. Geographic Coordinates: There is a right skew in the fine amount according to its geographical representation. The issue is the resolution is not enough to see individual data points only to see the skew. The geographic plots should be used for deeper analysis.
3. Paid Amount vs. Geographic Coordinates: Same as the fine amount.
4. Fine Amount vs. Paid Amount: The direct relationship between the fine amounts and the paid amounts make sense. With there being a few values where the paid amount is greater than the fine which is confusing. I assume this to be due to not paying.

### Limits of Spatial Datasets

1. Accuracy and Precision: The spatial data can only be as accurate as the report. If the location is reported without latitude and longitude, only reporting the ward for instance then it greatly reduces the resolution of the dataset.
2. Data Volume and Management: Large volumes of high-resolution spatial data present challenges in storage, processing, and analysis. The current dataset already suffers from the issue despite it being relatively light (Data is not larger than ram)
3. Complexity of Spatial Relationships: The relationships within spatial data can be intricate, with a multitude of influencing factors. I have attributed a regressive fine system as the root cause between the inverse relationship between per capita income and fine amount. This was found through only the spatial plots as models and correlation plots cannot currently capture such a complex relationship.
4. Dependency on Technology: Spatial data analysis relies heavily on technology. Errors or malfunctions in data collection or analysis technologies can introduce inaccuracies in the data, which can propagate through to the insights derived from it.
5. Spatial Autocorrelation: Spatial autocorrelation can be a limitation when analyzing spatial data because points near each other tend to have similar values. While this can be useful in some analyses, it can also lead to overestimations of patterns or relationships within the data, potentially skewing results. This is the source of the current lack of accuracy in predictive statistics. MPD-ROCN dominates the ISSUING\_AGENCY columns and that has autocorrelative effects as their specific citation will then dominate their own column. This is also true for the residuals as MPD Ward 1 will not have citations issued in any of the other wards.



**Figure 8 : Plot of memory usage by variable**

Correlation was calculated for fine amount and per capita income by ward and it was calculated as 29%, with a moderate negative correlation being found we can further test the relationship. In this example per capita income by ward was used almost as a proxy for the ward socioeconomic. I will also test the ward location with the x and y coordinates and try to find geospatial correlation with each axis to show that the creation of the wards is not only recommended, it is necessary to make this complex analysis. Correlation between fine amounts and longitude: -0.03. Correlation between fine amounts and latitude: 0.03. This is interesting as the proxy for the wards found a negative correlation of 29% but the individual axis show nothing.

# SMART Questions On Dataset

Smart Question One : Is there a regressive relationship between the ward and the FINE\_AMOUNT?

I choose this smart question because it helps answer the overarching goal of this research. What are the determinants of fine amounts and can they accurately be accounted for? If they can be identified and accounted for can they be used to predict the fine amount. I will make a model to attempt to capture the spatial component of the data, with my group member trying to capture other aspects of the fine amount.

Smart Question Two: Is there a regressive relationship between FINE\_AMOUNT and other features within the data?

I chose this question because it would allow the data to be used to to predict the FINE\_AMOUNT. This would help identify which features are important in determining the amount fined for the violations that occurred. The model that will be used will test different features and see how well they perform and predict the fine amount.

# Models Ran on The Dataset

## Initial Random Forest Model: Predicting the Ward

A random forest was run to attempt to predict the resultant ward that the citation was issued in. This was basically an attempt to predict the regressive relationship, but fell short because the dataset was strictly dominated. Preventing us from capturing and predicting significant variation in the dataset. The geospatial data was included in the model after the presentation as it captures more variation in the dataset. VIOLATION\_CODE must be added as it directly correlates with FINE\_AMOUNT, but later models should break it down into its spatial component.

This was an attempt to solve my smart question : Is there a regressive relationship between the ward and the FINE\_AMOUNT

**Variables Used in Random Forest**

Dependent Variable:

* FINE\_AMOUNT: Represents the amount of fine in the citation.

Independent Variables:

* ISSUE\_TIME: The time when the violation was issued, encoded as an hour of the day starting from August 1st midnight.
* WARD: The ward where the violation occurred, encoded categorically as 0-7.
* VIOLATION\_CODE: The code for the violation, which correlates directly with the type of violation and encoded categorically. It directly relates to the fines so it should take up a majority of the explained variance
* LONGITUDE: The longitudinal coordinate of the location where the violation occurred.
* LATITUDE: The latitudinal coordinate of the location where the violation occurred.

Model Results:

Test set MSE: 204.17

Test set RMSE: 14.29

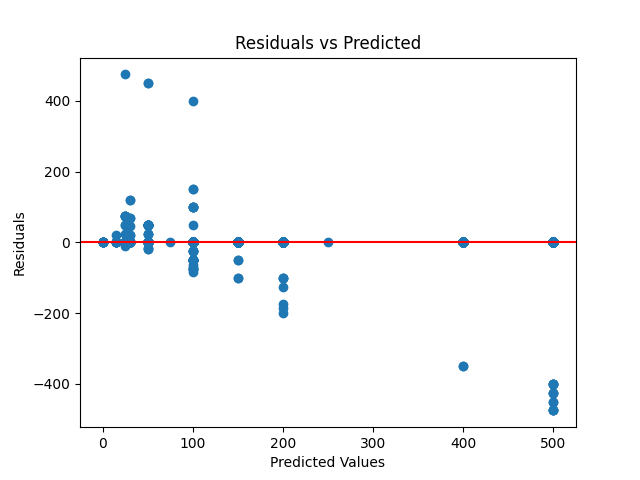
Test set MAE: 0.71

Test set R-squared: 0.92

This is not a bad result 92% of the fine amount variance is explained by those variables. Errors seem low but without another model to compare it with I have no idea it's a relative metric not an absolute one. MSE is 204 which with an average value of 117 means there is immense error in the dataset. There needs to be another iteration of the model that lowers MSE drastically if this model is to be formally applied/used by policy makers

With this model I can predict the fine amount with a high degree of accuracy. I need to break down the violation code into its geospatial components but this is a subject far beyond the aspects of this course.

Model Diagnostics:



**Figure 9 : Plot of the residuals**

The map of the residuals show there is a negative correlation present as they are not randomly distributed around zero. I do not know how to capture it. Maybe the interaction term between geography and ward would capture it but I simply needed to use a model not an iterated one.

Feature ranking:

1. feature VIOLATION\_CODE (0.825630812910029)

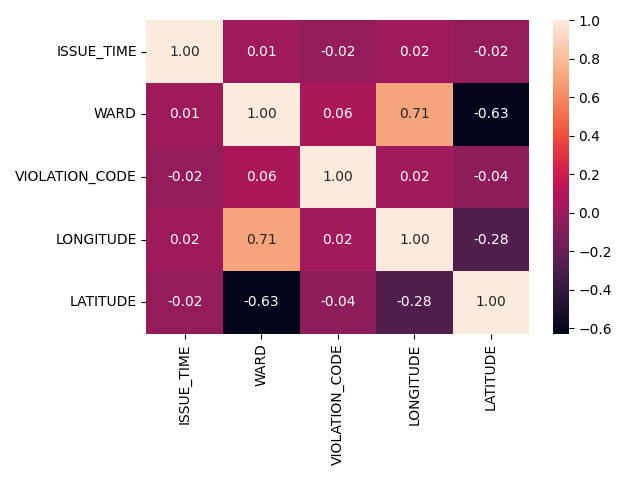
2. feature LATITUDE (0.08842850480110434)

3. feature LONGITUDE (0.0644694538140319)

4. feature WARD (0.02020711948941586)

5. feature ISSUE\_TIME (0.0012641089854187161)

It appears that the VIOLAITON\_CODE is an overwhelmingly important feature and this makes perfect sense as it directly relates to the FINE\_AMOUNT. The interaction between VIOLATION\_CODE and the other variables could help break down the variable into the components as there is interaction between it and the other variables I am sure. This is an exceedingly complex relationship.



**Figure 10 : Plot of the correlation matrix of the features**

A plot of the correlation matrix for the features showing the interaction term with ward and lat/long is recommended. It does not support the breakdown of the VIOLATION\_CODE into proxies that capture aspects of it. In the logistic regression we can make an attempt to add the interaction term and break VIOLATION\_CODE down to see if that helps.

## Initial Linear Regression: Predicting the Fine Amount

We were interested in observing how the fine amount issued was related and could be predicted within the dataset. A linear regression model was developed to observe the relationship of fine amounts alongside several key features we deemed important in the instances of the fine amount being issued. The key features that were seen as the most important after testing the linear regression model were, violation code, issuing agency name, issue time, issue date, x coordinates, and y coordinates. The coordinates were important in showing where the fine was given as well as the issuing agency in that area at the time.

The model performed well in predicting fine amounts based on the factors listed above. The mean squared error (MSE) was approximately 15.86 and the R squared value given was 0.994. Given these results the model shows high accuracy and able to explain the variance present within the data. The low MSE shows the average squared difference between the fine amounts and values predicted by the model. This indicates that the predictions actually are very close to the actual fine amounts within the data. The high R squared value indicates that the selected features were important in explaining the variability in the fine amounts. The value also indicates that the model fit the data well and can provide insight on how factors can influence the fine amounts. Removing any of these features caused the MSE and R squared to be drastically different and show very inaccurate scores.

However, there are limitations to this model as is the case with any model that could be fit to this data. There may be underlying factors that could also be affecting the fine amounts issued. Such factors, such as police presence in some wards of DC or crime rate within the wards, could influence how patrolled the areas and what types of violations are more readily booked. Further analysis of data that relates to DC and moving violations could prove important on how the model can be improved to predict fine amounts for violations.

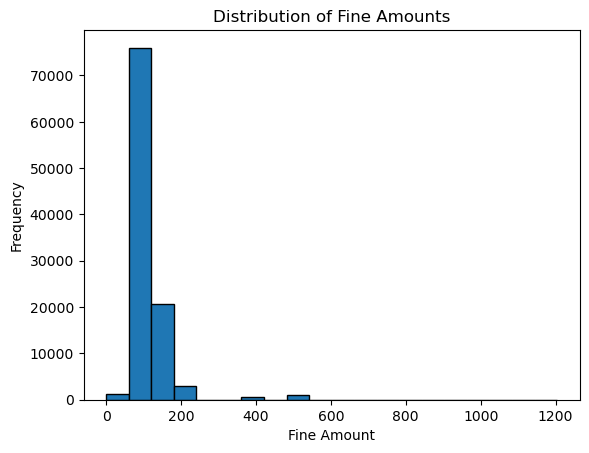
# Conclusions

The linear regression model developed to predict fine amounts based on key features within the dataset demonstrates robust performance and provides valuable insights into the factors influencing the issuance of fines. This was needed as we could not find the correlation between fine amounts and the amount being paid by violators. It was also easier to fit a model to the fine amount being predicted than to predict the types of violations occurring in different parts of the city. The fine amount was treated as a continuous variable to be predicted by certain key features. The inclusion of features such as violation code, issuing agency name, issue time, issue date, and geographical coordinates significantly enhances the model's predictive capability, as evidenced by the low Mean Squared Error (MSE) of approximately 15.86 and the high R-squared value of 0.994. These metrics underscore the model's high accuracy and its ability to explain the variance present in the fine amounts observed in the data. It is evident that these features play crucial roles in determining the fines imposed, highlighting the importance of considering various contextual factors in enforcement decisions. Overall, this model serves as a valuable tool for understanding and predicting fine amounts, offering opportunities for informed decision-making and policy development in traffic enforcement strategies.

The initial random forest regression is a good first attempt; the R Squared of .92 is fantastic. The lack of abstraction however in the VIOLATION\_CODE and its overwhelming presence in the feature rating is a cause for concern. It needs to be broken down to capture only the spatial aspect of the data so the research question: Is there a regressive relationship between the ward and the FINE\_AMOUNT. It was identified with the spatial triple plot earlier in the EDA of the data as there is a clear negative relationship between the ward and the FINE\_AMOUNT. This helps augment the logistic regression to paint a clearer picture of what aspects of the dataset can truly be used to accurately predict the fine amount.

This work shows that it is possible to predict the fine amount for DC traffic violations using the data present in the citations. Variables strictly dominating their columns, the presence of strong variables due to direct correlation between independent and dependent variables, and a failure to capture higher level effects such as the interactions between variables shows we still have a long way to go in predicting the fine amount. Additional data, such as criminal statistics, being included in the dataset may increase statistical power and help remove multicollinearity from the dataset. This represents a good first attempt but for further exploration of such a complex issue more detail is required.

# Supplemental Figures



**Supplemental Figure 1.** Distribution of fine amounts seem to center mainly around $100-$150.

# 

# Works Cited

1. <https://advisement.com/traffic-ticket-statistics/#:~:text=The%20average%20cost%20per%20traffic%20ticket%20is%20around,however%2C%20this%20amount%20can%20often%20exceed%20this%20average>.
2. District of Columbia. (2023). Moving Violations Issued in January 2023 (Version 1.0) [Data set]. Open Data DC.<https://opendata.dc.gov/datasets/0cbd79603f3e46bdaa9d8d2718d837e5_7/explore>