Analysis of Parking Violations in Los Angeles (2021–2025)

0.1 Introduction

Parking enforcement plays a critical role in managing urban traffic flow and generating municipal revenue. The City of Los Angeles provides open access to detailed parking citation records through its Open Data Portal. This analysis uses the LA Parking Citation dataset from 2021 to 2025. The dataset contains millions of parking citation records and detailed information such as the date and time of issuance, type of violation, amount of fine, make of the vehicle, and geographical location of the violation represented by latitude and longitude coordinates.

The research objective of this study is to explore structural patterns in parking violations over time. Specifically, this project aims to address the following research questions:

1. Are certain vehicle makes or body styles more likely to be linked to certain types of violations?
2. How has the number of parking violations changed from 2021 to 2025?
3. Do parking violations tend to cluster in certain geographic regions of the City of Los Angeles?
4. To what extent can parking violation types be predicted using temporal, vehicle, and spatial features?

Through the exploration of these research questions, this research seeks to help improve urban traffic management and give a better understanding of driver behavior in dense metropolitan areas.

0.2 Methods

0.2.1 Data Preparation and Cleaning

The data are obtained from the City of Los Angeles Open Data Portal via the Socrata API. Key variables are selected based on relevance to the research questions. Subsequently, records

with missing values are removed using listwise deletion to ensure data consistency. In addition, a series of basic plausibility checks are performed on the variables. Dates are constructed from year, month, and day fields, and the hour variable is verified to fall within the 24-hour range. Geographic coordinates are filtered using predefined latitude and longitude bounds corresponding to the Los Angeles region to remove implausible observations.

0.2.2 Exploratory Data Analysis Methods

Exploratory data analysis is conducted to summarize key patterns in the dataset. Histograms and boxplots are used to examine the distribution of fine amounts and identify potential outliers using the IQR method. Temporal patterns are analyzed by aggregating citation counts at the year–month level to construct a time series, and by hour of the day to examine daily variation in enforcement activity.

0.2.3 Statistical Analysis

A chi-square test of independence is conducted to evaluate the relationship between vehicle body type and violation type. A contingency table is constructed using observed frequencies across categories. The chi-square statistic and p-value are used to assess statistical significance. To quantify the strength of the association, Cramer's V is calculated as an effective size measure, providing an independent scale assessment of the practical relationship between the variables.

0.2.4 Geographic Analysis

Spatial patterns in parking violations are analyzed using latitude and longitude coordinates. Geographic data are separated into grid-based spatial bins, and citation counts are aggregated within each grid cell. Spatial visualization techniques are applied to identify areas with high concentrations of violations and to assess potential clustering in urban regions.

0.2.5 Predictive Modeling and Model Evaluation

To address the predictive research question, classification models are developed to predict violation types based on temporal, vehicle, and spatial features. A Decision Tree classifier is used as a baseline model, providing interpretability through recursive partitioning. An XGBoost classifier is then applied to build an ensemble of decision trees using gradient boosting to improve predictive performance.

The dataset is split into training, validation, and test sets using stratified sampling (60% training, 20% validation, 20% testing). The validation set is used for hyperparameter tuning. For the Decision Tree, tuned hyperparameters include maximum tree depth, minimum samples

required for a split, minimum samples per leaf, and the cost-complexity pruning parameter. For XGBoost, tuned hyperparameters include learning rate, maximum tree depth, subsampling ratio, and the number of boosting trees.

Model performance is evaluated on the test set using accuracy and macro-averaged F1-score. These metrics assess overall predictive accuracy and balanced performance across all violation categories.

0.3 Result

0.3.1 Data Summary

The initial dataset retrieved from the City of Los Angeles Open Data Portal contains 9,234,940 citation records and 23 variables, with a total memory usage of about 1.6 GB. After selecting relevant variables, the dataset is reduced to 13 variables and approximately 819 MB, resulting in a nearly 50% reduction in memory usage. To ensure data integrity, observations containing missing values are removed using listwise deletion. The dataset is reduced to 8,605,218 complete observations, corresponding to a retention rate of 93.18%. Although a small proportion of data is removed, the resulting dataset remains sufficiently large for analysis. The average fine amount is \$74.04, with values ranging from \$0 to \$1,100. The mean latitude and longitude are 34.06 and -118.33, respectively, indicating that most citations are geographically concentrated within the Los Angeles region. Fine amount distributions are examined to understand the overall structure of penalty values. The distribution is slightly right-skewed, with most observations concentrated within a narrow range.

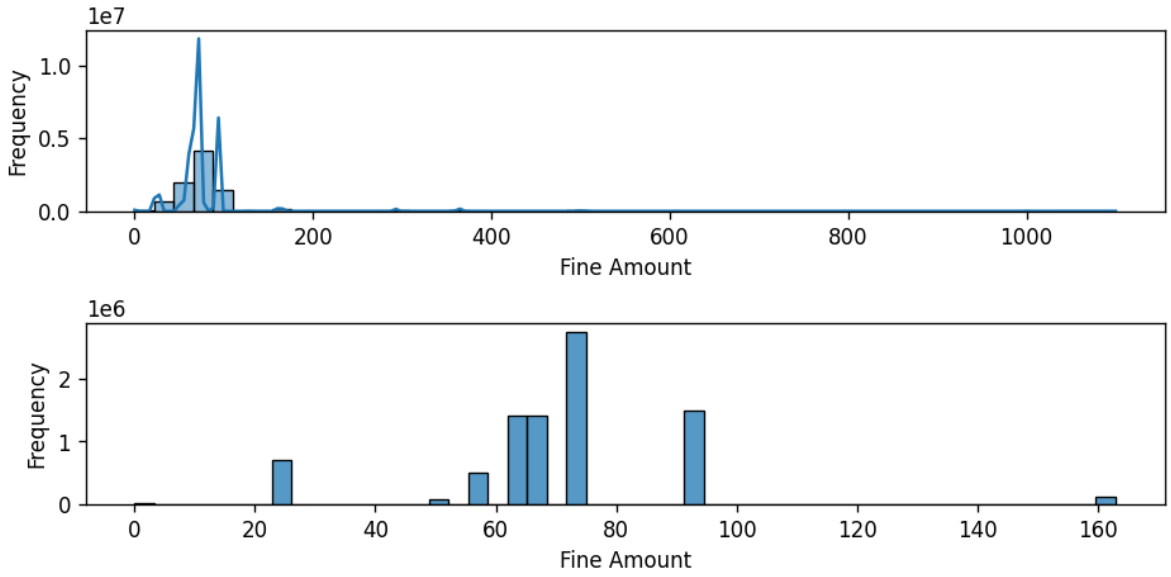

Figure 1: Fine amount distributions It shows both the overall distribution and the zoomed distribution below $200. The dominant concentration is between $60 and $80.

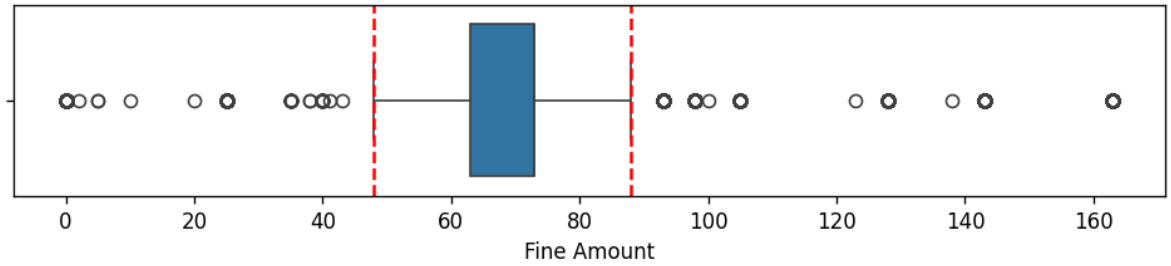

Figure 2: Boxplot of Fine Amount (<=200) with IQR Bounds It illustrates the interquartile range and the presence of higher value observation

0.3.2 RQ1: Are certain vehicle makes or body styles more likely to be linked to certain types of violations?

Vehicle characteristics are statistically associated with violation types, but the strength of this relationship is weak in practice. To assess this relationship, chi-square tests are conducted. The results indicate statistically significant associations for both body type (Chi-Square = 105568.83, $p < 0.001$) and vehicle make (Chi-Square = 208967.46, $p < 0.001$). However, the effect sizes are small, with Cramer's V equal to 0.064 for body type and 0.070 for vehicle make. This indicates weak practical associations. From a descriptive perspective, citation

activity is highly concentrated in a small number of violation categories. "NO PARK/STREET CLEAN" accounts for 28% of citations, followed by "METER EXP." (15%) and "RED ZONE" (12%), indicating that enforcement activity is focused on a limited set of recurring violations. Vehicle characteristics also exhibit strong structural patterns. Passenger cars account for the vast majority of citations, while the rest of each vehicle type represents smaller proportions. Among vehicle makes, Toyota, Honda, and Ford are the most frequently cited. It shows their prevalence in urban traffic. Overall, these results suggest that although differences in violation distributions exist across vehicle groups, vehicle characteristics are not strong predictors of violation types.

0.3.3 RQ2: How has the number of parking violations changed from 2021 to 2025?

Parking citation fluctuates substantially from 2021 to 2025. It suggests a seasonal pattern, with higher citation volumes tending to occur around the beginning of the year. The most noticeable peak occurs in early 2022, followed by repeated rises and declines in later years.

Monthly citation counts generally range between approximately 120,000 and 180,000 observations. Citation volumes remain relatively stable in 2021, followed by a noticeable increase in early 2022. After this peak, citation activity declines through late 2022 and reaches a local minimum in early 2023. From 2023 to 2025, citation counts continue to exhibit substantial variability, with recurring increases and decreases over time. Overall, these patterns suggest cyclical variation and may reflect seasonal effects rather than a sustained upward or downward trend.

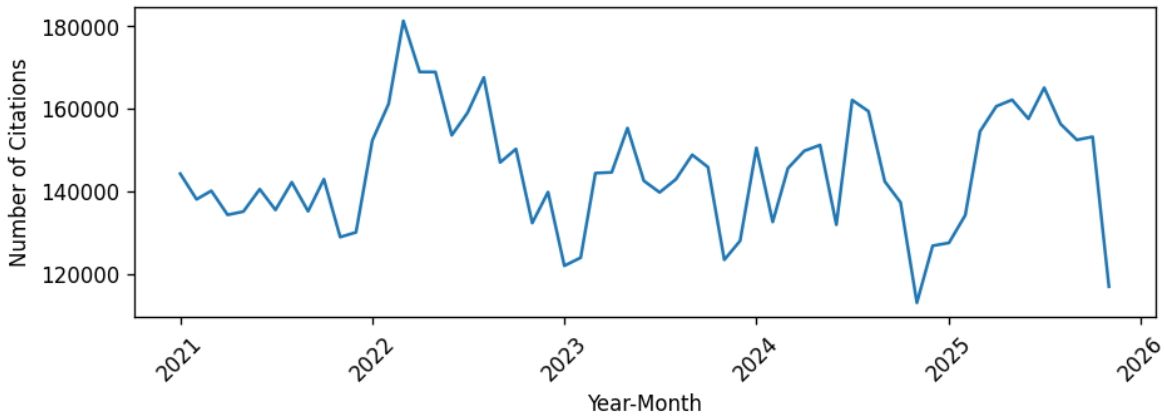

Figure 3: Monthly parking citation counts (2021–2025) The series exhibits substantial monthly variation, with a peak in early 2022.

0.3.4 RQ3: Do parking violations tend to cluster in certain geographic regions of the City of Los Angeles?

Parking violations are not uniformly distributed across the city and show clear spatial clustering in specific regions. High density clusters are observed in central urban areas, particularly around downtown Los Angeles. Additional localized clusters are visible in areas such as West Hollywood and Santa Monica, indicating concentrated enforcement activity beyond the city center. In contrast, areas farther from the center have much lower citation densities. This distribution suggests that parking violations are concentrated in high density urban and commercial regions rather than being randomly distributed across the city. Overall, the results indicate that parking violations tend to cluster geographically. It might reflects variations in traffic density, parking demand, and enforcement intensity across different parts of the city.

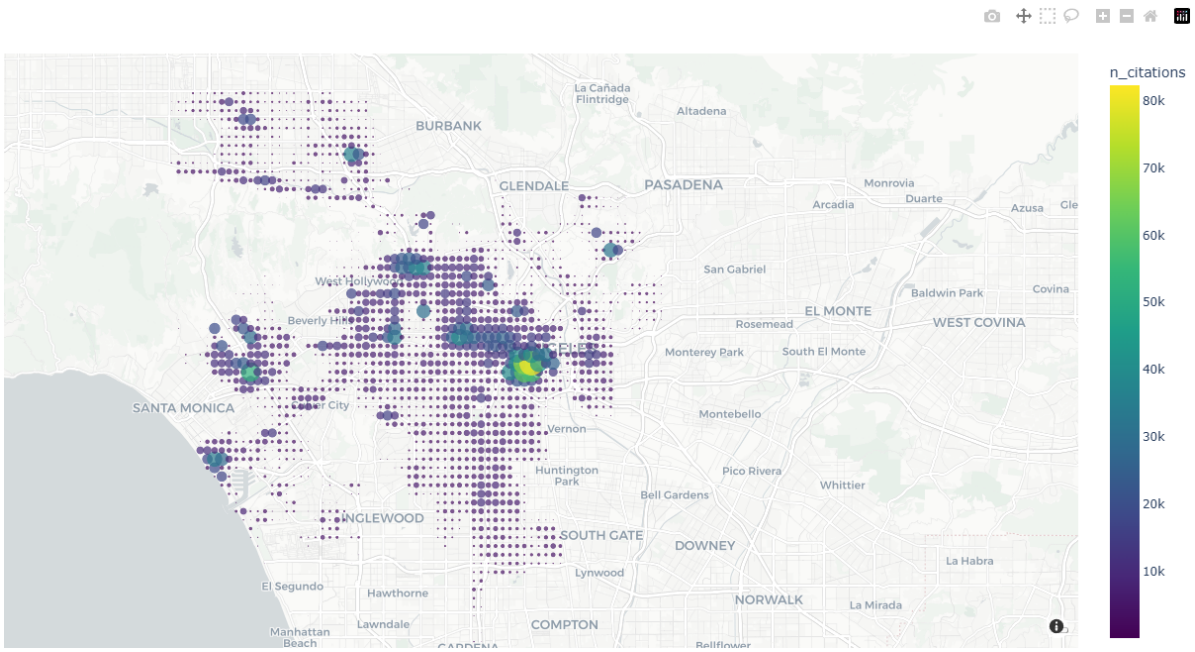

Figure 4: Gried-Based spatial concentration of Parking Citations in LA Core Area It presents a grid-based spatial distribution of citation counts across the Los Angeles core area.

0.3.5 RQ4: To what extent can parking violation types be predicted using temporal, vehicle, and spatial features?

Parking violation types can be predicted with moderate accuracy using temporal, vehicle, and spatial features. The pruned Decision Tree achieves a test accuracy of 0.549 and a macro-averaged F1-score of 0.457. In contrast, the final XGBoost classifier improves performance

to a test accuracy of 0.643 and a macro-averaged F1-score of 0.570. This indicates that the ensemble model captures stronger predictive patterns than the single-tree baseline.

The classification results also show that predictive performance varies substantially across violation categories. XGBoost performs best for "NO PARK/STREET CLEAN," with an F1-score of 0.83, and "PREFERENTIAL PARKING," with an F1-score of 0.76. It also performs reasonably well for "METER EXP.", with an F1-score of 0.68. In contrast, performance is much weaker for "DISPLAY OF PLATES," with an F1-score of 0.12. This indicates that this category is more difficult to distinguish using the available features. The confusion matrix further shows that some categories are frequently confused with "OTHER," especially lower-performing classes such as "DISPLAY OF PLATES" and "RED ZONE." This suggests that the available temporal, spatial, and vehicle related variables contain useful predictive information, but they do not fully separate all violation types. Feature importance from the XGBoost model indicates that hour is the most important predictor, followed by California plate status, body group, latitude, longitude, and distance from downtown. This suggests that temporal and spatial variables are especially important for predicting violation group, while vehicle characteristics also contribute to the classification task. Overall, the results show that parking violation types are predictable to a moderate extent. XGBoost improves substantially over the Decision Tree baseline, but the moderate macro-F1 score indicates that prediction remains challenging for less distinctive or more heterogeneous violation categories.

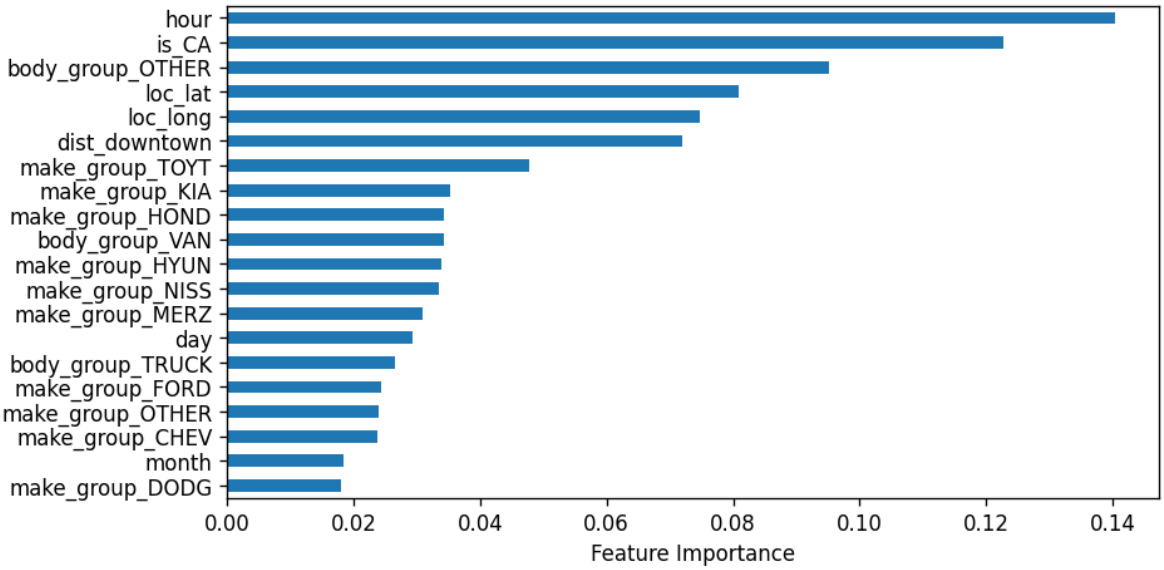

Figure 5: Top 20 feature importances from the XGBoost model Hour and location-related variables (latitude, longitude, and distance to downtown) are the most influential predictors, while vehicle-related features contribute less.

0.4 Conclusion and summary

This study analyzes parking citation data in Los Angeles from 2021 to 2025 and identifies clear structural patterns across time, space, and violation types. Citation activity follows both seasonal cycles. A small number of violation categories account for a large proportion of citations, and violations are geographically concentrated in dense urban areas such as downtown Los Angeles. Besides, the predictive modeling shows that violation types can be predicted with moderate accuracy using temporal, spatial, and vehicle features. The XGBoost model outperforms the Decision Tree, indicating the importance of nonlinear relationships. Overall, parking violations are not randomly distributed but reflect underlying urban dynamics, including traffic patterns, parking demand, and enforcement intensity. Several limitations should be noted. First, the data reflect enforcement activity rather than true violation behavior, which may introduce bias. Second, removing missing values may affect data representativeness. Third, the model only uses basic features and does not include external factors such as traffic or policy changes, which may limit predictive performance. Future work could incorporate additional contextual variables and more advanced modeling approaches to improve both prediction and interpretation.