FINAL REPORT

INTRODUCTION TO DATA MINING

ANALYZING URBAN TRAFFIC DYNAMICS IN NEW YORK CITY

ADAM SATTOUT | ALPER MERT
2220765061 | 2200765029

Introduction

Traffic congestion is a persistent challenge in modern cities, affecting productivity, sustainability, and quality of life. In New York City, these impacts are amplified by high density and strong variation in travel demand across time and location. In this project, we study historical NYC traffic volume measurements across many road segments and apply data mining and time-series analysis to extract interpretable patterns, including rush-hour behavior, weekday–weekend differences, longer-term trends, and similarities between locations through clustering, as well as anomaly-style shock detection.

Although the dataset is historical, the course requires a streaming component. We found that predicting volume from a very small recent window is not reliable without additional context, so streaming is treated mainly as a simulation constraint rather than the core method. We simulate real-time arrival using a Kafka producer/consumer pipeline, while the main conclusions are drawn from offline analysis of the accumulated data.

Data

This project uses the Automated Traffic Volume Counts dataset published by the New York City Department of Transportation (NYC DOT)**.** The dataset consists of around 2 million traffic volume measurements spanning 2000-2025 collected by automated counters across New York City, and each record represents an observation tied to a specific location (road segment and direction) at a particular time.

The raw data contains both temporal and spatial/road attributes. In our working dataset, the main fields include identifiers such as RequestID and SegmentID, a borough label Boro, and time components Yr, M, D, HH, and MM that jointly specify when the measurement was taken. The observed traffic volume is stored in Vol. The road segment's geometry is provided in WktGeom, while additional descriptive fields such as street, fromSt, toSt, and Direction help contextualize the measurement location and travel direction.

Method

Trend Analysis

Trend analysis was performed as a preliminary step to understand the temporal structure of traffic volume data. The objective of this analysis was to identify recurring patterns and long-term behaviors before applying segmentation or predictive modeling techniques.

Visualization techniques were used to detect patterns of the dataset.

Shock Events

We defined shock events as observations where traffic volume spikes abnormally compared to what is typically expected under similar conditions. We first constructed a proper timestamp from the dataset's separate date and time fields, then used historical data to form a simple baseline for "normal" volume by grouping records by comparable context (primarily location/segment and time-related fields).

After assigning each record an expected baseline value, we measured how far the observed volume deviated from that baseline using a difference-style anomaly score. We then ranked observations by this deviation to surface the largest spikes and aggregated the flagged spikes by day to identify dates that concentrate unusually high activity. Finally, we inspected the top shock-heavy days and used them as candidates for qualitative interpretation (for example, checking whether anomalies plausibly align with major disruptions such as unusual city conditions), while keeping the analysis descriptive rather than causal.

Clustering

Following the trend analysis, we focused on profiling individual road segments based on their temporal traffic behavior. Trend analysis showed us that the distribution is not homogeneous. So, rather than treating each observation independently, we aggregated traffic data at the segment level to extract representative behavioral features.

For each road segment, we computed several temporal features, including mean traffic volume, standard deviation of traffic volume, morning peak ratio, evening peak ratio, and weekend ratio. These features were designed to capture not only how much traffic a segment carries, but also when that traffic occurs. For example if there is lots of traffic in the mornings (because of work-school etc.) the morning ratio is high.

Using those features, we applied K-means clustering to group road segments with similar temporal usage patterns. We chose K-means due to its simplicity, interpretability and high performance.

We scaled the features with the standard scaler to prevent scale dominance.

Predictive Modelling

Predictive modeling was conducted to evaluate whether engineered temporal and segment-level features could be used to estimate traffic volume. Traffic volume was treated as a continuous target variable, and the task was formulated as a regression problem rather than a classification task (not discrete). So we used RandomForestRegressor as our main model and HistGradientBoostingRegressor to compare. Also, RandomForest is a suitable model to see feature importances.

Since the data has a lot of features and some of them are not informative, we used a subset of the features. The input features included temporal attributes (hour of day, day of week),

aggregated segment-level behavioral features (mean traffic volume, variability, temporal ratios), and additional engineered features capturing periodic traffic behavior.

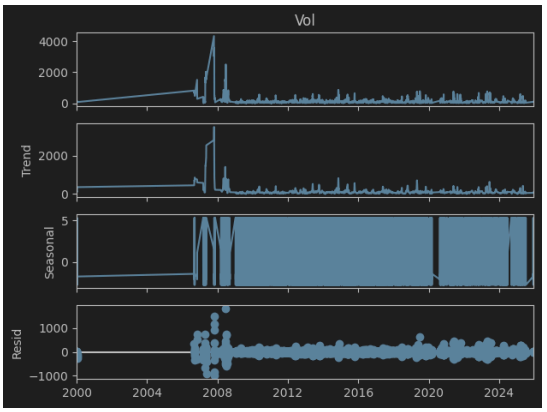
We used multiple regression models to compare predictive performance across different modeling approaches. Baseline models relied on only temporal features. Enhanced models were used with larger feature-set(behavioral features included).

We splitted the feature-sets as 0.8 train and 0.2 test. We avoided data leakage. To evaluate the results, we used Mean Absolute Error (MAE) and Coefficient of Determination (R^2), providing complementary measures of prediction accuracy and explanatory power.

Results

Trend Analysis

The trend analysis revealed strong and consistent temporal patterns in traffic volume across multiple time scales. Seasonal decomposition showed that the long-term trend
components remained relatively stable throughout the observation period, indicating sustained traffic demand in the urban environment.
The seasonal component exhibited clear and recurring fluctuations, confirming that traffic volume follows periodic patterns rather than random variation. Residual components were relatively small, suggesting that the majority of traffic variability can be explained by systematic temporal effects.



More of the plots and explanations were included in the .ipynb file.

Shock Events

date	
2021-05-26	74
2021-05-27	73
2021-05-25	73
2019-06-30	72
2019-04-27	42
2019-05-04	40
2019-04-28	37
2019-05-05	35
2019-04-29	33
2019-05-06	33

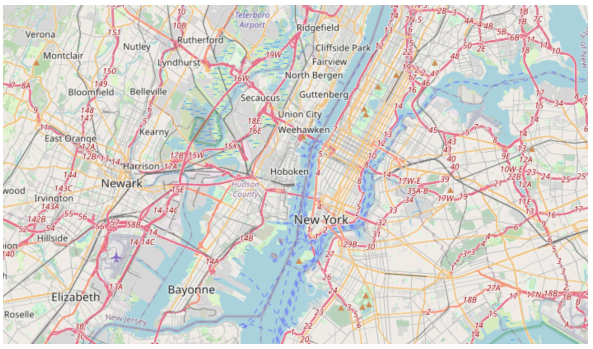
When we aggregated detected shock records by calendar date, the three most shock-heavy days were tightly clustered around May 25–27, 2021. The fact that the peak days are

consecutive suggests a citywide disruption rather than isolated sensor noise or a single-location incident. Cross-checking these dates against public event timelines, this cluster aligns with the one-year anniversary of George Floyd's death, when NYC saw rallies and marches that included road crossings and temporary traffic disruption (e.g., large gatherings in Brooklyn and marches over major crossings).

Beyond the anniversary cluster, another notable shock day in the top-ranked list was 2019-04-27. This date matches NYC DOT's Car-Free Earth Day (often described as "Car-Free Day"), where selected corridors were closed to vehicle traffic for several hours, which can plausibly create abrupt deviations from typical volume patterns near the affected routes.

Clustering

We used Silhouette Score to evaluate the success of clusterings. We tried different k values. Actually, when k=3, Silhouette Score was much better (0.52), but it was not very informative. However, when k=4, Silhouette Score was a bit low (0.31), but it was more informative and meaningful behavioral information. So we chose k as 4. The 4 clusters represent "Low-Traffic Weekday Local Roads", "High-Volume Major Arterials", "Morning Commute-Dominated Roads" and "Weekend-Oriented Leisure Roads". We then embedded the clusters into a map and visualized it.



These results demonstrate that clustering based on temporal features can effectively uncover meaningful road usage patterns beyond simple geographic classification.

Predictive Modelling

Model	Feature Type	Mean Absolute Error	R^2
RandomForestRegressor	Basic Feature-Set	23.99	0.84
RandomForestRegressor	Advanced Feature-Set	71.03	0.51
HistGradientBoostingRegressor	Advanced Feature-Set	71.04	0.51

We got the best score with RandomForestRegressor with the basic feature set. As mentioned, data has a lot of features and some of them are not informative, so adding more features didn't increase the results, even decreased the success. Trying another model(HistGradient) did not solve the problem. We also did feature importance analysis. SegmentID and Hour were the most effective features.

Discussion

The results confirm that urban traffic volume exhibits strong temporal regularities driven by human mobility patterns. Trend analysis highlighted systematic daily and weekly behavior, validating the use of temporal features for further analysis. Segment profiling and clustering demonstrated that traffic behavior can be meaningfully categorized beyond geographic location. The observed spatial coherence of clusters suggests that behavioral similarity often aligns with urban structure. Predictive modeling results indicate that temporal behavioral features enhance traffic prediction performance. However, variability across segments suggests that contextual factors may further improve accuracy.

Future Works

A natural next step is to enrich the problem setting with a dataset that contains more explanatory features beyond raw traffic counts. In particular, incorporating weather variables (rain, temperature, snow), holiday calendars, scheduled city events, and road work/closure information would make the analysis more "context-aware" and reduce ambiguity when interpreting shocks or modeling changes in volume. With such features, the predictive component could move from pattern extrapolation toward a more realistic forecasting setup that explains *why* volume changes, not only *how* it changes.

Methodologically, future work could improve anomaly detection by using robust baselines (median-based statistics, seasonal baselines per hour/day) and by separating "positive spikes" from "sudden drops," since disruptions can cause both. Clustering could be extended with richer representations such as weekly profiles per segment, dynamic time warping for shape-based similarity, or spatially constrained clustering that accounts for road connectivity and proximity. Finally, the predictive part could be expanded by comparing multiple forecasting families (classical seasonal models, gradient boosting with engineered temporal/exogenous features, and sequence models) and evaluating them under realistic horizons and deployment constraints.

Conclusion

This project examined historical traffic volume measurements from New York City to better understand how vehicle flow changes across time and location. Using a combination of data cleaning, exploratory analysis, and data mining methods, we showed that traffic volume is highly structured, with observable patterns and clear differences across road segments and boroughs. The shock-event analysis highlighted dates where traffic behavior deviated

strongly from typical conditions, and the top-ranked anomalies aligned with major citywide disruptions, Trend-focused analysis further supported that traffic dynamics are dominated by repeatable seasonality and longer-term variation, while clustering helped group segments that share similar behavior, suggesting that location-specific traffic profiles can be summarized into a smaller number of representative patterns. Finally, the predictive component served as a practical test of how much future volume can be inferred from historical structure, reinforcing that forecasting is most reliable when it leverages longer context rather than a short recent window alone.

References

[1] New York City Open Data. “Automated Traffic Volume Counts.” Dataset (Socrata platform). Accessed 2025.

[2] Socrata. “API Endpoints (SODA API) Documentation.” Developer documentation. Accessed 2025.

[3] Apache Software Foundation. “Apache Kafka Documentation.” Product documentation. Accessed 2025.

[4] Kreps, J., Narkhede, N., and Rao, J. “Kafka: A Distributed Messaging System for Log Processing.” (White paper / technical report), 2011.

[5] dpkp (community project). “kafka-python: Python client for Apache Kafka.” Software documentation / repository. Accessed 2025.

[6] Hunter, J. D. “Matplotlib: A 2D Graphics Environment.” Computing in Science & Engineering, 2007.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. “Scikit-learn: Machine Learning in Python.” Journal of Machine Learning Research, 2011.

[8] Waskom, M. L. “seaborn: statistical data visualization.” Journal of Open Source Software, 2021.

[9] python-visualization contributors. “Folium: Python Data. Leaflet.js Maps.” Project documentation / repository. Accessed 2025. GitHub

[10] New York City Department of Transportation (NYC DOT). “Car-Free Earth Day (NYC).” Official announcement / press materials, 2019.

[11] Gothamist. Coverage of New York City demonstrations on the one-year anniversary of George Floyd’s death, 2021.

[12] ABC7 New York (WABC-TV). Coverage of New York City protests/marches on the one-year anniversary of George Floyd’s death, 2021.