
Spatiotemporal Analytics and Data-Driven Pattern Mining in Urban Complaints

Rajat Rayaraddi

Department of Computer Science
The George Washington University
Washington, D.C. 20052
rajat.rayaraddi@gwu.edu

Aishwarya Sajjan

Department of Computer Science
The George Washington University
Washington, D.C. 20052
aishwarya.sajjan@gwu.edu

Abstract

This study applies spatiotemporal data mining and machine learning techniques to analyze New York City’s 311 service request data. A 10-million-record subset was extracted from the full 42-million-row dataset and cleaned, standardized, and feature-engineered. Exploratory data analysis revealed key trends in complaint distribution across boroughs, time, and socioeconomic factors. Advanced data mining techniques including geospatial KMeans clustering, contrast and sequential pattern mining, association analysis, and anomaly detection uncovered localized hotspots, co-occurring complaint behaviors, and event-driven surges. Time-series forecasting and XGBoost-based regression were used to predict complaint volume and resolution time. The findings demonstrate how large-scale civic data can inform proactive urban service planning.

1 Introduction

Large metropolitan cities generate vast amounts of civic data through public service platforms such as New York City’s 311 system. These systems allow residents to report non-emergency issues related to noise, sanitation, housing, transportation, public safety, and infrastructure. While such datasets offer an unprecedented opportunity to understand urban service demand; their scale, complexity, and variability make meaningful analysis challenging. As a result, many city agencies continue to rely on descriptive reporting and reactive decision-making, limiting their ability to anticipate problems and allocate resources efficiently.

Data mining and machine learning provide powerful tools for extracting actionable insights from large-scale civic datasets. Spatial analytics can reveal localized hotspots of recurring issues, temporal analysis can uncover daily and seasonal patterns, and pattern mining techniques can identify relationships between different types of complaints. Predictive models can estimate service resolution times and forecast future demand, enabling proactive planning rather than retrospective response.

In this work, we performed an end-to-end spatial-temporal analysis of New York City’s 311 service request data [1]. Using a subset of ten million records spanning November 2022 to November 2025, we apply data preprocessing, exploratory data analysis, and data mining techniques to uncover complaint patterns across geography, time, and demographics. Our approach integrates clustering, association and sequential pattern mining, anomaly detection, time-series forecasting, and supervised learning to model complaint behavior and resolution dynamics. The purpose of this study is to demonstrate how data-driven methodologies can support more efficient urban service management and inform policy-level decision-making in large cities.

Unique Key	Created Date	Closed Date	Agency	Agency Name	Complaint Type	Descriptor	Location Type	Incident Zip	Incident Address	Street Name	Cross Street 1	Cross Street 2
66778147	11/10/2025 01:50:51 AM	NaN	NYPD	New York City Police Department	Illegal Parking	Commercial Overnight Parking	Street/Sidewalk	10465.0	3286 RADIO DRIVE	RADIO DRIVE	GRISWOLD AVENUE	BEND
66780206	11/10/2025 01:49:45 AM	NaN	NYPD	New York City Police Department	Noise - Vehicle	Engine Idling	Street/Sidewalk	10025.0	2680 BROADWAY	BROADWAY	WEST 102 STREET	WEST 103 STREET
66777208	11/10/2025 01:47:32 AM	NaN	NYPD	New York City Police Department	Noise - Street/Sidewalk	Loud Music/Party	Street/Sidewalk	10040.0	157 NAGLE AVENUE	NAGLE AVENUE	ARDEN STREET	THAYER STREET
66779171	11/10/2025 01:47:12 AM	NaN	NYPD	New York City Police Department	Illegal Parking	Posted Parking Sign Violation	Street/Sidewalk	11204.0	1358 DAHILL ROAD	DAHILL ROAD	60 STREET	61 STREET
66779189	11/10/2025 01:47:09 AM	NaN	NYPD	New York City Police Department	Noise - Residential	Banging/Pounding	Residential Building/House	11221.0	1108 GATES AVENUE	GATES AVENUE	BROADWAY	BUSHWICK AVENUE

Figure 1: Example records from the NYC 311 service request dataset in its original, unprocessed form.

City	Landmark	Facility Type	Status	Due Date	Resolution Description	Resolution Action Updated Date	Community Board	BBL	Borough	Coordinate (State Plane)	Coordinate (State Plane)	Open Data Channel Type	Park Facility Name
BRONX	RADIO DRIVE	NaN	In Progress	NaN	NaN	NaN	10 BRONX	2.054140e+09	BRONX	1,034,564	246,662	ONLINE	Unspecified
NEW YORK	BROADWAY	NaN	In Progress	NaN	NaN	NaN	07 MANHATTAN	1.018740e+09	MANHATTAN	992,863	230,234	ONLINE	Unspecified
NEW YORK	NAGLE AVENUE	NaN	In Progress	NaN	NaN	NaN	12 MANHATTAN	1.021730e+09	MANHATTAN	1,004,751	253,034	MOBILE	Unspecified
BROOKLYN	DAHILL ROAD	NaN	In Progress	NaN	NaN	NaN	12 BROOKLYN	NaN	BROOKLYN	991,152	163,157	MOBILE	Unspecified
BROOKLYN	GATES AVENUE	NaN	In Progress	NaN	NaN	NaN	04 BROOKLYN	3.033390e+09	BROOKLYN	1,006,077	190,769	ONLINE	Unspecified

Figure 2: (Continued) Example records from the NYC 311 service request dataset in its original, unprocessed form.

2 Related Works

Previous research has explored the use of civic complaint data to understand urban dynamics and service efficiency. Studies analyzing 311 datasets have examined spatial distributions of complaints to identify neighborhood-level hotspots and service inequities, often linking complaint frequency to socioeconomic and demographic factors. Temporal analyses have further revealed strong daily and seasonal patterns in complaint behavior, particularly for noise, sanitation, and heating-related issues.

More recent work has applied machine learning techniques to predict service response times and forecast complaint volumes. Clustering methods have been used to group neighborhoods based on complaint composition, while association and sequential pattern mining have identified co-occurring and cascading complaint types. However, many existing studies focus on limited subsets of data or isolated techniques. This work extends previous research by integrating large-scale spatial, temporal, and predictive analytics within a unified end-to-end framework.

3 Methodology

This study follows an end-to-end data mining workflow encompassing data reduction, preprocessing, exploratory analysis, advanced pattern mining, and predictive modeling. The methodology is designed to handle large-scale civic data while extracting meaningful spatial, temporal, and behavioral insights.

3.1 Data Selection and Reduction

The original NYC 311 Service Request dataset contains over 42 million records spanning from 2010 to 2025. To ensure computational feasibility while preserving recent trends, the dataset was reduced to approximately 10 million records covering the period from November 24, 2022 to November 9, 2025. This subset captures complaint behavior and seasonal variations.

unique_key	created_date	closed_date	year	month	dayofweek	hour	agency	complaint_type	descriptor	borough	location_type	city
66778147	2025-11-10 01:50:51	NaN	2025	11	0	1	NYPD	Illegal Parking	Commercial Overnight Parking	BRONX	Street/Sidewalk	BRONX
66780206	2025-11-10 01:49:45	NaN	2025	11	0	1	NYPD	Noise - Vehicle	Engine Idling	MANHATTAN	Street/Sidewalk	NEW YORK
66777208	2025-11-10 01:47:32	NaN	2025	11	0	1	NYPD	Noise - Street/Sidewalk	Loud Music/Party	MANHATTAN	Street/Sidewalk	NEW YORK
66779171	2025-11-10 01:47:12	NaN	2025	11	0	1	NYPD	Illegal Parking	Posted Parking Sign Violation	BROOKLYN	Street/Sidewalk	BROOKLYN
66779189	2025-11-10 01:47:09	NaN	2025	11	0	1	NYPD	Noise - Residential	Banging/Pounding	BROOKLYN	Residential Building/House	BROOKLYN

Figure 3: Sample view of the NYC 311 service request data after cleaning and feature engineering.

incident_address	street_name	incident_zip	latitude	longitude	days_to_close	status	location
3286 RADIO DRIVE	RADIO DRIVE	10465	40.843562	-73.818153	NaN	In Progress	(40.84356202395745, -73.81815325101906)
2680 BROADWAY	BROADWAY	10025	40.798611	-73.968892	NaN	In Progress	(40.7986109485515, -73.96889161417334)
157 NAGLE AVENUE	NAGLE AVENUE	10040	40.861171	-73.925885	NaN	In Progress	(40.86117080770446, -73.92588494147341)
1358 DAHILL ROAD	DAHILL ROAD	11204	40.614502	-73.975140	NaN	In Progress	(40.61450200215088, -73.97514014922876)
1108 GATES AVENUE	GATES AVENUE	11221	40.690267	-73.921293	NaN	In Progress	(40.690266694713586, -73.92129345228891)

Figure 4: (Continued) Sample view of the NYC 311 service request data after cleaning and feature engineering.

3.2 Data Cleaning and Preprocessing

Data preprocessing was performed using Python libraries including pandas, NumPy, and scikit-learn. Column names were standardized by converting to lowercase and replacing spaces with underscores. Date-time fields were converted to proper timestamp formats, and new temporal features like year, month, day of week, and hour were extracted. ZIP codes were standardized as strings, and latitude and longitude fields were converted to numeric values.

A new feature, "days-to-close", was computed for complaints with valid closed dates to support resolution-time analysis. Approximately 20 low-information columns were removed to reduce dimensionality. Categorical missing values were imputed with the label "Unknown", while records missing important spatial information (latitude, longitude, or ZIP code) were dropped, representing less than 2 percent of the data. Categorical variables were label-encoded, and numerical features were scaled to support clustering and machine learning models.

3.3 Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted to understand complaint distributions across geography, time, and categories. This included analyzing top complaint types, borough-level complaint volume and per-capita rates, temporal trends by month and hour, and resolution time distributions. Correlation analysis was performed to examine relationships between complaint volume, median household income, and education levels.

3.4 Data Mining and Pattern Discovery

Several data mining techniques were applied to uncover hidden structures within the data. Geospatial KMeans clustering was used to group ZIP codes based on complaint-type frequency, with Principal Component Analysis (PCA) applied for dimensionality reduction and visualization. Contrast pattern mining using chi-square statistics and standardized residuals identified complaint types disproportionately represented across boroughs.

3.4.1 Co-Occurrence Patterns

To analyze co-occurring and sequential complaint behavior, a spatially constrained sequence mining approach was employed. Latitude and longitude values were rounded to four decimal places to

approximate city-block-level locations (approximately 100 meters), and a composite location-id was constructed by combining the rounded coordinates. Records were sorted by time to preserve temporal ordering. For each location, sequences of complaint types were generated to capture localized complaint evolution. Complaint pairs were treated as unordered to capture co-occurrence rather than directional dependence, and frequency counts were computed using a counting-based approach.

3.4.2 Anomaly Detection

To identify abnormal surges in complaint activity, an anomaly detection framework based on rolling statistical baselines was applied to daily complaint counts. Complaint creation timestamps were first standardized to datetime format and aggregated at a daily resolution to construct a univariate time series representing overall complaint volume. A rolling window of seven days was used to compute moving averages and standard deviations, capturing short-term temporal trends while smoothing day-to-day noise. An anomaly threshold was defined as the rolling mean plus 1.4 times the rolling standard deviation, and days exceeding this threshold were flagged as burst events.

$$\text{Threshold} = \mu_{\text{rolling}} + 1.4 \cdot \sigma_{\text{rolling}}$$

To detect category-specific anomalies, daily complaint counts were further computed separately for each complaint type. Rolling baselines were calculated independently per category to account for varying baseline frequencies, and standardized z-scores were used to identify significant deviations. This approach enabled the detection of both global and complaint-type-specific spikes while preserving interpretability and robustness to seasonal variation.

3.5 Predictive Modeling

Time-series forecasting was conducted using the Prophet library to model long-term trends and seasonal effects in complaint volume, both overall and for specific complaint categories.

For the prediction of complaint resolution time, the target variable days-to-close was modeled using a supervised regression framework. Only complaints with a valid closure date were retained, and extreme outliers with resolution times exceeding 100 days were excluded to reduce skewness and improve model stability. The feature set consisted of engineered temporal attributes (year, month, day of week, and hour), spatial coordinates (latitude and longitude), and selected categorical attributes, while identifier fields, raw datetime columns, and high-cardinality text fields were removed. The dataset was split into training and testing subsets using an 80/20 split. Numerical features were standardized using z-score normalization to ensure comparable feature scales across models. Multiple tree-based ensemble regressors including XGBoost, LightGBM, CatBoost, and Random Forest were trained and evaluated using RMSE and R^2 metrics.

4 Results and Discussion

This section presents the results of the exploratory data analysis and data mining techniques applied to the NYC 311 service request dataset, followed by a discussion of the insights derived from these analyses.

4.1 Exploratory Data Analysis

4.1.1 Top Complaint Types

The analysis of the top complaint categories revealed that Illegal Parking is the most frequently reported issue, followed by Residential Noise, Heat/Hot Water, and Blocked Driveways (Figure 5). This distribution highlights the prominence of traffic- and noise-related issues in dense urban environments and underscores their importance for municipal enforcement and quality-of-life management.

4.1.2 Complaint Volumes v/s. Boroughs

At the borough level, Brooklyn recorded the highest absolute number of complaints (Figure 6); however, when normalized by population, the Bronx exhibited the highest number of complaints per capita (Figure 7). This disparity suggests that while complaint volume scales with population,

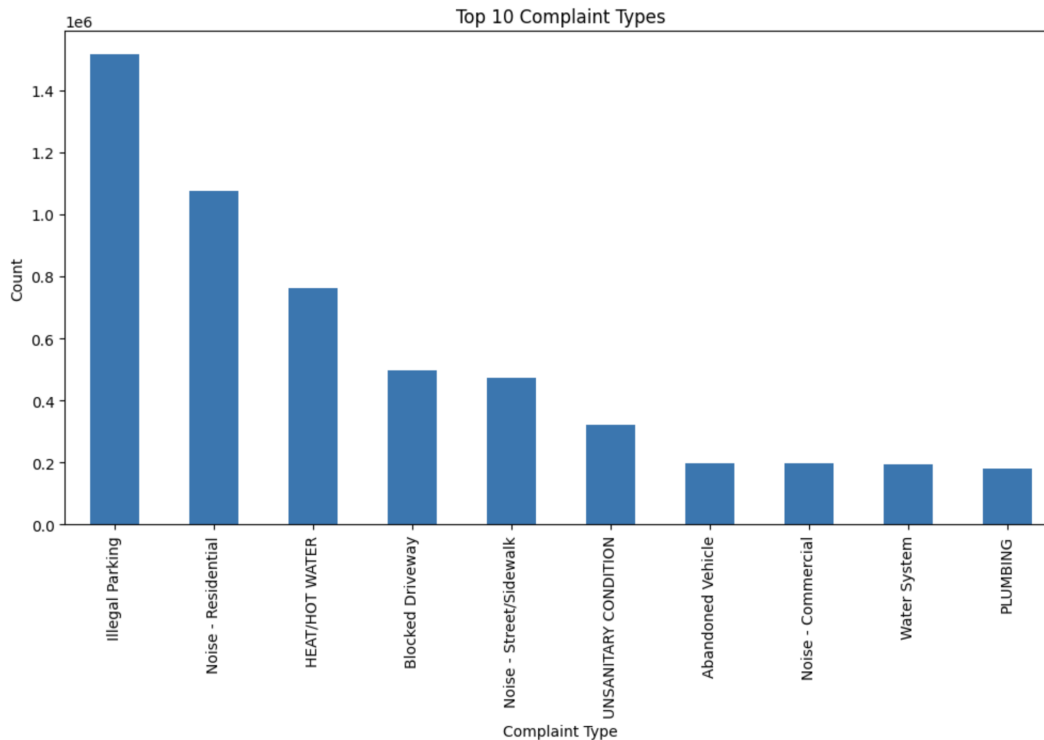


Figure 5: Top-10 complaint types.

service burden is disproportionately higher in certain boroughs, indicating potential inequities in infrastructure quality or service availability.

4.1.3 Correlation Between Complaint Volumes and Income and Education

Correlation analysis demonstrated a strong negative relationship between median household income and complaint volume (correlation coefficient -0.67), suggesting that lower-income areas tend to experience more reportable service issues (Figure 8). In contrast, education level exhibited a weaker positive correlation with complaint volume ($+0.26$), indicating that while education may influence reporting behavior, income remains the dominant socioeconomic factor. Population, income, and education statistics were obtained from publicly available sources [2, 3, 4].

4.1.4 Resolution Time by Complaint Types

Analysis of complaint resolution times showed that illegal parking and noise-related complaints are typically resolved immediately, reflecting efficient enforcement (Figure 9). In contrast, plumbing, heating, and unsanitary condition complaints required substantially longer resolution times, often several days, likely due to their reliance on infrastructure repairs.

4.1.5 Complaints by Hour of the Day by Agency

Temporal analysis revealed that complaints peak during late evening hours (9–11 PM), with the NYPD consistently handling a high volume of requests throughout the day (Figure 10). This finding aligns with known urban activity patterns and suggests the need for sustained nighttime enforcement resources. Agency abbreviations are specified in Appendix A.

4.1.6 Rodent Hotspot

Geospatial EDA identified a pronounced rodent complaint hotspot in the Upper West Side of Manhattan, with significantly higher complaint density than surrounding areas (Figure 11).

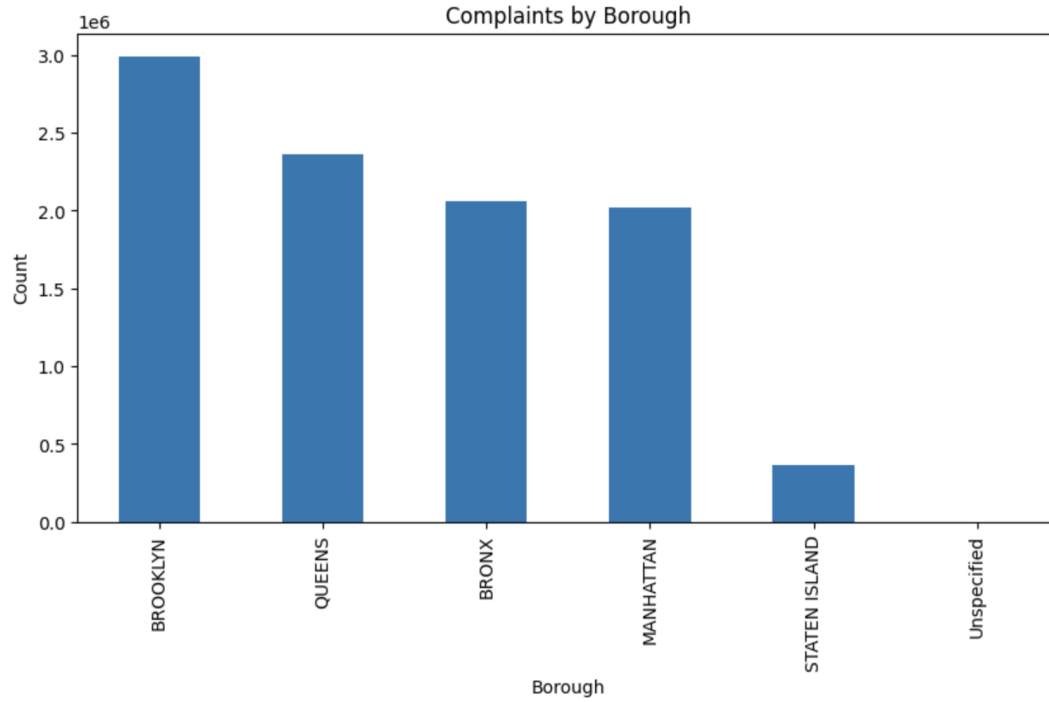


Figure 6: Complaint volumes by boroughs.

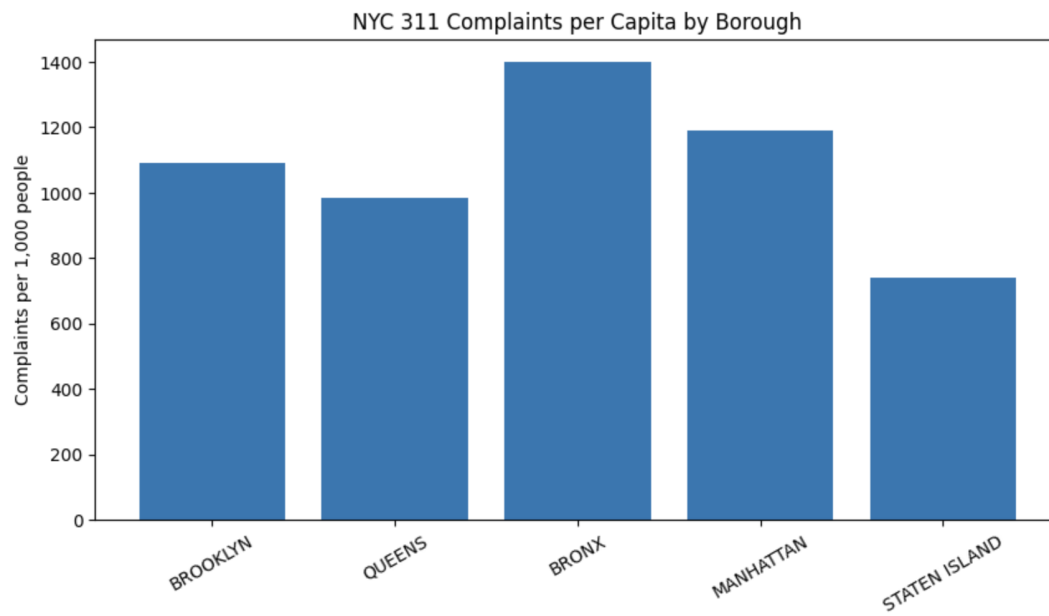


Figure 7: Complaint volumes per capita by boroughs.

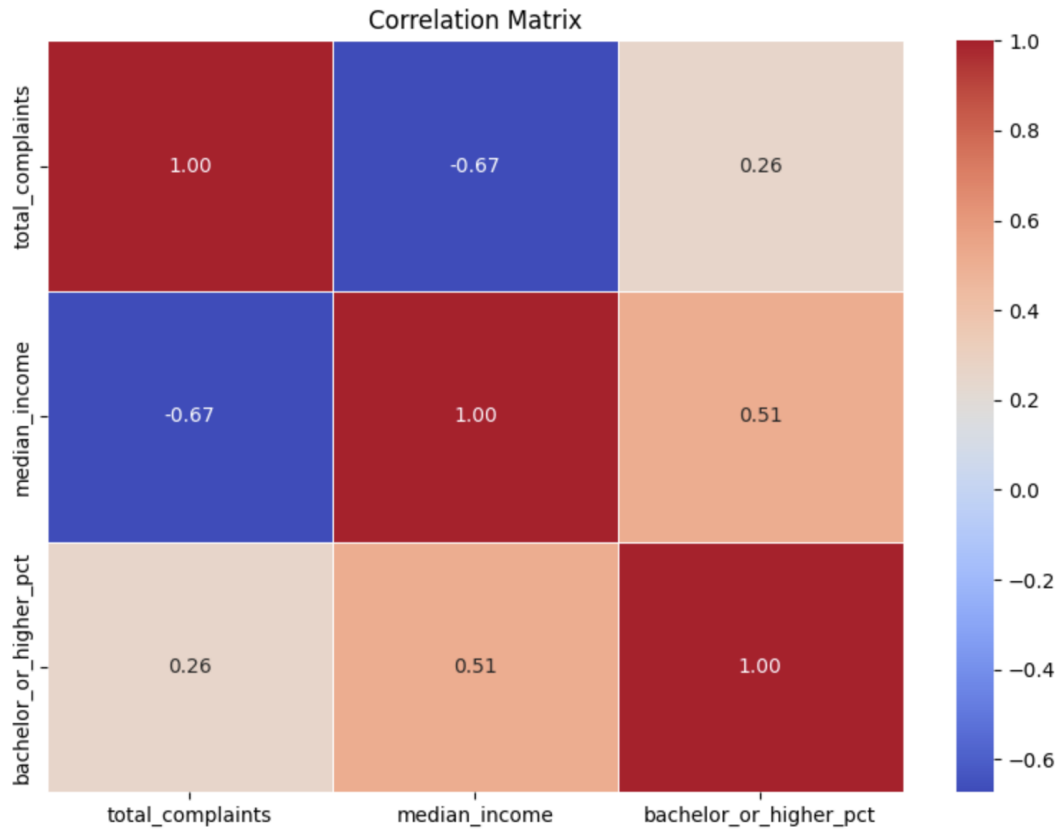


Figure 8: Correlation heatmap for how income and education levels affect complaint volumes.

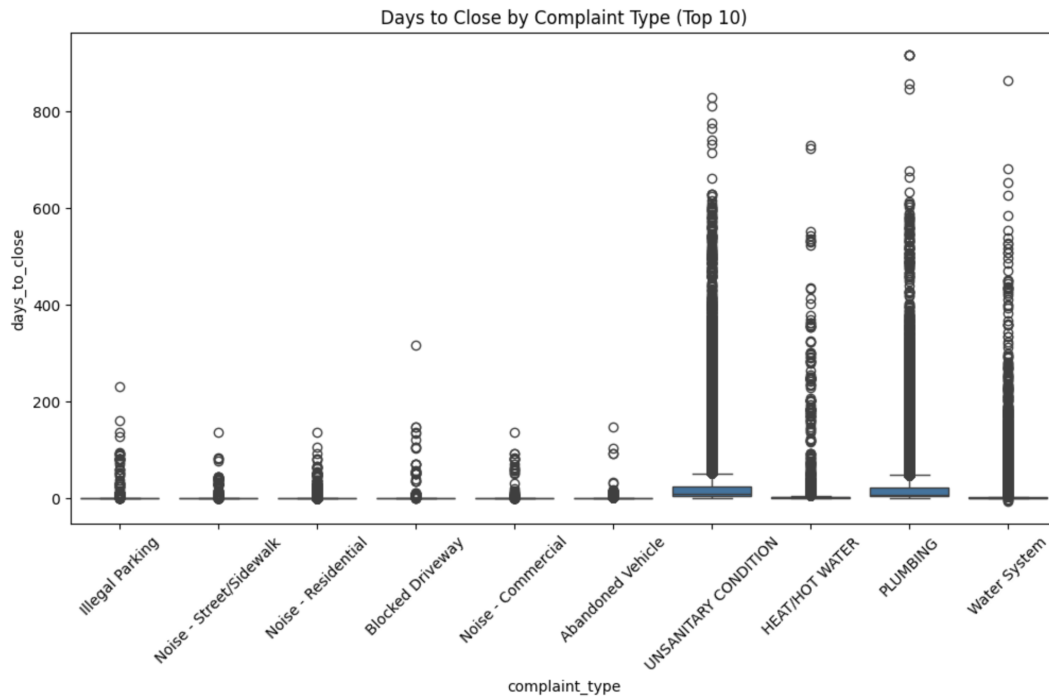


Figure 9: Distribution of resolution time across various complaint types.

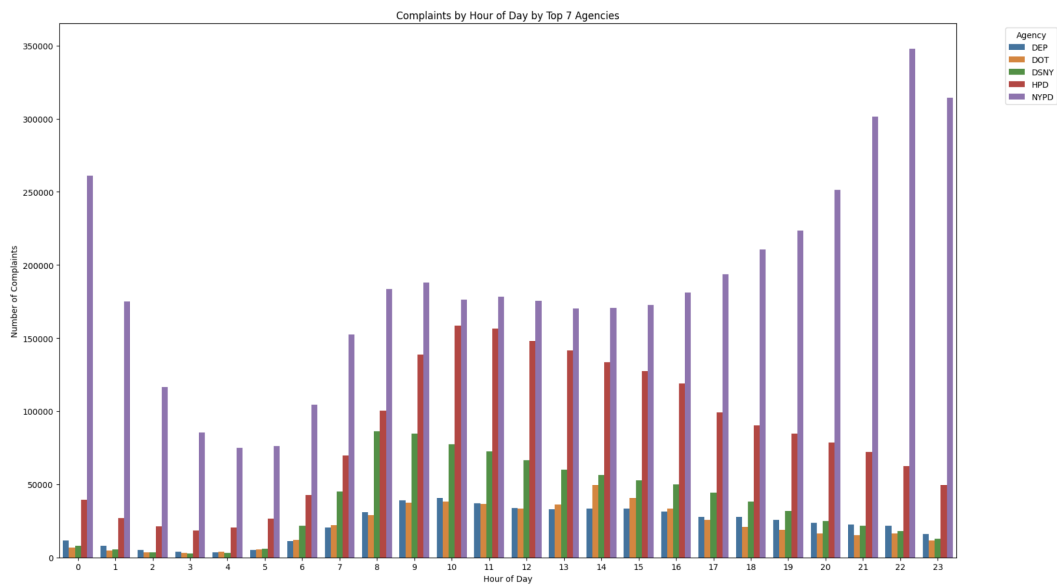


Figure 10: Distribution of complaints across the day by agencies.

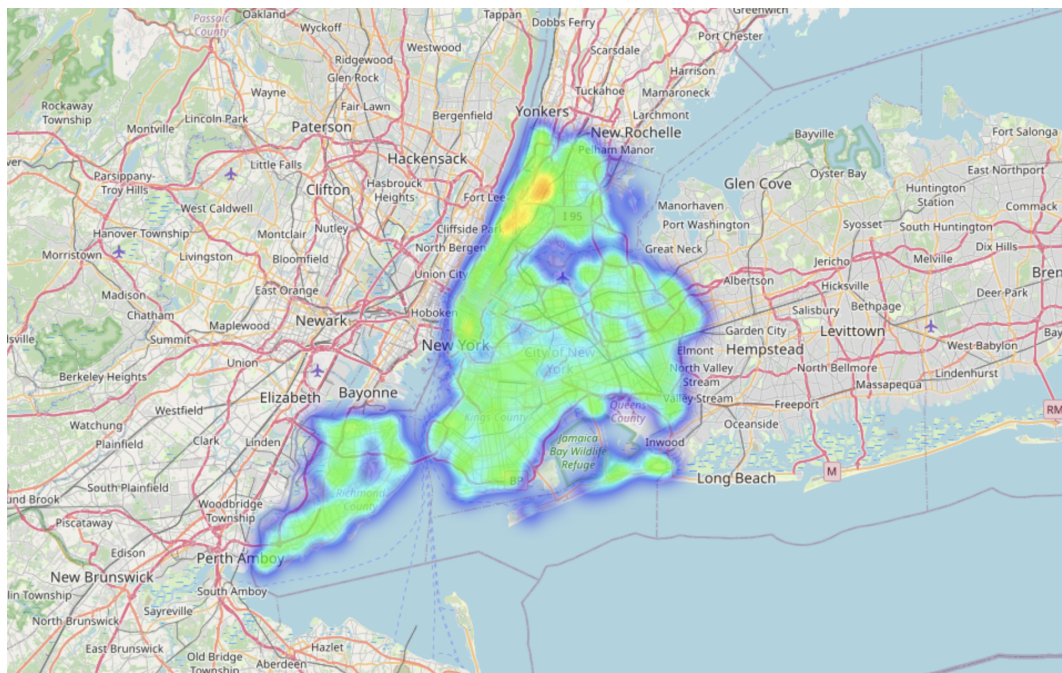


Figure 11: Heatmap of rodent hotspots across New York City.

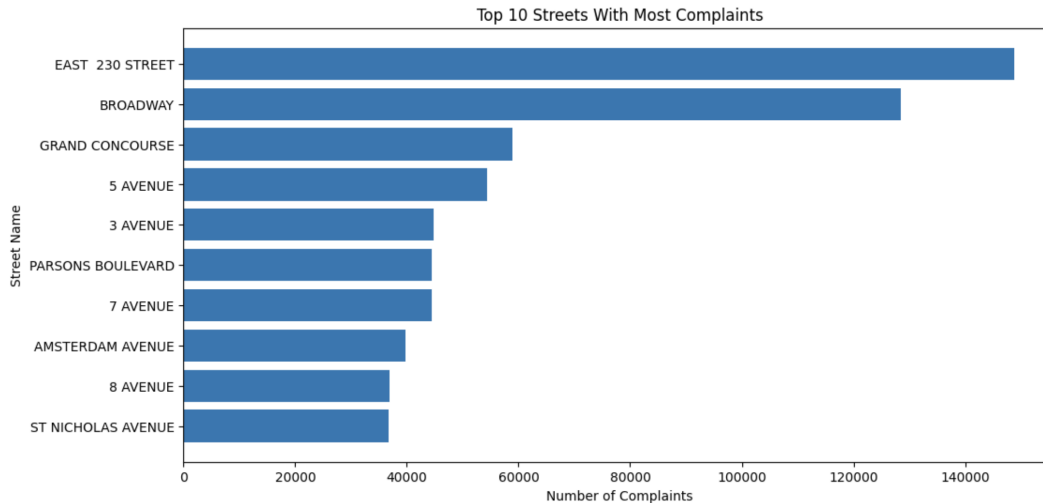


Figure 12: Streets with the highest number of complaints reported.

4.1.7 Streets v/s. Complaint Volume

Street-level analysis further showed that East 230th Street in the Bronx experiences an unusually high number of complaints, followed by major corridors such as Broadway and Grand Concourse, indicating localized infrastructure or quality-of-life challenges along these streets (Figure 12).

4.2 Data Mining and Advanced Analytics

4.2.1 Geospatial KMeans Clustering

Geospatial clustering using KMeans partitioned New York City into six distinct clusters, which closely resembled borough-level boundaries (Figure 13). This result suggests that complaint-type composition varies systematically by geography and that boroughs exhibit internally consistent complaint patterns.

4.2.2 Principal Component Analysis

Principal Component Analysis (PCA) of ZIP code–complaint type distributions revealed a triangular or funnel-shaped spread, indicating a small number of dominant complaint dimensions driving most variability (Figure 14).

This illustrates distinct complaint profiles, reinforcing the spatial heterogeneity of service needs. These findings support the use of cluster-based planning approaches for targeted resource allocation.

4.2.3 Time-Series Forecasting

Time-series forecasting using Prophet revealed a clear upward trend in overall complaint volume over time, along with high variability across months and seasons (Figure 15). This variability underscores the importance of dynamic forecasting models for anticipating service demand rather than relying on static historical averages.

4.2.4 Predicting Resolution Times

In predictive modeling of complaint resolution time, XGBoost achieved the best performance, followed by LightGBM, CatBoost, and Random Forest (Figure 16). However, all models exhibited relatively high RMSE values, and hyperparameter tuning produced limited improvement. To reduce the impact of extreme outliers that would disproportionately skew predictions, complaints with resolution times greater than 100 days were excluded from modeling. The remaining error suggests that resolution time may be influenced by latent factors not captured in the available features, such as contractor availability, building conditions, or policy constraints.

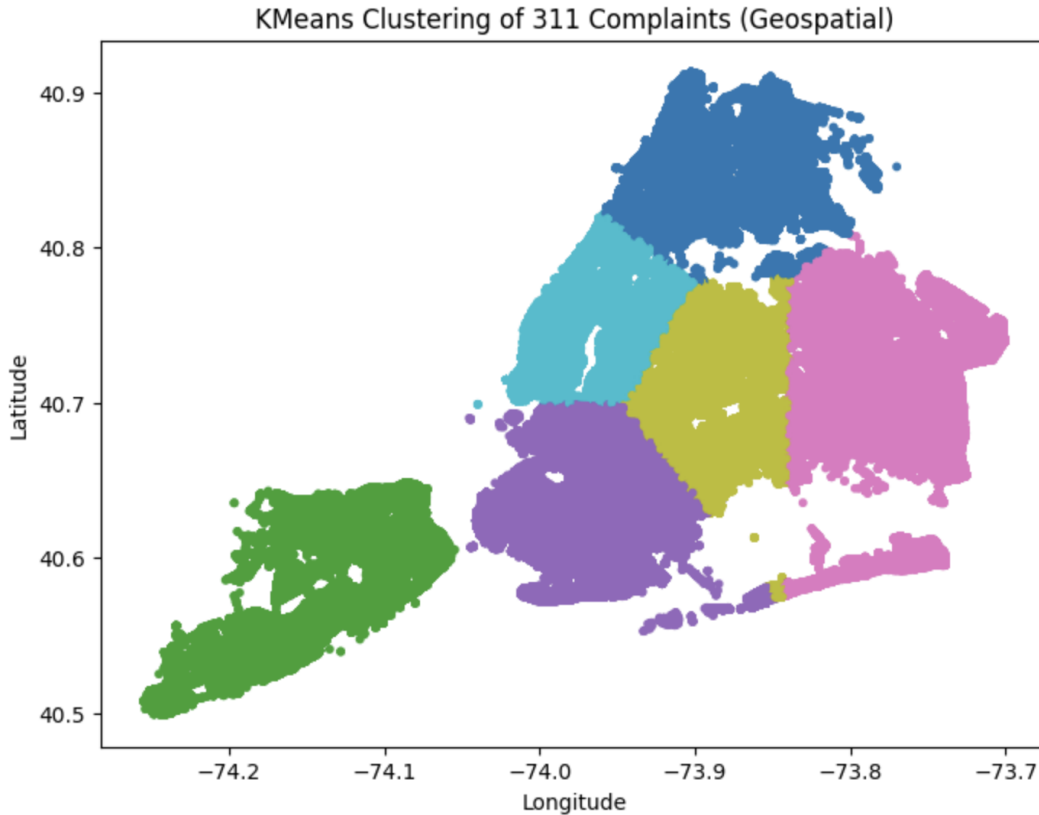


Figure 13: Geospatial KMeans clustering of complaints across the city.

4.2.5 Contrasting Complaint Types Across Boroughs

Contrast pattern mining using chi-square statistics revealed distinct borough-specific complaint signatures (Figure 17). Residential noise was disproportionately high in the Bronx, illegal parking dominated Brooklyn, encampment and homeless assistance requests were most prominent in Manhattan, drug-related complaints appeared more frequently in Queens, and missed collection issues were characteristic of Staten Island. These contrast patterns emphasize the need for localized, borough-specific policy interventions.

4.2.6 Co-occurring and Sequential Complaints Flows

Sequential and co-occurrence pattern mining uncovered strong relationships between complaint types (Figure 18). Illegal parking and blocked driveways frequently co-occur, as do residential and street noise complaints. Additionally, water leaks, plumbing issues, and heat/hot water complaints were closely linked to unsanitary conditions, suggesting cascading infrastructure failures.

4.2.7 Anomaly Detection in Complaint Volumes

Anomaly detection using rolling statistical thresholds identified numerous days with unusually high complaint volumes (Figure 19). This analysis was conducted on a reduced subset of one million records to ensure plot readability. Several anomalies aligned with real-world events, such as a nor'easter cyclone on 13th October 2025, validating the approach. Seasonal spikes in heat and hot water complaints were consistently observed during winter months, highlighting predictable demand surges related to weather conditions (Figure 20).

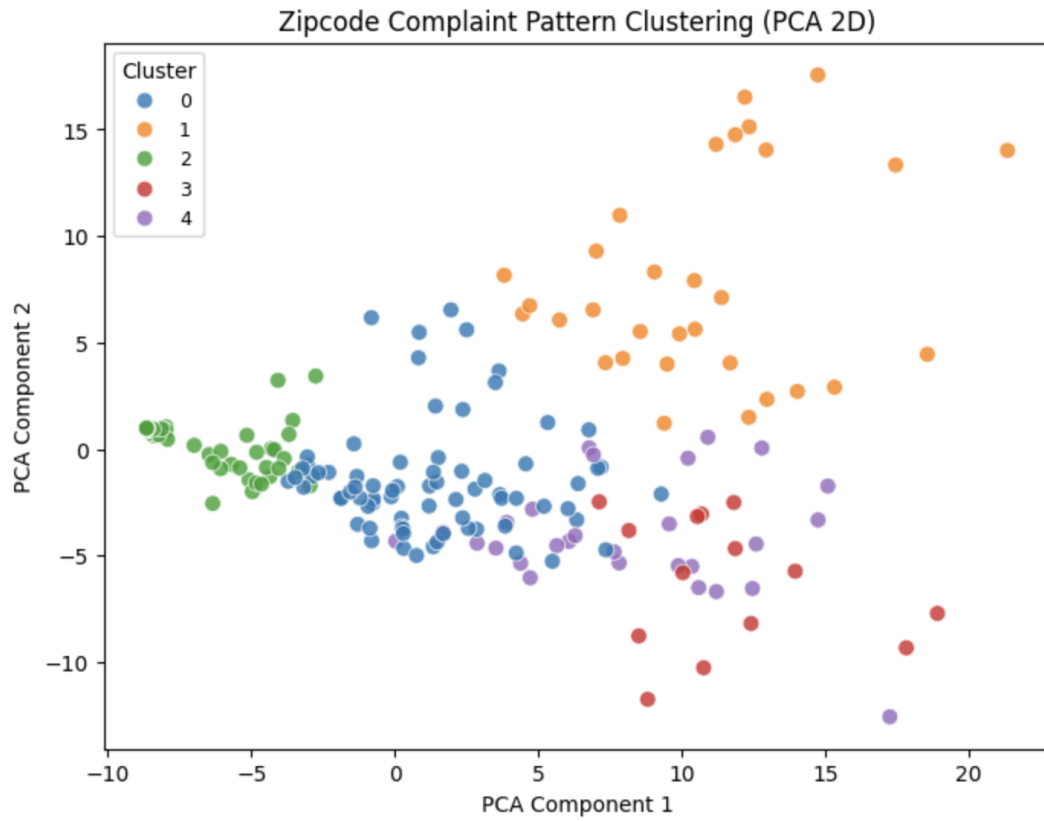


Figure 14: PCA of incident-zip and complaint type.

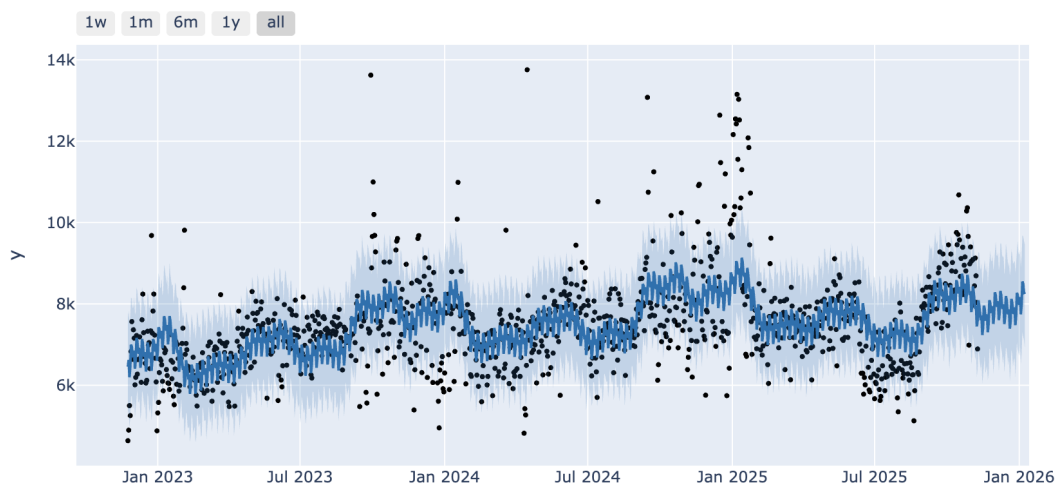


Figure 15: Time-Series forecasting using Prophet.

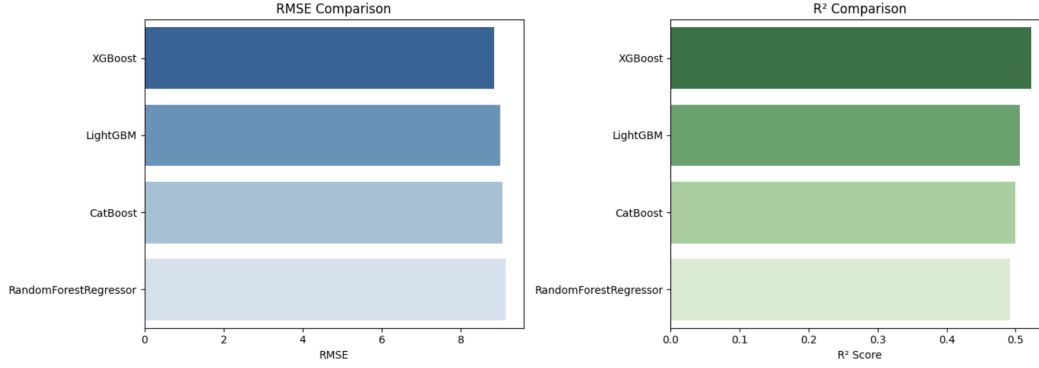


Figure 16: RMSE and R² comparison across different models.

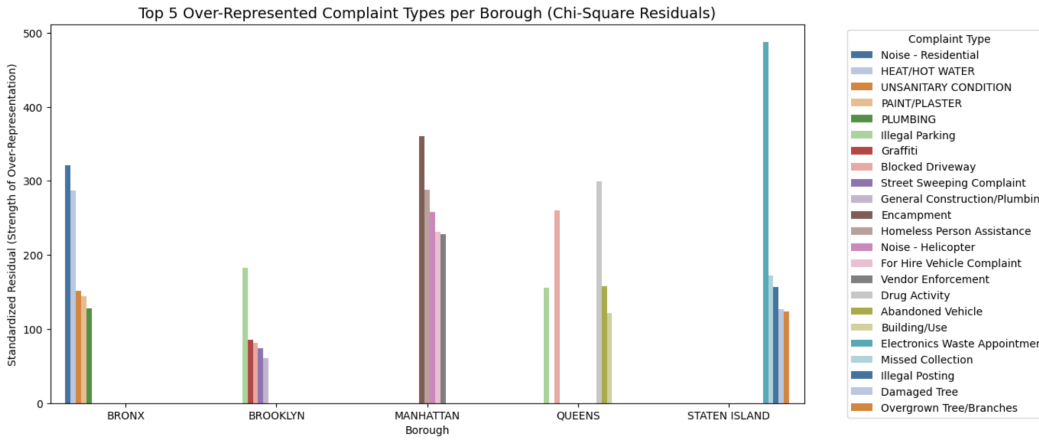


Figure 17: Distribution of overpowering complaint types across boroughs.

4.3 Discussion

Overall, the results demonstrate that NYC’s 311 data captures meaningful spatial, temporal, and behavioral patterns that can inform urban service management. The combination of exploratory analysis and advanced data mining techniques provides a comprehensive view of how complaints evolve across neighborhoods, time, and socioeconomic contexts. These insights can support proactive planning, targeted interventions, and data-driven policymaking in large metropolitan cities.

5 Limitations

While this study provides valuable insights into urban service complaint patterns, several limitations should be acknowledged. First, the analysis relies on 311 service request data, which reflects reported issues rather than actual incident prevalence. Reporting behavior may vary across neighborhoods due to differences in awareness, trust in public services, or access to technology, potentially introducing reporting bias.

Second, although the dataset is large, it lacks important contextual variables such as building age, property ownership, enforcement capacity, and contractor availability. The absence of these latent factors likely contributed to the relatively high error observed in resolution-time prediction models. Additionally, clustering and anomaly detection methods rely on parameter choices (e.g., number of clusters, threshold values), which may influence results.

Finally, the study focuses on a recent temporal subset of the data. While this improves computational feasibility, it may not fully capture long-term structural changes in complaint behavior across decades.

Complaint Pair Flow Diagram (Sankey Version)

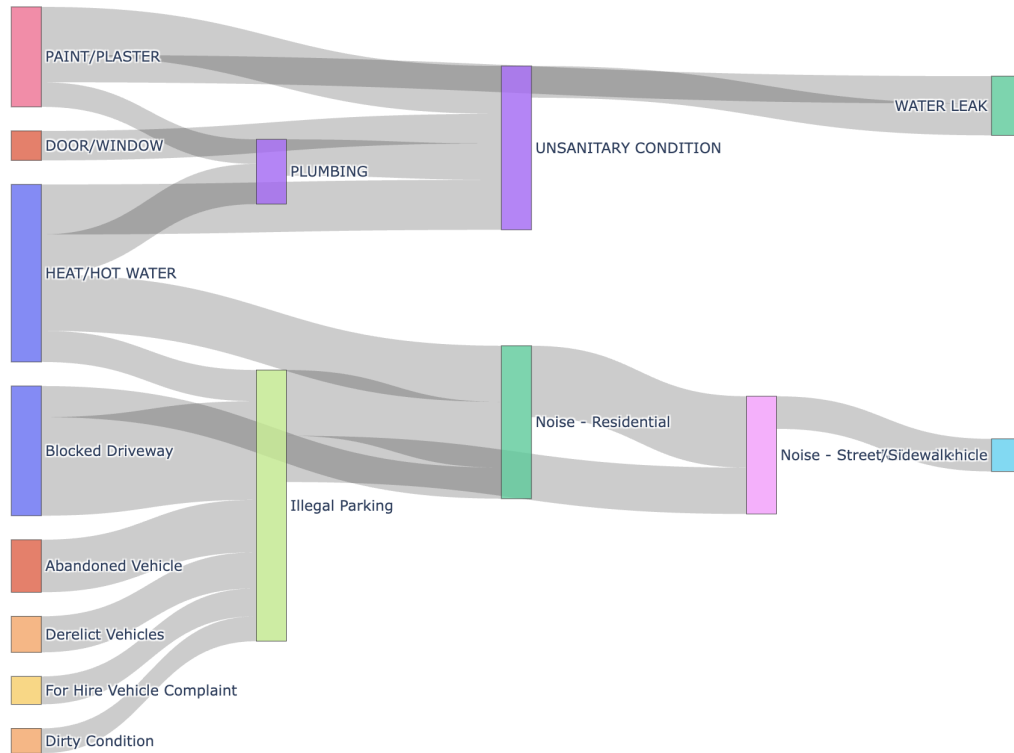


Figure 18: Sankey diagram representing sequential complaint flows.

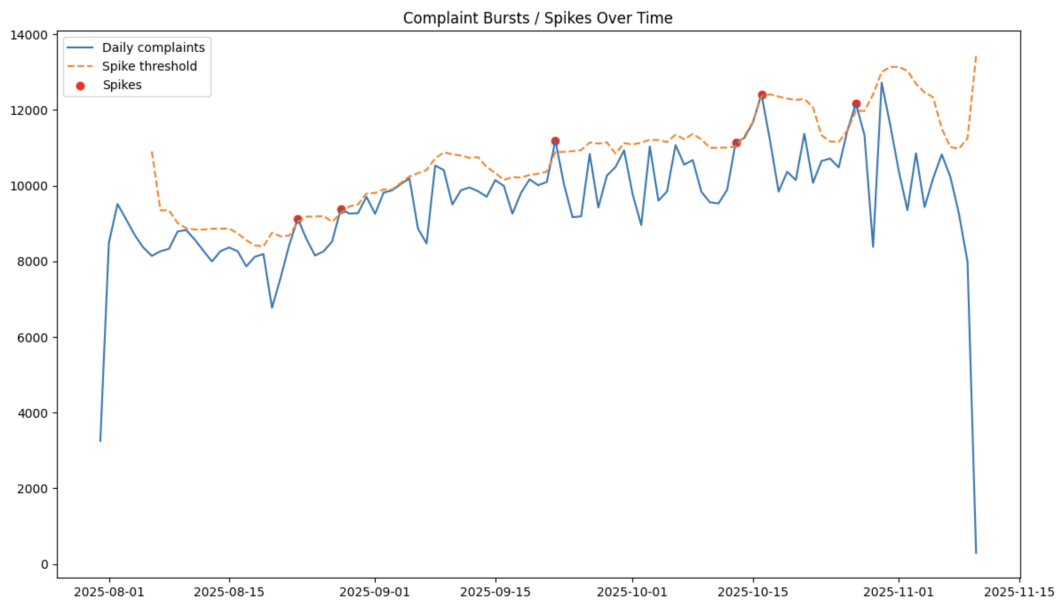


Figure 19: Detection of unusually high daily complaint volume spikes using rolling statistical thresholds.

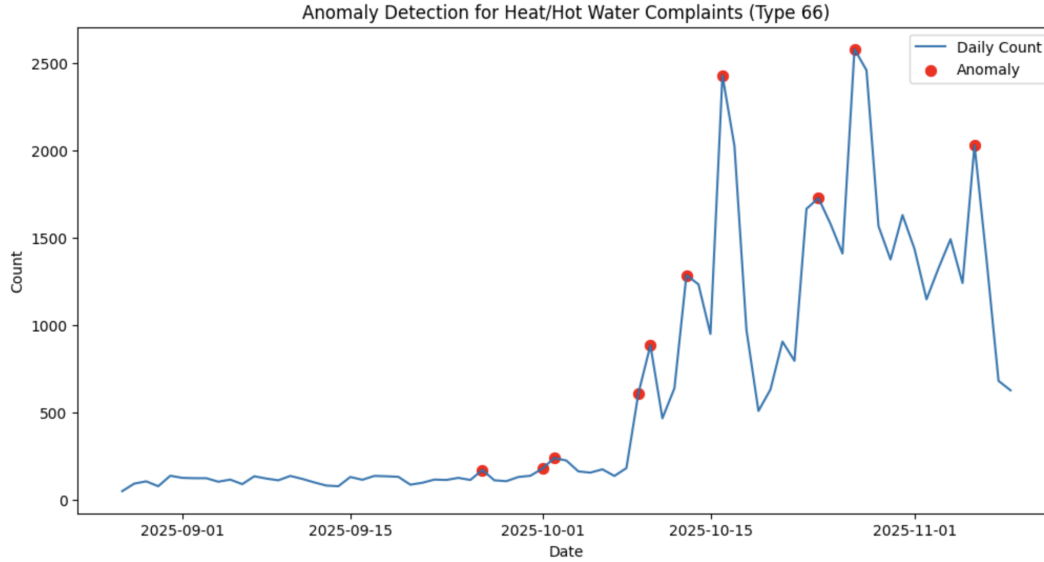


Figure 20: Seasonal increase in heat and hot water complaints as colder months approach.

6 Conclusion and Future Work

This study demonstrates the value of applying spatial-temporal analytics and data mining techniques to large-scale civic complaint data. By systematically cleaning and analyzing a 10-million-record subset of NYC’s 311 service requests, we uncovered meaningful patterns in complaint distribution, resolution dynamics, and borough-specific service challenges. Exploratory analysis highlighted strong relationships between complaint volume, time, geography, and socioeconomic factors, while advanced data mining methods revealed localized hotspots, co-occurring complaint behaviors, and event-driven anomalies. Predictive modeling further illustrated the potential of machine learning to estimate complaint resolution time, despite inherent data limitations.

Future work could focus on building more robust prediction and forecasting models to improve analytical accuracy and operational usefulness. Incorporating external datasets such as weather conditions, census demographics, and mobility patterns could enable richer correlations and stronger causal inference. Additionally, developing explanatory anomaly detection models would help distinguish routine seasonal variations from true abnormal events, enhancing the system’s ability to support proactive urban service management and data-driven policymaking.

Data and Code Availability

The dataset used in this study are publicly available through the NYC Open Data portal [1] (data.cityofnewyork.us). All code used for data preprocessing, analysis, and modeling is available at: <https://github.com/rajatrayaraddi/nyc-311-analytics>

Acknowledgment

The authors would like to express sincere gratitude to Paul Melby, Professor at The George Washington University, for his guidance, support, and insightful feedback throughout the duration of this project. His expertise and encouragement were instrumental in shaping the direction and depth of this work.

References

- [1] City of New York (2025). 311 service requests from 2010 to present. *NYC Open Data*.

- [2] New York City Department of City Planning (2024). Current population estimates for new york city.
- [3] Staten Island Advance / SILive (2025). Which nyc borough has the highest rate of college graduates?
- [4] U.S. Census Bureau (2023). Quickfacts: New york city and new york state counties.

A Appendix

Agency Abbreviations:

- DEP: Department of Environmental Protection
- DOT: Department of Transportation
- DSNY: Department of Sanitation of New York
- HPD: Department of Housing Preservation and Development
- NYPD: New York City Police Department