BA820 – Milestone 2

Clustering NYC 311 Service Requests for Smarter City Management

Mar 3, 2025

Section B1, Team 1:
Aastha Surana, Atishay Jain, Kendall Sims, Soham Sarvade

BOSTON
UNIVERSITY

1. Problem Statement

NYC 311 handles thousands of service requests daily, but identifying complaint patterns, response times, and agency involvement is challenging. Inefficient resource allocation and limited insights from resolution descriptions impact service quality.

Scope : The project analyzes NYC 311 data to identify complaint patterns, optimize resource allocation, and assess service effectiveness.

Objectives

1. Identify complaint patterns and response times to improve city services.
2. Detect geographic complaint clusters for better staffing and resource allocation.
3. Extract key themes from resolution descriptions to understand recurring issues.

2. Dataset, Exploratory Data Analysis (EDA) & Preprocessing

Dataset description

- **Source**: NYC Open Data – 311 Service Requests (2010–Present)
- **Size**: Over 28 million records, 41 features
- **Sampling:** We selected 1% of the total dataset using stratified sampling, ensuring proportional representation from each year.

Exploratory Data Analysis

Our analysis of NYC's 311 service request data identifies key patterns to improve city resource allocation.

Key Insights

- Complaints show **natural clustering** based on geography, time, and type.
- High-complaint zones and certain recurring issues indicate **target areas for intervention**.
- **Temporal Trends**: Peak: Monday-Friday, highest on Tuesdays; midnight spikes due to noise complaints.
- **Complaint Reporting & Categories**: Phone (49.1%) dominates over online (21.4%); noise, heating, parking are top concerns.
- **Agency Workload & Geospatial Trends**: NYPD, HPD, and DOT handle most complaints; Brooklyn & Queens lead in volume.
- **Service Efficiency & Next Steps**: Most cases resolve quickly, but some delays indicate inefficiencies needing optimization.

Data Cleaning & Preprocessing

1. **Handling Missing Values**: Address incomplete timestamps, location data, and complaint descriptors.
2. **Feature Engineering**: Extract time-based features, encode categorical variables, and construct new variables (e.g., complaint resolution time).
3. **Scaling and encoding:** We scaled numerical features with StandardScaler and encoded categorical variables (one-hot for multiple categories, label for ordinal/binary) for clustering and modeling.
4. **Dimensionality Reduction**: PCA will be applied to reduce noise while preserving variance in the dataset.

3. Analysis & Experiments

Model 1: Association Rules Analysis

Using the Apriori algorithm, we applied Association Rule Mining to uncover frequent complaint patterns based on boroughs, agencies, time trends, and complaint types, helping improve service efficiency and resource allocation.

Key Findings:

- Noise Complaints: Predominantly in Manhattan & Brooklyn, spiking on weekends.
- Parking Violations: Common in Queens & Brooklyn; 100% handled by NYPD.
- Heat/Hot Water Issues: Concentrated in Bronx & Brooklyn.
- Agency Assignments: NYPD (noise, parking), HPD (heating), DOT (street conditions).

Impact:

- Optimized Resource Allocation: Targeted interventions like increased weekend enforcement for noise or more heating specialists in winter.
- Better Response Strategies: Borough-specific prioritization—e.g., parking enforcement in Brooklyn & Queens, and improved coordination with streamlined agency collaboration.
- Public Awareness: Reduce misdirected complaints and improve resolution times.

Model 2: Clustering Analysis: Hierarchical & K-Means Clustering

We used Hierarchical Clustering and K-Means Clustering to segment complaints based on location and density, helping optimize resource allocation and improve response efficiency.

Criteria	Hierarchical Clustering	K-Means Clustering
Approach	Used latitude, longitude, and zip complaint density with Ward's method, identifying five clusters aligned with NYC's boroughs.	Used all numerical features, applied PCA (~90% variance retained), and sampled 5,000 rows for efficiency.
Challenges	Scalability issues made full dataset clustering infeasible. A 5,000-row sample was used to ensure feasibility.	K-Means assumes spherical clusters, which may not match real-world patterns, but PCA improved clustering accuracy.
Findings	High-density clusters in Brooklyn, Bronx, and Manhattan indicate persistent service demands, while Staten Island and parts of Queens show dispersed complaints.	Six clusters differentiated by location, complaint density, and resolution speed. High-density areas (Downtown Manhattan, Bronx) showed slower resolution times.
Impact	Helps city agencies prioritize high-complaint areas, improving response times.	Supports targeted interventions, enabling specialized response teams for noise, heating, and parking complaints.

Model 3: Text Mining: BoW, Topic Modeling, TF-IDF, Clustering, Sentiment Analysis

- **Expanded Methods from M1:** Previously, text mining was not included. In this phase, we applied BoW, Topic Modeling, TF-IDF, Clustering, and Sentiment Analysis to analyze patterns in the 'resolution_description' column.
- **Why These Methods Were Chosen:** As the only text-heavy column, all text mining techniques were applied here. BoW structured the data for pattern analysis, Topic Modeling identified key complaint themes to aid prioritization and clustering, and TF-IDF was tried (but not used) for clustering. Sentiment Analysis was conducted based on resolution time relative to complaint type.
- **Clearly Defined Goals:** Our objective was to extract key trends to improve response prioritization for 311 agencies by addressing:
  - Identify common themes in resolution descriptions.
  - Categorize complaints to assess urgency.
  - Analyze sentiment in relation to resolution time and complaint type.
- **Parameter Tuning & Adjustments:**
  - TF-IDF clustering resulted in overlapping groups with mean values near 0.
  - Topic Modeling yielded more distinct clusters, which we validated using K-Means. This output was then used for sentiment analysis.

- **Methods That Did Not Work:** In Topic Modeling, NFW errors led us to switch to LDA, which successfully assigned four distinct topics. Initially, Sentiment Analysis based on positive/negative references failed to classify complaints effectively. We pivoted to predicting sentiment based on resolution time relative to similar complaints, yielding better results.

4. Challenges, Dead Ends & Adjustments

Our analysis faced challenges across association rule mining, clustering, and text analysis, requiring optimizations to improve efficiency and extract meaningful insights.

Association Rule Mining

- Low lift values made rules weak despite high confidence. Adjusted support and confidence thresholds.
- High number of unique complaint types led to low-support patterns. Grouped similar complaints.
- Encoding issues with categorical data. Ensured proper formatting.

Clustering Challenges

- Hierarchical clustering was infeasible due to memory limitations. Sampled 5,000 rows using Ward's method.
- K-means struggled with initialization and cluster shapes. Applied PCA and optimized k using the elbow method.
- Despite these optimizations, full-scale hierarchical clustering and sentiment analysis on citizen feedback remained infeasible due to computational and data limitations.

Text Analysis and Topic Modeling

- Bag of Words (BoW) included many stop words, so we needed to filtered those out.
- NMF topic modeling was incompatible. Switched to LDA for clearer results.
- Sentiment analysis was limited due to lack of citizen feedback data. Analyzed official resolution descriptions instead.

5. Findings and Interpretations

- **Analysis Key Insights:** Our analysis of NYC 311 service requests identified key patterns in complaint types, borough distribution, and resolution efficiency:
  - Complaint Clustering: High-density complaint areas are concentrated in Brooklyn, Manhattan, and the Bronx, indicating where city services are most needed.

- Temporal Trends: Complaints peak on weekdays, especially on Tuesdays, with a notable midnight spike due to noise complaints.
- Complaint Categories: Noise, heating, and parking issues dominate service requests, accounting for over 50% of complaints.
- Resolution Efficiency: Response times vary by cluster—high-density areas take longer to resolve issues compared to low-density areas.

- **Comparison of Methods:**

Clustering	Approach	Insights
Association Rules	Apriori Algorithm	Noise (Manhattan, Brooklyn, weekends); Parking (Queens, Brooklyn, NYPD oversight); Heating (Bronx, Brooklyn, winter delays); Phone reports dominant, highlighting urgency
Clustering	Hierarchical, PCA, K-Means	Created 5 defined clusters based on location and resolution time
Text Mining	BoW, Topic Modeling, Clustering, Sentiment Analysis	4 key topics (General, Public Space, Urgent, Housing); 2013 saw a dip in positive sentiment, with winter months showing more negativity

- **Practical significance of Results:**
  - Faster Responses: Identifies high-complaint areas to allocate resources effectively.
  - Better City Planning: Recognizes persistent service demands (e.g., heating complaints in winter).
  - Improved Citizen Experience: Helps agencies prioritize complaints based on urgency.
- **Discuss how your findings apply to real-world problems or industry applications.**
  - Urban Planning & Public Policy: Optimizing resource distribution to underserved areas with high complaint densities or frequent urgent requests.
  - Customer Service Optimization: Using text mining techniques for automated complaint categorization. This will increase response time to alert the correct response agency so they can resolve the complaint as quickly as possible.
  - Crisis Management & Emergency Response: Utilizing complaint patterns to proactively deploy emergency services, such as dispatching more heating technicians before winter spikes or increasing noise patrols in nightlife-heavy areas.

Appendix

Contribution and Challenges

Member	Contribution	Challenges
Aastha	- Structured the Colab notebook - Compiled and structured the report - Conducted EDA - Implemented clustering techniques (Hierarchical, K-means) - Refined preprocessing steps - Engineered numerical features	- Resource Constraints: Hierarchical clustering caused RAM crashes, requiring multiple iterations and long rerun times. - Data Sampling Adjustments: Reducing the dataset from 5% to 1% altered EDA results, necessitating a pivot in analysis.
Atishay	- Preprocessing steps of the data - Association rules - Findings and interpretation of the project - Compilation and final check of the colab	- PCA implementation - Rules were taking too long to execute initially
Kendall	- Preprocessing of the data - Text mining (Bag of Words, Topic Modeling, TF-IDF, Clustering, Sentiment Analysis) - Findings and interpretation of the project - Compilation and final check of the colab -Refined the M1 report for an organized structure and clear format	- Filtering out many stop words to conduct BoW - Was not able to compute NMF topic modeling, had to use LDA instead - TF-IDF for clustering did not produce meaningful, well-separated clusters - Did not have a description column to analyze citizen sentiment directly - The large amount of data took a long time to complete each analysis
Soham	- Data Extraction from big query - Association rules - Created categorical features for association rule analysis - Featured engineering to create numerical features for the colab - Interpreted findings	-Many rules had low lift values, making them statistically weak despite high confidence -High number of unique complaint types led to complex, low-support patterns and redundant rules -Faced formatting errors with tuple-to-list conversions and incorrect categorical encoding

GitHub Project

- **Repository Link:** https://github.com/sohamss171/BA820_B1_Clustering_NYC_311_Service_Requests_for_Smarter_City_Management

References

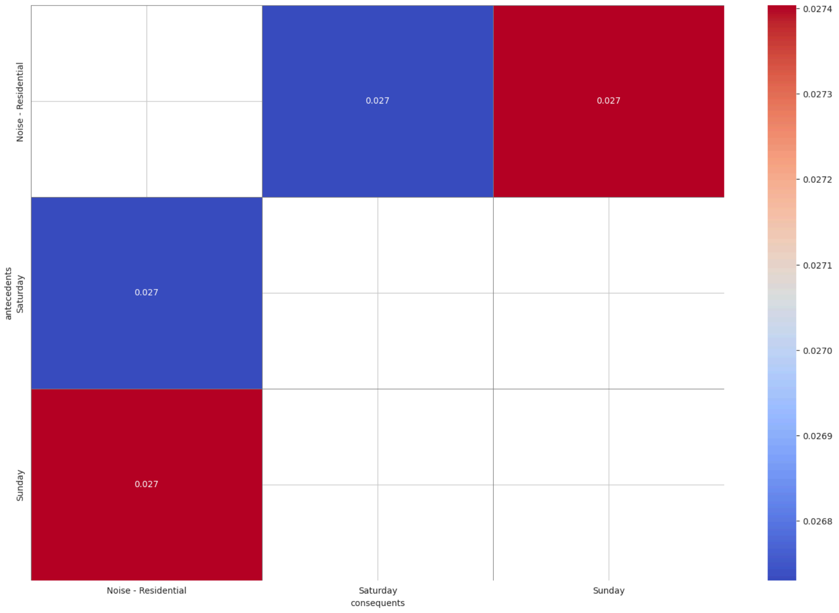
1. Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast Algorithms for Mining Association Rules." *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*. Vol. 1215. 1994. Apriori Documentation.
2. Mlxtend Mlxtend
3. Mulligan, Kerry, Celina Cuevas, Edwin Grimsley, Preeti Chauhan, and Erica Bond. "Justice Data Brief: Understanding New York City's 311 Data." Data Collaborative for Justice, March 2019 Justice Data Brief: Understanding New York City's 311 Data.
4. "The Value of 311 Data." *Hunter Urban Review* The value of 311 data - Hunter Urban Review
5. Harvard University. "Transforming Municipal Customer Service in Chicago" *Data-Smart City Solutions* Harvard University : Transforming Municipal Customer Service in Chicago

Timeline

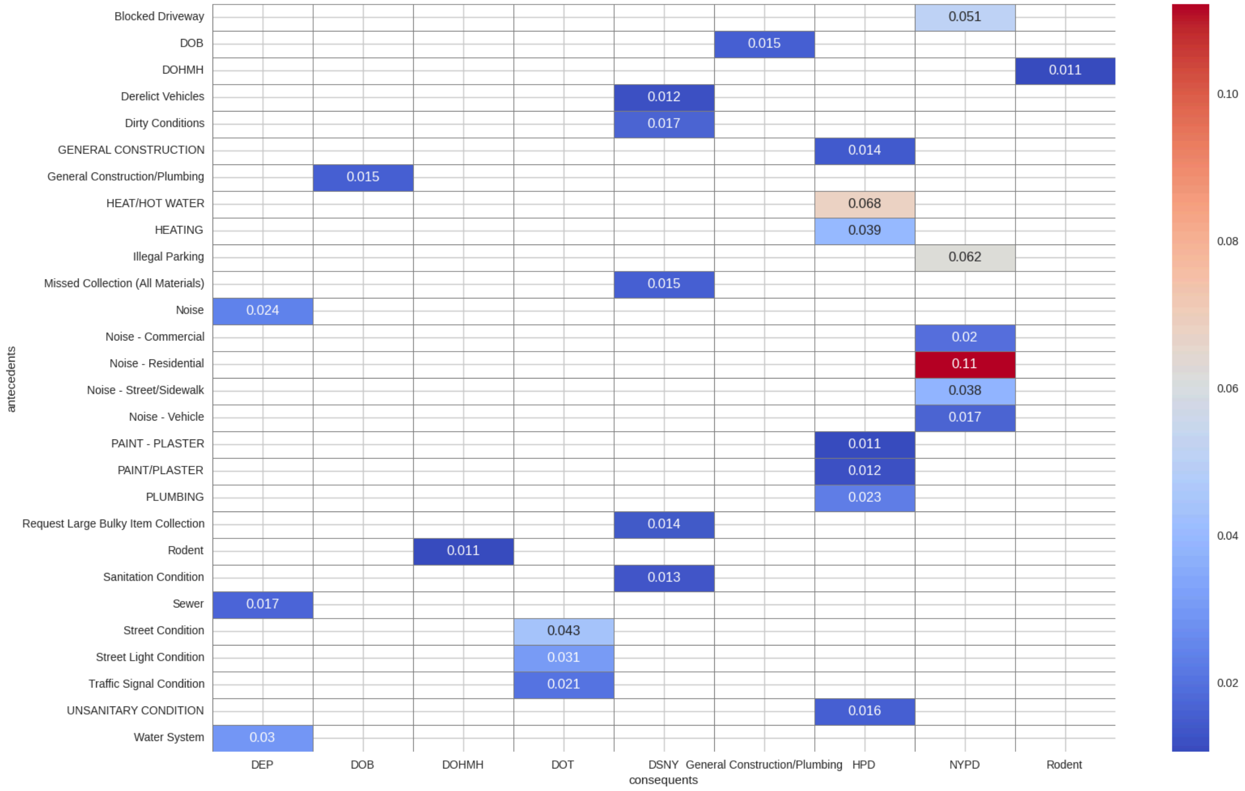
- **Proposal:** It took us one week to come up with the proposal.
- **Milestone 1:** After submitting our proposal our group conducted 4 meetings to distribute the workload, check in on progress and discuss our findings before M1 submission.
- **Milestone 2:** After going through the feedback received from our M1 submission we came up with alternative ideas and approaches to test out and explore. In a span of 2 weeks, we met 4 times as well to go over the progress and eliminate any issues.

Supplemental Data and Results

1. Association Rules



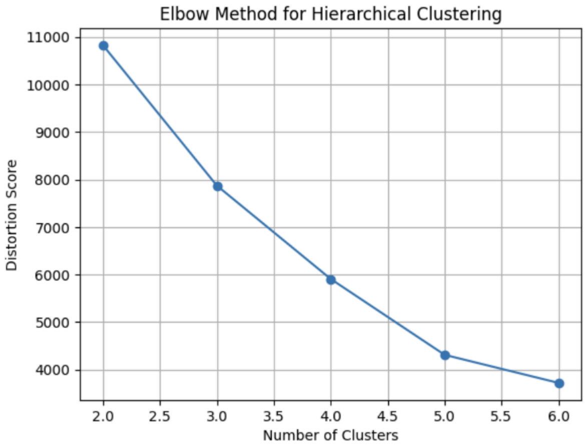
As we can see from the visualisation above, the association values (0.027) are consistent across the matrix, suggesting a uniform pattern. Residential noise tends to be reported more on weekends, as seen in the relationship between "Noise - Residential" and Saturday/Sunday.



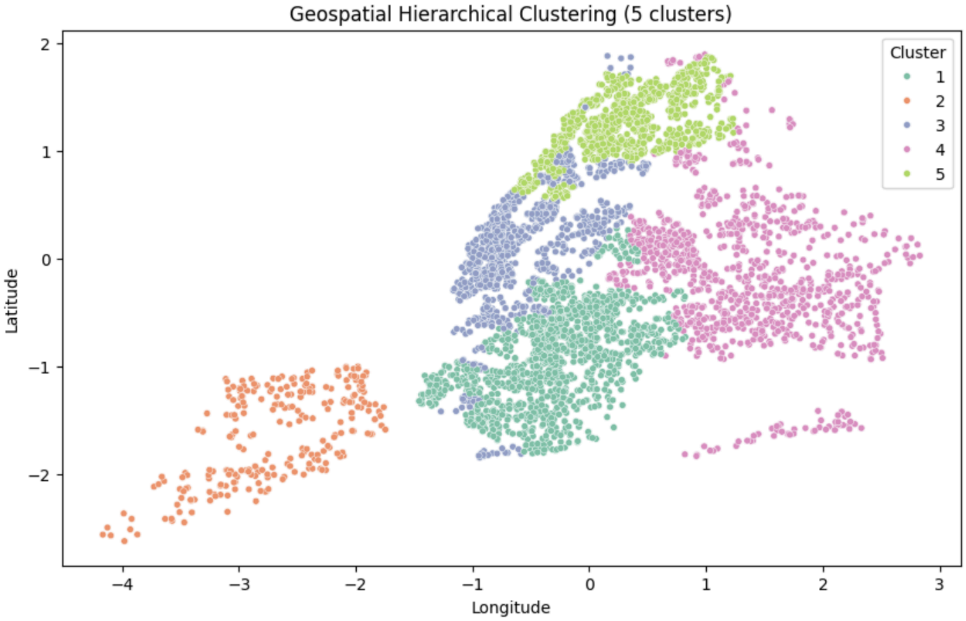
We can see in the visualisation above that:

- NYPD handles most Noise - Residential complaints (0.11)
- HPD sees frequent heat/hot water issues (0.068)
- DSNY struggles with missed waste collection (0.062)
- DOT faces infrastructure concerns (Street: 0.043, Lights: 0.031)
- HPD & DEP handle building and sewer issues

2. Hierarchical Clustering

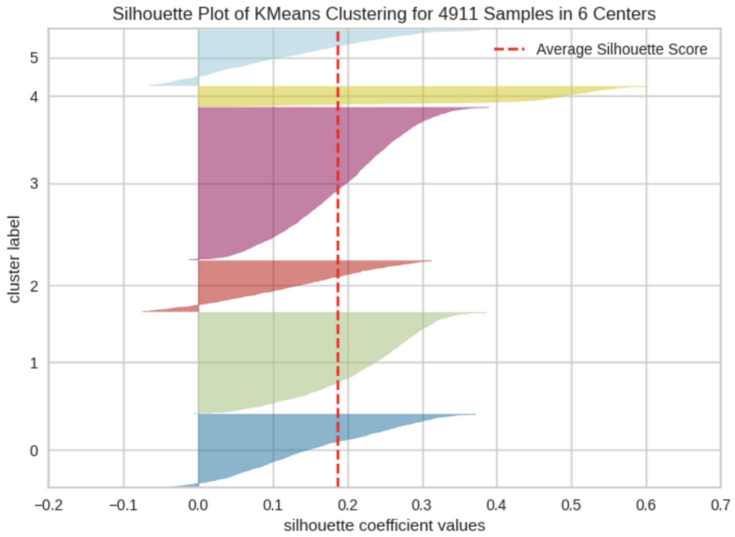


The silhouette score peaks at k=6, k=5 offers comparable clarity while avoiding unnecessary segmentation

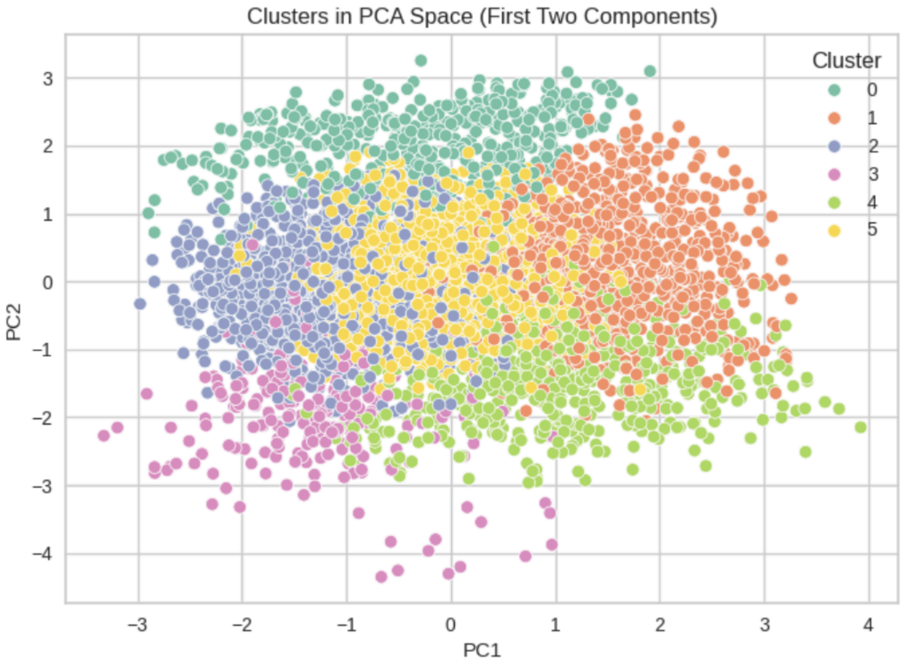


High-density clusters in the Bronx, Brooklyn, and Manhattan reflect urban issues driven by population density and service demand. In contrast, dispersed clusters in Staten Island and Queens align with lower complaint volumes in less populated residential areas.

3. K-means Clustering



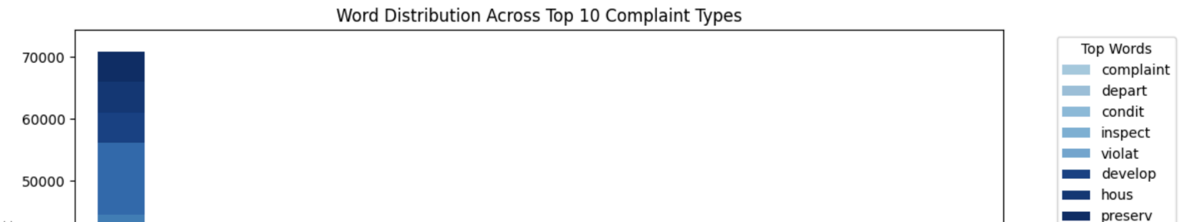
Silhouette Plot (k=6): Average coefficient around 0.20, indicating moderate cluster distinctiveness; each cluster band shows decent separation.



The plot shows six clusters projected onto the first two principal components, highlighting broad separation. While some clusters are well-defined, others overlap, indicating boundary blending. Full PCA space may reveal clearer distinctions.`

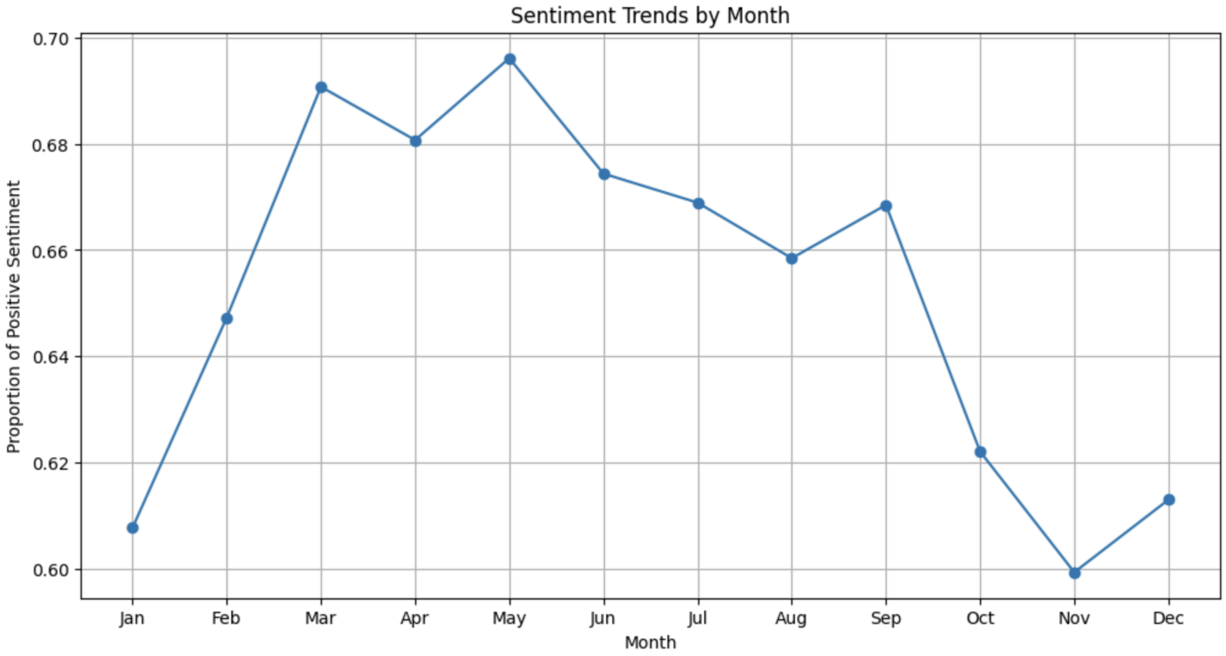
4. Text Mining

- Top BoW Words Associated with Complaint Types



We see that heating-related complaints have the highest word count, followed by Noise - Residential and heat/hot water issues. Frequent words like “complaint,” “department,” and “police” indicate high agency involvement, while "inspect," "violate," and "develop" suggest a focus on compliance and enforcement. Heating complaints and noise-related issues dominate 311 resolutions. Housing and law enforcement terms appear across many complaint types, emphasizing ongoing oversight and quick response times to these prevalent issues.

Complaint Resolution Sentiment Trends by Month



The line chart tracks monthly sentiment trends in NYC 311 complaint resolutions. Sentiment rises from January to a peak in April-May, suggesting improved resolutions or milder complaints. It then declines from June, hitting a low in November, possibly due to seasonal factors or resource constraints. A slight rebound in December indicates partial recovery. These insights can help agencies optimize resource allocation during periods of lower public satisfaction.