2) a short (250 - 500 word) technical description of the analyses you ran.

My analysis began by integrating three primary data sources: 1) New York City's 311 service request data for the year 2016, 2) demographic and socioeconomic attributes from the American Community Survey (ACS 2012-2016 5Y Estimate), and 3) spatial boundaries of census tracts (TIGER/Line shapefiles 2016) from the U.S. Census Bureau. The ACS data was merged with TIGER/Line shapefiles to assign demographic and income attributes to each census tract. Meanwhile, latitude-longitude coordinates provided in the 311 dataset were spatially joined with the same shapefiles, allowing each complaint to be assigned to a census tract.

Given that the raw 311 data contained 182 distinct complaint types, these were grouped into five broad categories—Noise, Housing and Building, Sanitation and Environmental, Street and Infrastructure, and Public Safety and Quality of Life—to simplify the analysis. The total complaint counts per category were aggregated at the census tract level. Because raw counts do not account for varying population sizes, the number of complaints was normalized by dividing by each tract's total population, then multiplied by 1000 resulting in complaint rates per 1,000 residents. Census tracts with extremely small populations (fewer than 500 residents) were filtered out to improve data reliability.

I then computed summary statistics and produced boxplots to understand the distribution of complaint rates. Additionally, correlation matrices (Pearson and Spearman) were computed to examine linear and monotonic relationships between complaint rates and demographic/ socioeconomic variables (e.g., median household income, racial composition percentages).

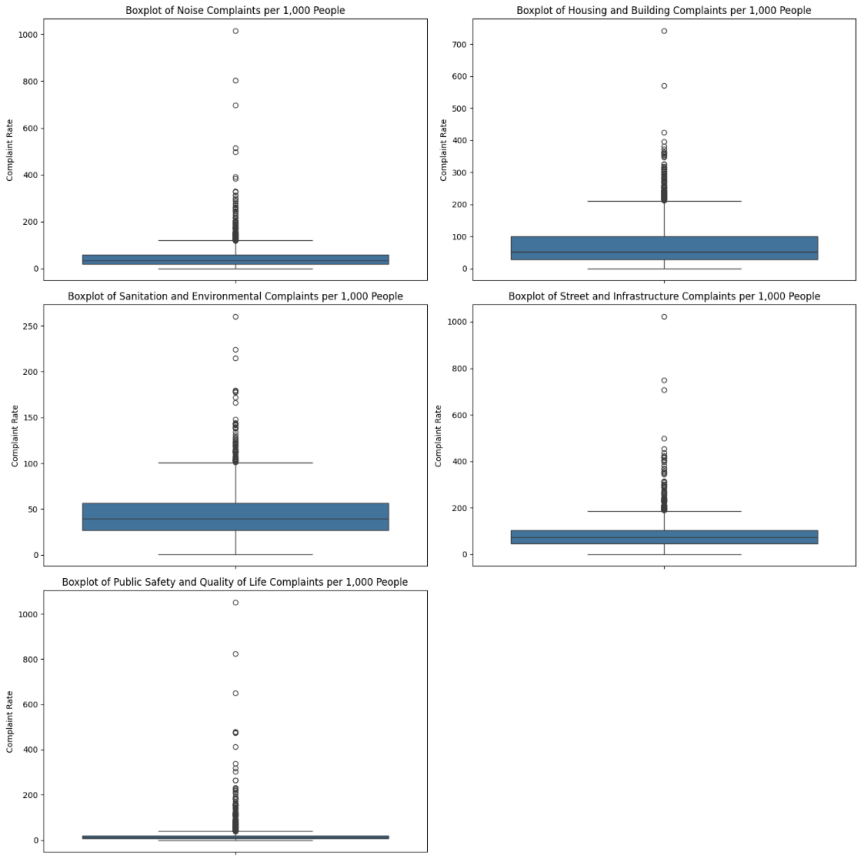
Next, linear regression models (Ordinary Least Squares, OLS) were fitted to examine how demographic and socioeconomic factors relate to complaint rates. Because urban data often deviates from normality and exhibit heteroscedasticity, dependent variables(complaint categories) were log-transformed (+1 added where zero counts occurred, after dropping all 0's) to stabilize variance and better approximate linear model assumptions. Robust standard errors using HC1 and HC3 covariance estimators were also employed to address heteroscedasticity and ensure findings had a valid computational backing.

Finally, Moran's I tests were conducted to detect spatial autocorrelation in complaint patterns, indicating whether residuals clustered geographically. Although spatial regression models were not implemented here, the identification of spatial autocorrelation suggests that future analyses could benefit from more advanced spatial econometric techniques, and this was included in the write-up.

In sum, the methodology involved data merging, spatial joins, normalization of complaint data, variable transformations, correlation analyses, OLS regressions with robust standard errors, and tests for spatial autocorrelation—each step chosen to refine insights into how demographic, socioeconomic, and geographic factors shape the distribution and intensity of NYC's everyday service requests.

Additional Explanation to Data:

Here, I noticed that some complaint categories had extreme outliers—areas with exceptionally high complaint rates that skewed the overall analysis. This prompted me to cut out areas with a total population below 500, as such sparsely populated areas often exhibit disproportionately high complaint rates due to their small population base. By excluding these outliers, I aimed to ensure that the analysis focused on neighborhoods with more stable and representative population sizes, allowing for fairer comparisons across the city. However, even still, there existed outliers; some boxplots to demonstrate:



This was particularly evident in categories like Public Safety and Quality of Life Complaints, where the bulk of the data was tightly compressed near the bottom of the scale, leaving the blue box in the boxplot barely visible. While important in highlighting exceptionally troubled areas, these outliers make it difficult to interpret broader trends across neighborhoods.

To address this, I decided to clip the values down to a maximum of 400 complaints per 1,000 people; this threshold was based on an examination of the data's distribution and quartile ranges, where the vast majority of neighborhoods fell well below this limit. By capping the extreme outliers, I aimed to create a clearer and more accurate picture of typical neighborhood dynamics, without letting a few exceptional cases dominate the narrative, especially when visualizing the data. This adjustment enabled a better reflection of experiences across the communities, making patterns in the complaint rates more interpretable across different complaint categories.