Quality of Life Analysis in San Antonio

Krishna Panthi, Micheaux Simmons, Frank Yang, AJ Garner
School of Computing, Clemson University
Clemson, SC, USA

Introduction

Urbanization and population growth present significant challenges for cities striving to maintain infrastructure, public safety, and quality of life. In San Antonio, non-emergency 311 service requests play a critical role in addressing diverse community issues, including animal control, graffiti removal, health concerns, sanitation, and property maintenance. These requests serve as a vital channel for residents to report concerns and seek municipal assistance. However, the sheer volume and diversity of these requests underscore the need for a more efficient and proactive approach to service delivery.

Many cities, including San Antonio, currently rely on reactive strategies to address 311 service requests. This can result in delays, inefficient resource allocation, and reduced public satisfaction. A data-driven approach offers the potential to transform service delivery by leveraging historical data to inform decision-making, improve resource allocation, and anticipate future needs.

The Better Future Institute (BFI), in collaboration with the city of San Antonio, seeks to address urban and social challenges related to quality of life through a data-driven strategy. The primary goal is to analyze patterns, correlations, and underlying factors associated with 311 service requests to enhance service delivery, allocate resources effectively, and improve residents' quality of life.

Efficiently addressing non-emergency service requests is crucial for maintaining a city's infrastructure, public safety, and resident satisfaction. Through the application of data science, San Antonio can transition from a reactive to a proactive approach in service delivery. This can lead to (i) Improved Resource Allocation: Identifying areas with high service request volumes can help prioritize resource development; (ii) Enhanced Service Delivery: Understanding the nature and frequency of requests can lead to optimized servicing strategies; (iii) Proactive Problem solving: Predictive modeling can anticipate future service needs, allowing for preventive measures; (iv) Data-Driven Policy Making: Insights obtained from the analysis can inform policy decisions related to urban planning, public health, and safety. The potential impacts extend beyond San Antonio, serving as a model for other rapidly growing cities facing similar challenges.

The primary objectives of this project are to
(i) Conduct a comprehensive analysis of San Antonio's 311 service request data in conjunction with socioeconomic and demographic information.
(ii) Develop predictive models to forecast 311 report trends, including frequency, type and location analysis.

(iii) Supplement the request data with demographic data from the US census bureau and do combined analysis to find relations between the service requests and demographics.
(iv) Visualize the results for effective understanding of the results by the stakeholders.

By achieving these objectives, the project will provide a data driven assessment of community needs and challenges, enabling stakeholders to identify and prioritise areas requiring immediate attention; Understand the relationship between socioeconomic factors and quality of life issues; Develop targeted interventions and policies; Monitor the effectiveness of service delivery and make necessary adjustments; Predict future service needs and allocate resources proactively; Provide an approximate resolve time to the service requester.

Methodology

Data Source and Description

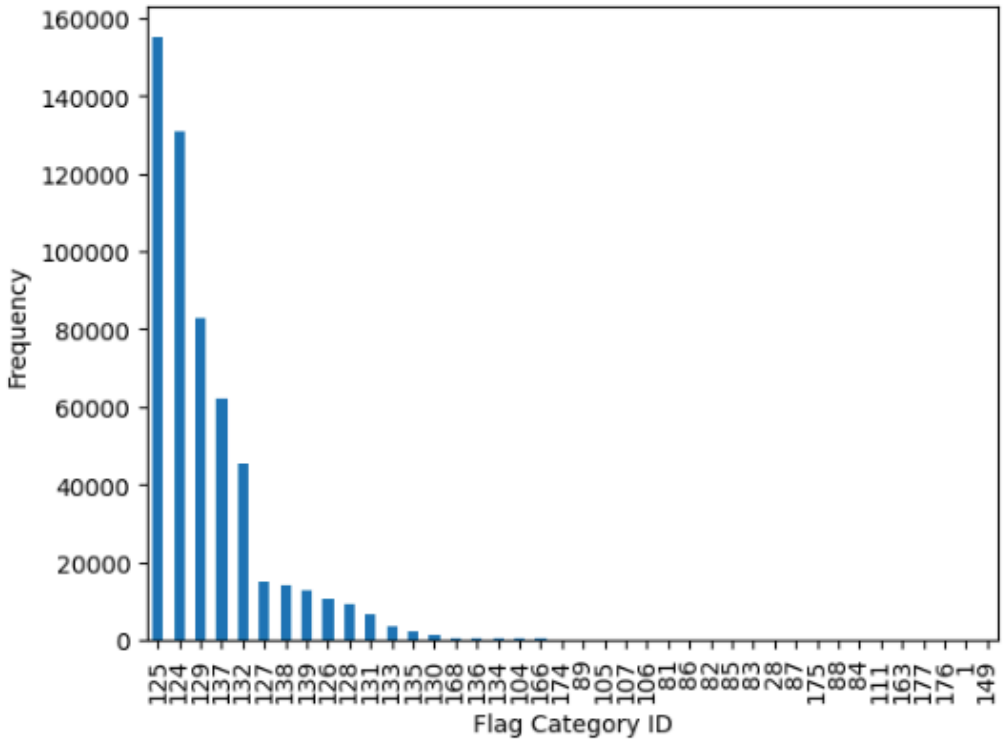


Figure 1. Graph showing frequency of service requests based on event category. We can observe that some events are far more frequent than others.

Primary Data Source

San Antonio is a city in South Texas and the county seat of Bexar County. The primary data source is the City of San Antonio's 311 Customer Service, which was provided to us by Better Future Institute (BFI). The data was provided as 14 individual CSV files, covering 14 months, from 2023-09-01 to 2024-10-29. Bexar County has 375 census tracts and majority of them lie in San Antonio. The service request data was supplemented with population, demographic and socio-economic data of Bexar County from the US census bureau. The population specific data was obtained per census tract. Overall there were 553,887 service request records with 44 columns. The Names of each of the columns are provided in the Appendix.

Secondary Data Source

We obtained population data, demographic data and socio-economic data per each census tract from the US census bureau. We obtained this information as per the suggestions made by the project stakeholders.

- No. College Educated people [2],
- Median Income Per Household [4],
- Total Population [5],
- No. of children aged less than 5 yr [5],
- No. of people aged 65 or higher [5],
- Old Age Dependency Ratio: Percentage of people dependent because of old age [5],
- Child Dependency Ratio: Percentage of dependent children [5],
- Age Dependency Ratio: Combination of above two [5],
- Median Age for the census tract [5],
- UnEmployment Rate in the census tract [3],
- Sex Ratio: No. of males per 100 females [5],
- Poverty rate in the census tract [3].

Data Preprocessing

All 14 csv files containing service request data were merged together to obtain a single combined csv file. Many of the columns in the primary data source were not useful. They were either incomplete or contained trivial information, such as having only a single unique value. Out of all the columns, the following were selected for analysis: Event Category (Flag Category ID), Event SubCategory, Date Created, Date Closed, and Location Coordinates.

Feature Engineering

From the Date Created column, year, month, weekday, and hour were extracted. Response Time in hours was calculated by subtracting the Date Created from the Date Closed. Additionally, we categorized the events into census tracts. Using the latitude and longitude coordinates for each event, we obtained

shapefiles containing census tracts for Bexar County (the county that includes San Antonio) from ArcGIS Open Data [1] as a GeoJSON file. We then mapped the event coordinates to the corresponding census tract coordinates to determine the census tract for each event. For regression analysis, we computed more features which include (i) Average response time by census tract; (ii) Average response time by category; (iii) Average response time by subcategory. Overall we have 40 categories, 296 subcategories and 375 census tracts to perform analysis on. Of all the events, about 28, 225 recent events don't have a closed date, so we drop those records when doing analysis related to response time.

Data Normalization

For columns number of college educated, number of children aged less than 5 yr, and number of people aged 65 or higher, the percentage of representation per total population are calculated for each census tract.

$$X = \frac{X}{Total\_Population}$$

Response Time and Median Income per Household are log normalized to reduce the effect of outliers.

$$X = log(X)$$

Weekday, hour of day and month of year are cyclic data, so cyclic encoding was used. Basically sine and cosine were calculated for each of them, using the following equations.

$$X = sine(\frac{2 \times \pi \times X}{max(X)}) + cosine(\frac{2 \times \pi \times X}{max(X)})$$

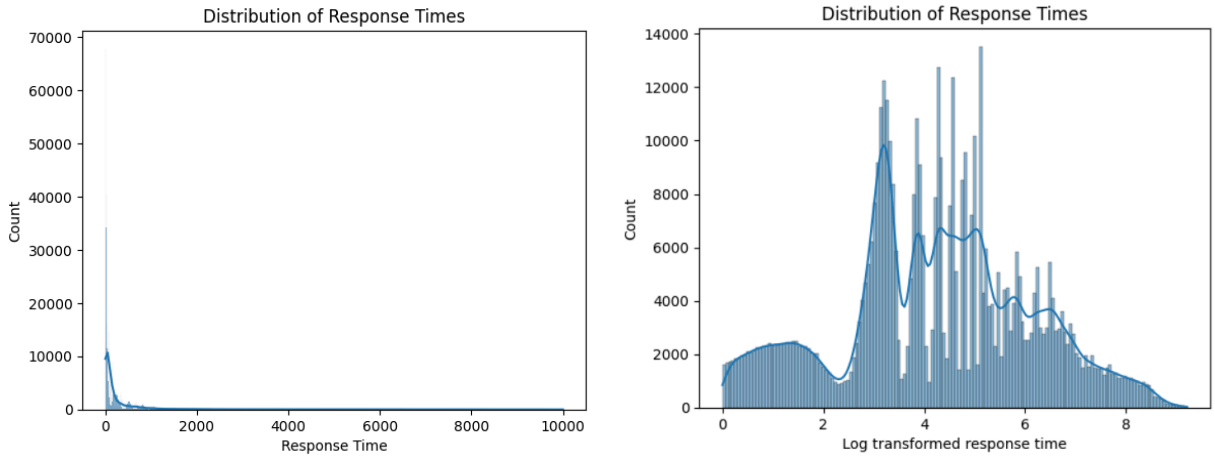


Figure 2. Response time before log transformation vs. after log transformation, showing how log normalization helps reduce effects of outliers.

Correlation Analysis

To find the relationships between the service requests and the population demographics, we calculate the correlations. We use different types of correlation methods.

1. Pearson’s correlation (finds linear relationship)
2. Spearman's Rank Correlation
3. Kendall's Tau Correlation
4. Mutual Information

We used multiple correlation methods because it’s possible that the relationship between variables might not be perfectly collinear as expected by say Pearson’s correlation. Spearman’s rank correlation is good if the relationship is non linear but monotonic (if one variable increases or decreases and so does the other). Mutual Information technique is useful if the relationship is not monotonic. Like correlation might exist after a certain threshold value. We also compute p-values for the first 3 correlation types. For analysis we consider the correlation type which has the highest value.

Regression Analysis

We perform regression analysis to predict response time based on various features. The target feature is response time. While the source features chosen were

1. Average response time by census tract
2. Average response time by category
3. Average response time by Subcategory
4. Weekday
5. Hour of the day
6. Month of the year
7. Percentage of College Educated for that census tract
8. Median Income Per household for that census tract

Response times are log normalized. Weekday, hour of day and month of year are cyclic data, so cyclic encoding is used. Rest of the variables are normalized using Min max normalization, i.e. they are mapped between 0 and 1. Note that response time was also log normalized. We used two models to perform regression analysis.

Linear Regression

The straightforward linear regression was used as the baseline method to predict response time after normalisation. We divided the data in the ratio of 80:20 for training and testing respectively.

We also implemented linear regression with Principal Component Analysis (PCA) just to see if features were correlated or not and to see how PCA affects the results. PCA was implemented while retaining 95% variance as well as with 90% variance.

Random Forest Regression

We used a random forest regressor as the alternative method to predict the log normalized response time, like in linear regression. Same source features were used. We selected random forest regression because it can take into account the non linear relationship between the features and it's also robust to overfitting. We identified there is a non linear relationship by observing the results obtained after correlation analysis. The data was split in the ratio of 80:20 for training and testing.

Following values of hyperparameters were tuned and tested using random search. All experiments were conducted keeping random seed at 42.

N_estimators (no. of trees): 50, 80, 100, 150, 200
Max_depth (depth of tree): 5, 10, 20, default
Min_samples_split (minimum samples required to split a node) : 5, 10, 15
Max_features: sqrt, log2, None
Min_samples_leaf (minimum samples required at a leaf node): 1,2,3,4,5
Bootstrap = True/False

Evaluation Metrics:

Mean Squared Error (MSE):

MSE measures the average squared difference between the predicted values and the actual values. It gives higher weight to larger errors due to the squaring operation.

$MSE = (1/n) * \Sigma(y_i - \hat{y}_i)^2$

n = number of data points
y_i = actual value
$\hat{y}_i$ = predicted value

Lower MSE values indicate better model performance, as it signifies smaller overall errors. It is sensitive to outliers because of the squared term.

Mean Absolute Error (MAE):

MAE measures the average absolute difference between the predicted values and the actual values. It treats all errors equally, regardless of their magnitude.

$MAE = (1/n) * \Sigma|y_i - \hat{y}_i|$

Lower MAE values indicate better model performance. MAE is more robust to outliers compared to MSE because it doesn't square the errors.

R-squared (R^2)/ Coefficient of Determination:

R-squared represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

$R^2 = 1 - (SS_{res} / SS_{tot})$
SS_{res} = Sum of squares of residuals (unexplained variance)
SS_{tot} = Total sum of squares (total variance in the data)

R^2 ranges from 0 to 1. An R^2 of 1 indicates that the model perfectly explains the variance in the data. An R^2 of 0 indicates that the model does not explain any of the variance. Higher R^2 values generally suggest a better fit, if the model is not overfitting.

Results

Correlation Results

On the first step we calculate correlation and p-values between all the demographic variables we have, and the total number of events reported per 1,000 people in each census tract. And the results are presented in a heat map in figure 3.

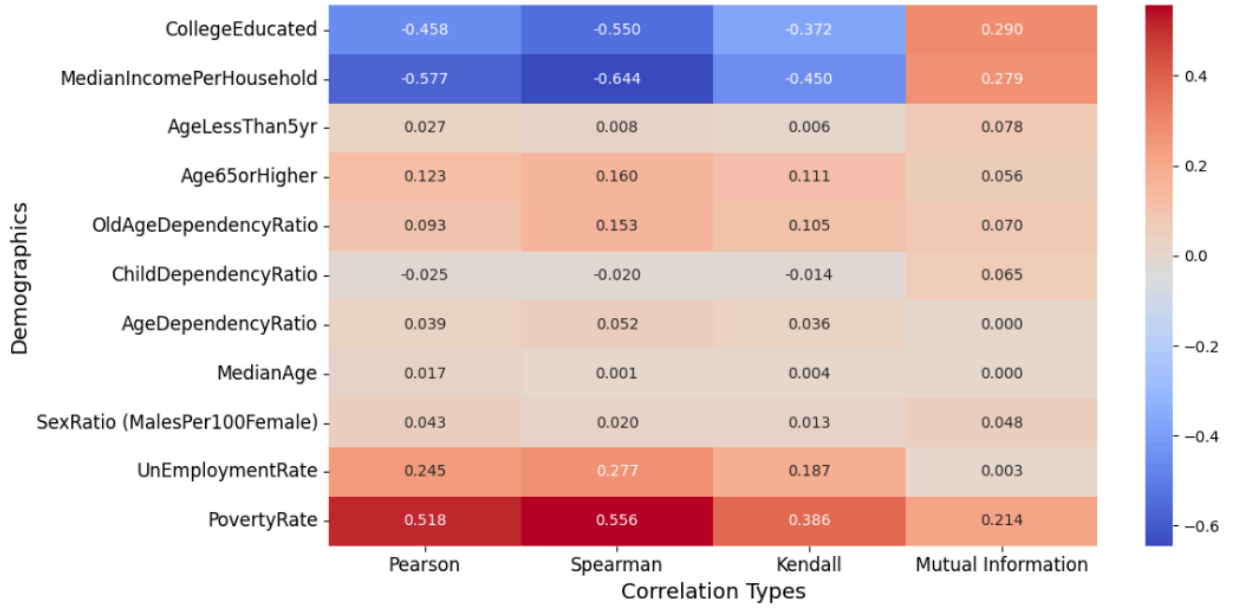


Figure 3 (a). Heatmap showing correlation between number of events reported per 1000 people and various demographic variables for each census tract.

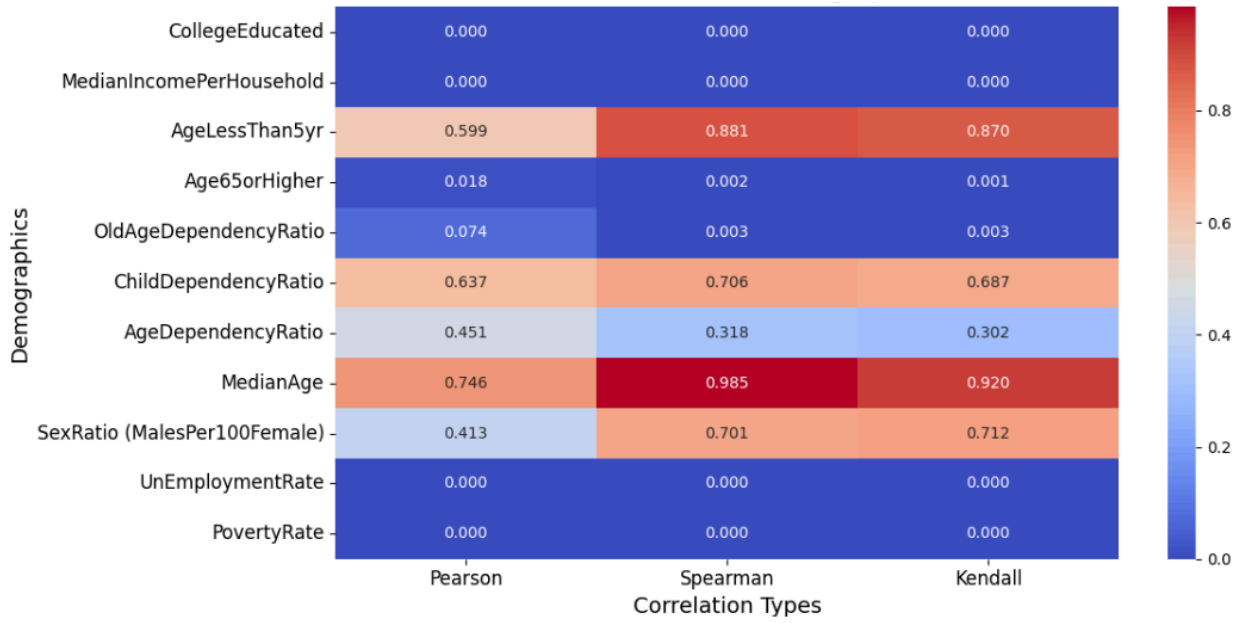


Figure 3 (b). Heatmap showing p-values for each of the demographic variables against the event count per 1000 people.

From figure 3, we observe a moderate negative correlation (-0.458, -0.550, -0.372) and near-zero p-values for CollegeEducated. Similarly, we observe a strong negative correlation (-0.577, -0.644, -0.450) and near-zero p-values in areas with higher household incomes. We observe strong positive correlation (0.518, 0.556, 0.386) and near-zero p-values in areas with higher poverty rates. There is some relationship (0.245, 0.277, 0.187) with the unemployment rate as well, as shown by near-zero p-values. There is also weak

positive correlation (0.123, 0.160) with the presence of an aging population (> 65 years of age). This is further supported by a p-value of 0.001, which means we cannot ignore the hypothesis that there is some relationship here.

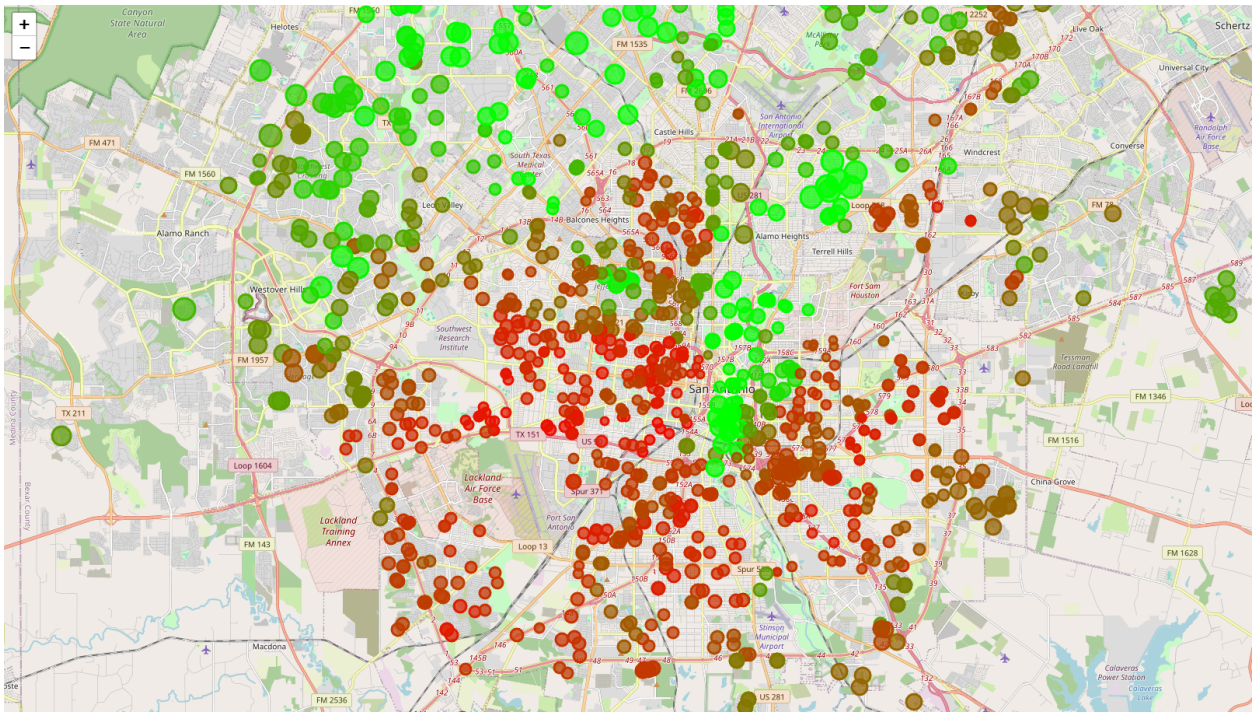


Figure 4. The map visualizes data about event creators in San Antonio, highlighting two key socioeconomic indicators: the percentage of college-educated individuals and the median income per household within each census tract. The data used to create this graphic has been truncated in order to provide better visual clarity and remove clutter.

In the appendix section, we provide similar results but for different categories and some sub categories of events.

Do population, demographics or socio-economic conditions influence response time?

In the next step, we investigate if there's bias when investigating or responding to these events. That is, is there a relationship between closing time of events and demographics.

First thing to consider is that, vast majority of events get resolved in less than a day. So, we calculate the response time in hours, normalize it using logarithm to decrease the effect of outliers. We then only consider the events that take more than 1 hour to resolve. We then calculate the mean logarithmic time for each census tract since we already know poor communities have a larger number of events reported than rich or educated communities. And finally we calculate correlation with population demographics. The results are presented in heatmaps in figure 5.

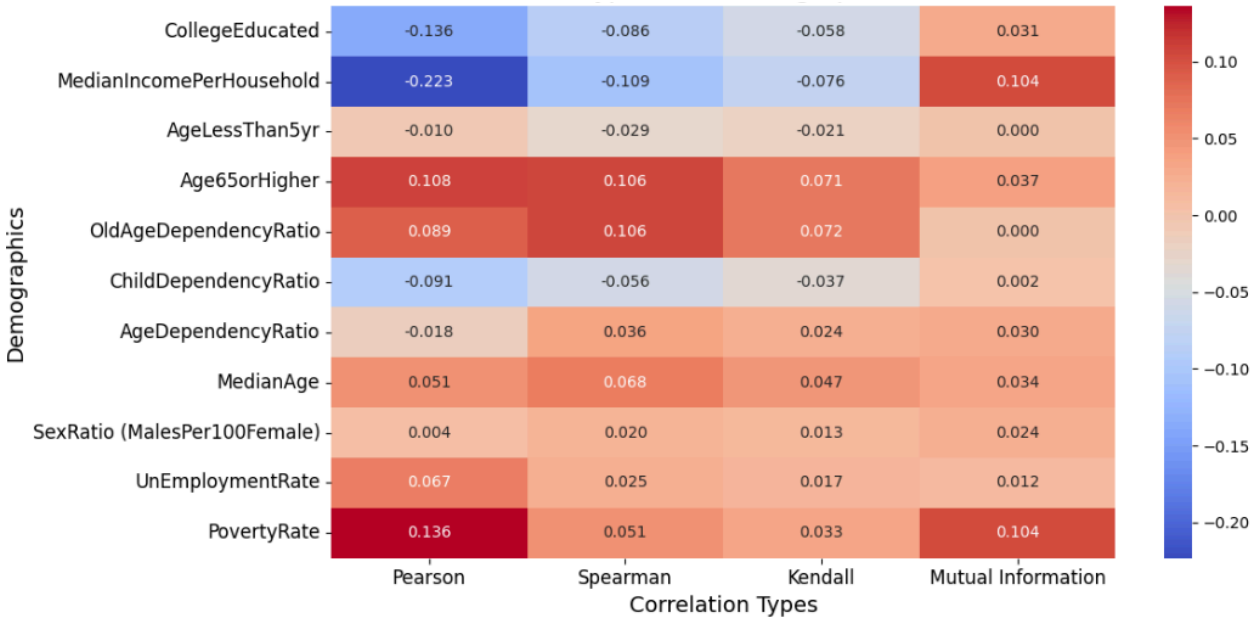


Figure 5 (a). Heatmap showing correlation between average response time per census tract and various demographic variables for each census tract.

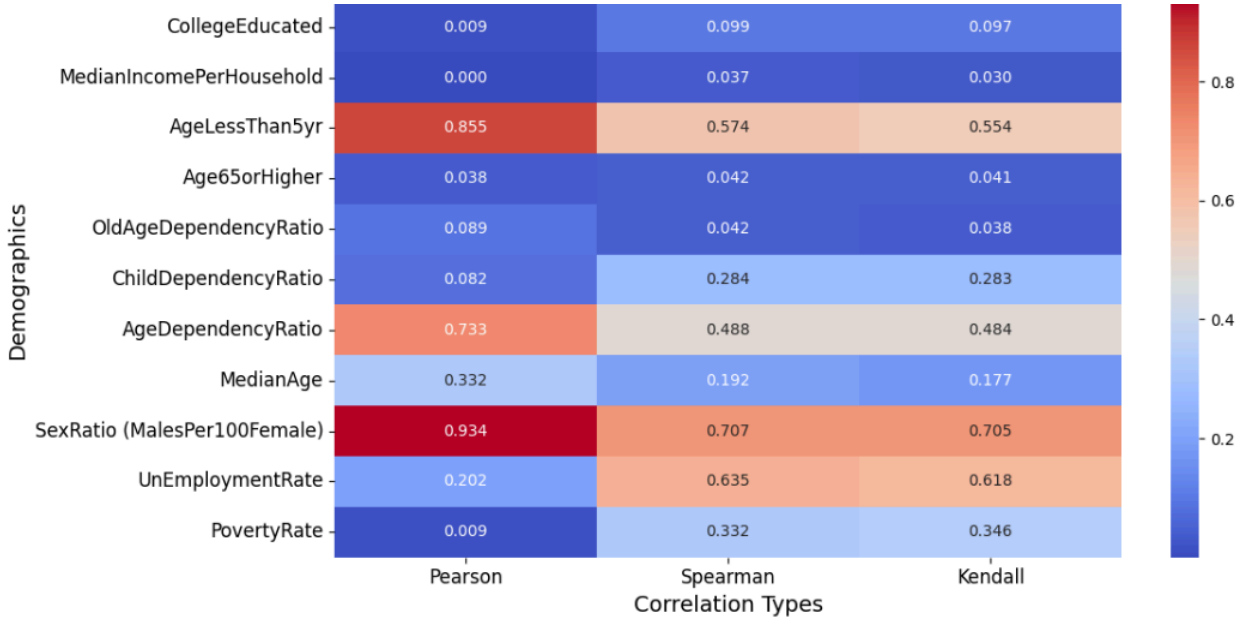


Figure 5 (b). Heatmap showing p-values for correlation between average response time per census tract and various demographic variables for each census tract.

The results from figure 5 show that there is slight negative correlation(corr = -0.223, p-value = 0.000) as given by pearson's correlation with Median Income Per Household. Similarly, there is a weak positive correlation with poverty rate (corr = 0.136, p-value=0.009) and response time.
The correlations, though statistically significant, are weak-to-moderate, suggesting that other factors (e.g., event type, population density, or reporting frequency) also might influence response time.

Regression Results

Linear regression results

Model Evaluation Metrics:
Mean Squared Error (MSE): 2.728
Mean Absolute Error (MAE): 1.2457
R-squared (R2): 0.3692

Coefficients for each feature

Feature	**Coefficient**
Avg by Census Tract	0.425924
Avg by Category	0.00425
Avg by Subcategory	0.993226
weekday_sin	-0.120438
weekday_cos	-0.117486
hour_sin	0.047899
hour_cos	0.081449
month_sin	0.149232
month_cos	0.103102
CollegeEducated	-0.06872
MedianIncomePerHousehold	-0.01103

The R-squared value of 0.3692 suggests that approximately 37% of the variance in log-normalized response time is explained by the chosen features. The model exhibits reasonable accuracy with an MAE of 1.2457, indicating that, on average, the model's predictions deviate by this amount from the actual log-normalized response times. The MSE of 2.728 provides a measure of the average squared difference between predicted and actual values.

Linear regression with PCA

In total we had 11 features. After applying PCA, features were reduced to 8 total. When linear regression was applied with 8 principal components following results were obtained.

PCA + Linear Regression Model Metrics:
Mean Squared Error (MSE): 2.7424

Mean Absolute Error (MAE): 1.2486
R-squared (R2): 0.3658

From the results, we can see that the R-squared value obtained after PCA is slightly lower than without using PCA.

Random Forest results

Best value of hyperparameters are presented below:

N_estimators (no. of trees): 200
Max_depth (depth of tree): default
Min_samples_split (minimum samples required to split a node) : 5
Max_features: sqrt
Min_samples_leaf (minimum samples required at a leaf node): 3
Bootstrap = False

Random Forest Regression Model Metrics:
Mean Squared Error (MSE): 2.08958
Mean Absolute Error (MAE): 1.00488
R-squared (R2): 0.5168

Feature Importance

Feature	**Importance**
Avg by Subcategory	0.460201
Avg by Category	0.103161
Avg by Census Tract	0.082445
CollegeEducated	0.071492
MedianIncomePerHousehold	0.071341
hour_cos	0.05081
hour_sin	0.038872
month_cos	0.038203
month_sin	0.035653
weekday_sin	0.029105
weekday_cos	0.018717

The random forest regression model, trained on an 80/20 train-test split and optimized using random search across various hyperparameters, demonstrates a moderate ability to predict log-normalized response time, achieving an R-squared of 0.5168. This value is much higher than 0.3692 we obtained from the linear regression model. We observed that increasing the number of trees to 200 yielded the best performance.

Discussion

Correlation Analysis

From correlation analysis we observe that census tracts with higher College Educated population have on average fewer events reported. Similarly, fewer events are reported from census tracts with higher median household incomes. It also suggests that more events are reported from census tracts with higher poverty rates and same goes with unemployment rate. A weak positive correlation suggests that more events than average are reported from census tracts with a strong presence of an aging population (> 65 years of age).

Additionally, we observe that Spearman's correlation has higher values than other correlation metrics, indicating that the relationship is non-linear and monotonic. The value of Mutual Information lies between 0 and infinity, showing that the percentage of college-educated individuals, median income per household, and poverty rate have a direct relationship with the number of events in a census tract.

Similarly, from correlation of response time and demographics we observe that if one reports an event from a census tract with higher median household income, the event gets resolved faster than average. Similarly if one reports an event from a census tract with a high poverty rate, it might take more time than average to resolve. Note that, there is weak correlation in this case, so other factors like number of events might also influence response time.

Regression Analysis

The linear regression model, designed to predict log-normalized response time, demonstrates a moderate level of predictive capability. Notably, 'Average Response Time by Subcategory' exhibits the strongest positive influence on response time, followed by 'Average Response Time by Census Tract'. Interestingly, socioeconomic factors like 'College Educated' percentage and 'Median Income Per Household' show a negative association, albeit weaker, suggesting higher socioeconomic status might correlate with slightly lower response times which was also shown by the correlation study above. The cyclic features (weekday, hour, month) display varying degrees of influence, with month showing a relatively larger impact compared to weekday and hour.

On applying PCA 8 components were retained out of 11 which means there's no significant multicollinearity in the data. We also applied PCA with 90% variance, results were similar while the

number of components reduced to 7. Going below 90%, R-squared value was affected as it started reducing.

We observed a much higher r-squared value from the random forest method compared to the linear regression model which means the random forest's ability to capture non-linear relationships is beneficial for this prediction task. Similarly, performance improved when the number of decision trees were increased suggesting that an ensemble approach improves predictive power. The model's feature importance analysis reveals that 'Avg by Subcategory' is by far the most influential predictor, followed by 'Avg by Category' and 'Avg by Census Tract'.

Limitations

The analysis is based on 14 months of data, which means we are not able to capture longer term based trends. We are able to find some relationships between reported events and demographic data, but it may not be enough to establish a causal relationship, so more analysis is needed. This analysis relies on the reported 311 events, which may not capture all quality of life issues in the city. There could be reporting biases based on access to technology which this analysis doesn't capture. We didn't have data about the resolution quality to properly assess the resident satisfaction. Similarly we don't have information about how the cases are handled to be able to identify possible resolution bottlenecks.

Model Recommendation

Based on the results, the random forest regression model is recommended for predicting response time due to its superior predictive accuracy (higher R2, lower MSE and MAE) compared to the baseline linear regression model. It is better suited to capture the non-linear relationships present in the data. For understanding the relationship between demographics and event reporting, the correlation analyses combined with geospatial visualizations is highly effective.

Conclusion

The analysis provides strong evidence that socioeconomic factors are significantly associated with the frequency of non-emergency 311 service requests in San Antonio. Areas with higher poverty rates, lower education levels, and lower median incomes tend to experience a higher volume of reported quality of life issues. There is also a weak but statistically significant relationship between demographics and response time, suggesting potential disparities in service delivery. Specifically, residents in higher-income areas might experience slightly faster response times, while those in poorer areas might experience slightly slower response times. The random forest model offers a promising approach for predicting response time. These findings can inform data-driven strategies to improve service delivery, allocate resources more effectively, and address quality of life disparities in San Antonio. Further research, incorporating more data sources and addressing the limitations mentioned, will be important for developing a more comprehensive understanding of the issues in the city.

References

[1] https://gis-bexar.opendata.arcgis.com/datasets/

[2] U.S. Census Bureau. "Educational Attainment." American Community Survey, ACS 5-Year Estimates Subject Tables, Table S1501, 2022, https://data.census.gov/table/ACSST5Y2022.S1501?q=san antonio&g=050XX00US48029$1400000. Accessed on November 19, 2024.

[3] U.S. Census Bureau. "Employment Status." American Community Survey, ACS 5-Year Estimates Subject Tables, Table S2301, 2022, https://data.census.gov/table/ACSST5Y2022.S2301?q=san antonio&g=050XX00US48029$1400000. Accessed on November 19, 2024.

[4] U.S. Census Bureau. "Median Income in the Past 12 Months (in 2022 Inflation-Adjusted Dollars)." American Community Survey, ACS 5-Year Estimates Subject Tables, Table S1903, 2022, https://data.census.gov/table/ACSST5Y2022.S1903?q=san antonio&g=050XX00US48029$1400000. Accessed on November 19, 2024.

[5] U.S. Census Bureau. "Age and Sex." American Community Survey, ACS 5-Year Estimates Subject Tables, Table S0101, 2022, https://data.census.gov/table/ACSST5Y2022.S0101?q=san antonio&g=050XX00US48029$1400000. Accessed on November 19, 2024.

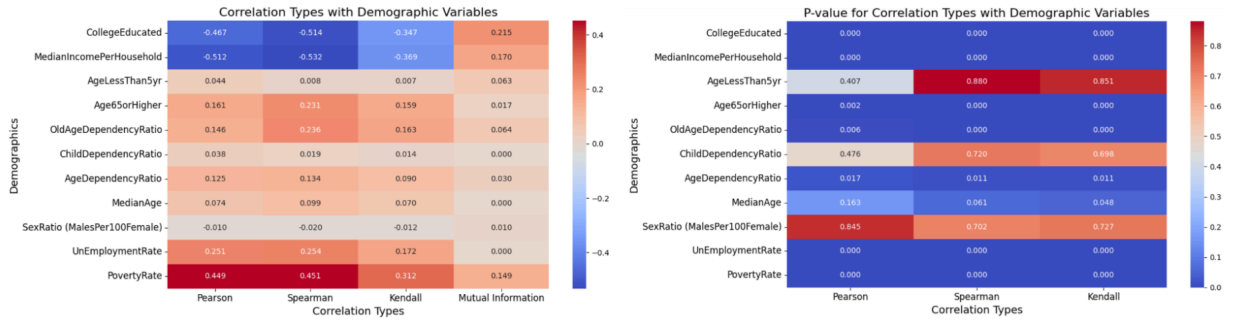
Appendix

All columns provided in the raw dataset are;

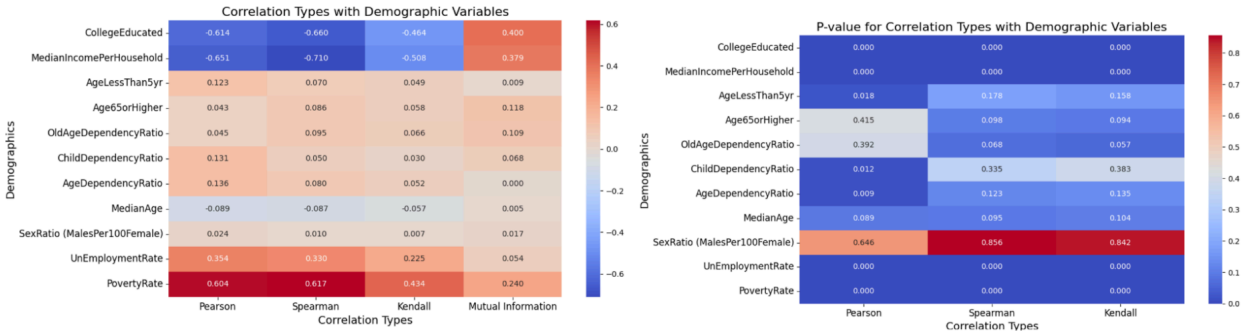
'ID', 'Flag Subcategory ID', 'Files Open', 'Priority', 'Status', 'Users Assigned Ids', 'Labels Text', 'Files Close', 'Folio', 'Organization ID', 'User Added ID', 'User Added Name', 'User Added Lastname', 'Flag Category ID', 'Flag Category Name', 'Flag Subcategory Name', 'Description', 'Location', 'Date Created', 'Users Assigned', 'V', 'Date Closed', 'User Closed ID', 'User Closed Lastname', 'User Closed Name', 'Location Address', 'Close Comment ID', 'Dislike Client Count', 'Dislike User Count', 'Lagan Details', 'Like Client Count', 'Like User Count', 'Location Address General', 'Spam', 'Survey Answers', 'User Added Type', 'Client Added ID', 'Client Added Name', 'Client Added Lastname', 'Location Geocode', 'Like Client Ids', 'Like User Ids', 'Old311Id', 'Migration'

Next, we do similar analysis for different categories of events and present some results

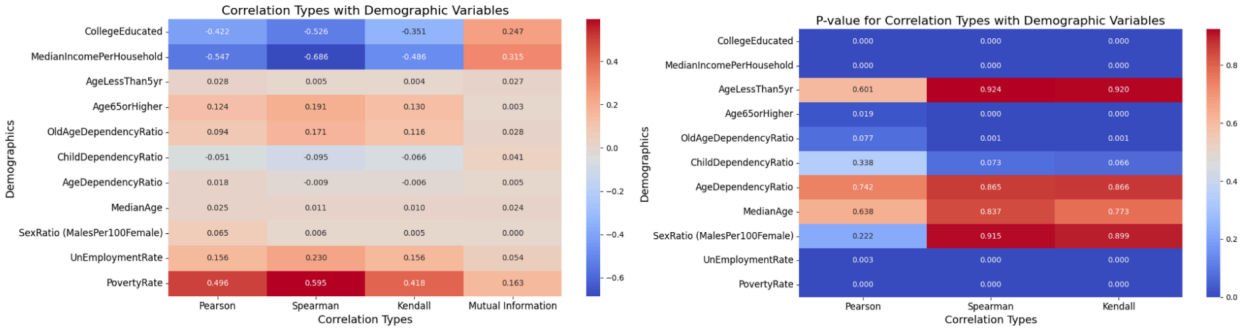
1. **Garbage/Dumping**



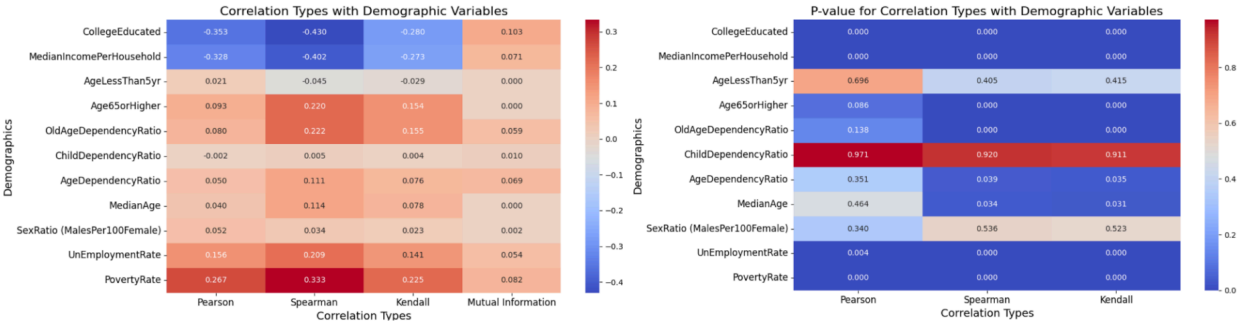
2. **Animals**



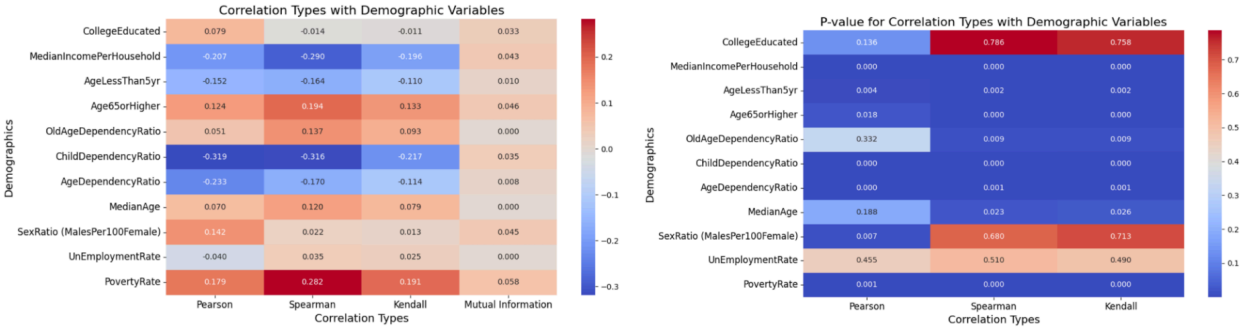
3. Property Maintenance



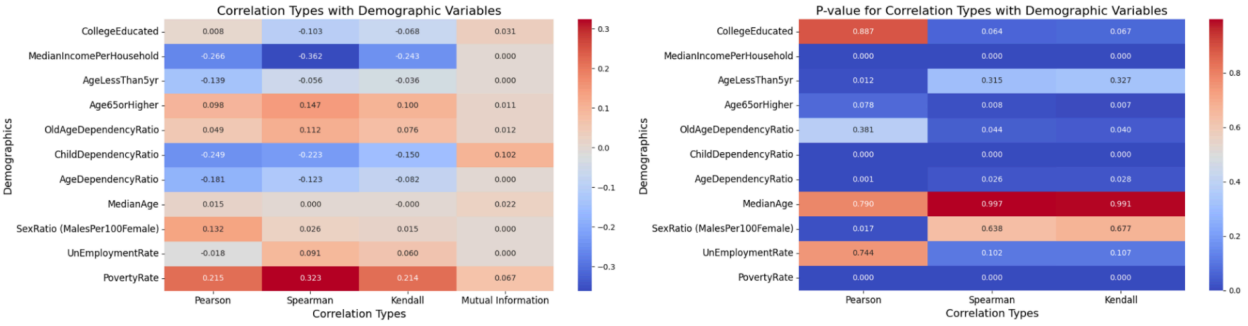
4. Solid Waste Services



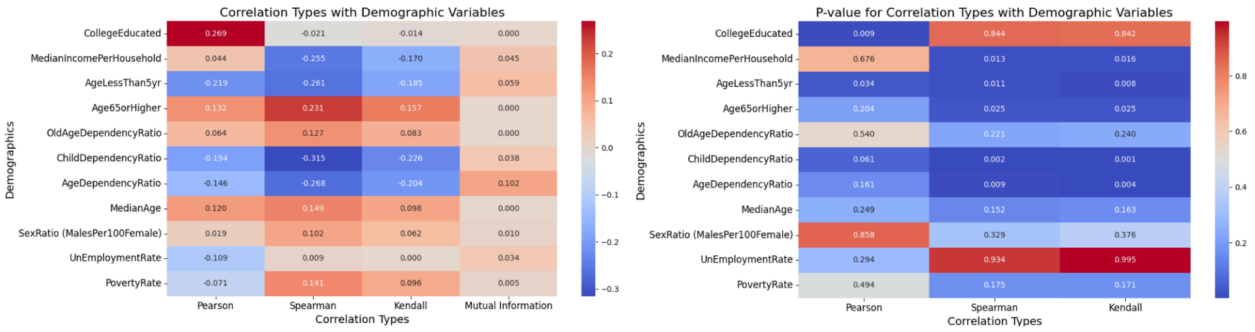
5. Traffic signals and signs



6. Heath and Sanitation



7. Scooters/Dockless vehicles



Discussion: In no. 5 and 6, there's a negative correlation with child dependency ratio which is a bit surprising. Similarly 7 has a strong positive correlation with the college educated population. These results are different from all other categories.

Conclusion of this section is that most issues arise in areas where people are poor and less educated. Also, many issues are reported from places where the number of aging population is higher.