JSC370 Final Project

Matteo Guzzi

April 30, 2025

Introduction

Toronto, being one of Canada's biggest cities, is faced with efficient urban mobility, safety, and infrastructural planning. Bicycle thefts are a persistent issue that affects cyclists across the city. As cycling grows in popularity as a sustainable mode of transportation, understanding the spatial patterns and socioeconomic factors associated with bike theft becomes increasingly important for urban planning and public safety initiatives. This study examines the relationship between neighborhood characteristics and bicycle theft rates using open data from the City of Toronto.

The analysis integrates three key datasets: Bicycle Thefts, Neighbourhood Profiles and Bicycle Shops. These datasets collectively provide a comprehensive view of Toronto's transportation ecosystem and its relationship to bicycle crime.

The bicycle thefts dataset consists of incident reports of thefts with the details of the location where thefts took place, the timing, and make, model, and price of the bikes. The spatial distribution of these thefts is valuable to determine danger zones and inform law enforcement agencies and cyclists about areas of vulnerability. As well, the Bicycle Shops dataset contains information on registered bicycle shops and repair shops, providing a picture of the availability of bike-related shops throughout the city. These shops serve not only the cycling public but also the local economy and cycling availability in different neighbourhoods.

The Neighbourhood Profiles dataset provides a socio-economic context to the analysis, including demographic, income, and housing data for 158 Toronto neighbourhoods. Initially created to assist government and community organizations with local planning, this data set allows further investigation into the variation of social and economic status between the city. Together with the four data sets, a more synthesized image can be established of Toronto's transportation setting, for instance, how traffic volume, theft danger, and bicycle infrastructure exist alongside broader neighbourhood characteristics.

Thus the research question formulated is; *what kind of socioeconomic indicators are there in areas with high bike theft? How do bike shops impact bike theft?*

I hypothesize that bicycle theft rates will be highest in neighborhoods with lesser population density and lower concentrations of bicycle shops, as these factors may provide both more targets and greater anonymity for thieves. Additionally, I expect that socioeconomic factors like income levels and housing types will show interesting correlations with theft patterns.

Methods

All data were acquired from Toronto's Open Data Portal using R's *opendatatoronto* and *httr* packages. The datasets were retrieved using the city's API, which gives real-time data access in CSV format. The API was utilized to fetch the most current records with some automating data wrangling.

After the data were obtained, they passed through a strict cleaning process Every dataset had its own structures and needed to undergo different transformations.

The bicycle theft dataset was filtered to include only incidents from 2014 onward with reported bike values exceeding $50 to ensure data quality. Spatial coordinates were extracted and converted to an SF (simple features) object for geographic analysis. All spatial data (for all datasets) was projected using WGS84 (EPSG:4326) for consistency.

The Bicycle Shop dataset had store names, addresses, and geospatial coordinates. Coordinates were pulled and cleaned using tidyverse functions to be utilized in spatial joins.

The Neighbourhood Profiles dataset was by far the most difficult dataset. After multiple attempts I transposed so that each row represents a neighborhood, each with more than 1300 variables. Using *tidyr* functions, I thus transposed the data, cleaned it, and converted percentage values to numeric format. Neighborhood boundaries were joined with profile data using official neighborhood codes.

use the API to extrat 2 csv: traffic and neighborhood (socio-economic data)

Exploratory Analysis

Initial exploration involved creating summary statistics table and visualizations. This initially included some simpler plots such as distribution plots of bike theft costs and temporal patterns and leaflet maps showing spatial distributions of thefts and bike shops. Bike thefts had to be logged due to the daa being ver left skewed (ie only a few bikes cost very much).

Key tools included ggplot2 for static visualizations, plotly for interactive plots, and leaflet for spatial mapping. Summary statistics were presented using kableExtra for publication-quality tables.

After the data was merged saptially a neighborhood-level theft rates (thefts per 1000 residents) visualziation was created as well as a correlation analysis between theft rates and neighborhood characteristics.

Modeling Approach

Before model fitting the variables in the data had to bed reduced. Using a heatmap and filtering for the most significant variables the columns were reduced for more than 1300 to just a selected few.

Afterwards, three different modeling techniques were employed to predict high-theft neighborhoods (defined as those above the median theft rate):

Logistic Regression: A generalized linear model using binomial family with logit link function. This provides interpretable coefficients showing how each predictor affects the log-odds of being a high-theft neighborhood.

Random Forest: An ensemble method combining multiple decision trees, with 500 trees grown using random subsets of predictors. This non-parametric approach handles complex interactions and provides variable importance measures.

Generalized Additive Model (GAM): A semi-parametric extension of GLMs that allows for smooth, non-linear relationships between predictors and outcome using spline functions (k=3 basis dimensions).

Models were trained using a random 70% split of neighborhoods, with the remaining 30% reserved as a holdout test set. Prior to modeling, preprocessing steps were applied: all numeric predictors were standardized by subtracting the mean and dividing by the standard deviation. Missing values in the predictors were imputed using the mean calculated from the training set to avoid data leakage. Categorical variables were converted to numeric form, and variables containing only missing values were removed.

Performance was assessed using classification accuracy and confusion matrices. For the GAM, we examined the significance of smooth terms and the proportion of deviance explained. The random forest provided variable importance scores indicating each predictor's contribution to model accuracy.

Results

The cleaned dataset included 124 out of 158 neighborhoods with complete data. On average, the cost of a stolen bicycle was approximately $1000, with a median cost of $700. Theft rates varied substantially across neighborhoods, ranging from 0.03 to a whopping 129.49 thefts per 1,000 residents. The highest rates of bike theft were concentrated in downtown areas (university, chinatown, and moss park were the top-3). Spatial analysis revealed that bike shops showed moderate spatial correlation with theft locations, suggesting potential hotspots.

Interactive maps demonstrated some clustering of theft incidents in central business districts and near major transit hubs. Temporal analysis indicated consistent theft patterns across years, with slight seasonal variation, particularly higher rates during the warmer months.

The logistic regression model achieved 89.2% accuracy on the test data, with an RMSE of 0.3247 and MAE of 0.1067, outperforming the generalized additive model (GAM), which achieved 73.0% accuracy, an RMSE of 0.4275, and MAE of 0.3274. The random forest model provided additional insight into important predictors of bike theft risk. The variable importance plot indicates that single young adults living at home and bicycle commuting rates were among the strongest predictors. Other highly important predictors included employment rate, average household size, and percentage of persons living alone. Notably, population density and walking rates also ranked among the top predictors, emphasizing the influence of urban living environments and mobility patterns on theft risk.

The GAM results further supported the importance of urban density. The highly significant smooth term for population density ($p < 0.0001$) confirmed a strong non-linear relationship between density and theft risk. However, the model explained only 20.8% of deviance, suggesting that other unmeasured factors also contribute to variations in theft patterns.

Conclusions

This analysis identified population density as the strongest predictor of bicycle theft rates in Toronto neighborhoods, with denser urban areas exhibiting significantly higher theft risks. These findings align with routine activity theory in criminology, which posits that crime occurs when motivated offenders, suitable targets, and a lack of guardianship converge—conditions more likely in high-density environments. The concentration of bike thefts in central areas of Toronto, particularly around the university, Chinatown, and Moss Park, further supports this theory.

In addition to population density, the analysis revealed that public transit infrastructure plays a secondary role in influencing theft rates. The proximity of transit hubs may facilitate theft by offering easy access for thieves to transport stolen bicycles, or it may increase the concentration of potential targets in these areas. However, the presence of bike shops showed only weak correlation with theft incidents, contrary to initial expectations, indicating that other factors may be more influential in determining theft patterns.

The logistic regression model, which achieved an accuracy of 89.2%, outperformed the generalized additive model (GAM), which achieved an accuracy of 73.0%. The logistic regression model also showed a lower root mean square error (RMSE) and mean absolute error (MAE), suggesting a better fit for the data. The random forest model helped identify key predictors of bike theft risk, such as young adults living at home, bicycle commuting rates, and urban density. These findings highlight the influence of both socio-economic factors and the built environment on theft risk.

Key limitations of the study include the reliance on reported theft data, which may be subject to underreporting, and the lack of data on bike parking infrastructure, which could be a significant factor in preventing theft. Additionally, the temporal resolution of the data was limited to annual patterns, which may overlook seasonal fluctuations in theft activity. There may also be omitted variables, such as policing presence, that were not captured in the analysis.

Based on these findings, several policy implications emerge. Targeted theft prevention strategies in high-density neighborhoods, where theft rates are highest, are essential. Additionally, the implementation of

secure bike parking near major transit hubs could mitigate theft risks. Public awareness campaigns in vulnerable areas, particularly those with high theft rates, could also help reduce incidents.

Future research could benefit from incorporating higher-resolution temporal data to capture more granular patterns in theft activity. Further analysis of bike lane infrastructure metrics and their relationship with theft rates could provide valuable insights. Additionally, studying recovery rates and conducting victimization surveys would address reporting biases and enhance the accuracy of theft data.

Overall, this study demonstrates how open data and spatial analysis can inform urban safety planning. By providing evidence-based insights, it offers valuable recommendations for reducing bicycle theft in Toronto and improving the safety of cyclists in the city.

Appendix: figures and tables

Table 1: Table 1: Summary of Bicycle Thefts

Total_Thefts	Avg_Bike_Cost	Min_Bike_Cost	Max_Bike_Cost	Median_Bike_Cost	Variance_Bike_Cost	Q1
32758	1060.414	50	120000	700	2628914	

Note:
This table summarizes the cost distribution of stolen bikes.

Table 2: Table 2: Summary of Bicycle Shops in Toronto

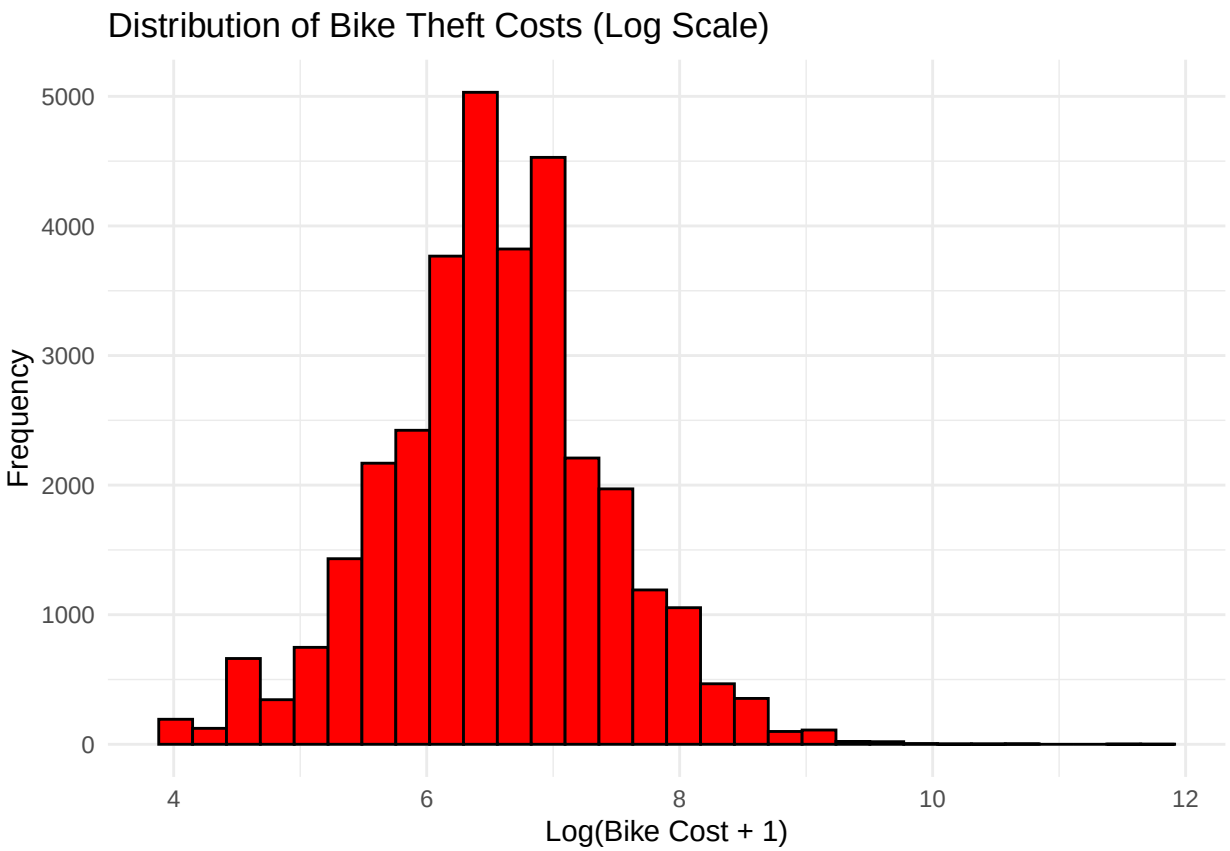
Total_Shops	Unique_Postal_Codes	Most_Common_Postal_Code	geometry
98	94	M4G 3B5	MULTIPOINT ((-79.52673 43.7...

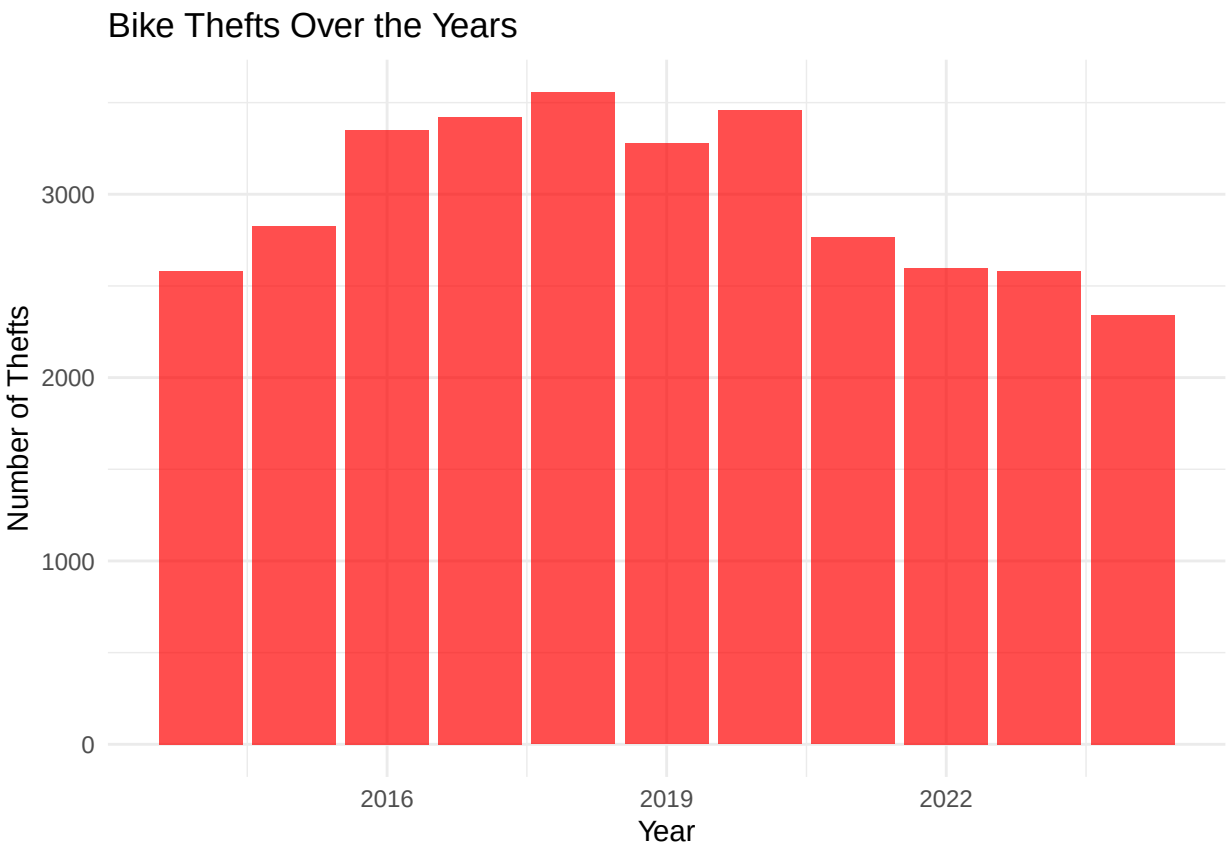
Note:
This table displays the number of bike shops and unique postal codes covered.

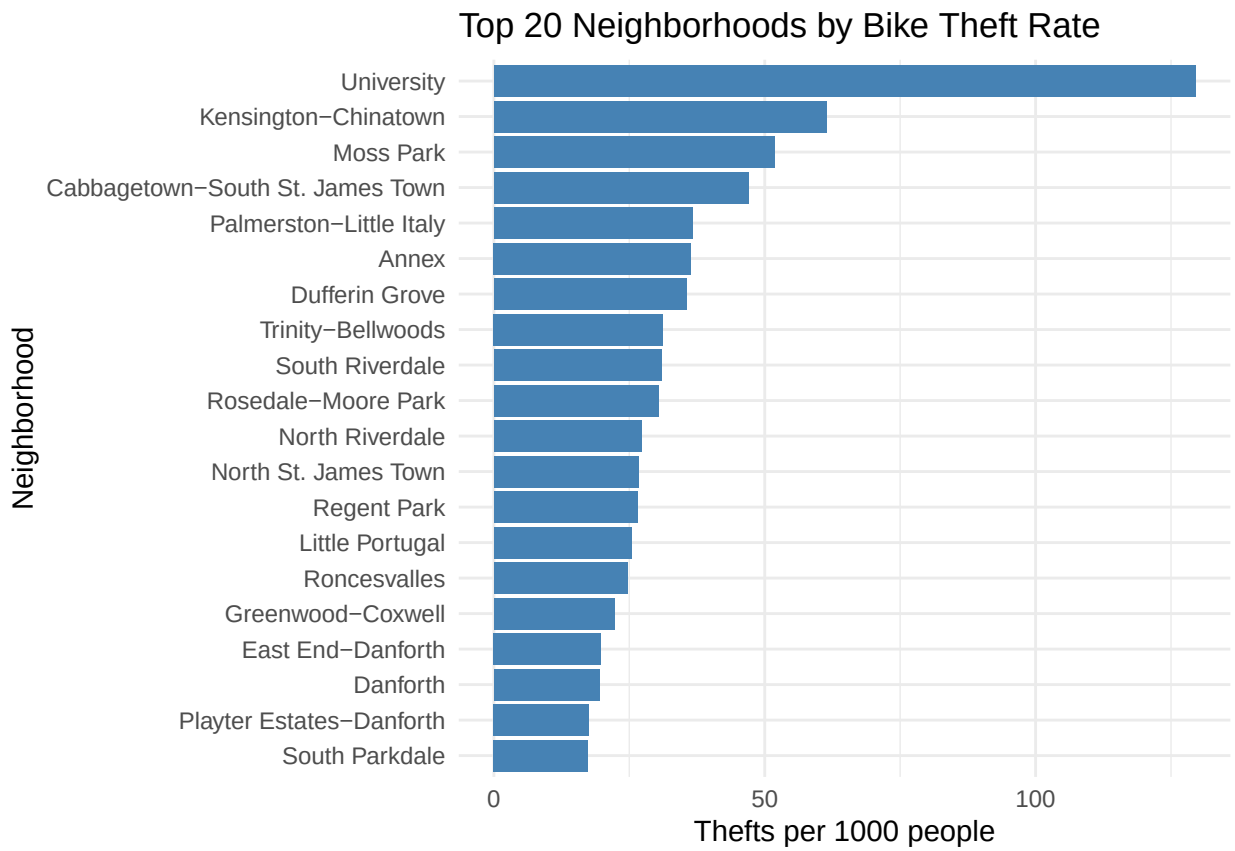
Table 3: Table 3: Neighbourhood Statistics (Count and Average Population)

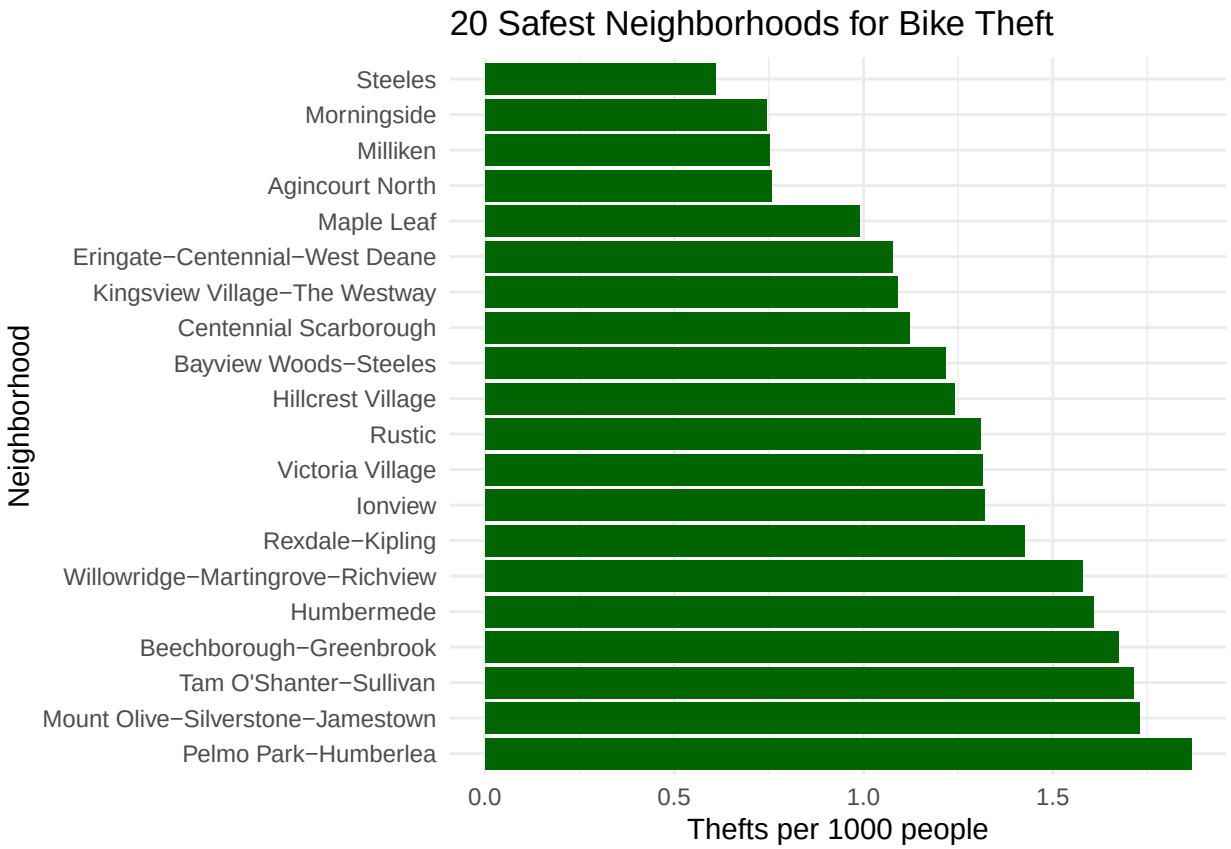
Number of Neighbourhoods	Overall Avg Population
142	19511

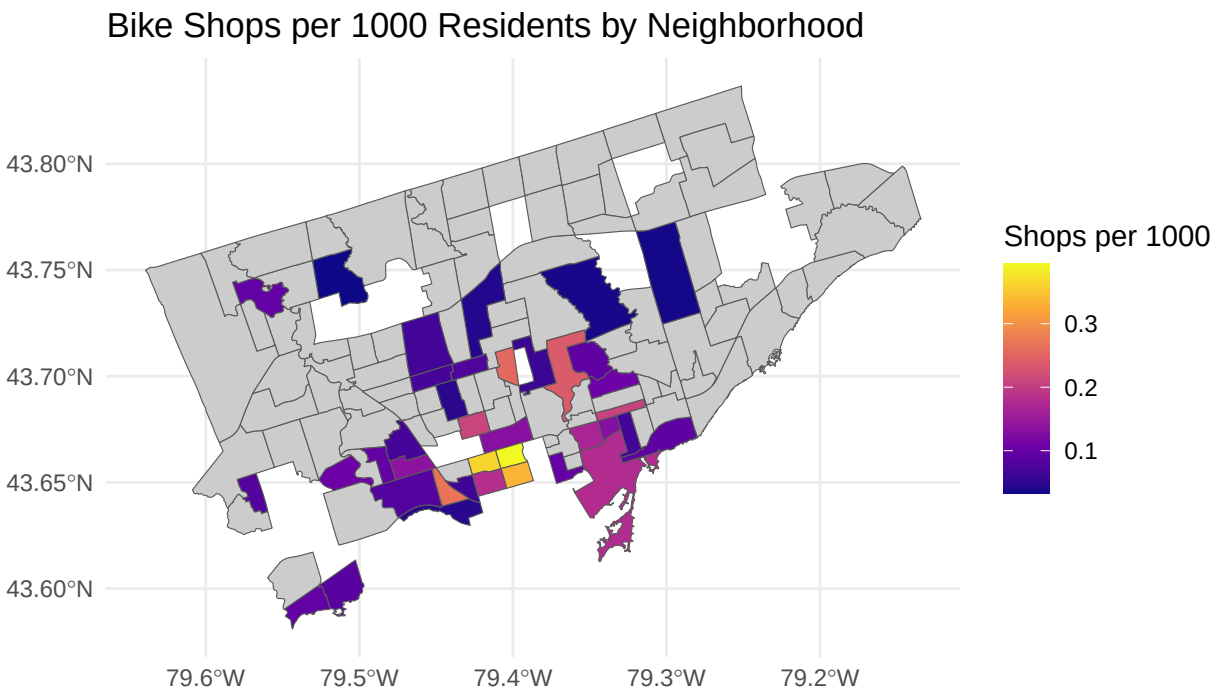
Note:
This table summarizes the number of neighbourhoods and their average population.

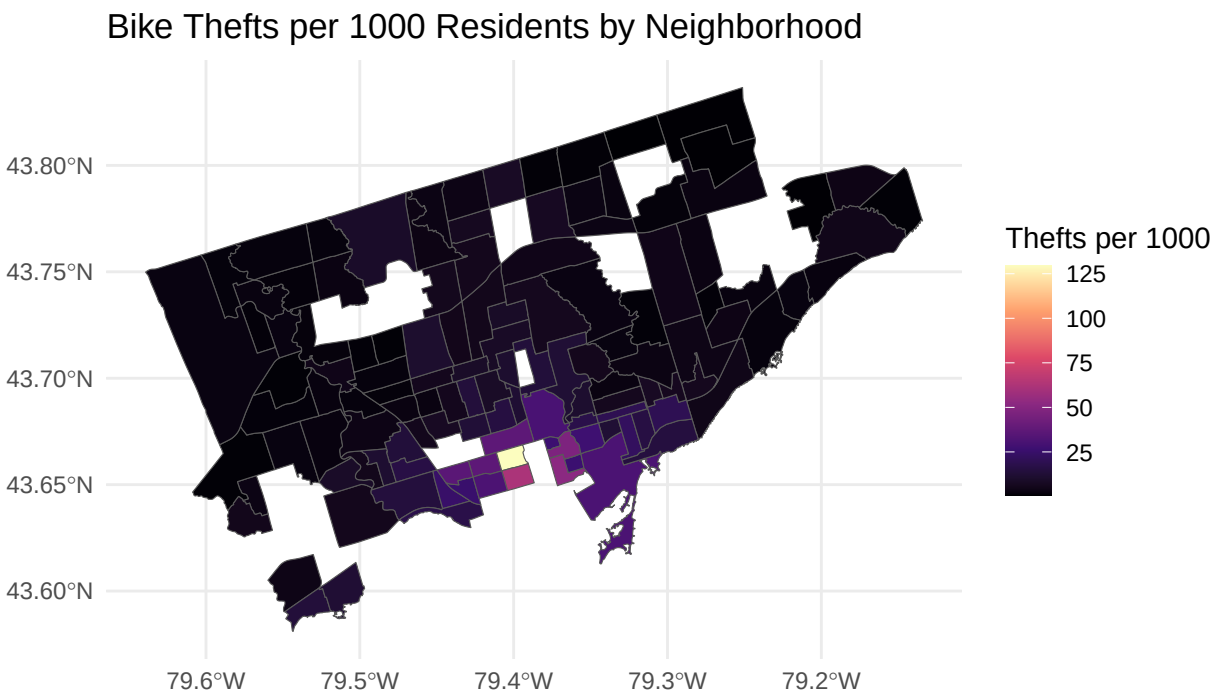

Distribution of Bike Theft Costs (Log Scale)
Frequency
5000
4000
3000
2000
1000
0
4
6
8
10
12
Log(Bike Cost + 1)

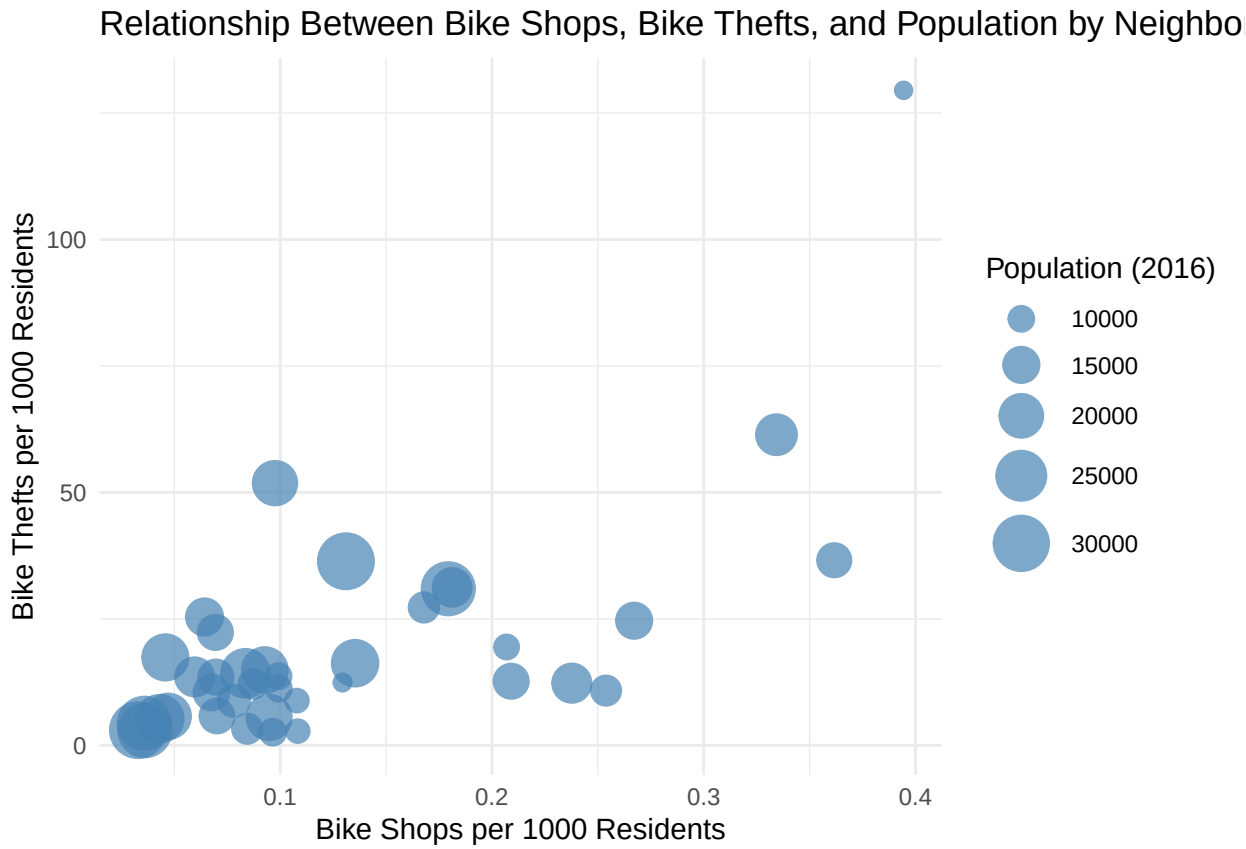

Bike Thefts Over the Years
Number of Thefts
3000
2000
1000
0
2016
2019
2022
Year

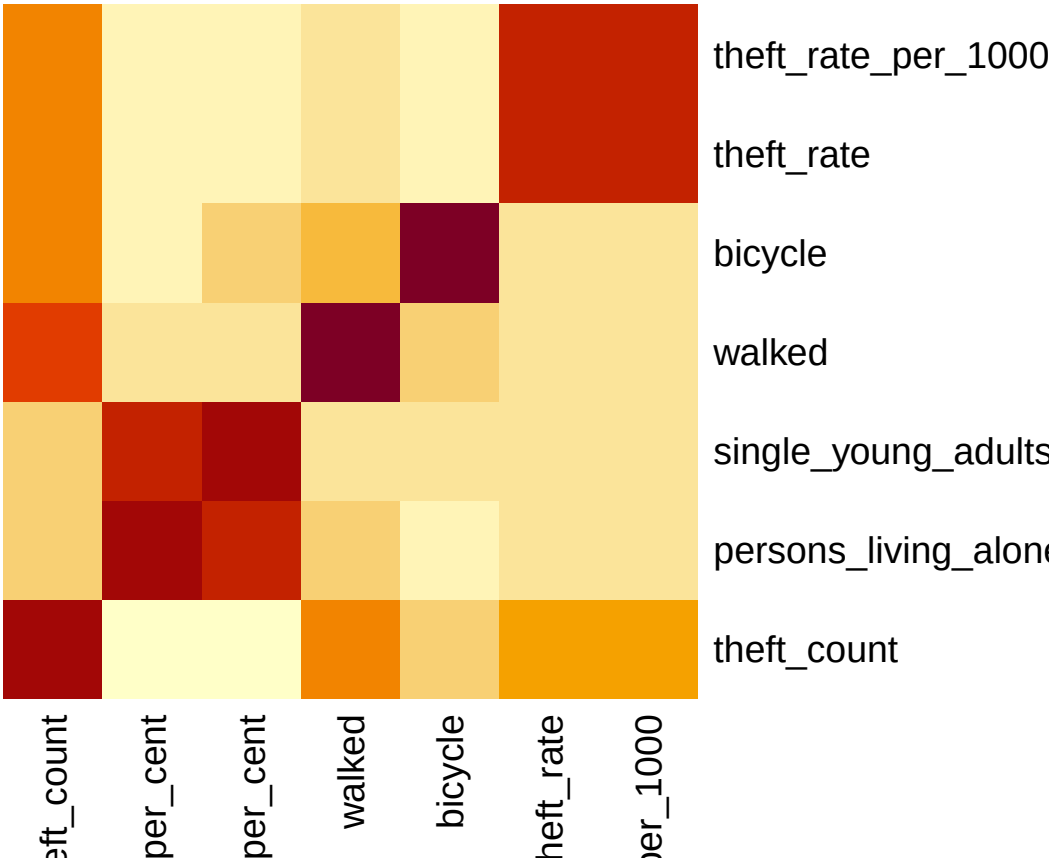

Top 20 Neighborhoods by Bike Theft Rate
Neighborhood
University
Kensington−Chinatown
Moss Park
Cabbagetown−South St. James Town
Palmerston−Little Italy
Annex
Dufferin Grove
Trinity−Bellwoods
South Riverdale
Rosedale−Moore Park
North Riverdale
North St. James Town
Regent Park
Little Portugal
Roncesvalles
Greenwood−Coxwell
East End−Danforth
Danforth
Playter Estates−Danforth
South Parkdale
0
50
100
Thefts per 1000 people


20 Safest Neighborhoods for Bike Theft
Neighborhood
Steeles
Morningside
Milliken
Agincourt North
Maple Leaf
Eringate−Centennial−West Deane
Kingsview Village−The Westway
Centennial Scarborough
Bayview Woods−Steeles
Hillcrest Village
Rustic
Victoria Village
Ionview
Rexdale−Kipling
Willowridge−Martingrove−Richview
Humbermede
Beechborough−Greenbrook
Tam O'Shanter−Sullivan
Mount Olive−Silverstone−Jamestown
Pelmo Park−Humberlea
0.0
0.5
1.0
1.5
Thefts per 1000 people


Bike Shops per 1000 Residents by Neighborhood
43.80°N
43.75°N
43.70°N
43.65°N
43.60°N
79.6°W
79.5°W
79.4°W
79.3°W
79.2°W
Shops per 1000
0.3
0.2
0.1


Bike Thefts per 1000 Residents by Neighborhood
43.80°N
43.75°N
43.70°N
43.65°N
43.60°N
79.6°W
79.5°W
79.4°W
79.3°W
79.2°W
Thefts per 1000
125
100
75
50
25


Relationship Between Bike Shops, Bike Thefts, and Population by Neighbo
Bike Thefts per 1000 Residents
100
50
0
0.1
0.2
0.3
0.4
Bike Shops per 1000 Residents
Population (2016)
10000
15000
20000
25000
30000



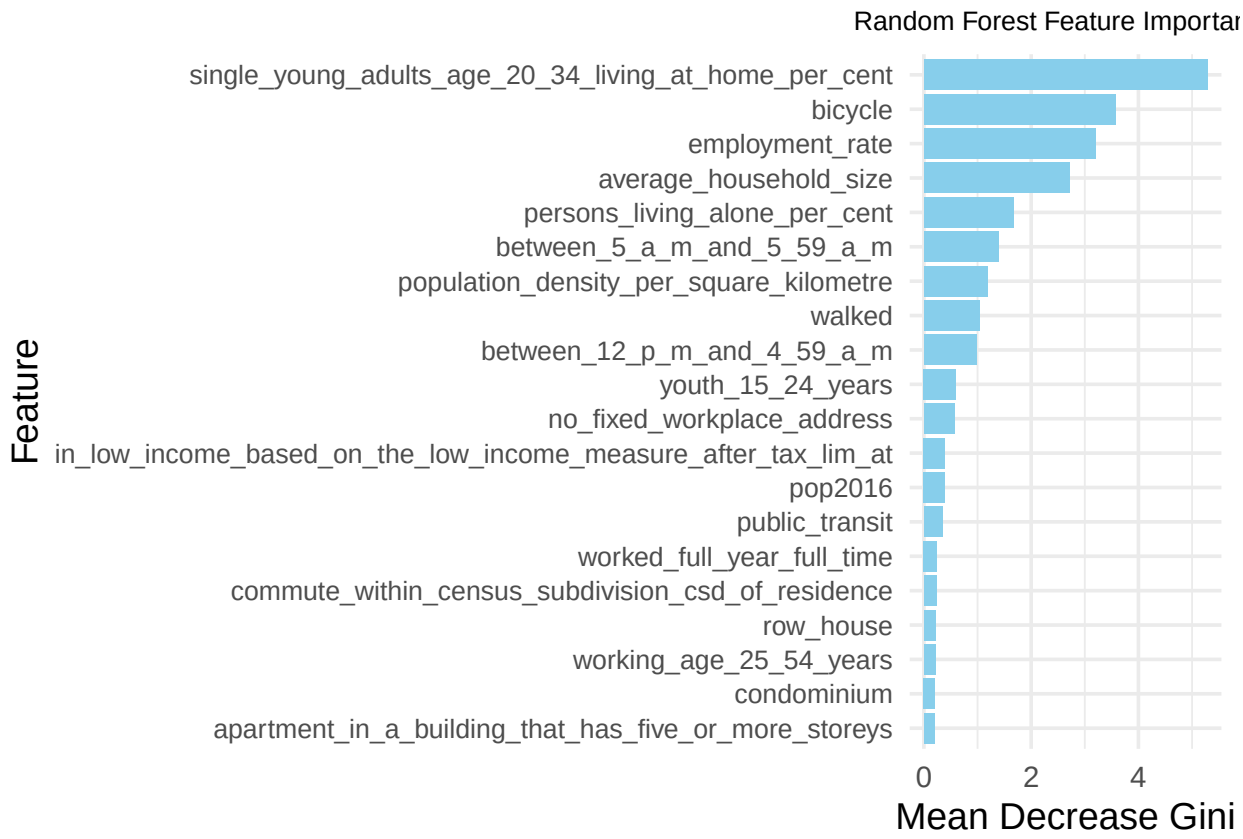

```
## NULL
```

Table 4: Table 4: Model Evaluation Results

Model	Accuracy	RMSE	MAE
Logistic Regression	0.8918919	0.3247399	0.1067478
GAM	0.7297297	0.4274723	0.3273676

```
## Warning in Ops.factor(rf_preds, 0.5): '>' not meaningful for factors
```

```
## Random Forest Accuracy: NA
```




```
##
## Family: binomial
## Link function: logit
##
## Formula:
## high_theft ~ s(population_density_per_square_kilometre, k = 3) +
##     s(public_transit, k = 3)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2582     0.2676   0.965    0.335
##
## Approximate significance of smooth terms:
##                                           edf Ref.df Chi.sq  p-value
## s(population_density_per_square_kilometre)   1      1  15.49 8.38e-05 ***
## s(public_transit)                            1      1   0.82    0.365
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.231   Deviance explained = 20.8%
## -REML = 48.154  Scale est. = 1         n = 87
```

link to repo: https://github.com/guzzim2022/jsc370-guzzi-finalproject