Exploring the Socio-Economic Factors of Auto Theft in Toronto

Rebecca Kong
17 March 2025

Github Link: https://github.com/rrjk76/auto_theft_socioecon_factors

Introduction

Auto theft is a prominent concern in urban environments, and the City of Toronto is no exception. Beyond being a property crime, auto theft is often intertwined to broader criminal activity, including organized crime, fraud, and illicit resale markets. For vehicle owners, the theft of a car can be more than just an inconvenience: it can disrupt their daily lives, limit access to work and essential services, and cause heavy financial strain due to insurance claims, replacement costs, and legal processes. From a more comprehensive perspective, high auto theft rates may indicate deeper socio-economic issues, such as income inequality, unemployment, or systemic vulnerabilities in certain neighbourhoods.

Toronto being Canada's largest city frequently reports cases of vehicle theft, often covered in the media. However, beyond simply the headlines, an important question that arises is whether there are identifiable patterns or relationships between auto theft rates and soco-economic factors. Do economic conditions, such as an individual's income, household income, or employment levels, correspond to higher vehicle-related crime? Thus, the goal of this analysis is to explore the relationship between auto theft rates in Toronto and various socio-economic factors, seeking to understand the question: Do certain socio-economic conditions serve as indicators for higher rates of vehicle-related crime?.

To explore the relationship between auto theft and socio-economic factors in Toronto, this analysis utilizes census datasets from the Toronto Open Data Neighbourhood Profiles. Since Canadian census data is collected every five years and released the following year, I have chosen to examine the census data published in 2016 and 2021, which correspond to socio-economic conditions in 2015 and 2020, respectively. These datasets provide a wealth of information, including demographic characteristics, income levels, employment and unemployment rates, and linguistic diversity across Toronto's neighbourhoods. In addition to the census data, this analysis incorporates the Toronto Police Service Auto Theft Open Data, which contains reported incidents of vehicle theft across the city. By integrating these datasets, this study aims to identify potential correlations between socio-economic conditions and auto theft rates, offering insights into whether certain neighbourhood characteristics may be indicative of higher rates of vehicle-related crimes.

Methods

The datasets used in this analysis were obtained from the Toronto Open Data Portal and the Toronto Police Open Data Service. Although the 2021 Neighbourhood Profiles dataset, which provides socio-economic information at the neighbourhood level, offered directly downloading an XLSX file, I decided to access the dataset using package identifiers and resource retrieval methods rather than direct download. Specifically, the `list_package_resources()` function was used to retrieve a list of available datasets, including package IDs for all census datasets dating back to 2006. From this list, I selected the 2021 census package ID. Using the `get_resource()` method, the dataset details were extracted, yielding a response containing all relevant census data. Both of the functions mentioned

above are from the Toronto Open Data Package system. The dataset itself was accessed through the hd2021_census_profile object and assigned to a data frame for further processing. Overall, this process involved querying metadata, identifying the correct dataset, and extracting structured data from the response.

Upon inspecting the 2021 data, it became evident that the data was structured in a long format, where neighbourhood names were set as column names, and attributes such as income, age, and employment status were stored as row names, as seen in *Table 1*:

Table 1: Sample of the Wide Dataset (First 5 Columns & Rows)

Neighbourhood Name	West Humber-Clairville	Mount Olive-Silverstone-Jamestown	Thistletown-Beaumond Heights	Rexdale-Kipling
Neighbourhood Number	1	2	3	4
TSNS 2020 Designation	Not an NIA or Emerging Neighbourhood	Neighbourhood Improvement Area	Neighbourhood Improvement Area	Not an NIA or Emerging Neighbourhood
Total - Age groups of the population - 25% sample data	33300	31345	9850	10375
0 to 14 years	4295	5690	1495	1575
0 to 4 years	1460	1650	505	505

This made it difficult to analyze. Thus, in order to address this, I transposed the data using the t() function, converting it from long format to wide format, and ensuring that each row corresponded to a neighbourhood, with relevant attributes stored in separate columns. As a result, the dataset, initially containing 158 variables and 2,604 observations, was transformed into 158 observations and 2,604 attributes.

The 2016 Neighbourhood Profiles dataset was acquired slightly differently. Originally, I tried acquiring the dataset by retrieving the package ID associated with the dataset. However, unlike the 2021 dataset, it lacked crucial socio-economic indicators, such as average and median household income at the neighbourhood level. To supplement this missing information, I identified an alternative data source, a Kaggle Notebook, that demonstrated an approach for deriving neighbourhood-level income estimates. I cloned this Kaggle Notebook to adapt the data deriving methodology, and I saved the output into a CSV file, which was then imported into R. This ensured that both the 2016 and 2021 census datasets had consistent column structures for direct comparison.

I also noticed that the 2016 census data only contained data for 140 neighbourhoods, while the 2021 census contained data for 158 neighbourhoods. To fix this inconsistency, I manually changed the data in the 2021 census to match the 140 neighbourhood system in the 2016 census data. Plus, since there are over 2000 columns/attributes for both the 2016 and 2021 census datasets, I used the select() function to choose only the key variables to keep for my analysis. *Table 2* outlines some of the important attributes I kept in each census dataset.

Table 2: Sample Output of Data Showing Some Key Variables

Neighbourhood Name	Neighbourhood Number	0 to 14 years	15 to 24 years	25 to 54 years	55 to 64 years	65 years and over	85 years and over	Average total income	Average after-tax income	Low Income Prevalence (LIM-AT) %
West Humber-Clairville	1	5060	5445	13845	3990	4980	615	31771	28066	15.8
Mount Olive-Silverstone-Jamestown	2	7090	5240	13615	3475	3560	300	26548	24122	27.9
Thistletown-Beaumond Heights	3	1730	1410	4160	1195	1880	350	32815	28842	17.8
Rexdale-Kipling	4	1640	1355	4300	1520	1730	300	34418	30201	18.6
Elms-Old Rexdale	5	1805	1440	3700	1255	1275	145	32012	28355	23.2
Kingsview	6	4240	3020	8635	2550	3585	575	36674	31447	24.9

Because the names of the two datasets were inconsistent, I used the rename() function to simplify and standardize the column names and ensure consistency for row binding the datasets later on. Once both datasets were cleaned, two additional columns were added:

- NEIGHBOURHOOD_140: A reformatted neighbourhood identifier that facilitated merging with the auto theft dataset.
- Year: A column specifying the corresponding census year (2015 for the 2016 census and 2020 for the 2021 census).

However, a key issue I encountered was that all values in the 2021 dataset were stored as character data, while the 2016 dataset contained numerical values (except for neighbourhood names). To resolve this inconsistency, all numeric variables in the 2021 dataset were converted to numeric data types using the as.numeric function, ensuring uniformity across both datasets. The two census datasets were then combined into a single dataset using rbind(), now containing socio-economic indicators for both 2015 and 2020.

Finally, the Auto Theft dataset was retrieved from the Toronto Police Open Data Service. This dataset contained information on vehicle theft incidents across different neighbourhoods in Toronto. It was merged with the combined 2015 and 2020 census dataset using two key attributes:

- NEIGHBOURHOOD_140 (to match neighbourhoods across datasets).
- Year and REPORT_YEAR (to ensure data alignment by year).

The final dataset, auto_theft_census.csv has 8117 observations and 66 columns. These pre-processing steps ensured that the final dataset was structured for further analysis, allowing for an investigation into the relationship between socio-economic factors and auto theft rates across Toronto's neighbourhoods.

Preliminary Results

To gain insights into auto theft trends in Toronto and their potential socio-economic correlations, I first created a summary statistics table. This table presents key values, including:

- Total auto thefts recorded in 2015 and 2020.
- Average total income in Toronto for each year.
- Unemployment rate (%), which reflects the percentage of unemployed individuals in the labor force.
- Low-income prevalence (%), indicating the proportion of the population classified as low-income.

Table 3: Summary Statistics for Auto Theft and Socio-Economic Factors

REPORT_YEAR	Total_Auto_Theft	Average_Income	Unemployment_Rate	Low_Income_Prevalence
2015	3244	49183.09	8.615351	19.60851
2020	4873	49446.71	14.548984	12.21200

In *Table 3*, we can see that there is an increase in the total number of auto thefts between 2015 and 2020. We can also see that there is a slight raise in income compared to 2016 and 2020, increasing from $49,183 in 2015 to $49,447 in 2020. We *do* see, however, that there is a surge in the unemployment rate, rising sharply from 8.62% in 2015 to 14.55% in 2020. This is likely influenced by the economic disruptions such as the COVID-19 pandemic, and I will take this into account for this analysis. Finally, *Table 3* shows us that the percentage of the population classified as low-income declined from 19.61% to 12.21%, suggesting potential improvements in economic conditions for some groups.

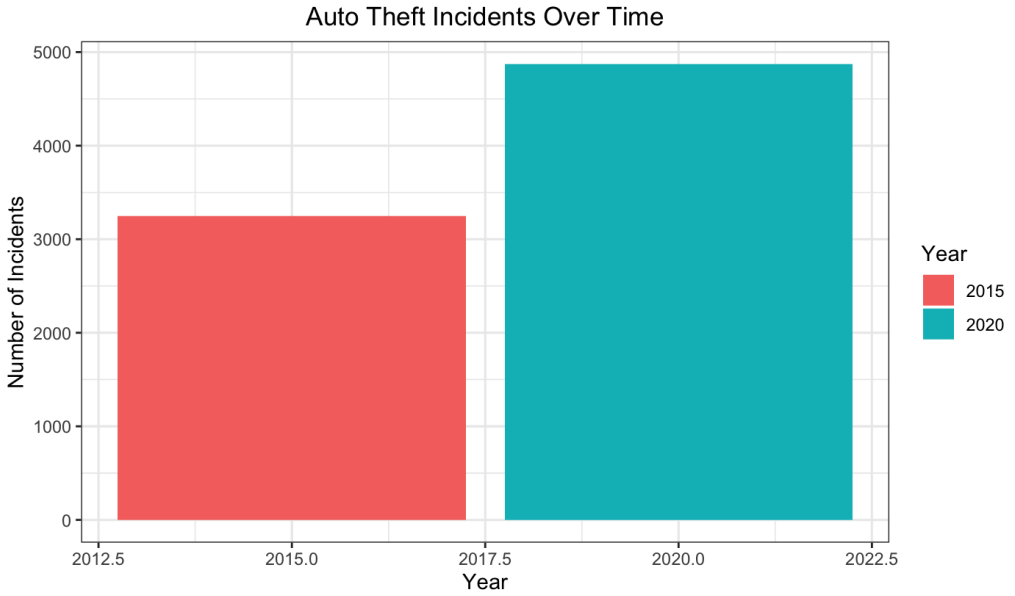

Figure 1: Bar plot of auto theft incidents from 2015 and 2020

Above in *Figure 1,* we can see visually that there is quite a large jump in auto theft numbers in 2020 compared to 2015, with 2015 having 3244 vehicles stolen, and 2020 having 4873 vehicles stolen.

Now, taking a closer look at income statistics:

Table 4: Summary Statistics for Average Household Income

REPORT_YEAR	Average_After_Tax_Income	Average_After_Tax_Household_Income
2015	49183.09	81495.00
2020	49446.71	98784.81

Table 4 shows the average after-tax income for individuals and households across two years. As mentioned, individual after-tax income remained fairly stable, with only a small increase from ~$49,183 (2015) to ~$49,447 (2020). However, household after-tax income increased significantly, from ~$81,495 in 2015 to ~$98,785 in 2020.

There is a key distinction in how income is calculated for individuals versus households, and that is:

- Individual Income (Average_After_Tax_Income) refers to total income statistics for an individual aged 15 years and over who are not in economic families or living in private households. It represents income at the personal level, including wages, government benefits, and investment earnings.
- Household Income (Average_After_Tax_Household_Income) refers to total income for all members of a household/people living together combined. It includes combined salaries, government benefits, and other earnings of everyone in that household, and hence why the overall values are much higher compared to individual income.

Both of these statistics are important for this analysis, as they give us a more comprehensive understanding of economic conditions and their potential relationship with auto theft trends. Individual after-tax income helps us understand the financial stability of single earners, particularly those not part of a family unit, while household after-tax income provides insight into the overall economic well-being of families and shared living arrangements.

Below, *Figure 2* gives us a great visual representation of the after-tax income trends, helping us visually see that individual income has remained rather stagnant while household income has increased quite a bit.

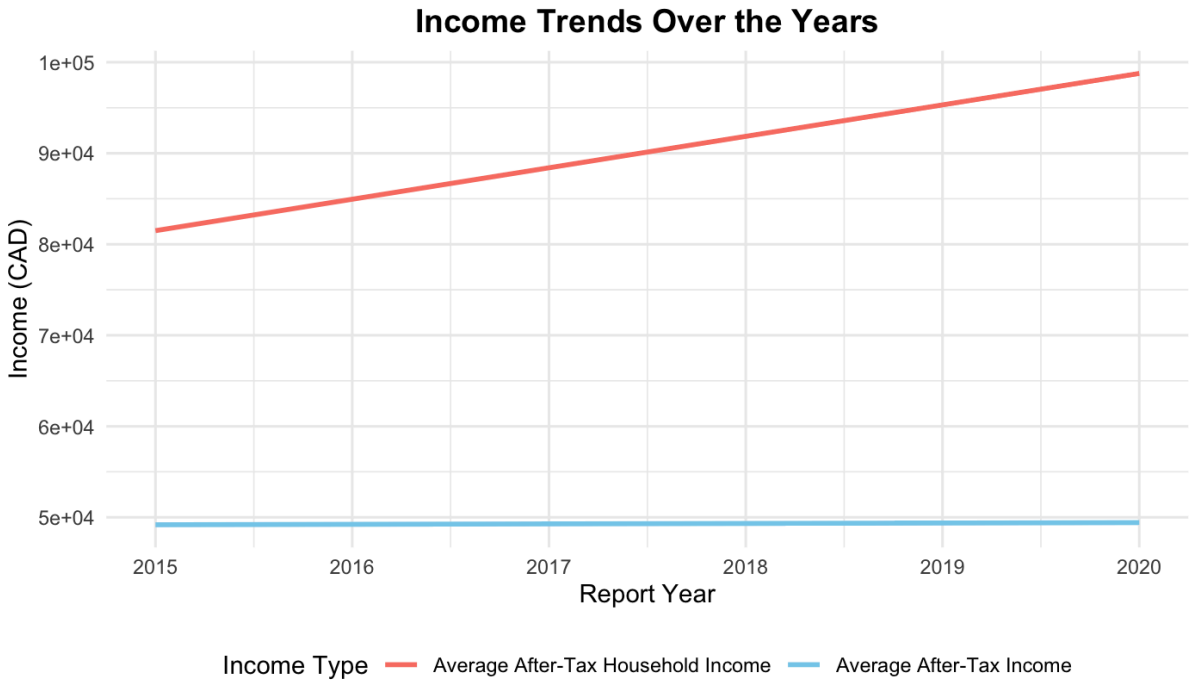

Figure 2: Line graph of after-tax income trends

Now, I want to also get a deeper understanding of where the most auto thefts are occurring, specifically which neighbourhood.

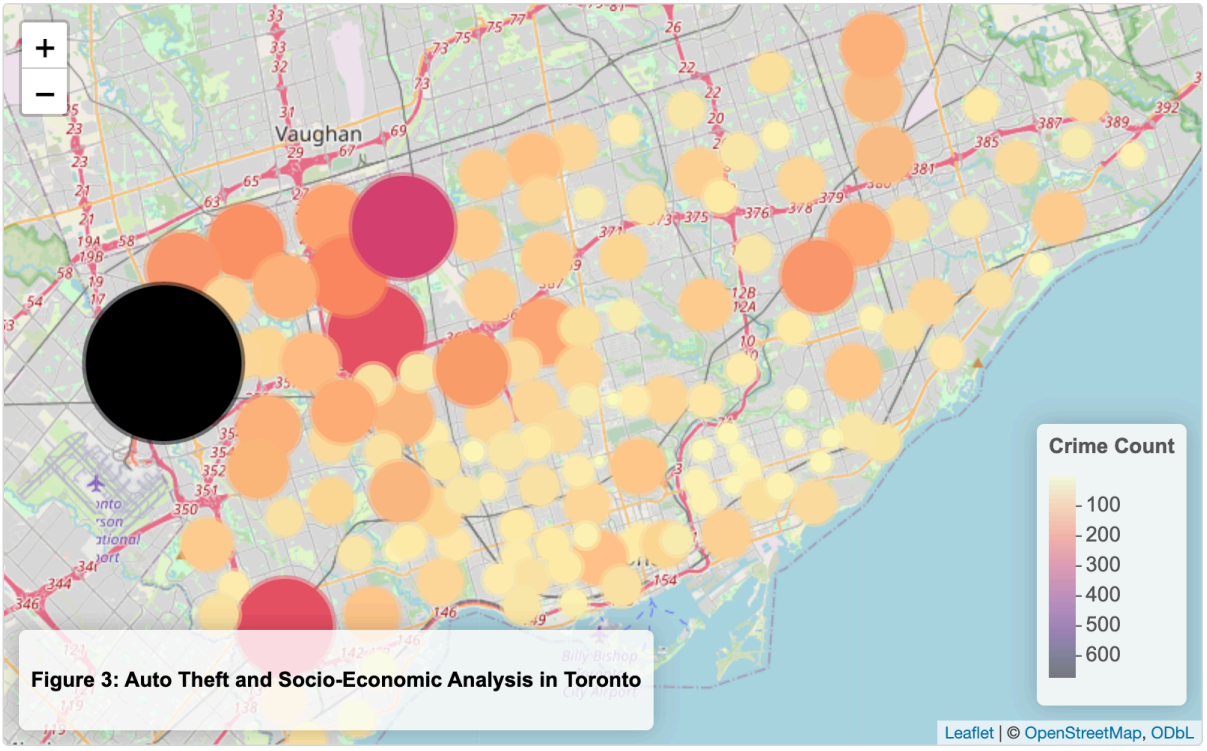

Figure 3: Auto Theft and Socio-Economic Analysis in Toronto

The *Figure 3* Leaflet map visualizes the spatial distribution of auto theft incidents across different neighborhoods in Toronto, providing an interactive way to examine crime density and making it easier to quickly identify areas that may require targeted interventions or additional resources. As seen, the large black circle represents the area with the most total auto thefts in 2015 and 2020.

To take a closer look at the comparison between 2015 and 2020, *Figure 4* displays the number of auto thefts of the top 20 neighbourhoods with the most reported cases of auto theft.

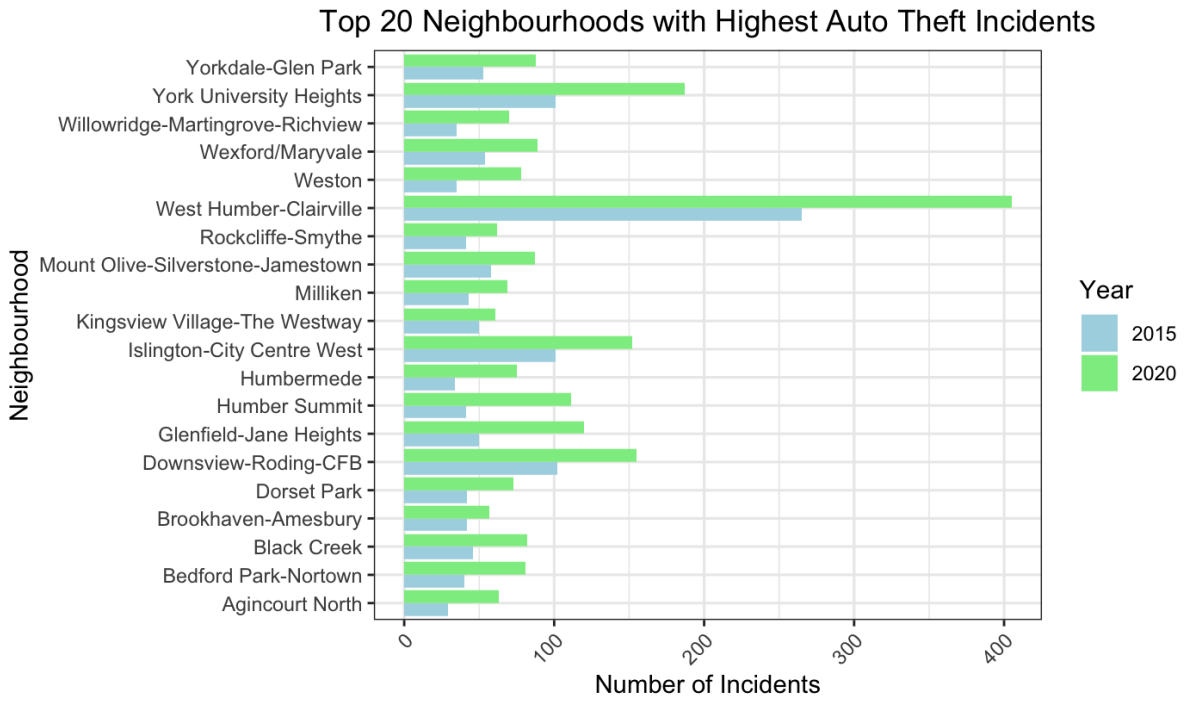

Figure 4: Double Bar Graph of the Top 20 Neighbourhoods

From *Figure 4* (and the interactive Leaflet map in *Figure 3*), we can see that West Humber-Clairville has the highest number of auto thefts in both 2015 and 2020, with a notable increase in 2020. To understand this trend, we need to further explore socio-economic factors such as average income and unemployment rates in that neighborhood specifically. Analyzing these factors alongside auto theft data will help determine if they are linked to the rise in crime.

Both *Figure 5* and *Figure 6* show the highest and lowest average individual incomes, while also including West Humber-Clairville for comparison. Since West Humber-Clairville had the highest auto theft counts in both 2015 and 2020, this allows for a deeper investigation into whether average individual income is a potential confounding factor. As seen in both figures, the average income of West Humber-Clairville is much closer to that of the lowest-income neighborhoods in both years. This suggests that lower income levels may be associated with higher auto theft

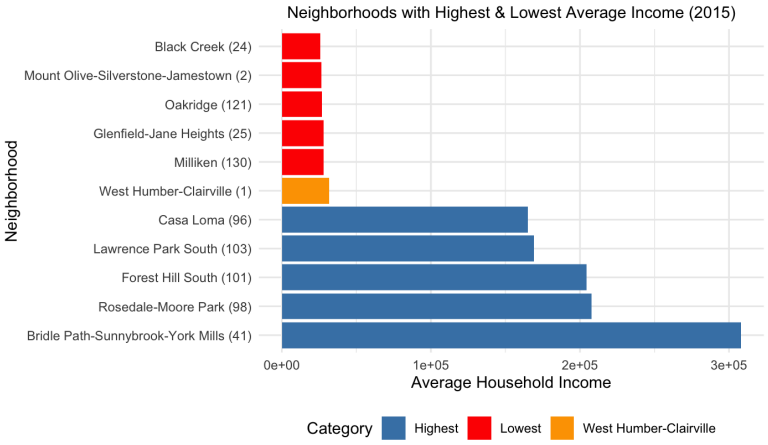

Figure 5: Flipped bar plot dipicting income extremes in 2015

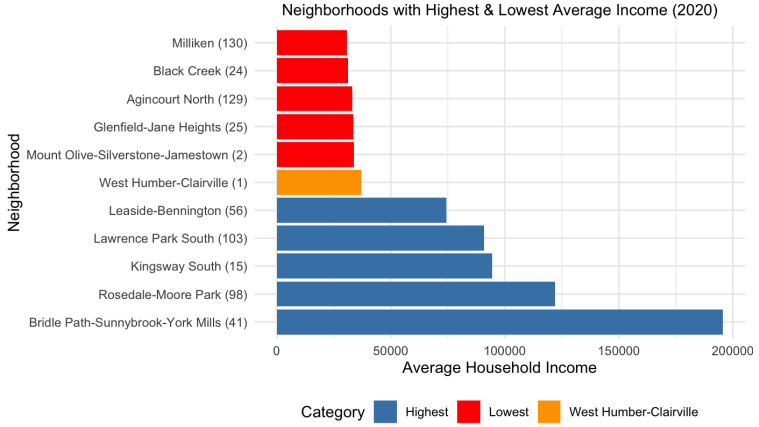

Figure 6: Flipped bar plot dipicting income extremes in 2020

rates, though other factors should also be considered.

To explore the impact of average income on auto thefts further, *Figure 7* presents a box plot that categorizes neighborhoods into five income brackets to examine whether lower average individual income correlates with higher auto theft rates. The figure illustrates that neighborhoods in the lower income brackets tend to experience greater variability in auto theft incidents, as indicated by the wider interquartile ranges and longer tails. In contrast, higher-income brackets exhibit lower variation, with flatter boxes and fewer extreme values. Notably, the highest recorded auto theft count—previously identified as West Humber-Clairville—falls within one of the lower income brackets in both 2015 and 2020, suggesting a potential relationship between neighbourhoods with lower income levels and higher auto theft rates.

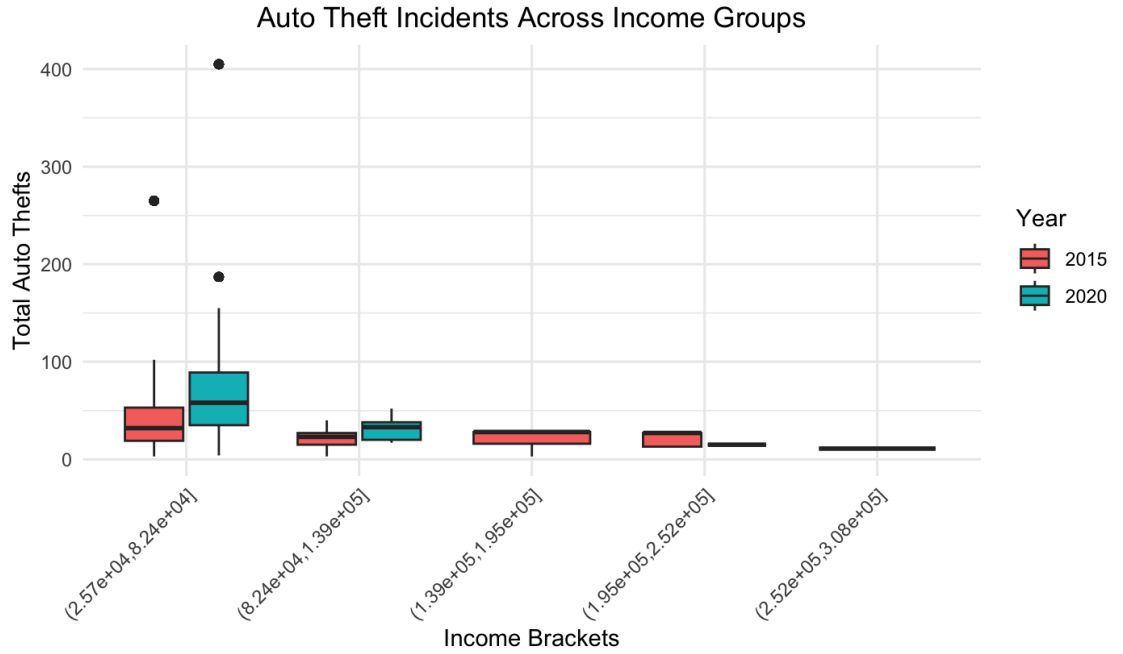

Figure 7: Boxplot using 5 breaks analyzing auto theft incidents across income groups

Given that there is a relationship between average income of a neighbourhood and auto theft rates, I believe that it will also be insightful to take a look at employment and unemployment rates.

Table 5: Correlation Matrix of Auto Theft and Socio-Economic Factors

	Average total income	Average after-tax income of household	Total_Auto_Thefts	Low Income Prevalence (LIM-AT) %	Employment rate	Unemployment rate
Average total income	1.0000000	0.4334165	-0.2185677	-0.3892399	0.3648976	-0.3553225
Average after-tax income of household	0.4334165	1.0000000	-0.1125420	0.1653809	0.4382516	-0.6692523
Total_Auto_Thefts	-0.2185677	-0.1125420	1.0000000	-0.2547223	-0.1084865	0.2282220
Low Income Prevalence (LIM-AT) %	-0.3892399	0.1653809	-0.2547223	1.0000000	-0.1007665	-0.1337259
Employment rate	0.3648976	0.4382516	-0.1084865	-0.1007665	1.0000000	-0.7692783
Unemployment rate	-0.3553225	-0.6692523	0.2282220	-0.1337259	-0.7692783	1.0000000

Above, *Table 5* presents a correlation table matrix, highlighting the relationships between several numerical social indicators. It's no surprise that individual income is positively correlated with household income, with a correlation of 0.43. This indicates that as individual income rises, so does household income, which is consistent with general economic expectations. Similarly, both average total income and average after-tax household income show a moderate positive correlation with the employment rate, suggesting that areas with higher income levels tend to have higher employment rates.

However, the more interesting part is that *Table 5* also reveals a pattern: as the employment rate increases, there is a negative correlation with the number of auto thefts. This indicates that areas with higher employment tend to experience fewer auto thefts. On the other hand, higher unemployment rates in neighbourhoods correlate with higher numbers of auto thefts. Although the correlation is weak (0.23), this positive relationship suggests a potential trend, implying that higher unemployment may be associated with an increase in auto theft incidents.

To investigate this further, we can examine the relationship between employment and unemployment rates with auto thefts using a scatter plot with a regression line, which will allow us to visually assess the strength and direction of these correlations.

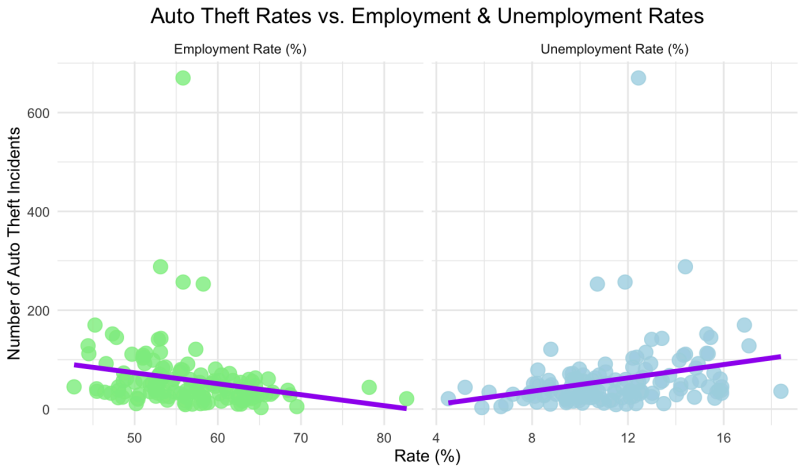

Figure 8: Faceted plots showing Auto Theft Rates vs. Employment and Unemployment Rates

Figure 8 is a faceted scatter plot with regression lines, illustrating the relationship between employment and unemployment rates and the number of auto theft incidents. The regression lines indicate an increasing slope with employment rates, suggesting that as employment rates rise, the number of auto thefts tends to decrease. Conversely, the regression line for unemployment rates shows a negative slope, implying that as unemployment increases, auto theft incidents tend to rise.

Now, most of the analysis I've done so far has been related to economic factors (as I've been analyzing data corresponding to individual income, household income, employment, and unemployment rates, etc.). Now, I want to dive deeper into some social aspects. Social aspects I will be analyzing are languages spoken in neighbourhoods and ages in neighbourhoods where auto thefts happen the most often.

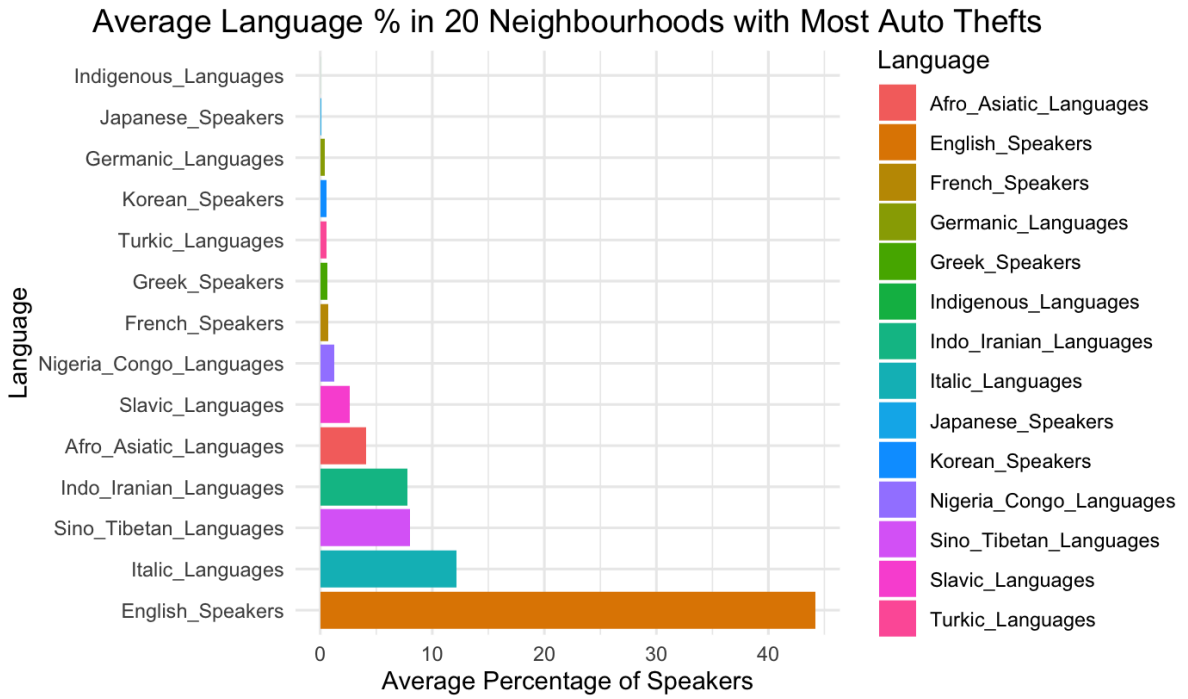

Figure 9: Bar plot depicting the most spoken languages in the 20 neighbourhoods with the most auto thefts

Above, *Figure 9* illustrates the mother tongue languages by percentage in the top 20 neighbourhoods. It's no surprise that English is the highest percentage, where just under 50% have English as their mother tongue. The second native language that is most spoken by percentage in the top 20 neighbourhoods with highest auto-theft count is Italic languages, which includes languages such as Italian, Portuguese, and Spanish. These languages collectively account for approximately 13% of the population in these neighborhoods. Following this, Sino-Tibetan languages, which include Mandarin and Cantonese, account for approximately 8% of the population. While this data does not establish a direct link between language demographics and auto-theft rates, it highlights the diverse linguistic landscape of these high-incident neighborhoods. Further analysis incorporating socioeconomic factors, population density, and law enforcement presence could provide deeper insights into potential correlations.

Now, analyzing the ages in the top five neighbourhoods with the most auto thefts:

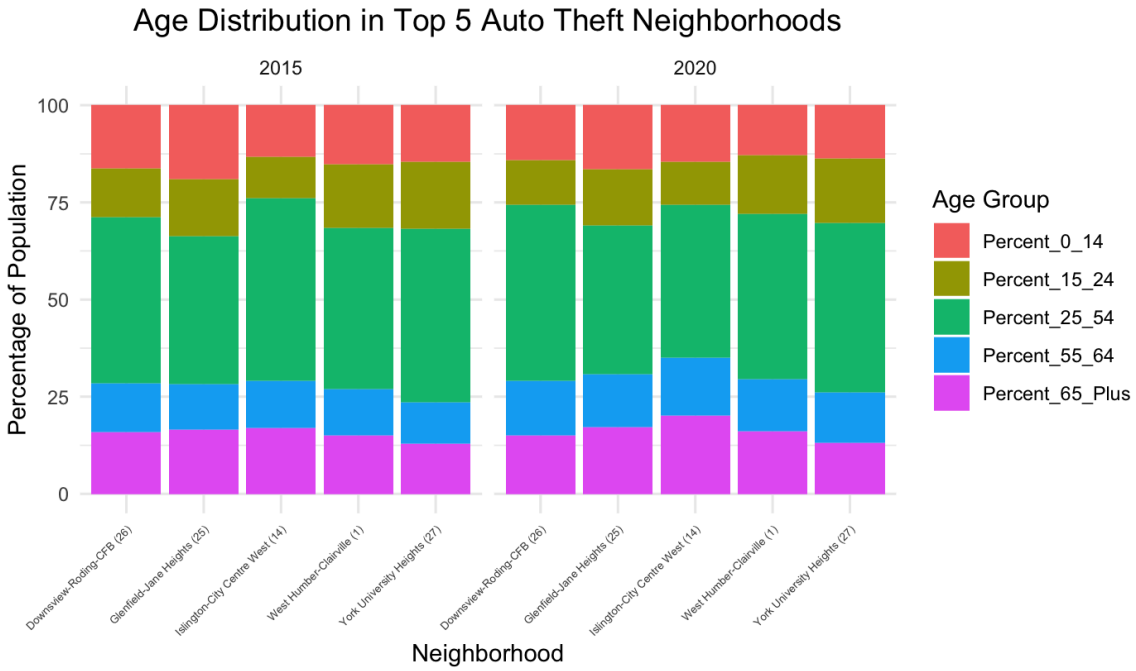

Figure 10: Stacked Bar Depicting Age Distribution in Top 5 Auto Theft Neighbourhoods

Table 6: Correlation Between Auto Theft and Age Groups

Corr_0_14	Corr_15_24	Corr_25_54	Corr_55_64	Corr_65_Plus
-0.5759993	0.2731887	0.067033	0.2704908	-0.0647428

Figure 10 and *Table 6* illustrate the age distributions across the neighborhoods. The correlation data in the table reveals how different age groups relate to auto theft incidents. Specifically, there is a negative correlation of -0.58 between auto theft incidents and the age group of 0 to 14, suggesting that neighborhoods with higher proportions of children tend to have fewer auto thefts. This can be an indication that younger populations are not typically associated with such criminal activities.

Conversely, the correlation with the 15-24 age group is 0.27, reflecting a weak positive relationship. This indicates that neighborhoods with a higher percentage of individuals in this age range tend to see a slight increase in auto theft incidents. I believe these numbers reflect the fact that this is the age group where many individuals begin driving, and hence, potentially leading to more thefts in areas with a higher proportion of young drivers.

The correlation for the 25-54 age group is relatively low at 0.07, suggesting minimal association between auto thefts and this group. But interestingly, the correlation with the 55-64 age group is slightly stronger at 0.27, though still weak, it indicates a mild positive trend. This could indicate that, in these areas, the types of vehicles owned or the frequency of car ownership might make them more susceptible to theft.

Finally, there is a negative correlation of -0.06 with the 65+ age group, implying that neighborhoods with higher proportions of elderly populations tend to experience fewer auto thefts, but the relationship is not very significant.

This data suggests that age distributions *do* play a role in auto theft trends. Although the correlations are generally weak, and other factors likely influence the occurrence of auto thefts more significantly, understanding the relationship with age groups still offers valuable context for further investigation.

Summary

My analysis from the preliminary results investigates trends of auto theft in Toronto and their potential relations with socio-economic factors. The comparison between the 2015 and 2020 census data revealed that there was a significant increase in auto theft incidents, rising from 3,244 cases in 2015 to 4,873 in 2020. While individual after-tax income remained stable, household after-tax income saw a notable increase, suggesting improvements in overall economic well-being. However, unemployment surged from 8.62% to 14.55%, likely influenced by the COVID-19 pandemic, while low-income prevalence declined.

Further analysis showed that neighborhoods with lower individual income levels tend to experience higher auto theft rates, as seen in West Humber-Clairville. My analysis revealed that West Humber-Clairville had the highest number of auto thefts in both years, and bar plots showed that its average income was aligned with lower-income neighbourhoods. Plus, box plots of income brackets further solidified this relationship, since there was greater variability in theft incidents among

lower-income neighborhoods. It is important to note, however, that higher-income areas have fewer data points, which may affect the results.

Next, my correlation analysis showed a weak negative relationship between employment rates and auto thefts, and a weak positive relationship between unemployment and thefts. This indicates that higher employment rates may be slightly associated with fewer auto thefts, while higher unemployment rates may be linked to a slight increase in auto theft incidents. However, the strength of these relationships is weak, suggesting that other factors may also play a significant role.

My analysis also explored social factors, including language and age demographics. I found that the most common mother tongues in high auto theft neighborhoods were English (just under 50% - though this was expected), followed by Italic languages (14%), which include Italian, Portuguese, and Spanish. Sino-Tibetan languages was close behind, accounting for around 8% in the neighbourhoods with the most auto thefts. In addition to language, age group correlations were examined, revealing that neighborhoods with a higher percentage of children (0 to 14 years) tended to have fewer thefts. Conversely, areas with a higher percentage of individuals ages 15 to 24 and 55 to 64-year-olds showed a weak positive relationship with auto theft, suggesting that age demographics may have an impact on theft patterns.

Overall, the preliminary results indicate that economic conditions (like individual average income in neighbourhoods), employment and unemployment rates, and age demographics, may influence auto theft trends. However, further modeling and additional factors need to be explored to confirm these relationships.

To be able to confirm these relationships, my plan for the final project is to leverage high-performance computing, machine learning, and interactive visualizations. I plan to try and find some more data for analysis, like other socio-economic factors, such as education levels, housing affordability, and crime rates for other offenses, to better understand their relationship with auto theft.

Thus, for HPC, I plan on using parallel computing to efficiently process large datasets, including auto theft reports and socio-economic indicators, by leveraging multi-core processing and distributed computing techniques. This will allow for faster data manipulation, modeling, and visualization, enabling a more comprehensive analysis of trends and correlations.

For ML analysis, I plan to Train Decision Trees, Random Forest, and XGBoost, comparing the output of these three models, and attempt to predict auto theft rates based on socio-economic features like, average income, unemployment rate, education level.

Finally, I plan on developing better geospatial heatmaps to visualize auto theft hotspots and their correlation with socio-economic factors. I also plan on implementing some interactive dashboards to be able to explore trends dynamically.

With the help of these techniques and interpretable ML models, this guarantees a data-driven investigation of the relationship between socio-economic factors and auto theft rates.

References

Auto theft open data. Toronto Police Service Public Safety Data Portal. (2025, January 22). https://data.torontopolice.on.ca/datasets/TorontoPS::auto-theft-open-data/about

Gasmi, M. (2019, November 24). *Battle of the neighborhood*. Kaggle. https://www.kaggle.com/code/servietsky/battle-of-the-neighborhood/notebook

Open data dataset. City of Toronto Open Data Portal. (2024, April). https://open.toronto.ca/dataset/neighbourhood-profiles/