Table Of Contents

Introduction

Urban cities, often thought of as hubs of innovation, cultural exchange, and economic activity, are also common places where disparities and crimes occur. A clear example of this is the Downtown and Downtown Eastside of Vancouver, each harbouring vastly different environments despite being a neighbourhood away. This project explores the effect of socioeconomic factors on the number of crimes in Vancouver as well as investigating the possibility of modelling other urban cities using Vancouver's population demographics.

Data

Two datasets are needed for our analysis. The Vancouver Police Records from VPD GeoDash Open Data is provided as a CSV file accounting for all crimes in Vancouver dating back to 2003. Each record is categorized by the type of crime, time it occurred, and location in X and Y coordinates, formatted in UTM Zone 10N[1]. This dataset contains 32,201 observations from 2021. Vancouver's 2021 Census of Population is retrieved from the CensusMapper API through the programming language R, saved as a GeoJSON file. The census contains information on Canada's population at many geographical levels, with geometry data included for each geographical location. For our purposes, we will be using census tracts[2] since larger or smaller aggregation levels will omit too many determining factors of crime rates. The data contains 128 census tracts for Vancouver.

To merge the two datasets together, we used the provided coordinates in the crime data and the geometry information included with the census data. We converted the X and Y coordinates to Points objects and converted the longitude and latitude geometries into X and Y polygon objects. For each crime record, we calculated its distance to all census tracts using GeoPandas's `geoSeries.distance()`. We then assigned the record to the nearest census tract, allowing us to find the number of crimes occurring in each tract.

The following variables are calculated in **data_processing.py**:

- `crime_rate`: Number of crimes per person
- `pop_density`: Number of people per square kilometre
- `dropouts_to_grads`: Proportion of high school dropouts to high school graduates

[1] UTM or Universal Transverse Mercator is the format VPD GeoDash uses for their locational data, which is a projection used for maps with the curvature of the earth in mind.

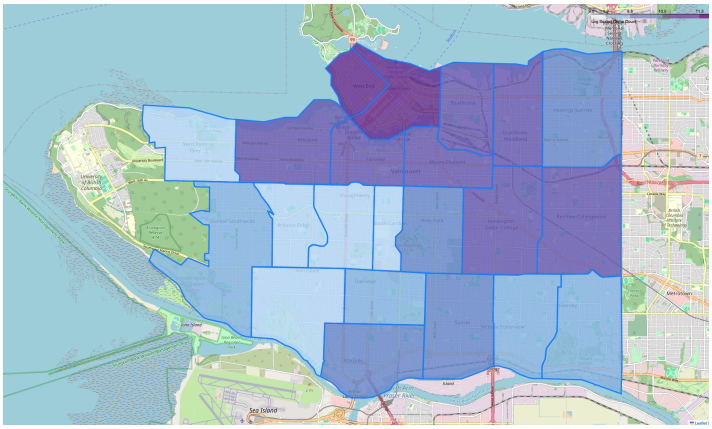
[2] Census tracts are defined as small, relatively stably geographic areas with populations usually between 2,500 and 8,000 people according to Statistics Canada.

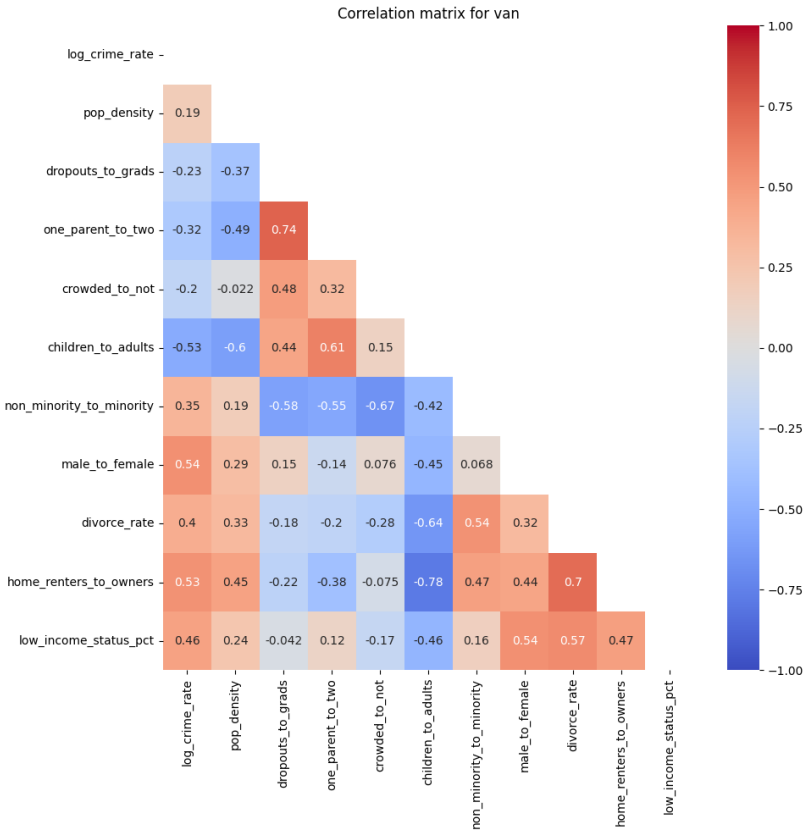
- `one_parent_to_two`: Proportion of single-parent families to two-parent families
- `crowded_to_not`: Proportion of households with more than one person per room, to households with less than one person per room
- `children_to_adults`: Proportion of children ages 0-14 to adults ages 15 and over
- `non_minority_to_minority`: Proportion of visible non-minority ethnicities to visible minorities
- `male_to_female`: Proportion of males to females
- `divorce_rate`: Proportion of divorced couples to the population ages 15 and over
- `home_renters_to_owners`: Proportion of private household renters to owners
- `low_income_status_pct`: Percentage of people with low income status after tax

We also obtained Toronto and Montreal crime data from their respective police service websites and merged them with the census data in a similar fashion. These were intended to be used for validating our machine learning models trained on Vancouver's dataset (the 3 cities' datasets were later combined together for the model). However, these crime datasets did not contain homicides or vehicle collisions, so Vancouver's crime records were filtered accordingly.

Analysis

A preliminary overview of the crime count data provided from the VPD in Vancouver for 2021 is shown visually in the choropleth map below. A logarithmic scale was chosen since the distribution of crime counts tended to be right skewed, leading to uneven colour assignment. The map shows as we move towards Downtown Vancouver, more crimes in the neighbourhoods are occurring.



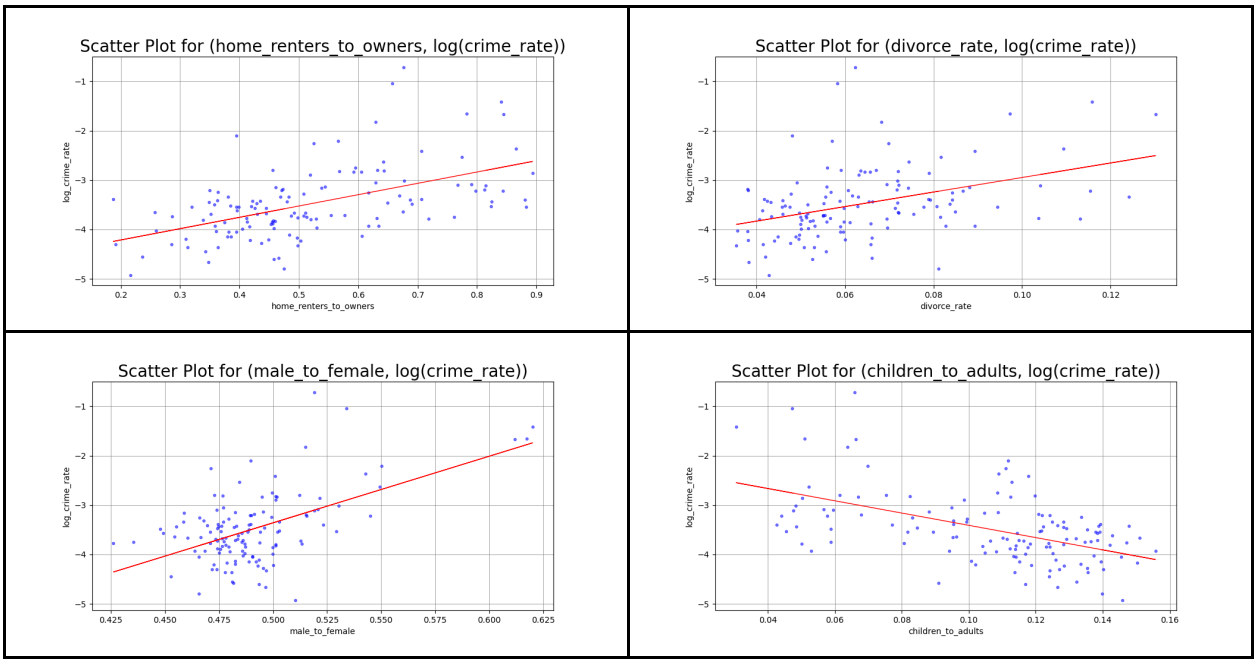


To better understand the relationships of our variables from the 2021 Canadian Census to crime rate, a correlation matrix was created to show the linear relationship of all pairwise combinations of variables. This allows us to closely examine our variables and detect possible multicollinearity, which occurs when variables are linearly related, in the context of estimating crime rates.

The correlation matrix presented some significant findings. Variables that present high positive correlation are `dropouts_to_grads` to `one_parent_to_two` and `home_renters_to_owners` to `divorce_rate`. Negative correlations that may pose collinearity are `home_renters_to_owners` to `children_to_adults` and possibly `non_minority_to_minority` to `crowded_to_not`. Many of these correlations may make intuitive sense, such as renting proportion to children proportion, where raising children requires financial stability, therefore they are more likely to own homes rather than rent in dense urban areas.

Scatter plots of `crime_rate` against the other demographic features were created to visually assess any non-linear relationships that the correlation matrix failed to capture. Similar to the choropleth map and correlation matrix, applying a logarithmic transformation to

`crime_rate` scales the data more effectively, and is therefore more suitable for analysis. Some notable plots are shown below.



These scatter plots show that `home_renters_to_owners`, `divorce_rate`, and `male_to_female` appear to have some positive linear relationship with `crime_rate`, whereas `children_to_adults` suggests a negative linear relationship to `crime_rate`. The other features that were plotted against `crime_rate` do not show any clear relationship, linear or otherwise.

Estimating Variable Effects

For the purpose of our OLS estimation and considering the correlation matrix, we will omit `divorce_rate` and `children_to_adults` as it can likely be explained with `one_parent_to_two` and `home_renters_to_owners`.

The model has an R^2 of 0.493, which could mean that the regression predicts about 49.3% of the variance in crime rate can be predicted with these variables. A noticeable impact to the crime rate is the proportion of males to females, where a 1% increase in proportion leads to a 10% increase in crime rate per person. Unexpectedly, population density has a coefficient of -0.00001938, meaning that a unit increase in population per square kilometre leads to a near 0% change in crime rate. The p-values shows that `pop_density`, `male_to_female`, and `home_renters_to_owners` are statistically significant with $\alpha = 0.05$, meaning that the

slope of these variables are not zero. Another way to interpret it is that these three variables likely have an influence on crime rate. However, this does not mean other variables are insignificant, but rather there is a lack of data to confidently state that it does.

OLS Regression Results

Dep. Variable:	crime_rate_log	R-squared:	0.493
Model:	OLS	Adj. R-squared:	0.459
Method:	Least Squares	F-statistic:	14.48
Date:	Wed, 31 Jul 2024	Prob (F-statistic):	1.33e-14
Time:	21:32:35	Log-Likelihood:	-95.206
No. Observations:	128	AIC:	208.4
Df Residuals:	119	BIC:	234.1
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
pop_density	-1.938e-05	7.44e-06	-2.605	0.010	-3.41e-05	-4.65e-06
dropouts_to_grads	-1.8084	1.413	-1.280	0.203	-4.606	0.989
one_parent_to_two	-2.0985	2.187	-0.960	0.339	-6.428	2.231
crowded_to_not	-2.8971	3.367	-0.861	0.391	-9.563	3.769
non_minority_to_minority	-0.1154	0.393	-0.294	0.769	-0.893	0.663
male_to_female	10.0645	2.507	4.014	0.000	5.100	15.029
home_renters_to_owners	1.2907	0.453	2.847	0.005	0.393	2.189
low_income_status_pct	0.0216	0.016	1.354	0.178	-0.010	0.053
one	-8.5268	1.229	-6.941	0.000	-10.959	-6.094

Omnibus:	26.208	Durbin-Watson:	1.287
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45.301
Skew:	0.934	Prob(JB):	1.46e-10
Kurtosis:	5.237	Cond. No.	9.31e+05

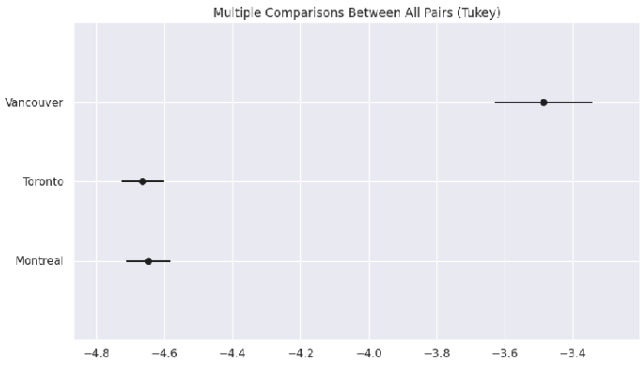
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.31e+05. This might indicate that there are strong multicollinearity or other numerical problems.

ANOVA

Before attempting to create a model for predicting crime rates in Toronto and Montreal, an ANOVA test was conducted. Upon preliminary review of the data distribution for `crime_rates`, it was found that the data was right skewed. To satisfy ANOVA's assumption of normality, $log(crime\_rate + 0.000001)$ was done to avoid $log(0)$. Using `scipy.stats.normaltest()` confirms the `crime_rate_log` of all three cities are now normal. The p-value of the ANOVA test was $8.35 * 10^{-40}$, meaning that we reject the null hypothesis of average crime rate between 3 cities being equal. From this we could then conduct a post hoc analysis using Tukey's HSD test.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Montreal	Toronto	-0.0176	0.9437	-0.1446	0.1095	False
Montreal	Vancouver	1.1625	0.0	0.954	1.371	True
Toronto	Vancouver	1.1801	0.0	0.9734	1.3867	True



The results of Tukey's HSD revealed that Toronto and Montreal had the same crime rate, and only Vancouver's crime rate was different. Taking a look at the distributions suggests that Vancouver's crime rate is greater than the other cities, which may cause crime rate predictions to be inflated when modelling.

Modelling

To train the machine learning model with the city data, feature engineering was done within the **data_processing.py** program to get the variables we wanted. Initially there were only 5 features being considered: average age, population density, proportion of high school dropouts to graduates, proportion of single-parent to two-parent families, and finally the proportion of households with more than one person per room, to households with less than one person per room.

We tried 4 different pipeline models, each using the `StandardScaler()` as its first step and then one of the following regressors as the second step: Gaussian Process, k-Nearest Neighbors, Random Forest, and Gradient Boosting. Then the models were trained and validated with the partitioned Vancouver data.

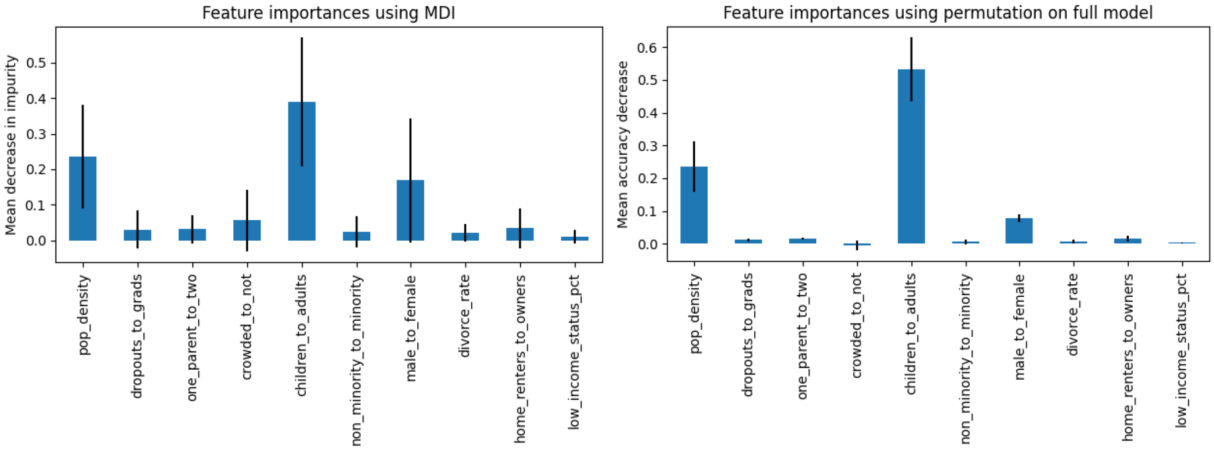
At first, these scores did not make sense; sometimes they were positive and sometimes they were negative. The first thing we did to try and remedy this problem was adding more features for the model to be trained on. Average age was replaced with the proportion of children to adults, and an additional 5 features were also added, giving us the 10 variables listed under the "Data" section of this report. However, these extra features did not seem to help improve the validation scores very much. We suspected that the negative scores were most likely due to insufficient data in Vancouver's dataset. As a result, we took the partitions of all 3 cities' datasets and combined them into a single training set and a single validation set.

```
Gaussian Regressor:           -0.090
kNN Regressor:                0.294
Random Forest Regressor:      0.248
Gradient Boosting Regressor:  -0.087
```

Training the models on this larger dataset yielded somewhat better results; although the Random Forest model had a low validation score ranging from 3% to 30% and averaging around 10%, it was now consistently above 0, whereas previously it had been fluctuating along with the other scores. The model was also validated separately with each city's data to see how it generally performed on a single city, but the scores were still inconsistent. In short, it seemed

that the Random Forest model performed the best out of the 4 machine learning models attempted.

Next, we examined the feature importance of the Random Forest model, trying feature importance based on mean decrease in impurity, and feature importance based on feature permutation. Both methods showed that the proportion of children to adults was consistently the most important feature for predicting crime rates, with population density and gender ratio also being somewhat important.



Conclusion

The analysis of crime in Vancouver provided several insights into its relationship with demographic features including: population density, gender ratios and other socioeconomic factors.

As we approach Downtown Vancouver, an increase in crime occurs in the surrounding neighbourhoods. The downtown area is characterized by a high population density and is a center of economic and commercial activity. This may indicate a need to allocate resources in such places to effectively combat crime.

The statistical analysis performed on the dataset revealed that `pop_density`, `male_to_female`, and `home_renters_to_owners` demonstrate statistical significance at 5%. In other words, these variables likely have a meaningful impact on crime rates and should be considered in future research and policy creation.

Finally, attempts to create a general model for predicting crime rates across urban cities proved to be challenging, suggesting that each city may have unique underlying factors to crime rates. Calculating feature importance revealed that the 3 most significant variables were `children_to_adults`, `pop_density`, and `male_to_female`. This supports our

findings from the statistical analysis, and demonstrates that other variables such as age proportion are also important predictors of crime rates.

Limitations

Firstly, the census data was limited to 2021 because Statistics Canada only surveys the population every 5 years. This meant that we could not analyze this year's crime data or data from more recent years. If more recent data was available, we would have used it.

Many locational data are omitted from the dataset as a means to protect personal data. For this reason, we could not assign the roughly 4,000 crime records of Vancouver to the appropriate census tract due to the unavailability of coordinates. Another fact to note is that the coordinates provided by Vancouver's Police were slightly offset, contributing to possible inaccuracies in crime counts. Despite these limitations, we were left with approximately 28,000 records of crimes throughout 2021.

Another challenge was determining what features should be used as indicators of crime rates. The census data had so many variables to choose from that it was difficult to decide which ones should be used, and with the given timeframe for completing the project, there was little time for further consideration of features. In addition, we could not be sure which features were actually good predictors for crime rates, so we had to guess what we believed were reasonable predictors of crime.

When processing the data, we realized that Vancouver was divided into 128 census tracts. Consequently, we did not have enough data points to train a well-rounded model to predict other cities. This led to the additional data from Toronto and Montreal being used as training data to ensure that our model was moderately consistent with its validation scores.

In retrospect, more data points could have been collected by combining surveys and their corresponding crime count from previous years in Vancouver. For every n surveys included, our dataset would have increased from 128 to $128*n$ data points. This would be valid if census information for every year was available, allowing a larger time frame for our analysis.

Project Experience Summary

Benley Hsiang

- Utilized feature engineering to create useful variables, helping to improve predictions on the machine learning models
- Improved the performance of machine learning models through rigorous testing and optimizations to parameters and features
- Communicated with group members in phrasing ideas, ensuring clear and accurate explanations of findings in the report

April Nguyen

- Developed a visualization of crime counts in Vancouver using Python's folium library to create a choropleth map segmented by neighbourhood boundaries, providing a geographic overview of the data.
- Depicted the relationship between crime rates and other demographic features by creating and customizing scatter plots and a correlation matrix with Python's matplotlib and seaborn libraries, providing a clearer understanding of the data
- Collaborated to write a comprehensive project report to document findings by compiling data and visualizations into detailed sections, effectively communicating the project's outcome

Hayden Mai

- Coordinated with group members on project direction, ensuring effective collaboration and timely executions
- Collected and unified crime and census data from 4 diverse sources, leveraging locational and geometrical information to estimate crime rates
- Developed and implemented statistical tests to estimate the impact of socioeconomic factors on crime rates, providing useful insights and data-driven recommendations