**Determining the best part of Toronto to migrate to**

Harjit Singh Gakhal

January 30, 2018

1. **Introduction**
   1. **Background**

House prices are affected by a wide range of factors of which some are more relevant than others. These different factors contribute to living standards and contribute to the overall value of the neighbourhood, hence make the area either more or less attractive. House prices generally tend to rise over time and small changes if factors can make a big difference in prices, for plans for a train line to pass through a particular neighbourhood, those house prices are likely to go up as there is more accessibility to public transport and thus leading to a better standard of living.

* 1. **Problem**

A number of factors affect the value of neighbourhoods and the houses within them. A more expensive house does not necessarily mean a better standard of living, and I will be exploring many data features related to neighbourhoods to decide some of the best areas to move to in Toronto.

Initially I was going to compare the data between New York and Toronto, however Toronto had some features specific to the city. In addition, it would have been difficult to compare two sperate cities, so I decided to focus specifically on Toronto.

* 1. **Interest**

Many old buildings are being turned into real estate and factors affecting the area such as accessibility and crime rate greatly influence the value of the property, thus making it an interesting topic.

1. **Data acquisition and cleaning** 
   1. **Data sources**

I found a large portion of data on the Toronto open data portal [here](https://open.toronto.ca/catalogue/?sort=last_refreshed%20desc) and the crime data from the Toronto public safety data portal [here](http://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-boundary-file-/data?geometry=-81.356%2C43.368%2C-77.401%2C44.063&selectedAttribute=TheftOver_Rate_2018). The crime rates are for the year 2018, however the rest of the data is for the year 2016, however I feel only a small amount would have changed in that time frame, and hence I believe the data is still relatively accurate.

* 1. **Data Cleaning**

There are 140 neighbourhoods in Toronto, and I sorted each dataset by the index, and used this as a key to combine the data sets in excel. The data sets were found is xlsx (Excel) format, hence it was relatively straight forward to combine the data. I dropped columns that were of little relevance or similar to other columns. For example for crime rates, there was a column for 2018 which was the number of crimes in that neighbourhood, and a column for 2018 rate, which was the crime rate as a proportion of that neighbourhood which was much more useful, and thus I got rid of the other crime rate columns. I dropped the “Social housing waiting list” column as it was similar to the “Social assistance receipts”, and the Social assistance receipts had a better correlation with other variables.

In regard to my question, some areas clearly had very poor stats, but I have kept these neighbourhoods within the data so that we can compare the spread of some of the features. There are some data which we may class as outliers, but these are just more deprived areas, or areas with a unusually high rate of crime, thus I have kept this data, but removed some of it while plotting to see the spread of variables more clearly.

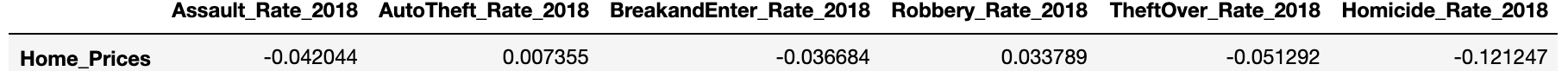
My data set consists of 140 columns and 23 rows.

1. **Exploratory Data Analysis**
   1. **Testing of target variables**

I calculated the correlation of numerous variables and how they correlate to house prices, to see if they were good predictors of a house, and potentially provide a better association with a good neighbourhood. In addition, I also tested the correlation different variables to see how they relate to each other to see how one indicator may be and indicator of another. I will discuss this in more depth in my analysis. I used regression plots to visualise the correlation between variables and used box plots to see the spread of the variables across the neighbourhoods.

* 1. **Relationship between crime and house prices**

It is commonly known that areas with higher rates of crime generally tend to be more deprived areas, and thus house prices are likely to be lower.



We can see from our data that the data was quite diverse with nearly no correlation between higher rates of crimes, suggesting that crime rate is not the best indicator for house prices and thus may not be as significant as we thought it was. Nevertheless, this is not to say that crime rate is not significant as we will still take it into account when making our conclusion.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Looking at some of the crime data we can see that there are some extreme outliers, so we will ignore these values.

A screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generatedA screenshot of a social media post

Description automatically generated

We can see that all the crimes have a positive skew, with indicates more areas have a crime rate lower than the average crime rate, suggesting that crime may not be a major problem in the neighbourhoods of Toronto.

* 1. A close up of a map

     Description automatically generatedA close up of a map

     Description automatically generated**Relationship between Debt risk score and home prices**

A debt risk score is the ability for a homeowner to meet their next 3 house payments, so it is no surprise that neighbourhoods that have homes of higher value attract people that can afford these homes and thus are more likely to pay them back. This graph had a moderately strong correlation of 0.63.

Furthermore, if we compare to debt risk score to Social Assistance Recipients which is essentially the number of people that receive homes from the council there is a strong negative correlation of -0.7 which is understandable as those who require social assistance can’t afford homes.

A screenshot of a cell phone

Description automatically generated

A debt risk score of 700 + is generally considered a good score, and we can see that a vast majority of the neighbourhood residents have a Debt risk score of over 700. In particular, only 15 people had a credit score of less than 700, meaning that 89 % of the neighbourhoods had a good credit score. This may explain why crime rate did not have as much of a influence a I previously thought as there are not as many deprived areas as I previously imagined.

* 1. **Relationship between education and Debt Risk Score**

A close up of a map

Description automatically generated

There was a positive correlation of 0.65 between the proportion of students that went to university from each neighbourhood, and a good Debit Risk Score, suggesting that those neighbourhoods were more likely to have better schools in the area. This is subsequently supported by the positive correlation of 0.5 between Proportion of university of home applicants and home prices.

* 1. **Relationship between Tree cover and Walk score**

**A close up of a map

Description automatically generated**

This is quiet an interesting graph. The relationship between Tree Cover and Walk Score has a moderately negative correlation of -0.47. If we put this into reality, areas with more Tree cover and more green area are likely to be nearer the countryside or more outer city. The walk score shows how accessible nearby facilities are such as the ability to walk to the supermarket, so rural areas are more likely to require a vehicle to travel around and hence a lower walk score. Alone these two variables have little correlation with house prices, which I did not expect, and I thought a low walk score would have a much bigger impact on house prices, but the correlation was only 0.21, and the correlation between tree cover and house prices was only 0.16

I was not able to obtain the data for the transit score which is the accessibility to public transport and how convenient it is to get to other parts of the city. For example, in London you have the underground which connects the city and makes it easy to travel from one side to the other. I can imagine the transit score would have a similar score, where more dense urban areas have more accessibility to other areas as opposed to rural areas.

* 1. **Relationship between Population and traffic collisions**

**A close up of a map

Description automatically generated**

Another interesting observation was that there was a strong correlation between the number of traffic collisions and population of 0.68. This could simply be that more people means there are more cars on the road, and hence a high probability of accidents. This may also be in the case of inner-city neighbourhoods where there is a larger proportion of people and business activity.

Furthermore, those living inner city are likely to have better public transport (a higher transit score) which may appeal to some people and not appeal to others so the higher number of collisions is expected more in the city, but people may not notice this as much as you’d think so may not be a deciding factor.

* 1. **Neighbourhood Equity score**

The Neighbourhood Equity score combines ratings for economic opportunity, social development, health, participation in decision-making and physical surroundings.

A screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generated

We saw that house prices increased as debt score increases, but there is also a strong correlation between house priced and Neighbourhood Equity Score of 0.74, meaning this may be a good factor to determine the standard of living of a neighbourhood. Furthermore, there was a very strong correlation of 0.85 between Neighbourhood Equity Score and Debt Risk Score so this is also another factor to take into account.

1. **Predictive Modelling**

A screenshot of a cell phone

Description automatically generatedI built some models to see how good the features were at predicting the house value, which could have meant that they were a good measure of predicting the best neighbourhood in the area. I took some of the features and built single feature linear models to predict house values. My results are below.

A screenshot of a map

Description automatically generated

A picture containing screenshot

Description automatically generated

A picture containing screenshot

Description automatically generated

As we can see, the predictive models using single features are quite inaccurate so we will look at the multi feature linear regression models.

1. Using Catholic University Applicants, Social Assistance Recipients, Neighbourhood Equity Score, and Debt Risk Score

A picture containing screenshot

Description automatically generated

## Using Catholic University Applicants and Debt Risk Score

A picture containing screenshot

Description automatically generated

We can see that this model with only two variables is more accurate than our previous model with 4 variables.

## Using Catholic University Applicants, Social Assistance Recipients, and Debt Risk Score. This graph looks similar to the previous one.

**A picture containing screenshot

Description automatically generated**

#### **Results and discussion**

#### We can see from our data past the house value of $1,000,000 it is very difficult to predict if very expensive houses necessarily mean an even better standard of living, and not just a more luxurious home. In addition, we see that although not very accurate the above features are good factors to take into account for deciding what neighbourhoods are better than others.

I have decided to split the data into smaller segments based on the value on the house, and then will used some of the features that we have evaluated throughout this task to see which neighbourhood is the best depending on that price range.

I have split the data into 6 categories:

1. Up to $400,000
2. $400,000 - $500,000
3. $500,000 - $600,000
4. $600,000 - $900,000
5. $900,000 - $1,200,000
6. $1,200,000 +

For the data where house prices were up to $400,000

* We can see that west hill had a low relative rate of crime and average number of university applicants. It also had a decent walk sore and a Debit score of 706 which is good so this may be one the better areas to move to within this budget. Another similar area is Woburn with an even lower crime rate but slightly lower university score and higher number of social assistance requests. At the higher end of this range we also Milliken with a higher rate of crime but above average number of University applicants score of 0.82 as well as a neighbourhood equity score of 56.85, much higher than the other two areas

For the data where house prices were between $400,000 and $500,000

* We can see that Rogue has one of the lowest rates of crime, an above average University Applicant score, and a below average Neighbourhood Equity score. Around a similar house price, we have the Waterfront Communities-The Island with low crime as well as an above average University applicant score, high Neighbourhood Equity score and a walk score of 99 making it very ideal.

For the data where house prices were between $500,000 and $600,000

* We can generally see that even though house prices are higher University Applicant scores aren’t that much higher, although Willowdale east was mid-range and had a low rate of crime, and a high College applicants score of 62.50 and a high neighbourhood score of 69.97 making it the ideal place to live in this price range.

For the data where house prices were between $600,000 and $900,000

* It is becoming noticeable that more expensive areas also experience high rates from crime, and one area with a low crime rate is Banbury-Don Mills which also has an above average number of University applicants score of 61.67, and a high neighbourhood score 73.19 and is also in the lower end of this price range, making it the ideal area in this price range.

For the data where house prices were between $900,000 and $1,200,000

* As we reach this much more expensive price range the Neighbourhood Equity Score is consistently high, but crime rates are also high and the number of neighbourhoods with a above number of University applicants score has dropped significantly. From this I would say Yonge-Eglinton is the best area in this price range with a low crime rate, a just above average University Applicants score of 42.86, and very high Neighbourhood Equity Score of 88.11

For the data where house prices were greater that $1,200,00

* Here Rosedale Moore Park has the lowest crime rate and a high Neighbourhood Equity Score of 83.78, but only a slightly above average University applicants score of 46.15. Bridle Path Sunnybrook York mills is the most expensive neighbourhood in all of Toronto, but has an average crime rate, high University Applicants score of 66.77, and Neighbourhood Equity Score of 83.04

To summarise the results, I have put my recommendations into a table alongside their average crime rate, University applicants score, Neighbour Equity Score and Price.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Neighbourhood | Average Crime Rate | University Applicants Score | Neighbourhood Equity Score | Price |
| Up to $400 k | Milliken | 345.625 | 52.67 | 52.85 | $387,879 |
| $400k - $500k | Waterfall Communities The island | 86.55 | 54.55 | 80.99 | $416,759 |
| $500k - $600k | Willowdale east | 49.175 | 62.50 | 69.97 | $540,911 |
| $600k - $900k | Banbury-Don Mills | 47.925 | 61.67 | 73.13 | $613,647 |
| $900k - $1.2 M | Yonge-Eglinton | 96 | 42.86 | 88.11 | $975,449 |
| $1.2 M + | Rosedale Moore Park | 127.875 | 46.15 | 83.78 | $1,265,389 |

1. **Conclusion**

I have learnt that crime rates are common in expensive areas than previously imagined, and that a more expensive home does not necessarily mean that it will have a higher number of University applicants or a higher Neighbourhood Equity Score. If I was able to get data for the transit score, I would have been able to more accurate decisions on which locations are convenient for those who travel a lot, and homes near stations are generally more expensive, so it would have been interesting to see this.