Applied Data Science Capstone Project

GREATER TORONTO AREA REAL-ESTATE VALUE ESTIMATION

--- Arvind Subramanian ---

(04/05/20)

INTRODUCTION

Purchasing a home is truly a pivotal moment, starting a new chapter in life for buyers and their families. It is also one of the most financially critical decisions we make, leaving most people in debt for the foreseeable future. As such, it is vital that home buyers are armed with as much knowledge as possible before taking the plunge, so as to ensure that their hard-earned money is invested wisely.

Unfortunately, most home buyers do not have more than surface-level knowledge of the real-estate market in their cities. They go in with a handful of considerations and preferences in mind. Thereafter, they are at the mercy of real-estate listings or advertisements that are purely meant to market homes to buyers rather than provide an objective review of their value. It is also too cumbersome to trawl through hundreds of historic sales records to determine if a new listing represents fair value. As a result, most buyers succumb to the sales pitch of realtors, which could lead to overpaying and also exposes them to unnecessary risk.

My project aims to address this problem using a two-step approach. First, it will provide home buyers in the Greater Toronto Area (GTA) with an accurate price estimate for their dream home. This will be achieved by analysing real-estate sales data in the GTA for the year 2019, to obtain insights on the various considerations involved and their correlation with the sale price. The information will then be used to model a price calculator that accurately predicts home prices with customisable configurations.

Second, buyers will also be equipped with information on other factors, such as crime rates and fire safety in the GTA. Such data is not readily available to home buyers and since these factors do not directly affect price, they may be even be overlooked. However, they heavily influence the physical and financial safety of buyers and their families. Therefore, being aware of the implications could make or break the decision to purchase a particular home.

Armed with this knowledge, home buyers can venture into the GTA real-estate market with greater confidence and make decisions with peace of mind.

DATA

Data was wrangled from five different datasets provided by various sources for this project. The datasets are as follows:

1. Real-Estate Sales. Canadian law requires all real-estate companies to release their annual sales figures. The complete record of all houses sold in the GTA for 2019 is therefore available from Zoocasa, a popular real-estate listings website in Toronto. The dataset already included most of the required information such as the price, floor area, configuration, coordinates and neighbourhood of each house sold. Additional information, namely the 'Distance from City-Centre' and 'Price per Square Foot' parameters, were calculated using this existing information and appended to the dataset. Some cleaning was also required to fill-in empty values and remove unwanted data. The correlations between parameters in this dataset were primarily used to model and develop the price calculator.

2. GTA Neighbourhoods. The names and postal codes of the various Toronto neighbourhoods were scrubbed from a Wikipedia page. Their coordinates where obtained using the GeoData API. This information was then fed to the Foursquare API to obtain the 'Nearby Venues' dataset. In addition, a GeoJSON file demarcating the locations and area boundaries of each neighbourhood was obtained from GitHub. This was used to produce a Chloropleth Map of the GTA to provide a visual representation of the various districts in the GTA.

3. Nearby Venues. The GTA Neighbourhoods dataset was fed to the Foursquare API using the 'get venues/explore' function. This returned the nearest venues for each neighbourhood within a pre-determined radius. These venues were classified according to category and the Top 3 categories for each neighbourhood were evaluated. The neighbourhoods were then clustered and plotted on the Map. This will provide buyers with an appreciation of the amenities available around their home as well as the general 'vibe' of the neighbourhood.

4. Crime Rates. Historic crime data from was obtained from the Toronto Metropolitan Police database. Data from 2019 was specifically sliced out to ensure that the crime trends observed were up-to-date. The coordinates and type of each incident were extracted from this data and used to generate a HeatMap overlaid on the Map. This will provide buyers with a visual representation of the crime hotspots and average annual crime rates in each neighbourhood.

5. Fire Incidents. Historic fire occurrence data was obtained from the Toronto Fire Department database. Data from 2019 was specifically sliced out to ensure that the fire incident trends observed were up-to-date. The coordinates of each fire incident were extracted from this dataset and used to generate a HeatMap overlaid on the Map. This will provide buyers with a visual representation of the fire-prone areas in their neighbourhoods and also alert them to potential fire damage that might have occurred at or around a particular listing.

Using this combination of datasets, I will provide buyers with both quantitative and qualitative assistance in their decision-making process.

METHODOLOGY

To provide quantitative and qualitative guidance to home buyers, I first put myself in their shoes. I was then able to conceptualise their thought process and any problems that they may experience. The relevant data could then be extracted from the 5 datasets. A two-prong approach was used to obtain and analyse this data.

Quantitative

There are several variables that factor into the price of a home, including:

- Type (Apartment, Townhouse etc.)
- Floor Area
- Configuration
- Proximity to City-Center
- Neighbourhood Median Income

The variables listed above can be broadly categorised as quantitative. They have a direct and tangible impact on the price. These variables could be found in the Real-Estate Sales data. As the dataset was very large, a slice of 1000 listings was used.

The general distribution of price per square foot (p/sqft) was first projected using a Chloropleth map. This provided a visual representation of the price range for each neighbourhood in the GTA. The data was then pre-processed and the missing variables were incorporated. This was followed by a dissection of the data to evaluate the correlation of each of the variables with p/sqft. Bar charts, box plots and regression graphs were used to visualise this relationship. This piecemeal analysis illustrated the relative impact of each variable and the peculiarities of the GTA real-estate market.

Qualitative

There are also qualitative variables that are not measurable or obviously stated in real-estate listings, which can still influence price. These include:

- Common Venues
- Crime Rates
- Fire Incidents

Foursquare location data comes in handy to obtain the common venues in each neighbourhood. The GTA Neighbourhoods dataset was fed into the Foursquare API with the 'get venues/explore' function, to generate the Foursquare Nearby Venues dataset. The top 10 venues for each neighbourhood were collated. Next, the machine learning tool of K-means clustering was employed to group the neighbourhoods into clusters of similar venue types. The clusters were then plotted onto the earlier Chloropleth map of p/sqft distribution in the GTA. This provides a birds-eye view of the various price ranges and the common venues found in each neighbourhood. Buyers can clearly discern the 'vibe' of a certain neighbourhood as well as its amenities, to determine if it will suit their lifestyles.

The coordinates of each incident in the crime rates dataset and the fire incidents dataset were used to generate two HeatMaps of crime distribution and fire prevalence respectively. These were overlaid on the original Chloropleth map. This allows buyers to evaluate the general safety of a particular neighbourhood over 2019.

While past occurrences are no guarantee of future issues, it is unlikely that such statistics would change drastically year-to-year. As such, they are good indicators of the dynamics of a neighbourhood which are difficult to spot.

Based on the quantitative and qualitative analyses above, buyers should be able to shortlist a few homes they find real-estate listings. Now it is time to ascertain if the asking price stated on these listings represents fair value. To determine this, the disparate trends used in the quantitative analysis needed to be evaluated collectively to estimate the price for a given set of variables.

First, the real-estate sales dataset was split into training (80%) and testing (20%) sets. Next, several models were built, including:

- Multiple Linear Regression
- Pipe Method
- Ridge Regression
- Polynomial with Ridge Regression

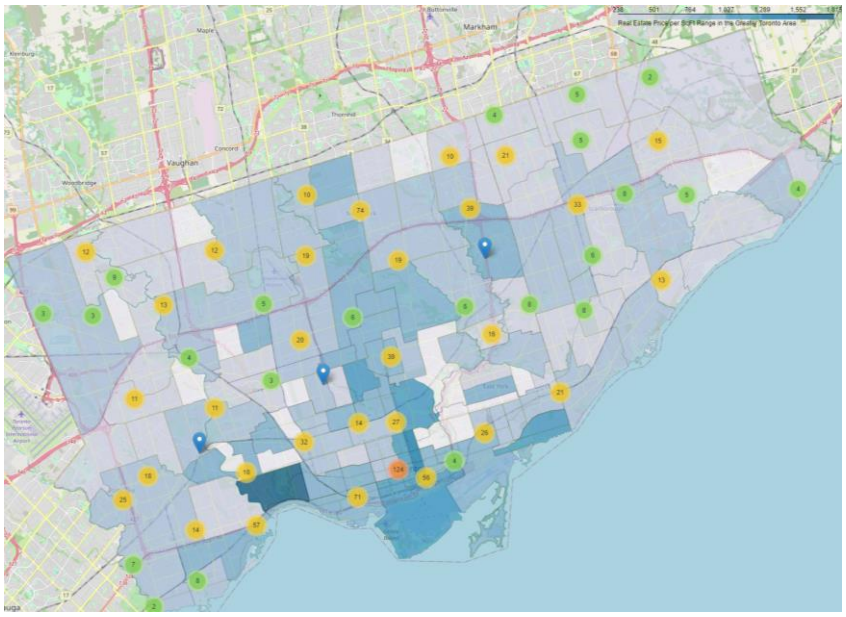
Each model was trained on the training set and tested on the testing set to mimic the models' performance with new data. R-Score and Means-Squared-Error (MSE) were calculated for each model to compare their performance. The most accurate model was chosen and the visualised using a regression graph. To further ascertain the robustness of the model, a separate slice of 1000 listings from the real-estate sales data was obtained. This simulated out-of-sample testing and ascertained the model's performance with real-world data.

After some refinement, the most appropriate model was used to build the price calculator. It requests user input for each of the 5 quantitative variables and returns an estimated price. This allows buyers to accurately estimate the fair price for any listings they find. In addition to this estimate, based on quantitative variables that were modelled into the calculator, buyers can also analyse the 3 qualitative variables to adjust their expected final price accordingly.

This holistic approach to determining the price ensures that buyers do not overpay for a house, find a suitable neighbourhood and minimise their physical and financial risk.

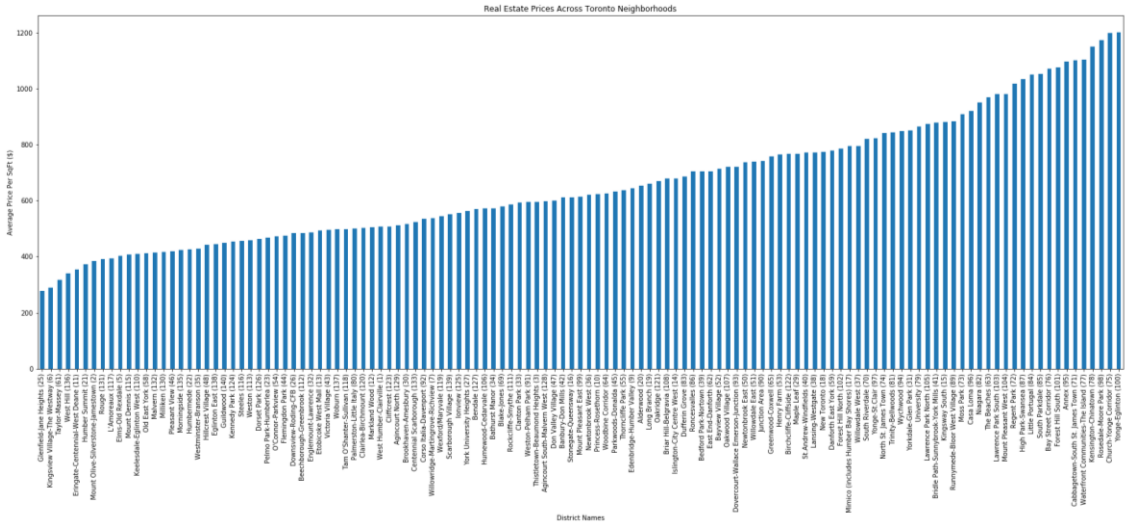
RESULTS

The P/SqFt distribution across GTA neighbourhoods was first plotted on a Chloropleth Map.



Buyers now have a bird's eye view of real-estate sale prices across the GTA. This will give them a rough idea of the P/SqFt to expect in each neighbourhood. We can now dive deeper into each individual consideration / variable.

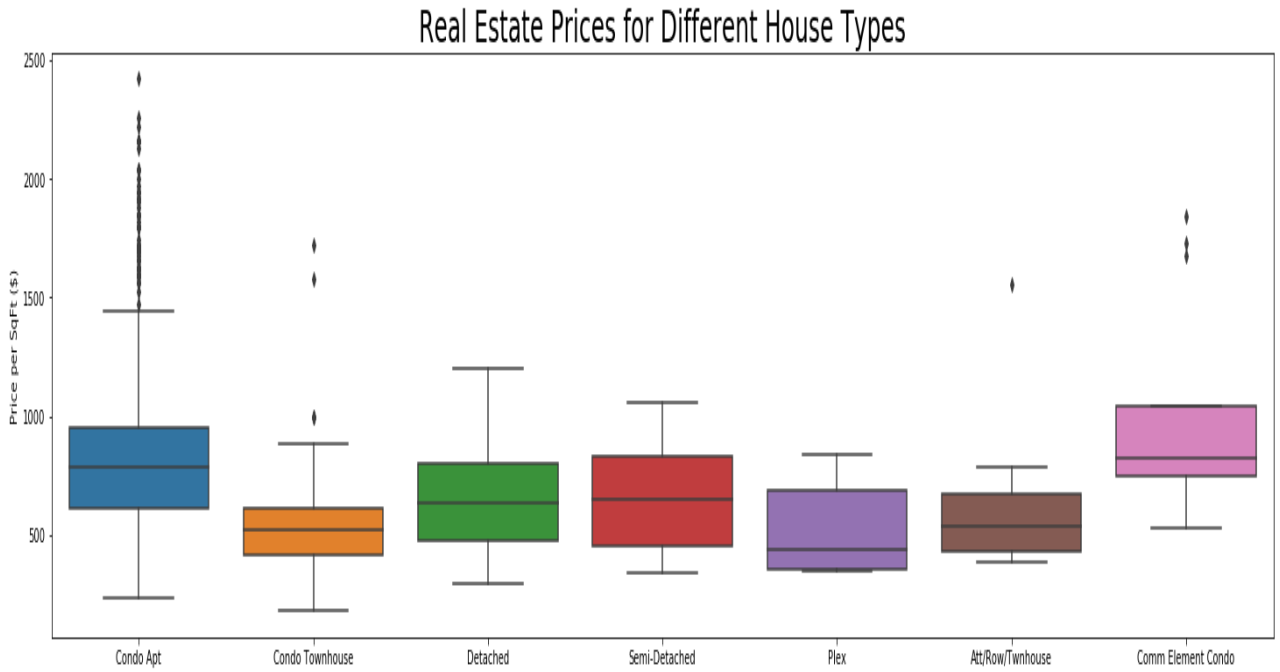
1. <u>Neighbourhood Average P/SqFt Distribution across GTA</u>



This bar chart provides a more quantifiable view of the real-estate prices and allows buyers to pinpoint similarly priced neighbourhoods.

Now that we have observed the price trends, let us examine some typical considerations of buyers and how they affect the prices.

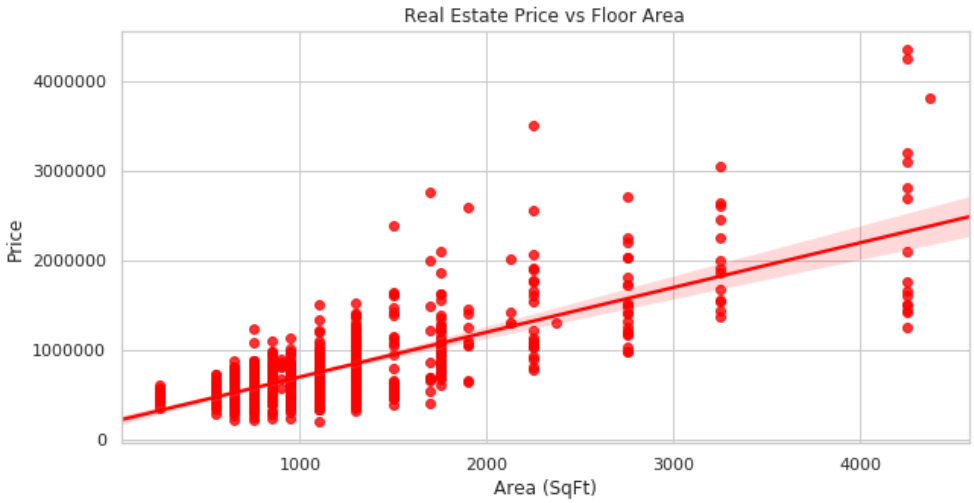
2. House Types



Buyers now have an estimate of the median prices of the various house types across the GTA and their degree of variance with neighbourhood.

3. Floor Area

	final_price	sqft
final_price	1.000000	0.759203
sqft	0.759203	1.000000



We can observe a good correlation between price and area of the houses (0.76). Therefore, size is likely to have the biggest impact on price across all neighbourhoods.

4. Configuration (Number of Bedrooms)

	ppsqft	beds
ppsqft	1.00000	-0.45322
beds	-0.45322	1.00000

This indicates that houses become more 'worth the money' with more bedrooms, up till 3.

5. Configuration (Number of Bathrooms)

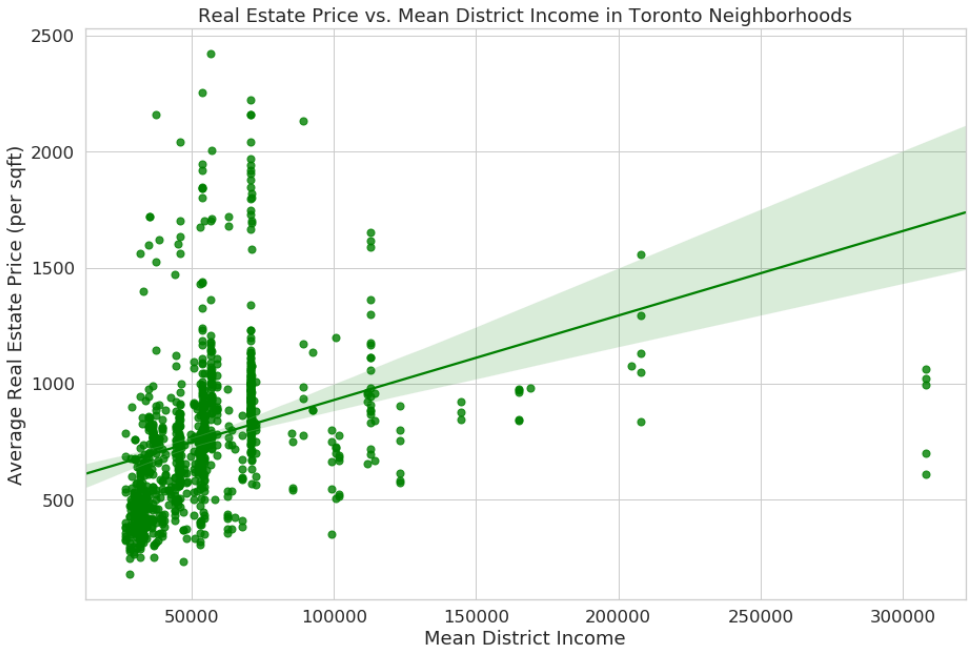
	ppsqft	baths
ppsqft	1.000000	-0.291102
baths	-0.291102	1.000000

This indicates that houses become slightly more 'worth the money' with more bedrooms, up till 4.

The last 2 graphs combined suggest that a 4 Bed, 4 Bath home would provide the best P/SQFT or value for money.

6. Neighbourhood Median Income

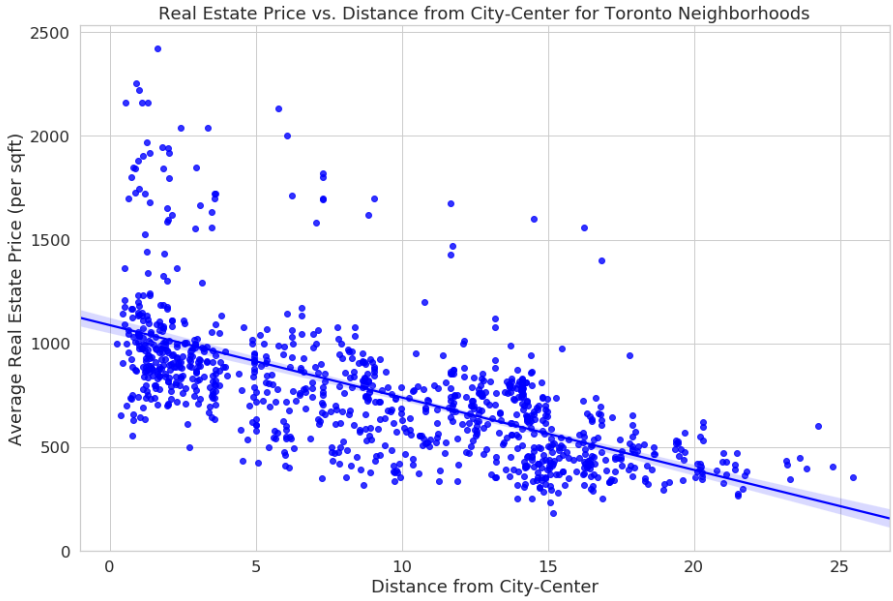
	ppsqft	mean_district_income
ppsqft	1.000000	0.329119
mean_district_income	0.329119	1.000000



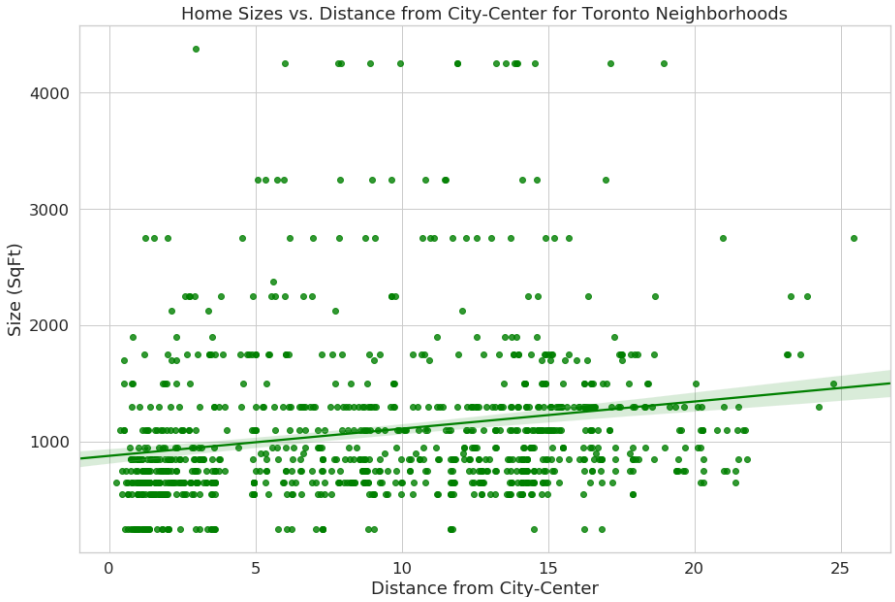
The correlation between mean district income and real-estate prices is positive but poor (0.33).

7. Proximity to City-Centre

	ppsqft	distcc
ppsqft	1.000000	-0.620397
distcc	-0.620397	1.000000



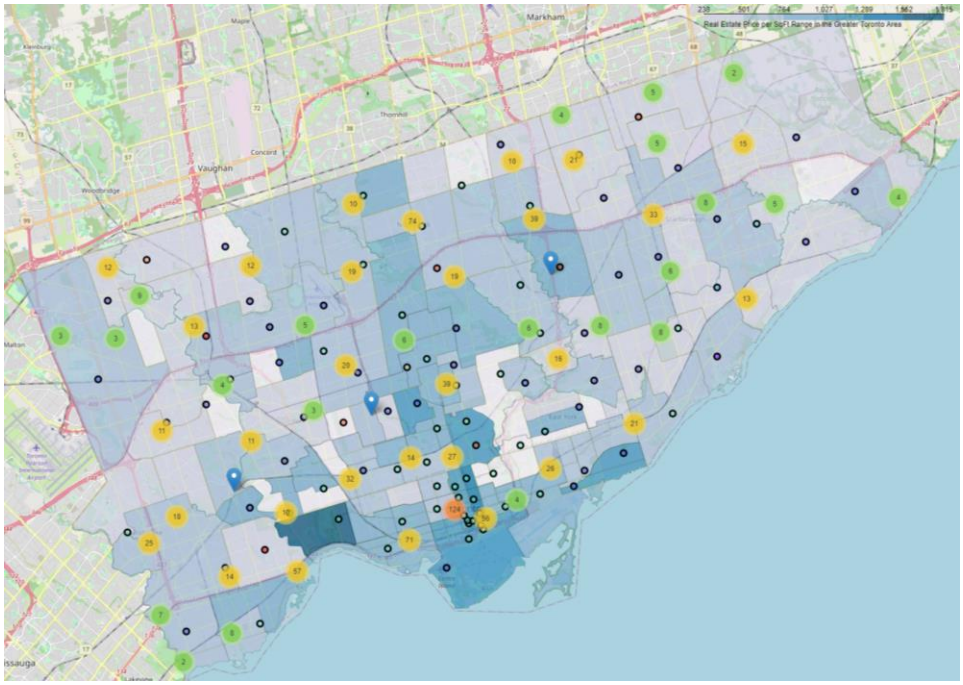
	sqft	distcc
sqft	1.000000	0.204865
distcc	0.204865	1.000000



The results indicate a moderate correlation between proximity to the city-centre and real-estate prices (-0.62). However, the size of houses also does not seem to be greatly affected by proximity to the city-centre.

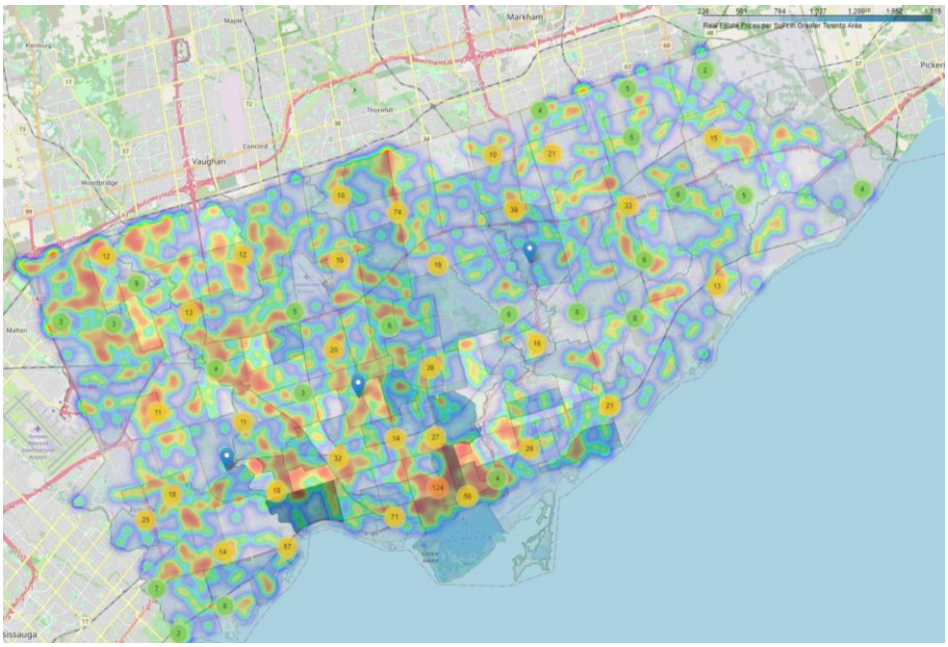
Based on this information, we can conclude that if buyers were to prefer living near the city-centre, they do not need to sacrifice much space but will need to pay a moderate premium.

8. Most Common Venues



Buyers are able to use the above diagram to visualize the culture and 'heartbeat' of a particular neighbourhood, to ascertain if it would meet their daily needs and fit in with their lifestyle.

9. Crime Rate



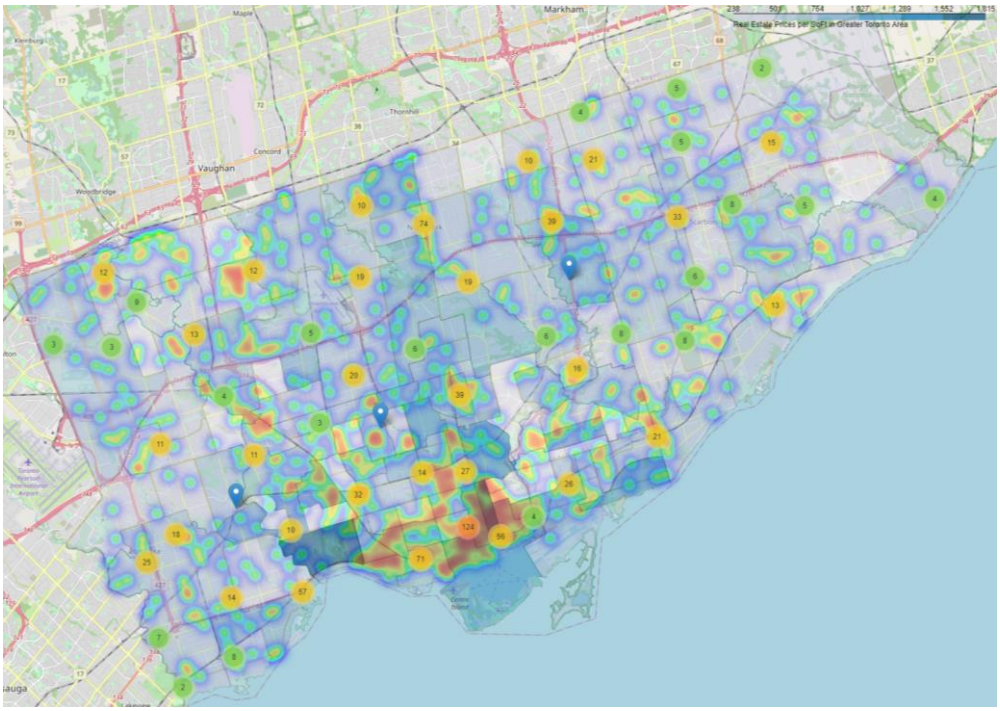
This gives buyers a clear overview of the regions of higher crime in the GTA which they may wish to avoid. You can zoom in on the map to see the individual plots for each crime. This is very useful as such information is not readily apparent or quantifiable for buyers. Let us now examine the effect of crime density on real-estate prices.

	ppsqft	Average
ppsqft	1.00000	-0.14165
Average	-0.14165	1.00000

This regression plot shows that crime rates in a neighbourhood do not significantly affect real-estate prices.

Nevertheless, this information concerns safety and may make or break a buyer's decision to purchase a home in a particular neighbourhood.

10. Fire Safety



The Heatmap displays the prevalence of fire incidents in each neighbourhood. This can alert buyers to the possibility of fire damage at or near a property they are considering. It also encourages them to ensure that fire safety systems are in place and operational.

11. Price Modelling

Model	R-Score	MSE (x 10^{10})
Multiple Linear Regression	0.687	6.57
Pipe Method	0.763	4.99
Ridge Regression	0.687	6.57
Polynomial with Ridge Regression	0.763	4.99

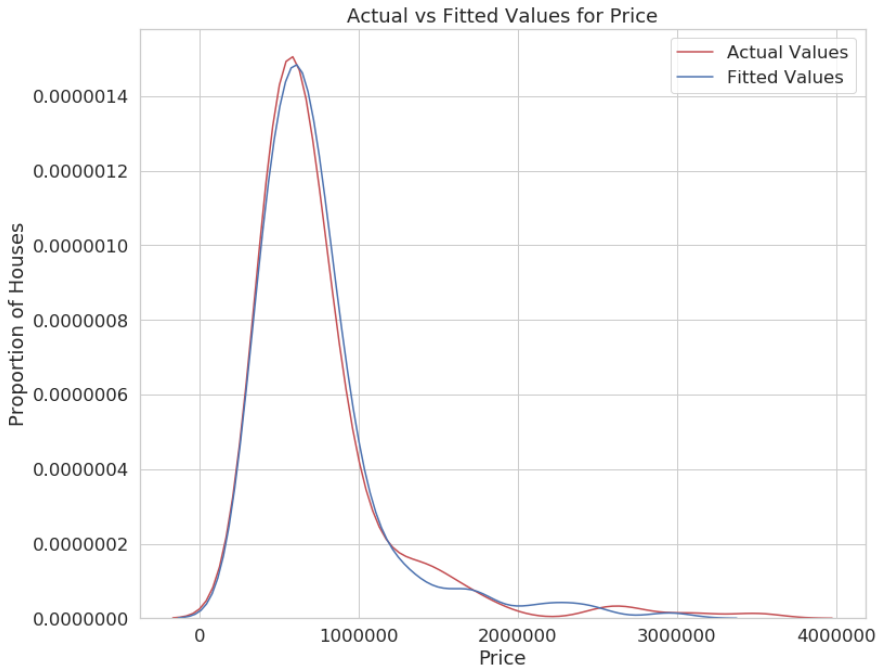
Based on the above tests, it can be observed that Polynomial Regression (degree=2) with Ridge Regression model and the Pipe method producing the best results (R-Score of 0.763, MSE of 4.99 x 10^10). Polynomial with Ridge Regression manages to edge out Pipe when more decimal places are used. As such, it will be the model of choice.

To further confirm the viability of this model, we can train it using the current dataset and test it on completely new data (i.e. a completely different slice of the original dataset) for an 'out-of-sample testing'. This will help to simulate the real-world performance of our model even better.

Model	R-Score	MSE (x 10^{10})
Polynomial with Ridge Regression	0.829	4.27

As shown by the new R-Score and MSE, the model performs even better on real-world data.

Let us visualize its performance on a graph.



The fitted and actual values are very close together up till a price range of about 1,200,000 with slight divergence at the $600,000 region. They are still reasonably close thereafter. This is a good fit.

Therefore, this model can be considered reliable for use in the calculator!

12. Price Calculator

```
Type the number of bedrooms:
 3
Type the number of bathrooms (add 0.5 for half-bathrooms):
 2
Type the floor area in sqft (only digits):
 1300
Parking (1 for yes, 0 for no):
 1
Type the mean annual income of families in the district (only digits):
 120000
Type the distance from the city-center in km (only digits):
 0.5
The estimated price for your selection is $ 1,333,713!
```

The image above displays a sample instance of the price calculator in action. It has taken in user input for the 6 variables and output the estimated price.

DISCUSSION

The following insights can be drawn from analysing the variables.

1. House Type. Most of the various house types are somewhat similarly priced, with condo apartments and common element condos being well ahead of the rest with a median price range at or above the 75th percentile of the others. Condo townhouses have the smallest inter-quartile range, meaning they are the most consistently priced. The cheapest option is the Plex. Condo apartments also have several high outliers, indicating that there were other factors that could have hiked up the price of specific units.

2. Floor Area. We can observe a good correlation between price and area of the houses (0.76). Therefore, size is likely to have the biggest impact on price across all neighbourhoods.

3. Configuration. The p/sqft generally decreases with increasing bedrooms and does not change much with varying numbers of bathrooms. A 4 Bed, 4 Bath home would provide the best p/sqft or value for money.

4. Neighbourhood Median Income. This does not seem to significantly affect p/sqft, with a correlation of jus 0.33.

5. Proximity to City-Centre. The results indicate a decent correlation between proximity to the city-centre and real-estate prices (-0.62). The size of houses also does not seem to be greatly affected by proximity to the city-centre (-0.20). Based on this information, we can conclude that if buyers were to prefer living near the city-centre, they do not need to sacrifice much space but will need to pay a moderate premium.

6. Nearby Venues. The main clusters closer to the city-centre tend to consist mostly of cafes, coffee shops and small restaurants/bakeries. They are meant to cater to a working crowd. For busy singles/working adults, this is ideal as food is readily available. However, for families that prefer to cook and spend time outdoors, consider moving further out where the clusters contain more grocery stores, family restaurants and parks.

For common venues, crime rates, fire safety and the price calculator, the tools presented are interactive. Home buyers need to run through the code in the Jupyter Notebook themselves and play with the features themselves to obtain information on specific listings.

CONCLUSION

This project provides excellent quantitative and qualitative analysis to guide home buyers in the GTA. For quantitative analysis, a price calculator has been modelled to accurately predict future prices based on user input. For qualitative analysis, machine learning techniques have been used on Foursquare location data to place our finger on the 'pulse' of a neighbourhood. In addition, valuable information on crime rates and fire incidents has been provided for each neighbourhood to clearly assess its overall safety.

I believe that the methodology used for this project is sound. However, it could benefit from having more variables in the real-estate sales dataset to account for more of the outliers that were not captured here. Examples include 'Proximity to Train Stations', 'Renovations', 'Views', 'Floor Level' etc. These variables also affect prices directly and may become more impactful as we move closer to the city-centre. This also explains the numerous outliers that were found for the Condo Apartments boxplot, which could have been for premium listings in downtown Toronto.

Having said that, it is worth noting that Data Science always involves an iterative process. As newer and more relevant data comes to light, this project can be updated to make it even more accurate and useful to home buyers in the GTA.