2020

# IBM Applied Data Science Capstone Project Report

## EVALUATING HOUSING PRICES FOR TORONTO NEIGHBOURHOODS UTILIZING REAL ESTATE AND NEAR BY VENUES DATA

TARIQ PERVEZ

Table of Contents

# Introduction

## Background

Toronto is the capital hub of Ontario province with enormous employment and business opportunities. These recompenses have gained attention of several local Canadians and international immigrants, aspiring them to settle and lead a good lifestyle in the city of Toronto. But with all these amenities, there lies Toronto's real estate market hype. For the last few years, significant influx of locals from other provinces of Canada and immigrants from worldwide have outpaced housing demand in comparison to supply. This has lead prices inch steadily upward developing a bubble in Toronto real estate market.

## Business Problem

Immigrants and locals moving into Toronto city founds selection of house challenging for them and their families due to hot housing market. Affordability, accommodation features and nearby facilities always remain qualifying parameters for selection of an appropriate place to live for them.

Price variation with respect to housing attributes for different Toronto neighbourhoods in conjunction with nearby venues needs to be addressed to resolve the encounters faced by any individual during selection of a place. The project seeks to explore real estate data to get an insight of property price variation in combination with its traits and near services available based on location data along all the neighbourhoods to establish relationship between them, which can be used as a recommendation while selecting a place to live.

## Audience & Stakeholders

The target audience for this report are:

- Locals planning to settle from other provinces of Canada to Toronto city
- Immigrants across the globe planning to settle in the city of Toronto
- Potential house sellers and buyers residing in Toronto city who can optimize their advertisements
- Potential business investors setting up their businesses in Toronto city and aims to accommodate their workforce
- City planning authorities to set up more amenities in neighbourhoods with less venues

# Data

This section describes the data sourced for this project.

## Data Sources

This project integrates data sources such as Home finder Canada, Google Map, Wellbeing Toronto as well as Foursquare data. This section describes each of these data sources and provides examples of the data.

### *Home finder Toronto Data*

Home finder Canada is a leading real estate website covering details of commercial and residential properties available for sale. The website consists of around more than 4000 registered properties. In view to the scope of the project, family living options such as Detached, Condo & Town Houses in Toronto city are considered. This

data aims to provide details such as property type (Detached/Condo/Town House), price, details such as number of bedrooms & bathrooms with their respective addresses. Since the data is available on homefinder.ca website, the data have been web scraped from the following link;

https://www.homefinder.ca/homes-for-sale/47618-toronto-real-estate?utf8=%E2%9C%93&sort_order=suggested&bedrooms=Any&advanced_filters%5Blifestyle%5D%5Badult_living%5D=Adult+Living&advanced_filters%5Blifestyle%5D%5Bfamily%5D=Family&advanced_filters%5Blifestyle%5D%5Byoung_professional%5D=Young+Professional&advanced_filters%5Blifestyle%5D%5Bspacious_living%5D=Spacious+Living&commit=Filter+Search&property_types%5Bcondos%5D=Condos&property_types%5Bdetached%5D=Detached&property_types%5Btowns_semi%5D=Towns+%26+Semi

*Home Finder Toronto Data Sample*

Following is a sample of the scraped data extracted from homefinder.ca website showing the property price, details, type and address:

| | Property Price | Property Details | Property Type | Property Address |
|---|---|---|---|---|
| 0 | $624,900 | 1 + 1 beds \| 1 bath | Condo | 5 Marine Parade Dr Unit 211, Toronto |
| 1 | $729,900 | 2 + 1 beds \| 1 bath | Detached | 5 Hatfield Cres, Toronto |
| 2 | $899,900 | 3 beds \| 2 baths | Detached | 534 Rouge Hills Dr, Toronto |
| 3 | $849,900 | 3 + 1 beds \| 2 baths | Detached | 6 Bolger Pl, Toronto |

*Google Map API Data*

Since Home Finder data sample doesn't have latitude and longitude data related to each property, which will be a pre-requisite for visualizing data on choropleth map using folium. Moreover, it will also help in finding out nearby venues using Foursquare API. In this regard, google map API is used to extract latitude and longitude of every property

*Home Finder Toronto with Google Map API Latitude & Longitude Data sample*

Following is a sample of the Home Finder data with latitudes and longitudes extracted utilizing google map API:

| | Property Price | Property Details | Property Type | Property Address | Property Latitude | Property Longitude |
|---|---|---|---|---|---|---|
| 0 | 624900 | 1 + 1 beds \| 1 bath | Condo | 5 Marine Parade Dr Unit 211 Toronto | 43.6299 | -79.4756 |
| 1 | 729900 | 2 + 1 beds \| 1 bath | Detached | 5 Hatfield Cres Toronto | 43.7137 | -79.5463 |
| 2 | 899900 | 3 beds \| 2 baths | Detached | 534 Rouge Hills Dr Toronto | 43.7996 | -79.1344 |
| 3 | 849900 | 3 + 1 beds \| 2 baths | Detached | 6 Bolger Pl Toronto | 43.7248 | -79.5734 |

*Wellbeing Toronto Neighbourhood Boundaries Data*

Wellbeing Toronto (WbTo) provides neighbourhood-level datasets about Toronto services, facilities, and well-being. The datasets are segmented by community indicators grouped under 11 categories, including

demographics, civics, and health as examples. While primarily a mapping application, and underlying datasets are downloadable.

For this project, the dataset of interest includes latitude and longitude data of each neighbourhood and polygon co-ordinates showing limit for each of Toronto's 140 neighbourhood. These data are downloadable in csv/shape/geojson/geopackage format file. Wellbeing Toronto is publicly accessible at https://open.toronto.ca/dataset/neighbourhoods/.

*Wellbeing Toronto Neighbourhood Boundaries Data Sample*

Following is a sample of the neighbourhoods boundaries shape file which includes geometry:

| | geometry |
|---|---|
| 0 | POLYGON ((-79.43592 43.68015, -79.43492 43.680... |
| 1 | POLYGON ((-79.41096 43.70408, -79.40962 43.704... |
| 2 | POLYGON ((-79.39119 43.68108, -79.39141 43.680... |
| 3 | POLYGON ((-79.50529 43.75987, -79.50488 43.759... |
| 4 | POLYGON ((-79.43969 43.70561, -79.44011 43.705... |

Following is a sample of the neighbourhoods boundaries csv file which includes neighbourhood names with their latitude, longitude and geometry;

| | _id | AREA_ID | AREA_ATTR_ID | PARENT_AREA_ID | AREA_SHORT_CODE | AREA_LONG_CODE | AREA_NAME | AREA_DESC | X | Y | LONGITUDE | LATITUDE | OBJECTID | Shape__Area | Shape__Length | geometry |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 5461 | 25886861 | 25926662 | 49885 | 94 | 94 | Wychwood (94) | Wychwood (94) | NaN | NaN | -79.425515 | 43.676919 | 16491505 | 3.217960e+06 | 7515.779658 | {u'type': u'Polygon', u'coordinates': (((-79.4... |
| 1 | 5462 | 25886820 | 25926663 | 49885 | 100 | 100 | Yonge-Eglinton (100) | Yonge-Eglinton (100) | NaN | NaN | -79.403590 | 43.704689 | 16491521 | 3.160334e+06 | 7872.021074 | {u'type': u'Polygon', u'coordinates': (((-79.4... |

Above mentioned both datasets are joined using AREA_NAME which is actually neighbourhood name, LONGITUDE, LATITUDE with the geometry in shape file. In this way, now we have a new dataset of all the neighbourhoods with their defined latitude, longitude and boundaries in terms of polygon which is as follow;

| | AREA_NAME | LONGITUDE | LATITUDE | geometry |
|---|---|---|---|---|
| 0 | Wychwood (94) | -79.425515 | 43.676919 | POLYGON ((-79.43592 43.68015, -79.43492 43.680... |
| 1 | Yonge-Eglinton (100) | -79.403590 | 43.704689 | POLYGON ((-79.41096 43.70408, -79.40962 43.704... |
| 2 | Yonge-St.Clair (97) | -79.397871 | 43.687859 | POLYGON ((-79.39119 43.68108, -79.39141 43.680... |
| 3 | York University Heights (27) | -79.488883 | 43.765736 | POLYGON ((-79.50529 43.75987, -79.50488 43.759... |
| 4 | Yorkdale-Glen Park (31) | -79.457108 | 43.714672 | POLYGON ((-79.43969 43.70561, -79.44011 43.705... |

This data set aims to supplement Home Finder data set by assigning each property their respective neighbourhood name, latitude and longitude by checking property latitude and longitude from Home Finder data set that whether it lies within the specified neighbourhood polygon co-ordinates provided in above attached table. Moreover, neighbourhood boundaries Geojson file is utilized for visualizing neighbourhood extents.

### *Foursquare Data*

Foursquare provides a mobile app that allows users to search for near-by venues and see information and reviews. Users also feed information back to Foursquare both passively, as the app tracks users' locations, and actively as users enter venue names, locations, and reviews.

Since 2009, users have provided Foursquare with location data on over 105 million venues, with over 75 million tips from local experts. As one of the largest sources of location-based venue data, the company describes itself as a technology company that uses location intelligence to build meaningful consumer experiences and business solutions.

This project will access Foursquare venue data for all neighbourhoods. The Foursquare venue data will particularly seek to identify venues that have significant impact on property prices. These data will then be used for subsequent comparison and categorization to provide insight to the business problem.

The Foursquare venue data are accessible via application programming interface (API). A free developer account is used to access the data from https://developer.foursquare.com/places-api.

### *Foursquare Data Sample*

Following is a sample of the imported data showing particularly the venues (by name) and the respective venue categories for each neighbourhood.

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | The Beaches | 43.676357 | -79.293031 | Glen Manor Ravine | 43.676821 | -79.293942 | Trail |
| 1 | The Beaches | 43.676357 | -79.293031 | The Big Carrot Natural Food Market | 43.678879 | -79.297734 | Health Food Store |
| 2 | The Beaches | 43.676357 | -79.293031 | Grover Pub and Grub | 43.679181 | -79.297215 | Pub |
| 3 | The Beaches | 43.676357 | -79.293031 | Upper Beaches | 43.680563 | -79.292869 | Neighborhood |
| 4 | The Danforth West, Riverdale | 43.679557 | -79.352188 | MenEssentials | 43.677820 | -79.351265 | Cosmetics Shop |

## Data Gathering & Cleaning

### *Load web scraped real estate data from homefinder.ca*

Data acquired after web scraping utilizing Beautiful Soup library had several inconsistencies which were resolved in following mentioned steps;

(i) Property Details column had empty cells which were replaced by NAN and then dropped from data frame df.

(ii) Data frame df was checked if any missing value exist but found none.

(iii) Descriptive statistics of the data frame df showed that there was duplication of addresses in Property Address column. Duplication of addresses were counted and then dropped from the data frame.

(iv) Property Price column contained some outliers like prices less than $50000 leading to even $0.00 which was removed from data frame to develop a consistent dataset.

(v) $ sign in the Property Price column was removed and the datatype of the column was converted from object to float, as property price values were going to be utilized for analysis purpose.

(vi) Lastly the Property Details column was split into two columns of Bedroom and Bathroom, as the aim of the project is to establish relationship of property price with every individual attribute of the property.

After cleaning of Data frame df, it was utilized for extracting property latitude and longitudes with the help of GOOGLE MAP API. So uptill now data frame df contained features including Property Price, Property Type, Property Address, Bedroom, Bathroom, Property Latitude & Property Longitude. Property Latitude & Longitude was converted from object type to float type.

*Load Neighbourhood Boundaries data from Wellbeing Toronto*

Neighbourhood boundaries data from Wellbeing Toronto were downloaded as shape file (shx) and comma separated values (csv) file and saved locally for subsequent access. Shape file contained only polygon co-ordinates showing limit for each of Toronto's 140 neighbourhoods, whereas, csv file contained relevant parameters for the project like AREA_NAME, which is the neighbourhood name, LATITUDE & LONGITUDE for neighbourhoods. In addition to this, csv file also contained several irrelevant parameters which weren't required for the project. In order to make this data a useful asset for the project following steps were taken;

(i) Shape file was imported using geo pandas library and was defined a name as df_polygon.

(ii) CSV file of neighbourhood boundaries was imported using pandas library and was defined a name as df_shapedata.

(iii) Irrelevant parameters like '_id', 'AREA_ID', 'AREA_ATTR_ID', 'PARENT_AREA_ID', 'AREA_SHORT_CODE', 'AREA_LONG_CODE', 'AREA_DESC', 'X', 'Y', 'OBJECTID', 'Shape__Area', 'Shape__Length', 'geometry' were dropped from df_shapedata.

(iv) After above mentioned steps, data frame df_polygon and df_shapedata were concatenated which resulted in a new data frame df_shapepolygon with feature including geometry, AREA_NAME, LONGITUDE & LATITUDE.

*Merger of Data frame df & df_shapepolygon*

As the project objective is also to establish relationship between property prices and nearby venues based on neighbourhoods. In this regard, property addresses with their relative latitude and longitude positions needs to be assigned under which neighbourhood they lie. As a result, to generate this idea following steps were taken;

(i) A function was defined in conjunction with shapely library which reads property latitude & longitude from data frame df and check if the property lies within specified polygons coordinates mentioned in data frame df_shapepolygon.

(ii) As a result of any property latitude & longitude occurring within a specified polygon coordinates; neighbourhood name, neighbourhood latitude, longitude & polygon was added in data frame df against each specific address.

(iii) Some property latitudes & longitudes didn't fall in any of the specified polygon coordinates due to which none value was generated for them against neighbourhood name, neighbourhood latitude, longitude & geometry column.

(iv) Rows with none values were dropped and data frame df indexes was reset
(v) Neighbourhood column contained numbers with bracket which was removed
(vi) Final data frame df contained 11 columns & 2902 rows.

*Foursquare venues for Toronto neighbourhoods*

The Foursquare application programming interface (API) was accessed to obtain the venues for neighbourhoods in data frame df. This project also aims to understand any relation between neighbourhood nearby venues with respect to average property price. In this regard a data frame was grouped based on neighbourhood with average property price was developed in following steps;

(i) A new data frame by the name of df_neighbourhood was generated.
(ii) Neighbourhood, Neighbourhood Latitude, Neighbourhood Longitude & Property Price column from data frame df was inserted in data frame df_neighbourhood.
(iii) Data frame df_neighbourhood was grouped based on Neighbourhood, Neighbourhood Latitude & Neighbourhood Longitude and was assigned to a data frame by the name of df_neighbourhoodgrp.
(iv) Property Price was averaged based on grouped Neighbourhood, Neighourhood Latitude & Neighbourhood Longitude column.
(v) Averaged property price was added to the same data frame df_neighbourhoodgrp by the name of Average Property Price.

After developing data frame df_neighbourhoodgrp, venue data using Foursquare API was acquired as follow;

(i) Client ID, Client secret & version was entered in order to access foursquare venue data
(ii) A function was developed which accessed neighbourhood name, its latitude & longitude from data frame df_neighbourhoodgrp and extracted venues based on that.
(iii) Search radius was defined as 1000 m from respective neighbourhood latitude & longitude values with a defined limit of 100 venues per neighbourhood.
(iv) Extracted venues data with neighbourhood name, latitude & longitude was appended to a new data frame toronto_venues.
(v) Toronto_venues dataframe contained features including Neighbourhood, Neighbourhood Latitude, Neighbour Longitude, Venue, Venue Latitude, Venue Longitude & Venue Category.

# Methodology

This section describes the data exploration, inferences and machine learning approach that was applied and how they relate to the original business problem of gaining data insights specifically to identify property price variation in combination with its attributes and near services available based on location data along all the neighbourhoods and develop relation between them.

The methodology includes exploratory data analysis with the aid of histogram, boxplots & barplot, data visualization using choropleth mapping, regression analysis to investigate the influence of property attributes & nearby venues on property prices, as well as the choices and considerations within the methods.

## Exploratory Data Analysis

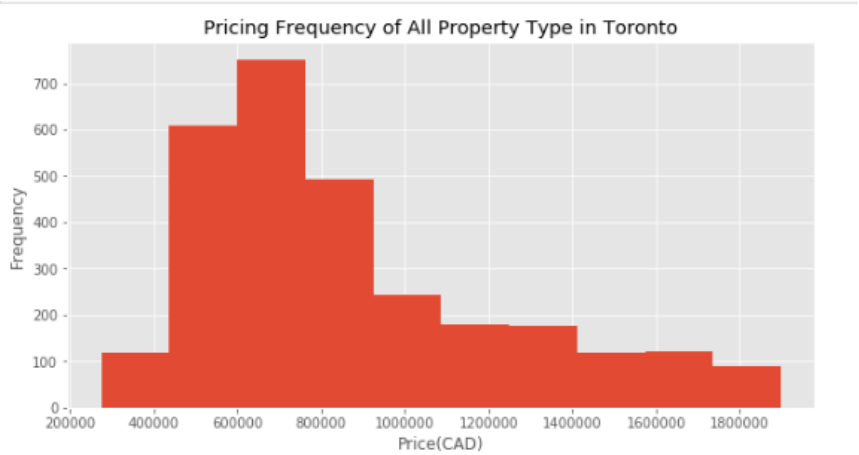
Exploratory data analysis was conducted utilizing histogram, boxplots & barplot.

### Histogram

This exploration of data present four histograms for comparison of property prices. The first histogram presents pricing frequency for all property types, the second histogram presents pricing frequency for condos, the third histogram presents pricing frequency for detached houses and lastly, the fourth histogram presents pricing frequency for Town/Semi houses.

#### 1. Histogram for All Property Types

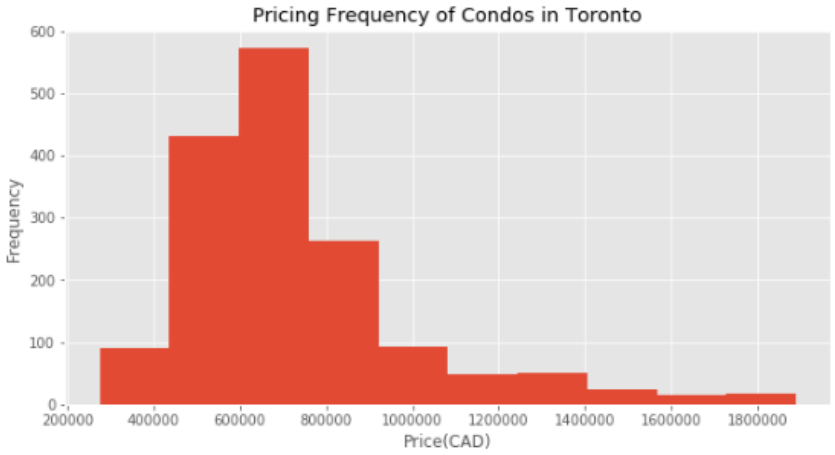
This histogram is generated utilizing data frame df.




This histogram illustrates that mostly property prices lies in the range of CAD 450,000 – 920,000.

#### 2. Histogram for Condos

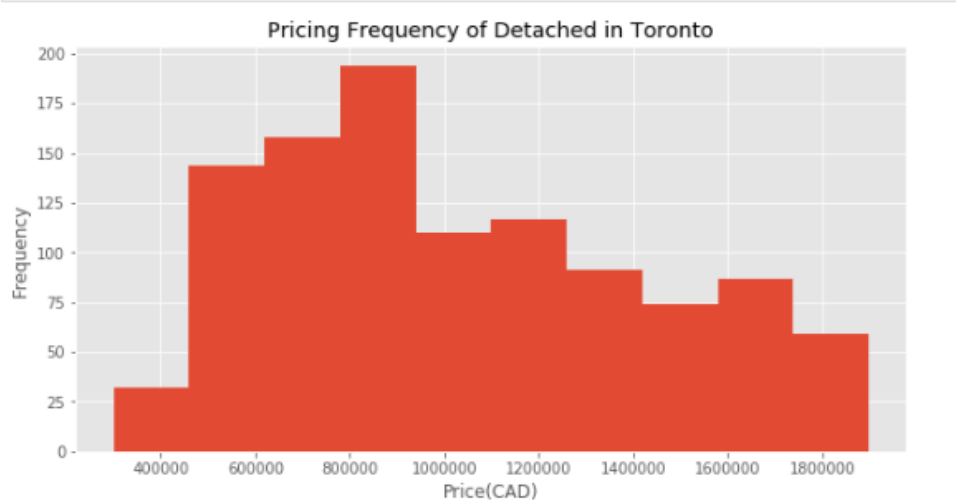
In order to generate this histogram a new data frame df_condo was generated with removal of other property types including detached & town/semi houses.



This histogram illustrates that mostly condo prices lies in the range of CAD 450,000 – 750,000.

### 3. Histogram for Detached Houses

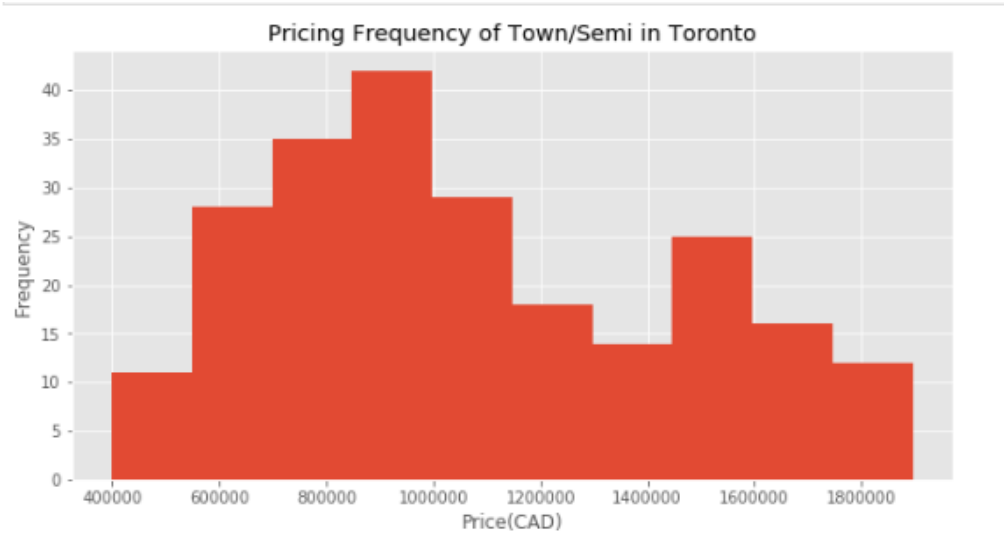
In order to generate this histogram a new data frame df_detached was generated with removal of other property types including condos & town/semi house.




This histogram illustrates that mostly detached houses prices lies in the range of CAD 450,000 – 1,250,000.

### 4. Histogram for Town/Semi Houses

In order to generate this histogram a new data frame df_townsemi was generated with removal of other property types including condos & detached houses.



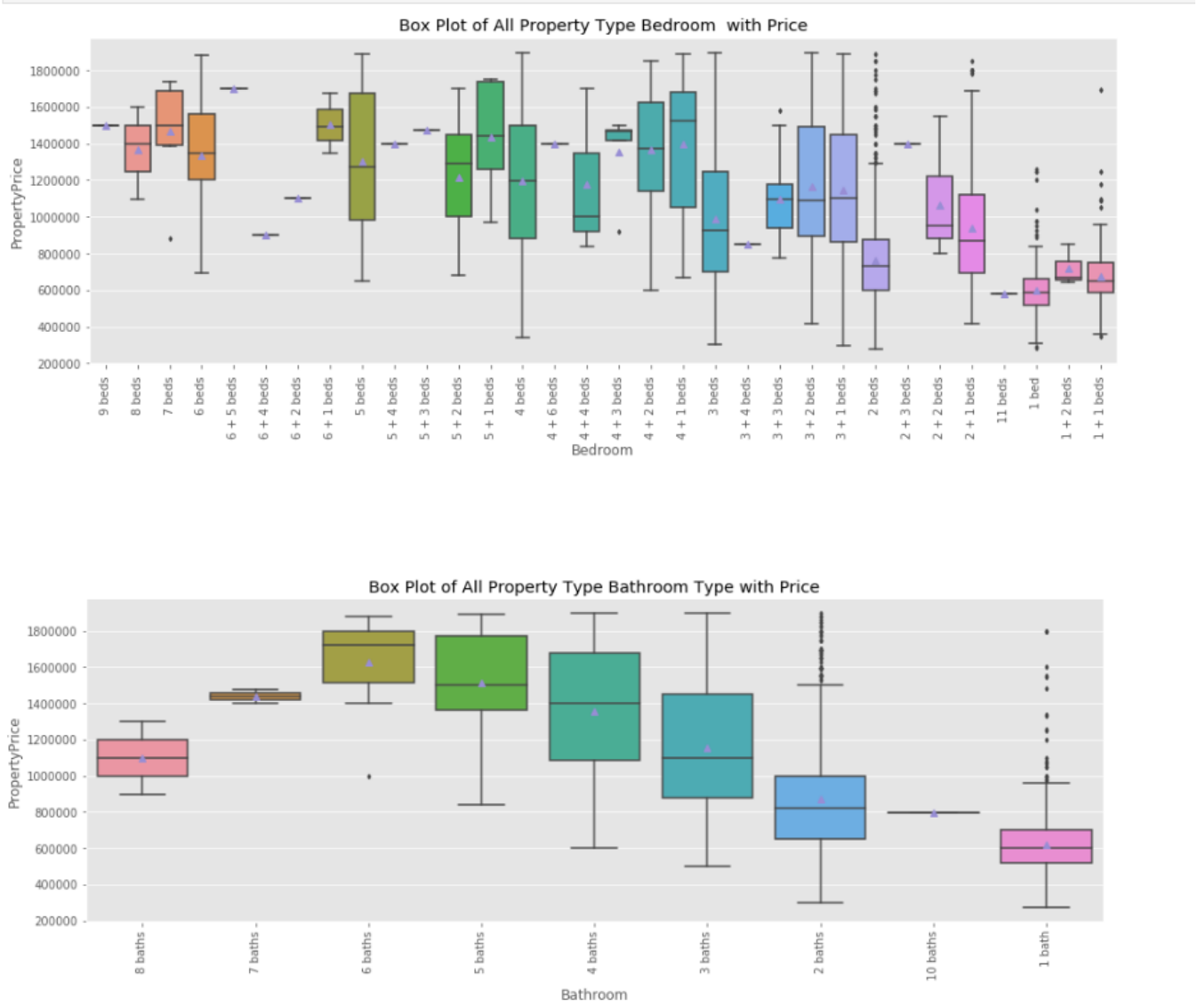

This histogram illustrates that mostly town/semi houses prices lies in the range of CAD 550,000 – 1,150,000. Further observations and comparisons will be discussed in the Results section.

## Boxplot

This exploration of data presents eight boxplots in order to investigate the distribution of property price with respect to any property number of bathrooms & bedrooms. The first two boxplot is for all property types, third & fourth is for condos, fifth & sixth is for detached houses & seventh & eight is for town/semi houses.

### 1. Boxplot for All Property Types

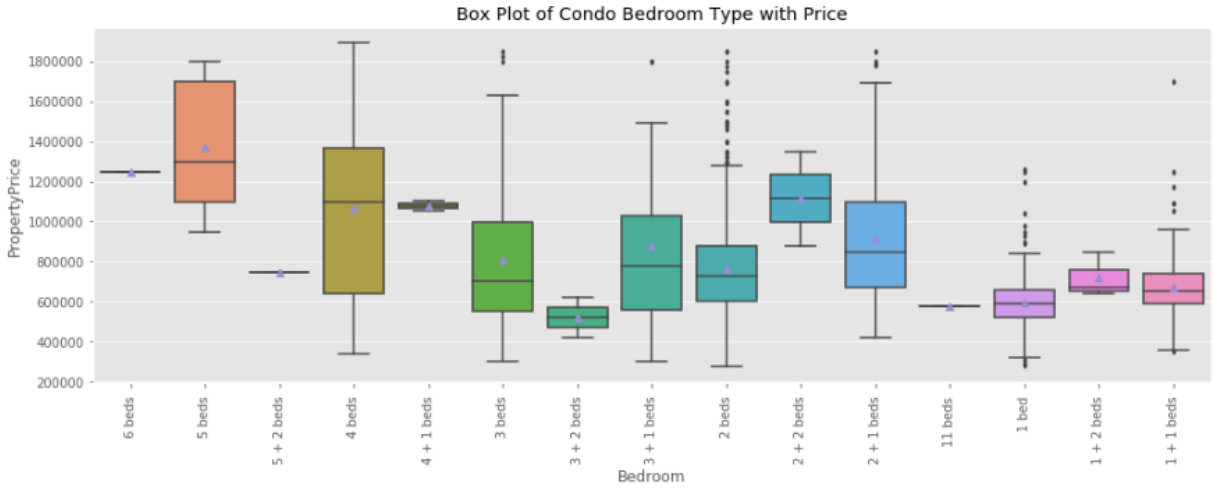
This boxplot is generated utilizing data frame df.

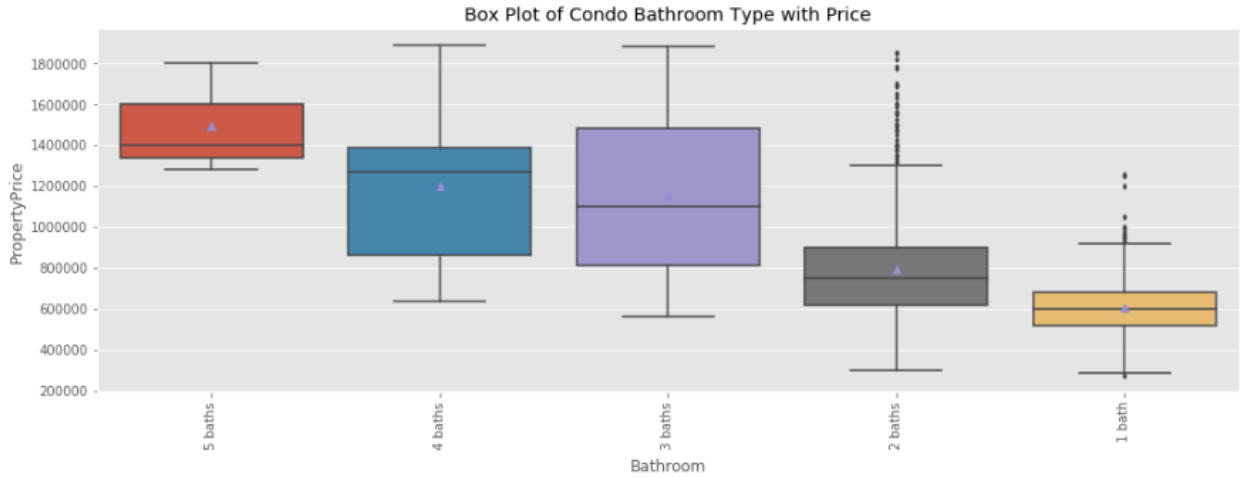



In view to above generated boxplot, prices for all property types seems to have more association with the number of bathrooms in any property.

2. Boxplot for Condos

This boxplot is generated utilizing data frame df_condo.

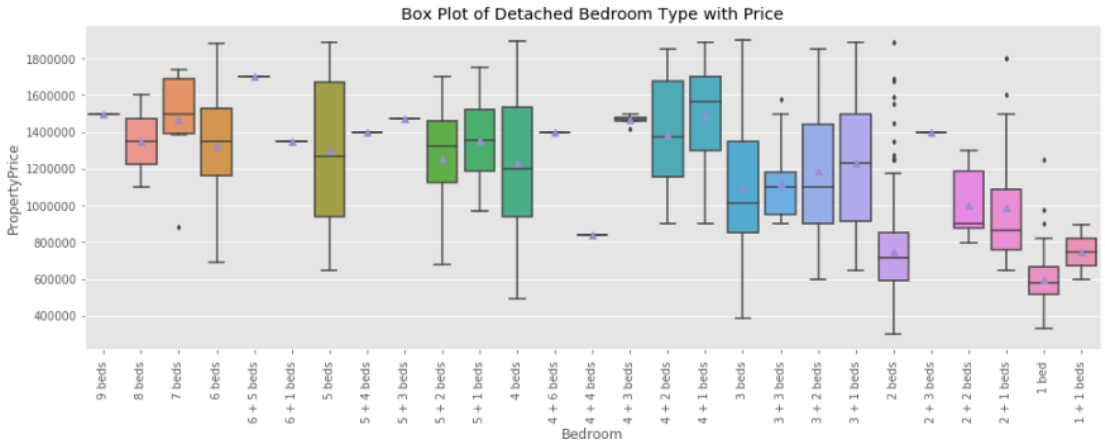


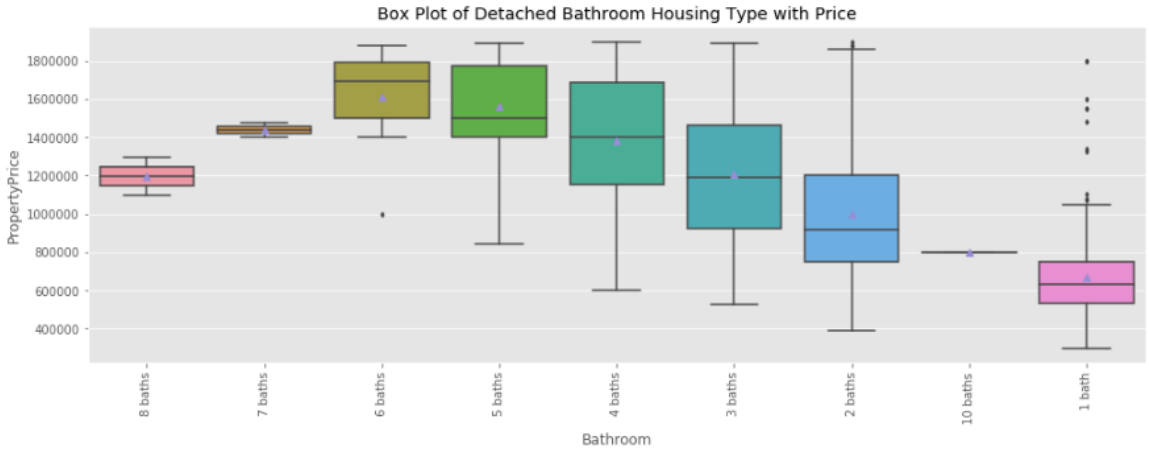




In view to above generated boxplot, prices for all condos seems to have more association with the number of bathrooms in any condo.

3. Boxplot for Detached Houses

This boxplot is generated utilizing data frame df_detached.

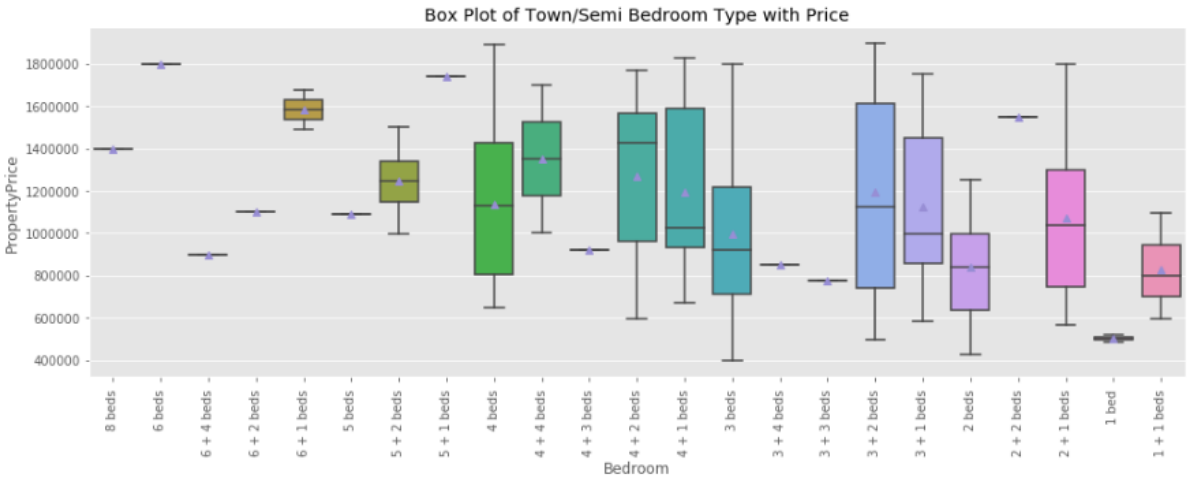


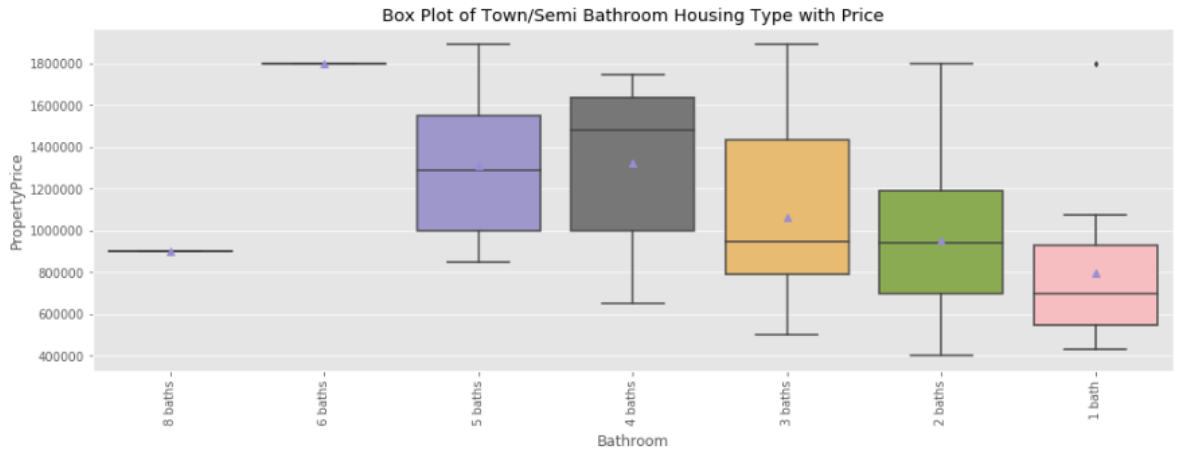




In view to above generated boxplot, prices for all detached houses seems to have more association with the number of bathrooms in any detached house.

## 4. Boxplot for Town/Semi Houses

This boxplot is generated utilizing data frame df_townsemi.




In view to above generated boxplot, prices for all town/semi houses seems to have more association with the number of bathrooms in any town/semi house.

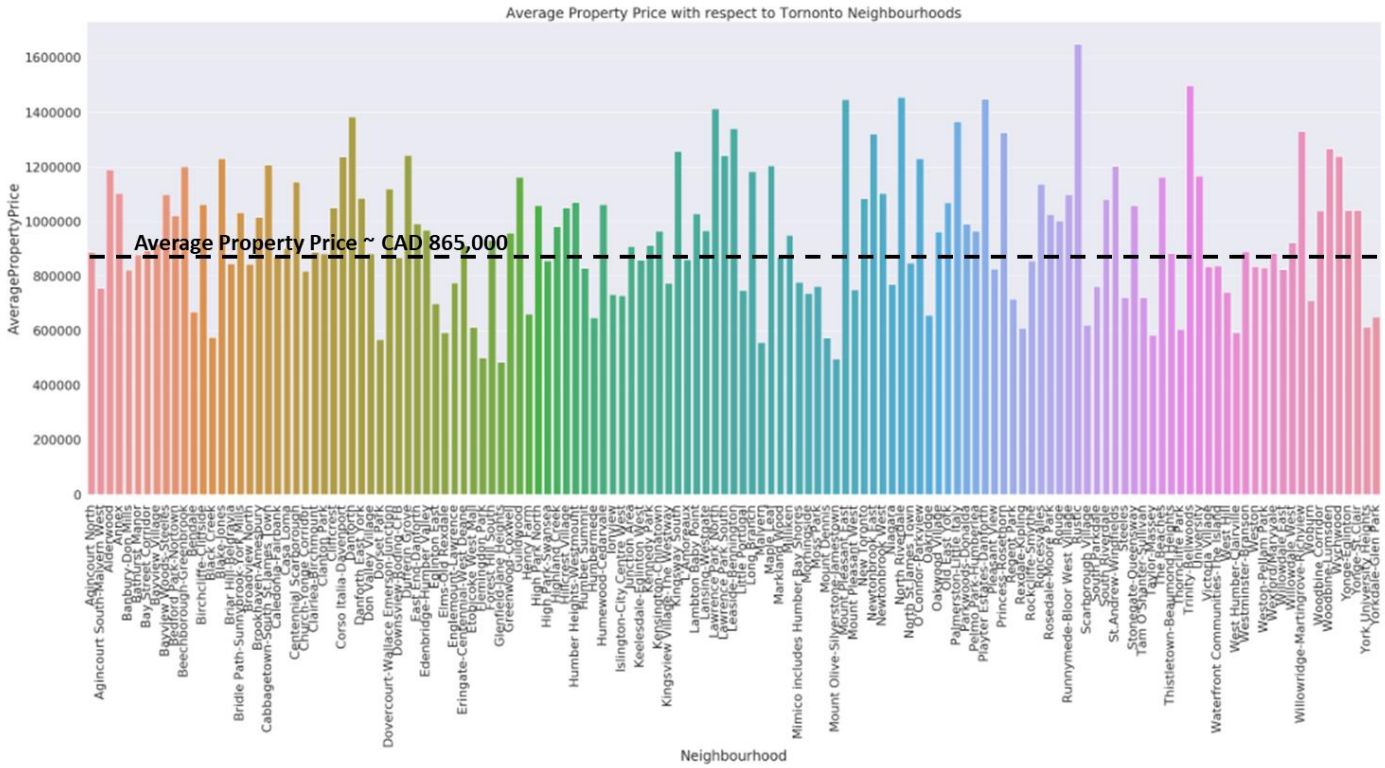
In order to measure the strength and direction of the relationships between property price and number of bedrooms & bathrooms, spearman's rank order correlation is utilized. Below is the tabulated summary of the correlation results;

| Cases | Rs coefficient (Strength of correlation) | P-value (Probability) |
|---|---|---|
| All Property Type Bedroom | 0.54 | 4.72E-223 |
| All Property Type Bathroom | 0.64 | 0.00 |
| Condo Bedroom | 0.24 | 5.88E-23 |
| Condo Bathroom | 0.49 | 1.42E-98 |
| Detached houses Bedroom | 0.59 | 8.37E-100 |
| Detached houses Bathroom | 0.66 | 2.23E-133 |
| Town/Semi houses Bedroom | 0.19 | 0.004 |
| Town/Semi houses Bathroom | 0.37 | 1.07E-08 |

Based on above results computed through spearman's rank correlation a moderate to strong correlation exist between bathroom and price for all the property types, whereas in case of all property type & detached houses a weak to moderate correlation also exist between bedroom & property price. But in terms of correlation coefficient, cases with number of bathrooms lead due to more Rs value in comparison to bedroom ones.

## Barplot

This barplot is generated utilizing data frame df.



Based on above generated bar plot, most of Toronto's neighbourhood lies above the average property price i.e ~ CAD 865,000 with **Rustic & Trinity Bellwoods** in leading position.

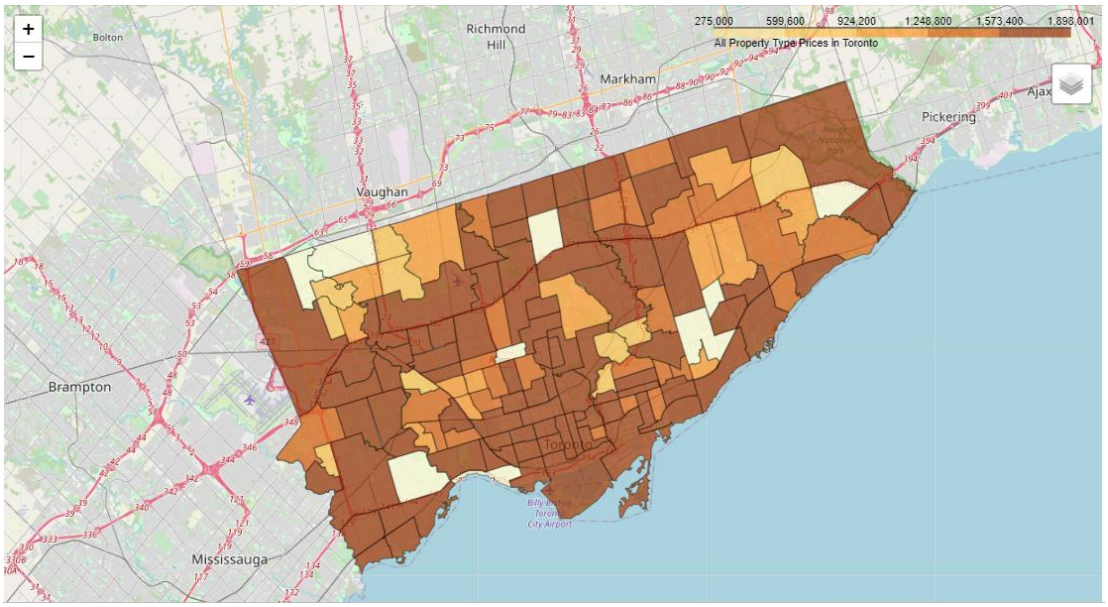
## Data Visualization

This section will cover the visual exploration of the data utilizing Folium Chloropleth map to get an insight of the following attributes;

(i) Property price for all property type
(ii) Property price for Condos
(iii) Property price for Detached Houses
(iv) Property price for Town/Semi Houses
(v) Count of Condos
(vi) Count of Detached Houses
(vii) Count of Town/Semi Houses
(viii) Venue Count
(ix) Venue Category Count

### (i) Map for Property price of all proper type

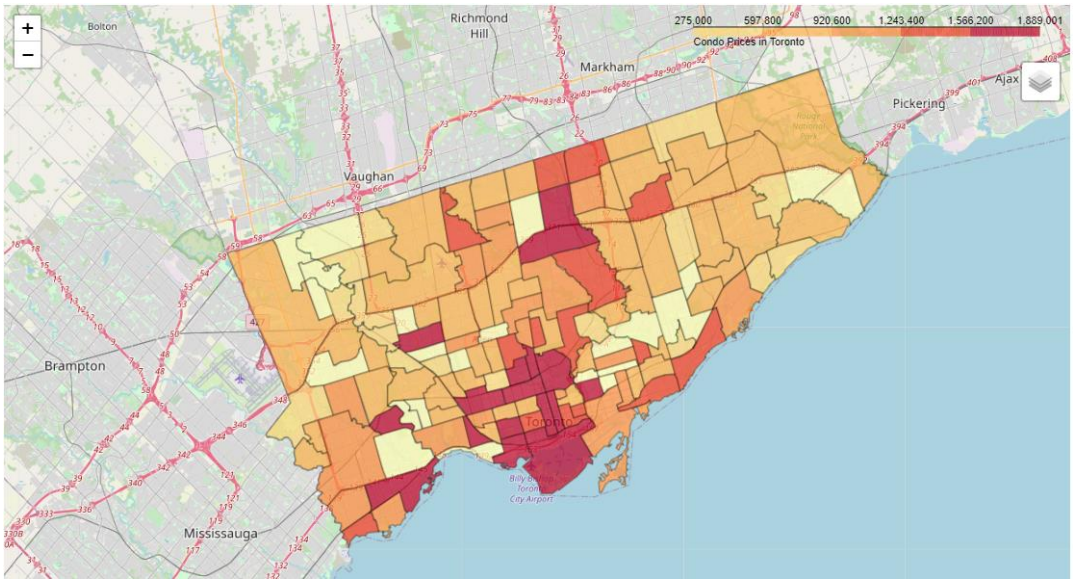
This map is generated utilizing data frame df.




This choropleth map illustrates the price for all property types among all the neighbourhoods in Toronto. The darker areas indicate neighbourhood with price in the range of CAD 1,573,000 – 1,898,000 while the lightest areas have prices less than this. This choropleth map illustrates that the majority of neighbourhoods fall under the highest bracket of property price.

### (ii) Map for Property price of Condos

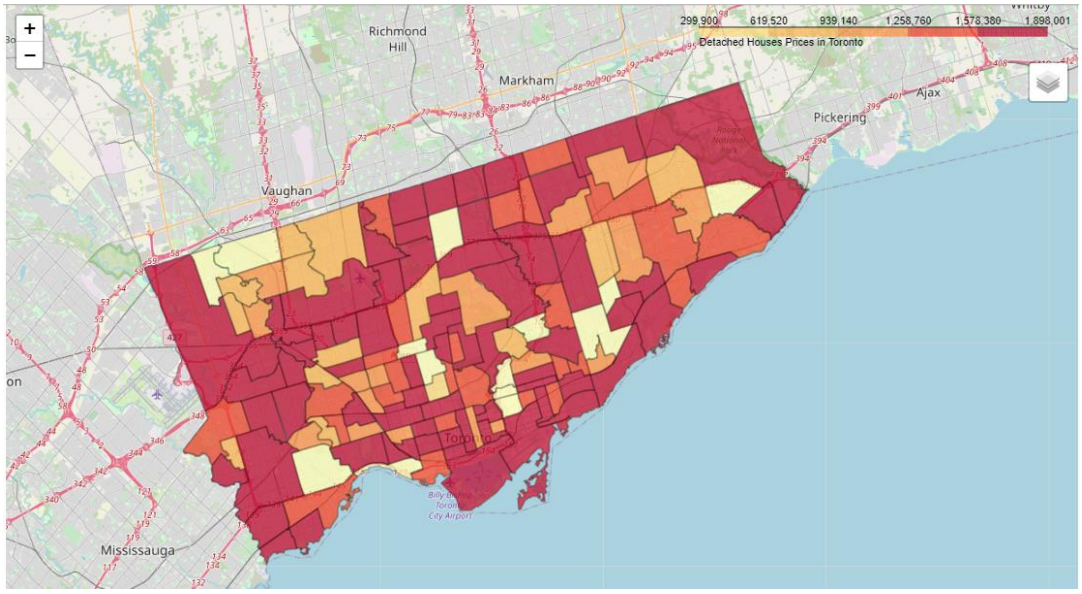
This map is generated utilizing data frame df_condo.




This choropleth map illustrates the price for condos among all the neighbourhoods in Toronto. The darker areas indicate neighbourhood with price in the range of CAD 1,573,000 – 1,898,000 while the lightest areas have prices less than this. This choropleth map also illustrates that condos located in the downtown and midtown Toronto falls in the highest price bracket.

### (iii) Map for Property price of Detached Houses

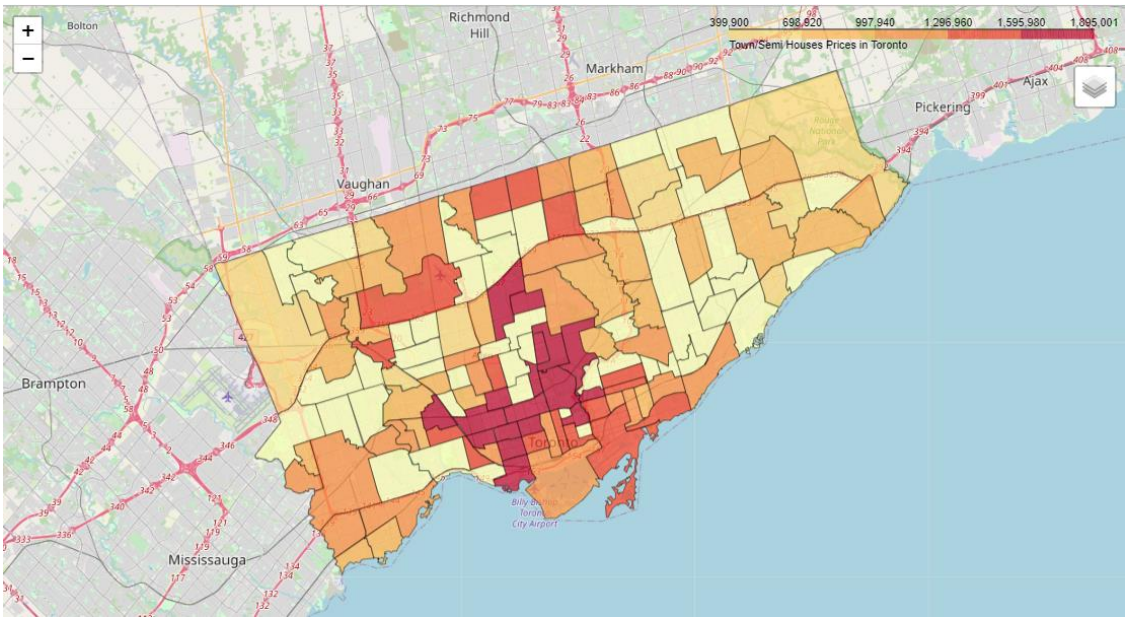
This map is generated utilizing data frame df_detached.



This choropleth map illustrates the price for detached houses among all the neighbourhoods in Toronto. The darker areas indicate neighbourhood with price in the range of CAD 1,573,000 – 1,898,000 while the lightest areas have prices less than this. This choropleth map also illustrates that detached houses among all the neighbourhoods in Toronto falls under the highest price bracket, except few neighbourhoods which are located near to boundary of Toronto city and the ones which are in the central part of Scarborough district.

### (iv) Map for Property price of Town/Semi Houses

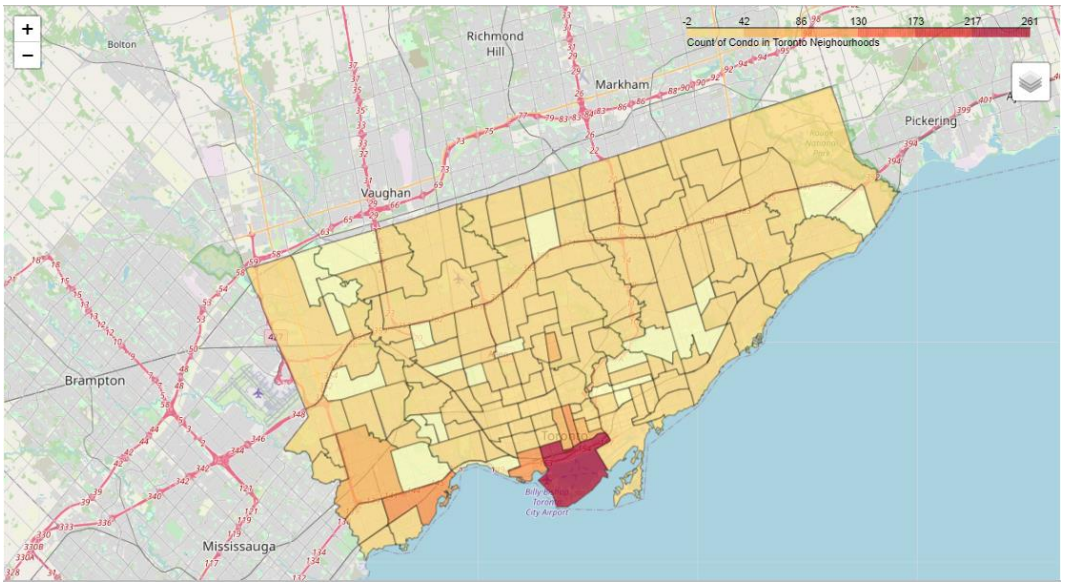
This map is generated utilizing data frame df_townsemi.




This choropleth map illustrates the price for town/semi houses among all the neighbourhoods in Toronto. The darker areas indicate neighbourhood with price in the range of CAD 1,573,000 – 1,898,000 while the lightest areas have prices less than this. The choropleth map also illustrates that most of town/semi houses falls under lower to mid-price bracket except few ones which are in midtown & west end district.

### (v) Map for Condos Count

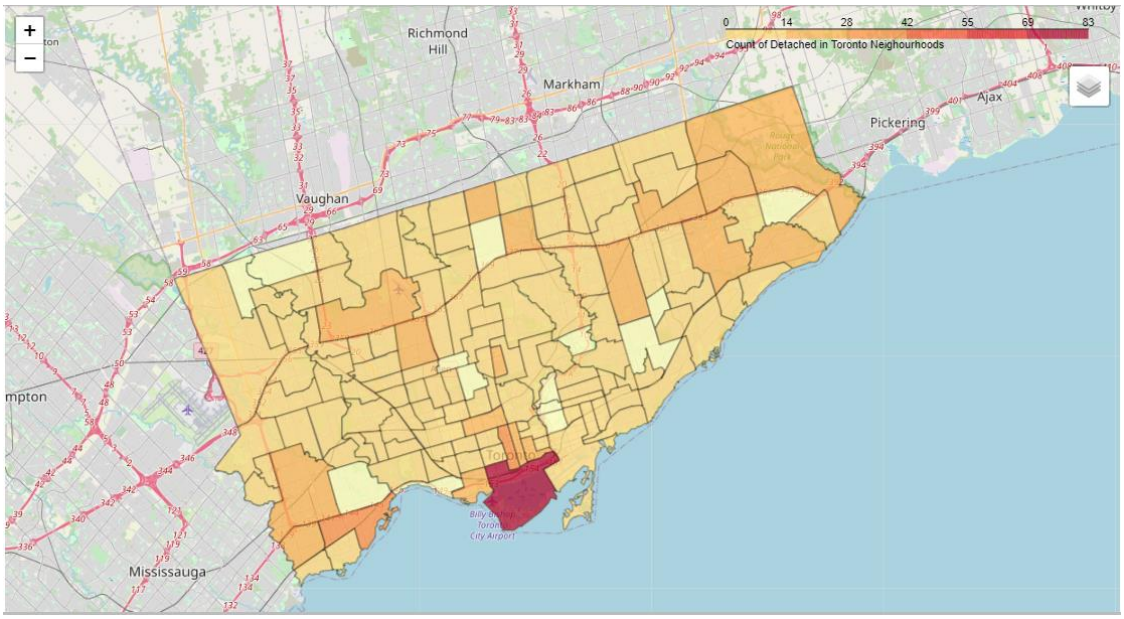
In order to generate this map data frame df_condo was grouped based on Neighbourhood, Neighbourhood Latitude & Neighbourhood Longitude column to get count of condos against each neighbourhood and defined to a new data frame df_condo_group which was utilized for mapping.




This choropleth map illustrates the count of condos among all the neighbourhoods in Toronto. The darker areas indicate maximum number of condos in the range of 217 – 280 while the lightest have count less than this. This choropleth map also illustrates that maximum number of condos are concentrated in downtown Toronto in Waterfront Communities-The island.

### (vi) Map for Detached Houses Count

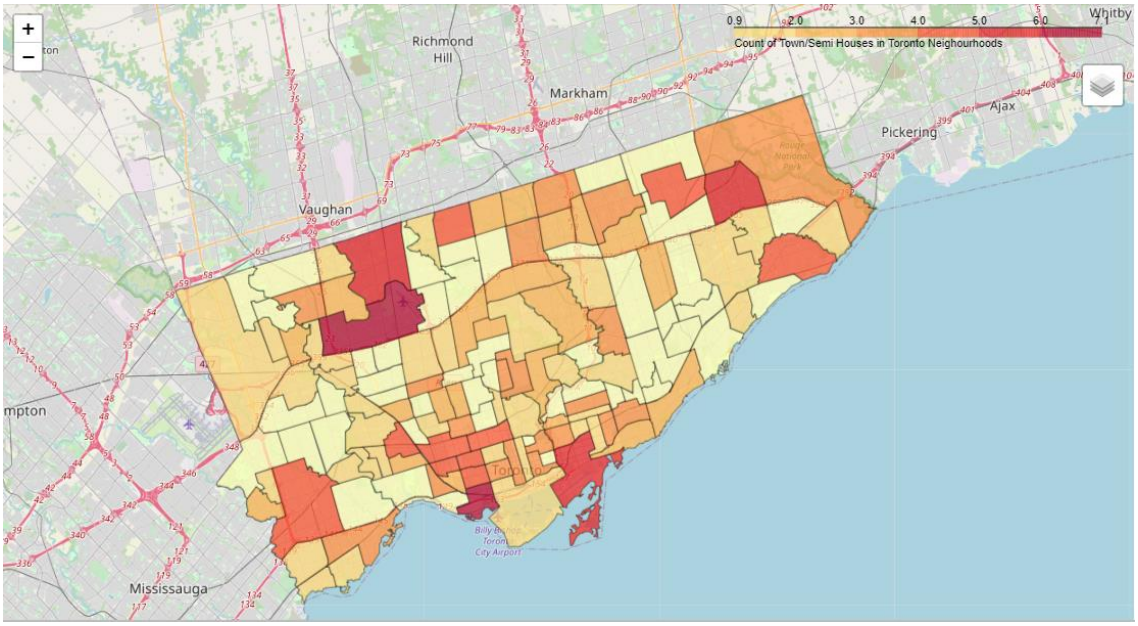
In order to generate this map data frame df_detached was grouped based on Neighbourhood, Neighbourhood Latitude & Neighbourhood Longitude column to get count of detached houses against each neighbourhood and defined to a new data frame df_detached_group which was utilized for mapping.



This choropleth map illustrates the count of detached houses among all the neighbourhoods in Toronto. The darker areas indicate maximum number of detached houses in the range of 70 – 83 while the lightest have count less than this. This choropleth map also illustrates that maximum number of detached houses are concentrated in downtown Toronto in Waterfront Communities-The island.

### (vii) Map for Town/Semi Houses Count

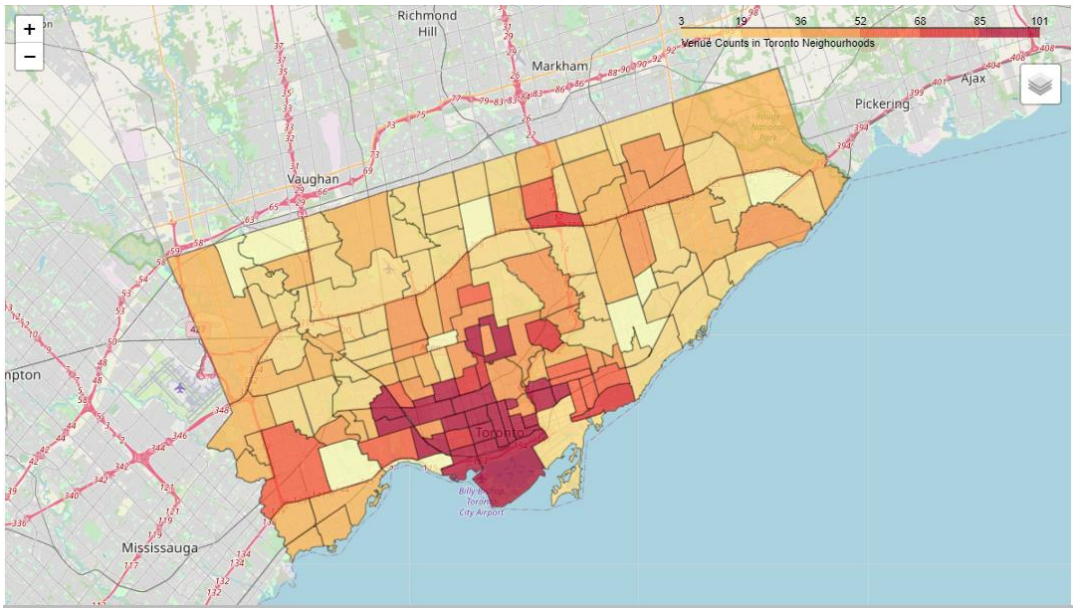
In order to generate this map data frame df_townsemi was grouped based on Neighbourhood, Neighbourhood Latitude & Neighbourhood Longitude column to get count of town/semi houses against each neighbourhood and defined to a new data frame df_townsemi_group which was utilized for mapping.




This choropleth map illustrates the count of town/semi houses among all the neighbourhoods in Toronto. The darker areas indicate maximum number of town/semi houses. This choropleth map also illustrates that neighbourhoods including Downsview Roding-CFB & Niagara have maximum number of town/semi houses leading to a count of six, else all remaining neighbourhoods are less than this.

### (viii) Map for Venue Count

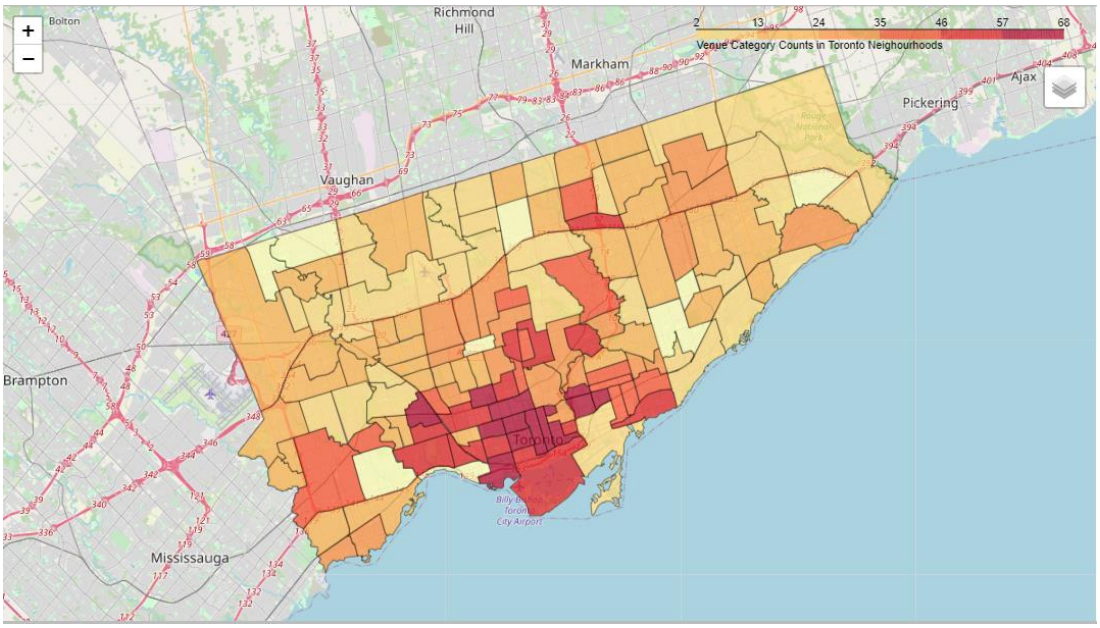
In order to generate this map data frame toronto_venues was grouped based on Neighbourhood column to get count of venues against each neighbourhood which was defined as a new data frame df_venuecounts. Neighbourhood Latitude, Neighbourhood Longitude and Average Property Price from df_neighbourhoodgrp was inserted into df_venuecounts which was utilized for mapping.




This choropleth map illustrates venues count among all the neighbourhoods in Toronto. The darker areas indicate neighbourhoods with maximum number of venues. This choropleth map also illustrates that maximum number of venues are concentrated in Midtown & Downtown Toronto including neighbourhoods such as Waterfront Communities-The Island, Niagara, Kensington-ChinaTown, Trinity Bellwoods, Palmerston-Little Italy, University, Baystreet Corridor, Church-Young Corridor, Moss Park, Annex, Dovercourt-Wallace Emerson-Juncti.

### (ix) Map for Venue Category Count

In order to generate this map data frame toronto_venues was grouped based on Neighbourhood column to get count of unique venues category against each neighbourhood which was defined as a new data frame df_venuecategorycounts. Neighbourhood Latitude, Neighbourhood Longitude and Average Property Price from df_neighbourhoodgrp was inserted into df_venuecategorycounts which was utilized for mapping.




This choropleth map illustrates venue category count among all the neighbourhoods in Toronto. The darker areas indicate neighbourhoods with maximum number of unique category venues. This choropleth map also illustrates that maximum number of unique category venues are concentrated in Midtown & Downtown Toronto including neighbourhoods such as Waterfront Communities-The Island, Niagara, Kensington-ChinaTown, Trinity Bellwoods, Palmerston-Little Italy, University, Baystreet Corridor, Church-Young Corridor, Moss Park, Annex, Dovercourt-Wallace Emerson-Juncti.

## Model Development & Evaluation: Machine Learning Approach

Machine learning approach regression was chosen because of its simplicity and with the aid of Sklearn library implementation of model is quick and easy which is perfect to start the analyzing process. Regression approach was used to develop model for following dependent & independent variable;

(i) Number of bedrooms versus price for all property types
(ii) Number of bathrooms versus price for all property types
(iii) Number of bedrooms versus price for Condos
(iv) Number of bathrooms versus price for Condos
(v) Number of bedrooms versus price for Detached Houses
(vi) Number of bathrooms versus price for Detached Houses
(vii) Number of bedrooms versus price for Town/Semi Houses
(viii) Number of bathrooms versus price for Town/Semi Houses
(ix) Neighbourhood venue count versus average property price
(x) Neighbourhood venue category count versus average property price
(xi) Neighbourhood venue category versus average property price

Regression approach is not limited to Linear type, it encompasses Polynomial, K Nearest Neighbour and Random Forest Regression which is specially used for feature importance. Below is regression analysis performed for above mentioned relations with mentioned regression types.

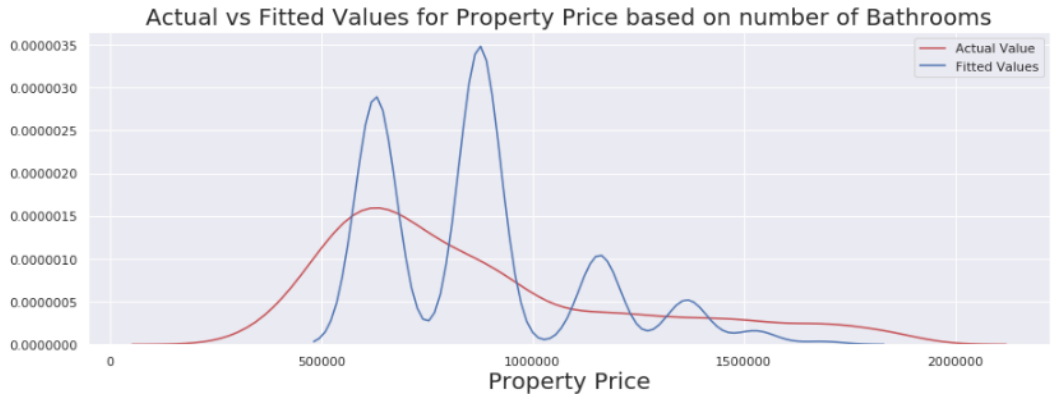
## Linear Regression

### (i) Number of bedrooms versus price for all property types

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of ~ -1.41E+24 and mean squared error (mse) of ~ 1.81E+35.

### (ii) Number of bathrooms versus price for all property types

In this case Linear regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of ~ 0.41 and mean squared error (mse) of ~ 7.59E+10. Below is the distribution plot for actual versus fitted values for property price based on number of bathrooms.




### (iii) Number of bedrooms versus price for Condo

In this case Linear regression was performed utilizing data frame df_condo, which resulted in regression coefficient (R^2) of ~ 0.15 and mean squared error (mse) of ~ 5.89E+10. Below is the distribution plot for actual versus fitted values for condo type property price based on number of bedrooms

### (iv) Number of bathrooms versus price for Condo

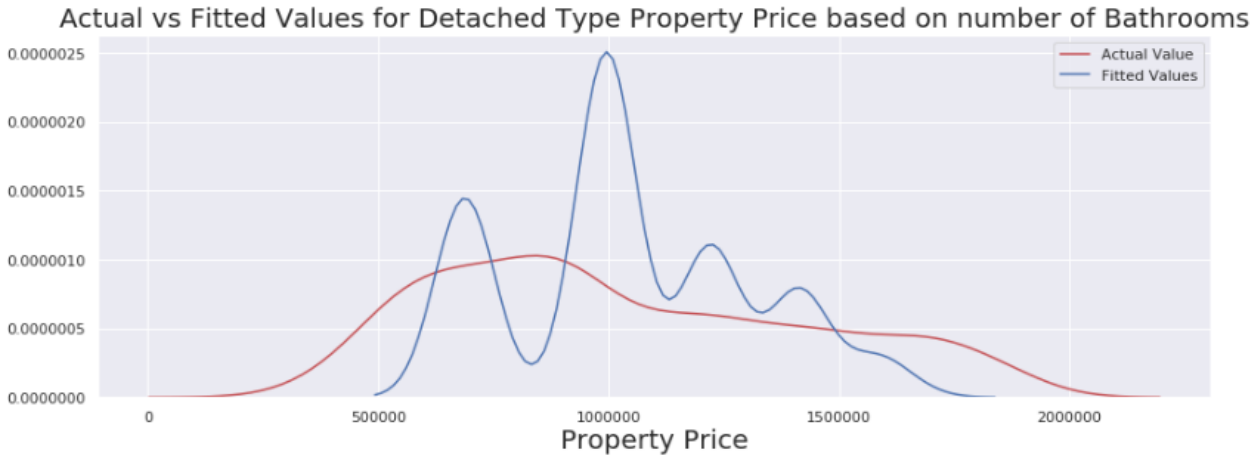
In this case Linear regression was performed utilizing data frame df_condo, which resulted in regression coefficient (R^2) of ~ 0.26 and mean squared error (mse) of ~ 5.15E+10. Below is the distribution plot for actual versus fitted values for condo type property price based on number of bathrooms


### (v) Number of bedrooms versus price for Detached Houses

In this case Linear regression was performed utilizing data frame df_detached, which resulted in regression coefficient (R^2) of ~ -1.23E+26 and mean squared error (mse) of ~ 1.98E+37.

### (vi) Number of bathrooms versus price for Detached Houses

In this case Linear regression was performed utilizing data frame df_detached, which resulted in regression coefficient (R^2) of ~ 0.45 and mean squared error (mse) of ~ 8.96E+10. Below is the distribution plot for actual versus fitted values for detached houses type property price based on number of bathrooms.

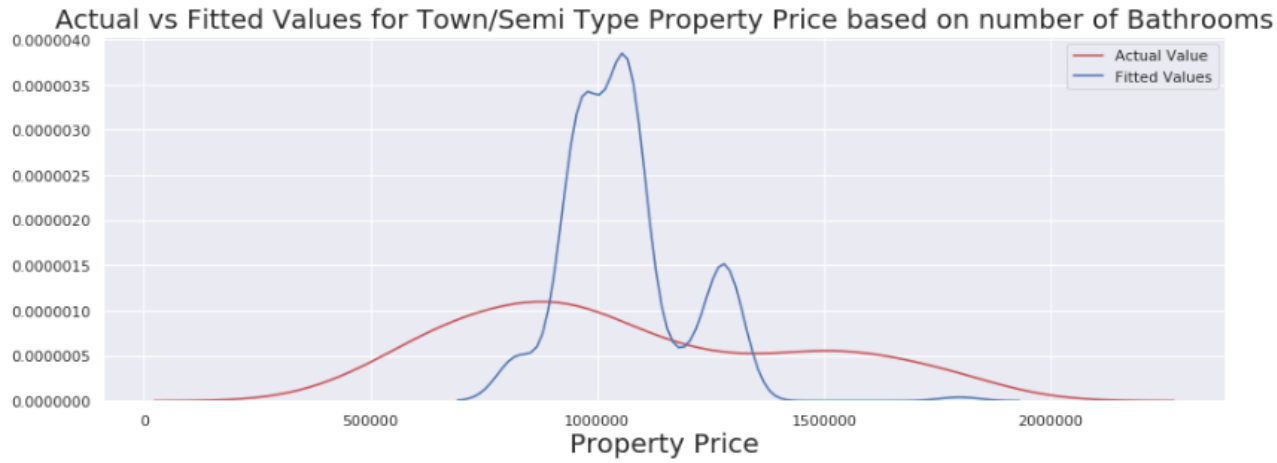



### (vii) Number of bedrooms versus price for Town/Semi Houses

In this case Linear regression was performed utilizing data frame df_townsemi, which resulted in regression coefficient (R^2) of ~ 0.09 and mean squared error (mse) of ~ 1.61E+11. Below is the distribution plot for actual versus fitted values for town/semi houses type property price based on number of bedrooms.

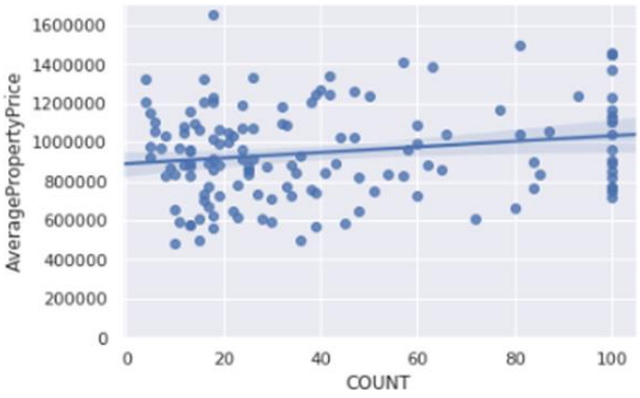

### (viii) Number of bathrooms versus price for Town/Semi Houses

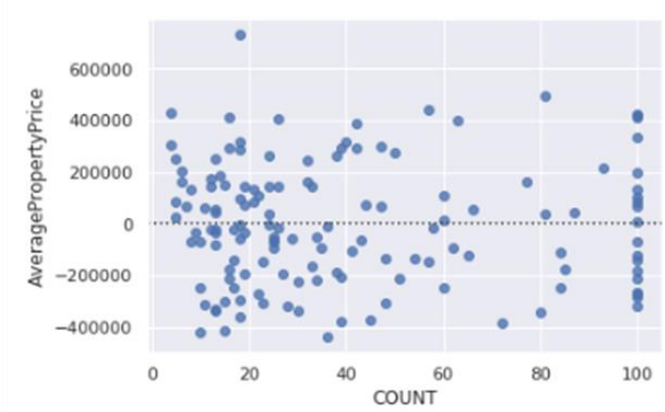
In this case Linear regression was performed utilizing data frame df_townsemi, which resulted in regression coefficient (R^2) of ~ 0.25 and mean squared error (mse) of ~ 1.34E+11. Below is the distribution plot for actual versus fitted values for town/semi houses type property price based on number of bathrooms.




### (ix) Neighbourhood venue count versus average property price

In this case Linear regression was performed utilizing data frame df_venuecounts, since venue count & average property price were in float type, in this regard seaborn and residual plot was generated to investigate if linear relationship exists between venue count and average property price. Below are the seaborn and residual plots;

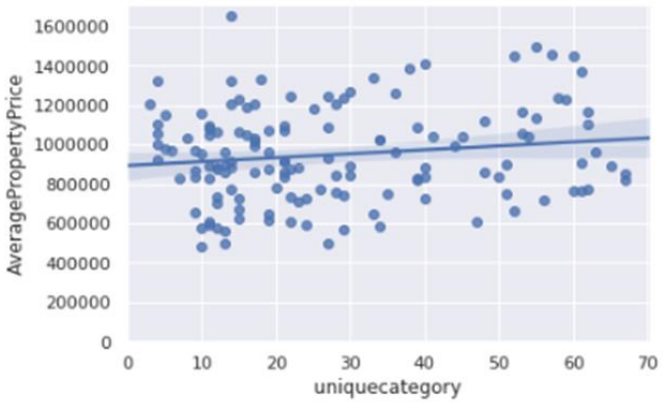


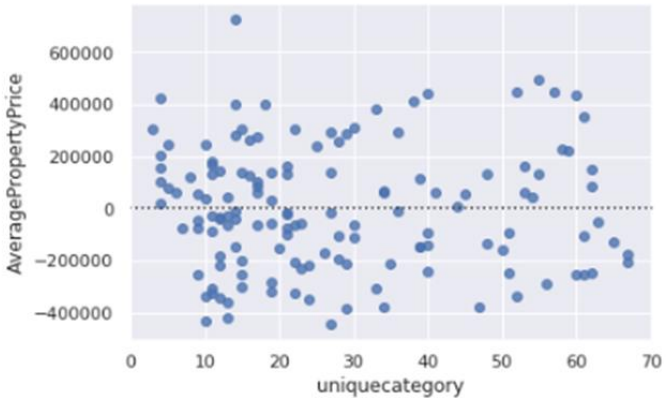




Seaborn and residual suggested poor relationship between venue count & average property price which was supplemented by linear regression model which computed regression coefficient(R^2) of ~ -0.02 & mean squared error of ~ 6.04E+10.

### (x) Neighbourhood venue category count versus average property price

In this case Linear regression was performed utilizing data frame df_venuecategorycounts, since venue category count & average property price were in float type, in this regard seaborn and residual plot was generated to investigate if linear relationship exists between venue category count and average property price. Below are the seaborn and residual plots;




Seaborn and residual suggested poor relationship between venue category count & average property price which was supplemented by linear regression model which computed regression coefficient(R^2) of ~ -0.02 & mean squared error of ~ 6.04E+10.

### (xi) Neighbourhood venue category versus average property price

In order to establish relationship between neighbourhood venue category versus average property price, venue category from data frame toronto_venues were converted into binary format 0 & 1 using one hot encoding which was defined as a new data frame by the name of venues_type_onehot. Neighbourhood from toronto_venues was inserted into data frame venues_type_onehot. Venue category in venues_type_onehot was grouped & summed based on Neighbourhood and was defined to a new data frame venue_count_df. Average property price

from data frame df_neighbourhoodgrp was inserted in venue_count_df which was utilized for regression modelling which resulted in regression coefficient (R^2) of ~ 2.29 and mean squared error (mse) of ~ 1.96E+11.

Below is the tabulated summary of Linear Regression results for all cases;

| | Linear Regression Results | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All Property Types | | Condos | | Detached Houses | | Town/Semi Houses | | Venue Count | Venue Category Count | Venue Category |
| | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | | | |
| Regression Coefficient (R^2) | -1.41E+24 | 0.41 | 0.15 | 0.26 | -1.23E+26 | 0.45 | 0.09 | 0.25 | -0.02 | -0.02 | -2.29 |
| Mean Squared Error (MSE) | 1.81E+35 | 7.59E+10 | 5.89E+10 | 5.15E+10 | 1.98E+37 | 8.96E+10 | 1.61E+11 | 1.34E+11 | 6.04.E+10 | 6.04E+10 | 1.96E+11 |

Linear regression results based on regression coefficient and mean squared error suggests that property attribute bathroom to be the most relatable parameter when comparing with property price. Models developed between bathroom and property price seems to show some predictability specially in the case of detached houses and all property types.

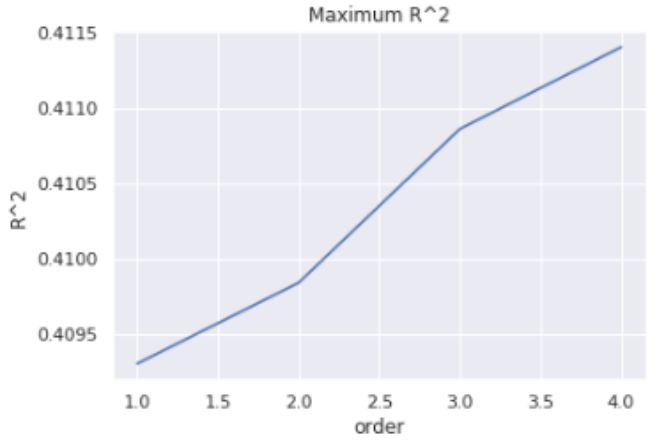
## Polynomial Regression

### (i) Number of bedrooms versus price for all property types

In this case Polynomial regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of ~ -1.10E+23 and mean squared error (mse) of ~ 1.42E+34. Polynomial degree of 2 was used in generating the model.

### (ii) Number of bathrooms versus price for all property types

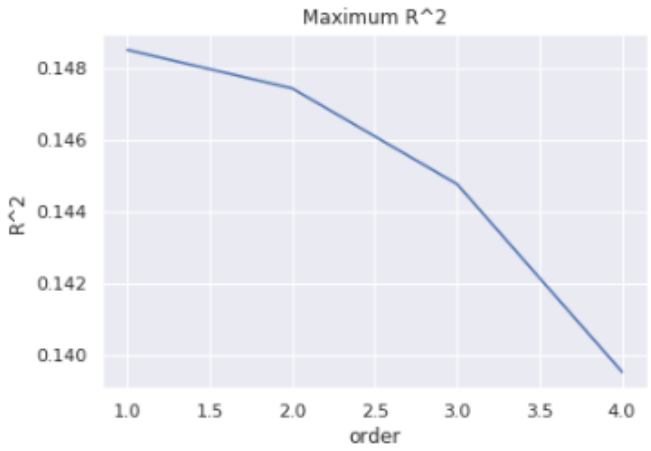
In this case Polynomial regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of ~ 0.41 and mean squared error (mse) of ~ 7.63E+10. Polynomial degree of 4 was used in generating the model which resulted from the sensitivities which were ran in order to optimize R^2 value. Below is the plot showing R^2 value with respect to different polynomial degrees.




### (iii) Number of bedrooms versus price for Condos

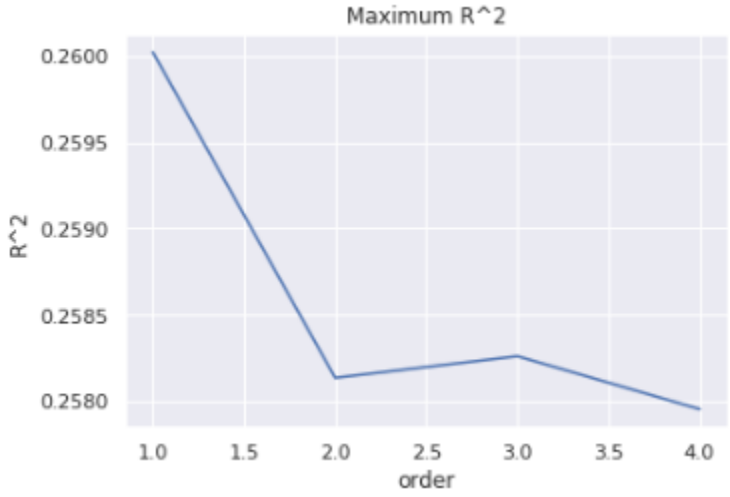
In this case Polynomial regression was performed utilizing data frame df_condo, which resulted in regression coefficient (R^2) of ~ 0.15 and mean squared error (mse) of ~ 5.89E+10. Polynomial degree of 1 was used in

generating the model which resulted from the sensitivities which were ran in order to optimize R^2 value. Below is the plot showing R^2 value with respect to different polynomial degrees.



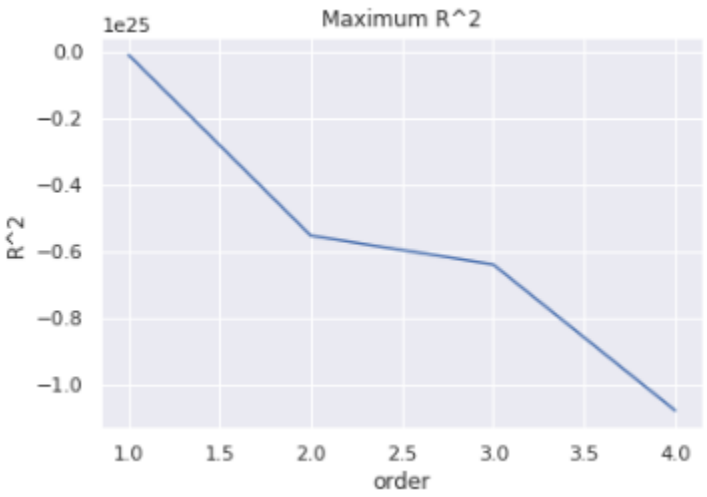

### (iv) Number of bathrooms versus price for Condos

In this case Polynomial regression was performed utilizing data frame df_condo, which resulted in regression coefficient (R^2) of ~ 0.26 and mean squared error (mse) of ~ 5.12E+10. Polynomial degree of 1 was used in generating the model which resulted from the sensitivities which were ran in order to optimize R^2 value. Below is the plot showing R^2 value with respect to different polynomial degrees.



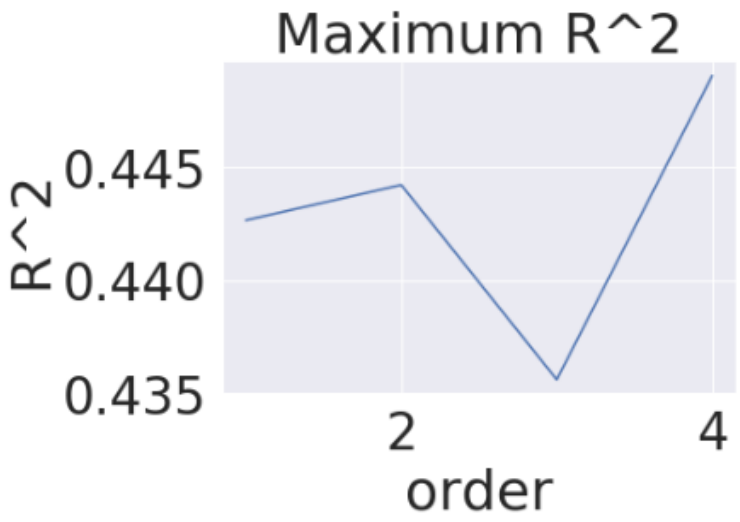

### (v) Number of bedrooms versus price for Detached

In this case Polynomial regression was performed utilizing data frame df_detached, which resulted in regression coefficient (R^2) of ~ -1.14E+23 and mean squared error (mse) of ~ 1.85E+34. Polynomial degree of 1 was used in generating the model which resulted from the sensitivities which were ran in order to optimize R^2 value. Below is the plot showing R^2 value with respect to different polynomial degrees.




### (vi) Number of bathrooms versus price for Detached

In this case Polynomial regression was performed utilizing data frame df_detached, which resulted in regression coefficient (R^2) of ~ 0.44 and mean squared error (mse) of ~ 8.98E+10. Polynomial degree of 2 was used in generating the model which resulted from the sensitivities which were ran in order to optimize R^2 value. Below is the plot showing R^2 value with respect to different polynomial degrees.

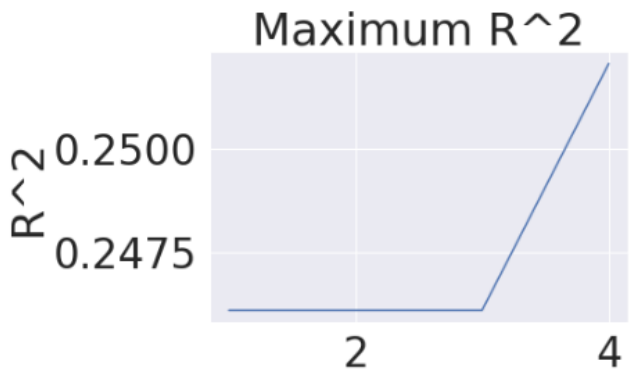



### (vii) Number of bedrooms versus price for Town/Semi Houses

In this case Polynomial regression was performed utilizing data frame df_townsemi, which resulted in regression coefficient (R^2) of ~ -1.45E+26 and mean squared error (mse) of ~ 2.58E+37. Polynomial degree of 2 was used in generating the model.

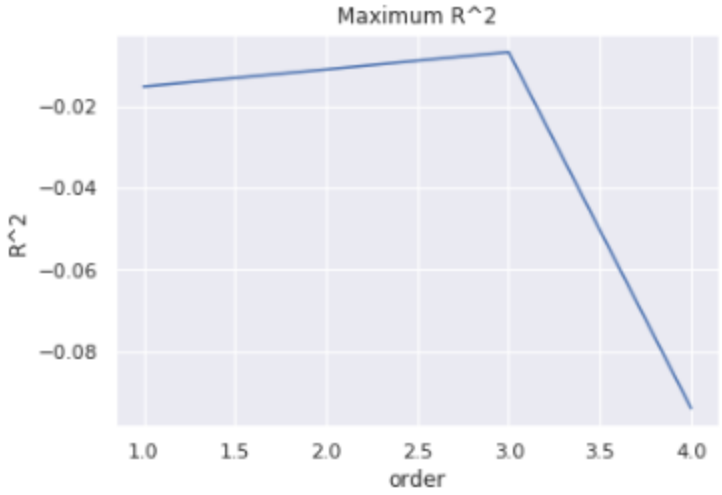
### (viii) Number of bathrooms versus price for Town/Semi Houses

In this case Polynomial regression was performed utilizing data frame df_townsemi, which resulted in regression coefficient (R^2) of ~ 0.25 and mean squared error (mse) of ~ 1.33E+11. Polynomial degree of 4 was used in generating the model which resulted from the sensitivities which were ran in order to optimize R^2 value. Below is the plot showing R^2 value with respect to different polynomial degrees.



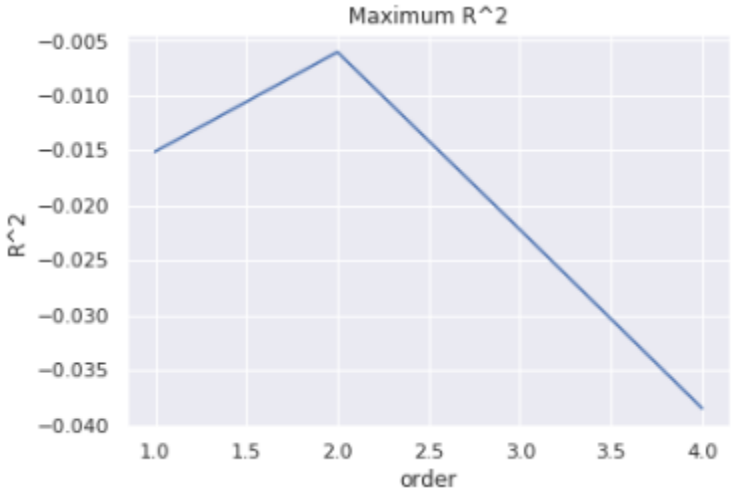

### (ix) Neighbourhood venue count versus average property price

In this case Polynomial regression was performed utilizing data frame df_venuecounts, which resulted in regression coefficient (R^2) of ~ -0.01 and mean squared error (mse) of ~ 5.99E+10. Polynomial degree of 3 was used in generating the model which resulted from the sensitivities which were ran in order to optimize R^2 value. Below is the plot showing R^2 value with respect to different polynomial degrees.




### (x) Neighbourhood venue category count versus average property price

In this case Polynomial regression was performed utilizing data frame df_venuecategorycounts, which resulted in regression coefficient (R^2) of ~ -0.01 and mean squared error (mse) of ~ 5.98E+10. Polynomial degree of 2 was used in generating the model which resulted from the sensitivities which were ran in order to optimize R^2 value. Below is the plot showing R^2 value with respect to different polynomial degrees.




### (xi) Neighbourhood venue category versus average property price

In this case Polynomial regression was performed utilizing data frame venue_count_df, which resulted in regression coefficient (R^2) of ~ -1.31 and mean squared error (mse) of ~ 1.37E+11. Polynomial degree of 2 was used in generating the model.

Below is the tabulated summary of Polynomial Regression results for all cases;

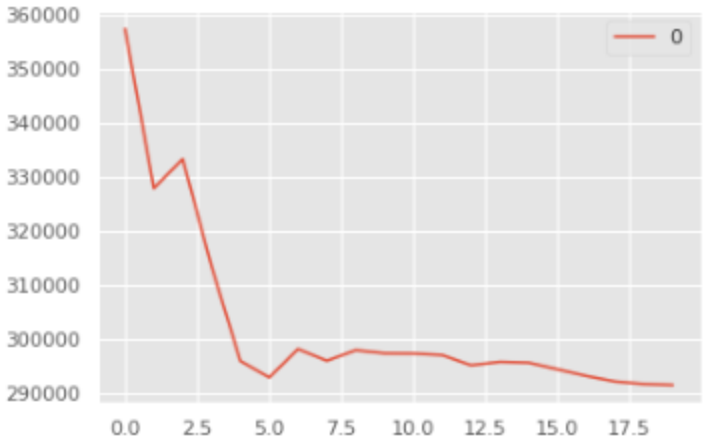
| | Polynomial Regression Results | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All Property Types | | Condos | | Detached Houses | | Town/Semi Houses | | Venue Count | Venue Category Count | | Venue Category | |
| | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | | | | | |
| Regression Coefficient (R^2) | -1.10E+23 | 0.41 | 0.15 | 0.26 | -1.14E+23 | 0.44 | -1.45E+26 | 0.25 | -0.01 | -0.01 | | -1.31 | |
| Mean Squared Error (MSE) | 1.42E+34 | 7.63E+10 | 5.89E+10 | 5.12E+10 | 1.85E+34 | 8.98E+10 | 2.58E+37 | 1.33E+11 | 5.99E+10 | 5.98E+10 | | 1.37E+11 | |

Polynomial regression results based on regression coefficient and mean squared error also suggests that property attribute bathroom to be the most relatable parameter when comparing with property price. Models developed between bathroom and property price seems to show some predictability specially in the case of detached houses and all proper types.

## K Nearest Neighbour Regression

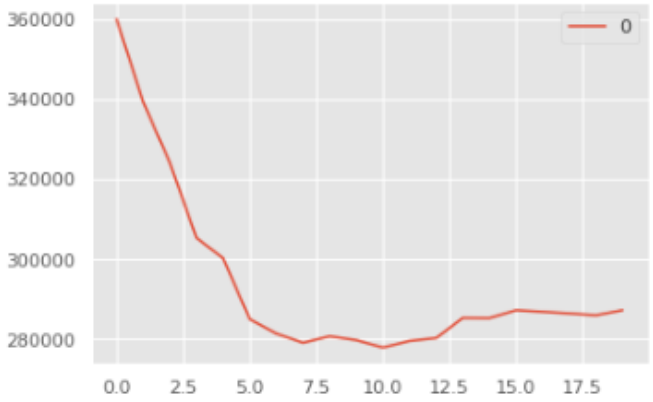
### (i) Number of bedrooms versus price for all property types

In this case KNN regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of ~ 0.34 and mean squared error (mse) of ~ 8.94E+10. K value of 9 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover, elbow plot was also generated to find optimum K value to be used for the model which is as follow;



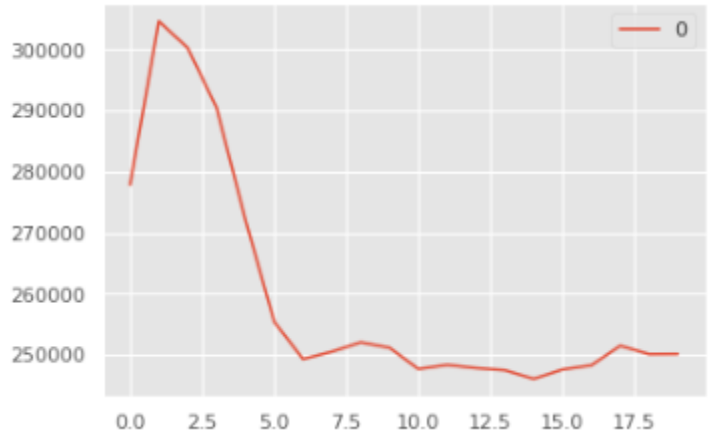

### (ii) Number of bathrooms versus price for all property types

In this case KNN regression was performed utilizing data frame df, which resulted in regression coefficient (R^2) of ~ 0.37 and mean squared error (mse) of ~ 8.16E+10. K value of 9 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover, elbow plot was also generated to find optimum K value to be used for the model which is as follow;



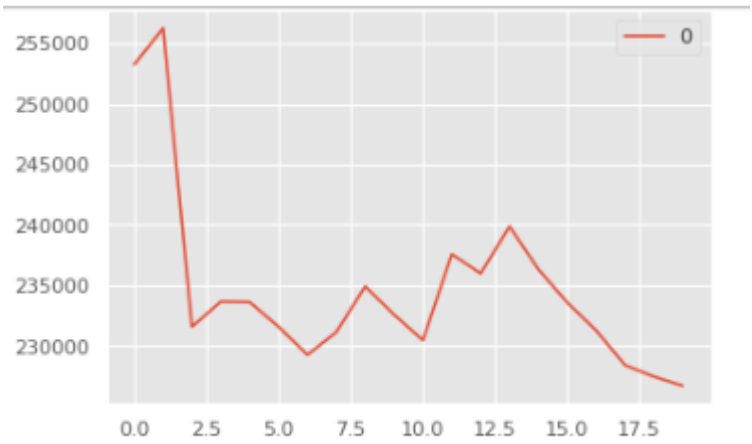

### (iii) Number of bedrooms versus price for Condos

In this case KNN regression was performed utilizing data frame df_condo, which resulted in regression coefficient (R^2) of ~ 0.10 and mean squared error (mse) of ~ 6.26E+10. K value of 7 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover ,elbow plot was also generated to find optimum K value to be used for the model which is as follow;



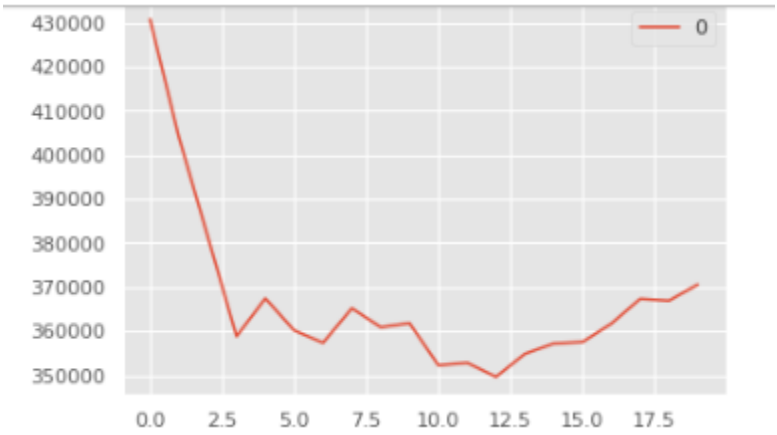

### (iv) Number of bathrooms versus price for Condos

In this case KNN regression was performed utilizing data frame df_condo, which resulted in regression coefficient (R^2) of ~ 0.26 and mean squared error (mse) of ~ 5.14E+10. K value of 8 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover, elbow plot was also generated to find optimum K value to be used for the model which is as follow;



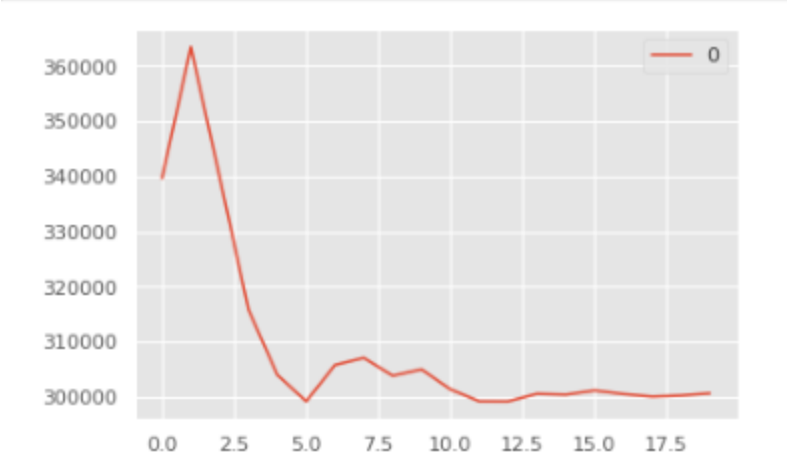

### (v) Number of bedrooms versus price for Detached Houses

In this case KNN regression was performed utilizing data frame df_detached, which resulted in regression coefficient (R^2) of ~ 0.26 and mean squared error (mse) of ~ 5.14E+10. K value of 7 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover, elbow plot was also generated to find optimum K value to be used for the model which is as follow;



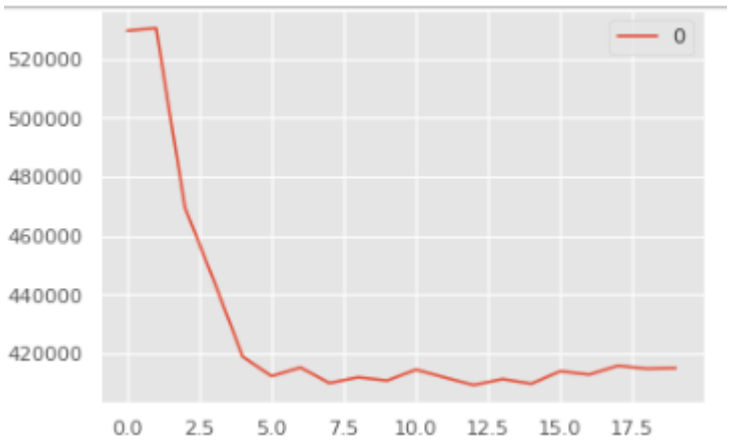

### (vi) Number of bathrooms versus price for Detached Houses

In this case KNN regression was performed utilizing data frame df_detached, which resulted in regression coefficient (R^2) of ~ 0.44 and mean squared error (mse) of ~ 9.04E+10. K value of 6 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover, elbow plot was also generated to find optimum K value to be used for the model which is as follow;



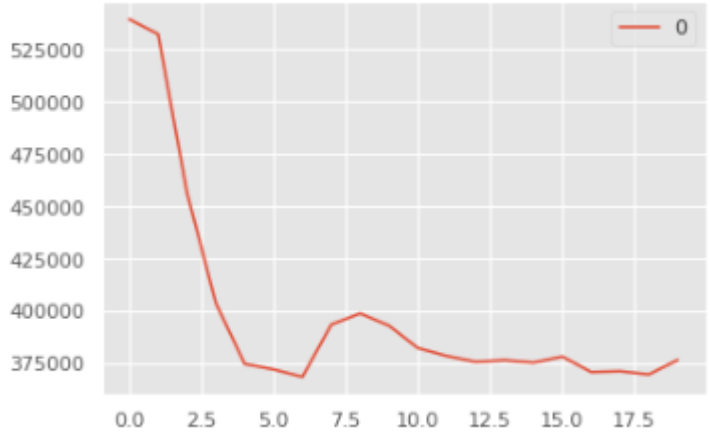

### (vii) Number of bedrooms versus price for Town/Semi Houses

In this case KNN regression was performed utilizing data frame df_townsemi, which resulted in regression coefficient (R^2) of ~ 0.03 and mean squared error (mse) of ~ 1.72E+11. K value of 9 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover, elbow plot was also generated to find optimum K value to be used for the model which is as follow;



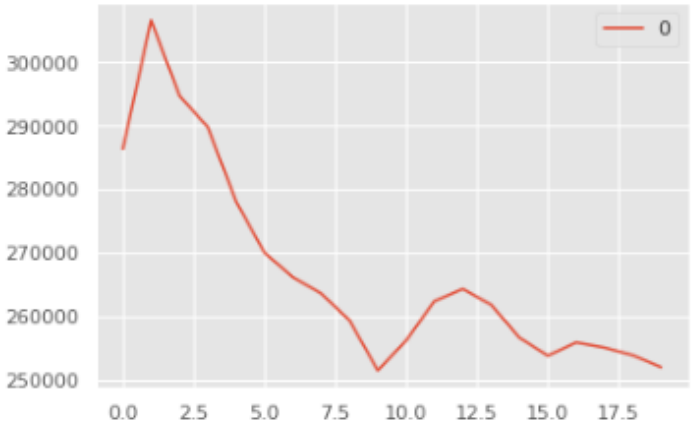

### (viii) Number of bathrooms versus price for Town/Semi Houses

In this case KNN regression was performed utilizing data frame df_townsemi, which resulted in regression coefficient (R^2) of ~ 0.20 and mean squared error (mse) of ~ 1.42E+11. K value of 7 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover, elbow plot was also generated to find optimum K value to be used for the model which is as follow;



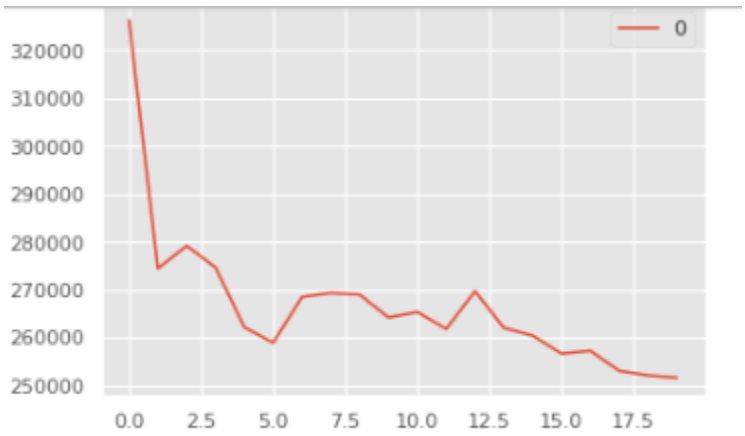

### (ix) Neighbourhood venue count versus average property price

In this case KNN regression was performed utilizing data frame df_venuecounts, which resulted in regression coefficient (R^2) of ~ -0.07 and mean squared error (mse) of ~ 6.35E+10. K value of 6 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover, elbow plot was also generated to find optimum K value to be used for the model which is as follow;



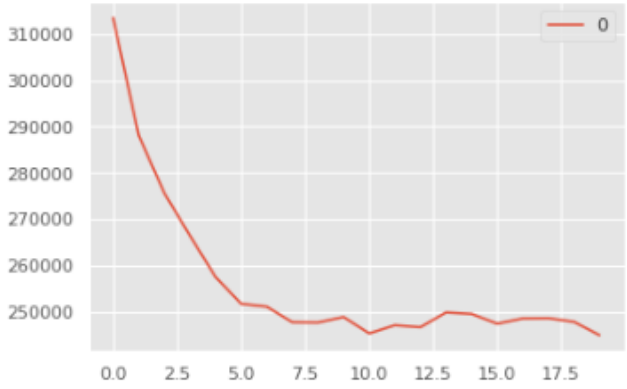

### (x) Neighbourhood venue category count versus average property price

In this case KNN regression was performed utilizing data frame df_venuecategorycounts, which resulted in regression coefficient (R^2) of ~ -0.06 and mean squared error (mse) of ~ 6.33E+10. K value of 9 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover, elbow plot was also generated to find optimum K value to be used for the model which is as follow;




### (xi) Neighbourhood venue category versus average property price

In this case KNN regression was performed utilizing data frame venue_count_df, which resulted in regression coefficient (R^2) of ~ -0.01 and mean squared error (mse) of ~ 5.99E+10. K value of 9 was used in generating the model which resulted from Grid Search CV to optimize R^2 value. Moreover, elbow plot was also generated to find optimum K value to be used for the model which is as follow;




Below is the tabulated summary of K Nearest Neighbour Regression for all cases;

| | K Nearest Neighbour Regression Results | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All Property Types | | Condos | | Detached Houses | | Town/Semi Houses | | Venue Count | Venue Category Count | Venue Category |
| | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | | | |
| Regression Coefficient (R^2) | 0.34 | 0.37 | 0.10 | 0.26 | 0.15 | 0.44 | 0.03 | 0.20 | -0.07 | -0.06 | -0.01 |
| Mean Squared Error (MSE) | 8.49E+10 | 8.16E+10 | 6.26E+10 | 5.14E+10 | 1.37E+11 | 9.04E+10 | 1.72E+11 | 1.42E+11 | 6.35E+10 | 6.33E+10 | 5.99E+10 |
| oot Mean Squared Error(RMSE | 2.91E+05 | 2.86E+05 | 2.50E+05 | 2.27E+05 | 3.71E+05 | 3.01E+05 | 4.15E+05 | 3.77E+05 | 2.52E+05 | 2.52E+05 | 2.45E+05 |

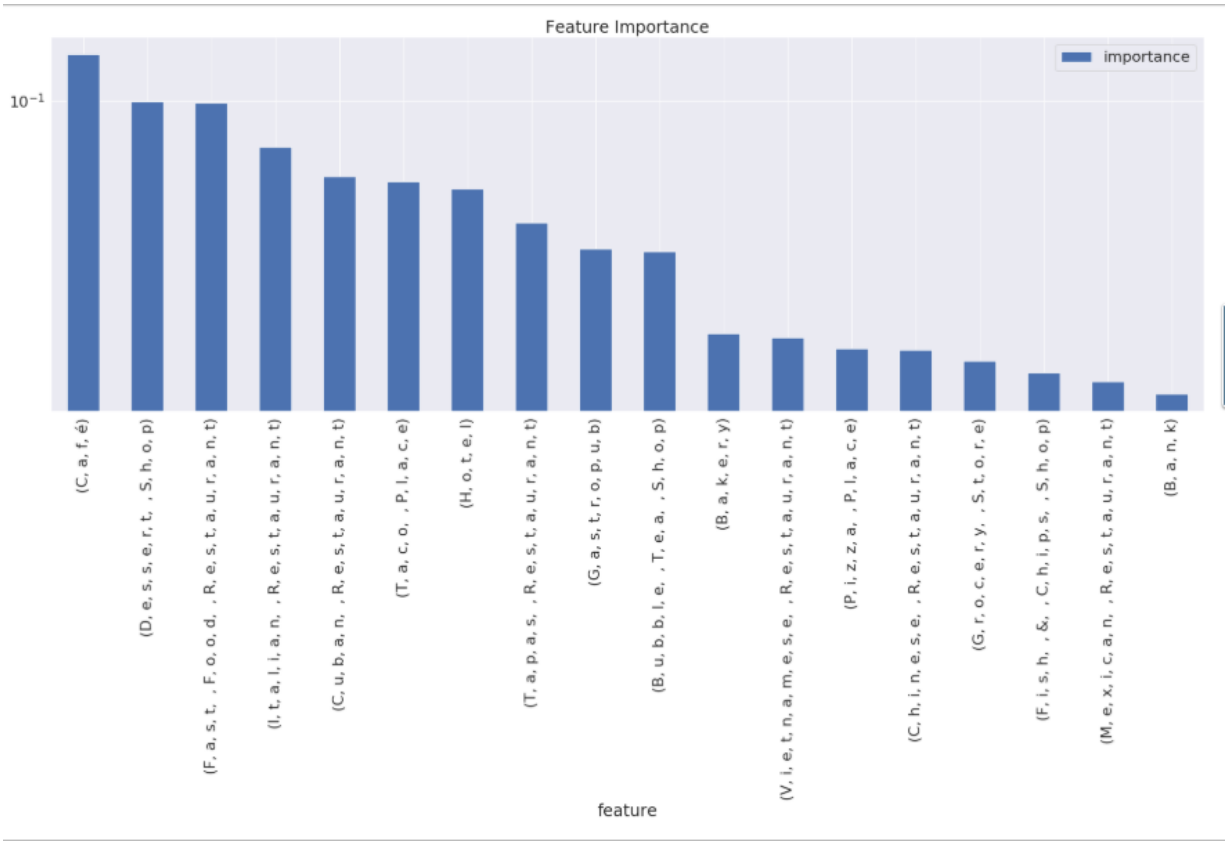
K Nearest Neighbour regression results based on regression coefficient and mean squared error also suggests property attribute i.e. bathroom to be the most relatable parameter when comparing with property price. Models developed between bathroom and property price seems to show predictability specially in the case including detached houses and all proper types. Moreover, K Nearest Neighbour shows improvement in the relationship between bedroom and property price for case of all property type and detached houses in terms of regression coefficient.

Since based on above modelling bathroom count seems to show predictability with respect to price, but in terms of nearby venues above developed models show poor relationship. Therefore, based on this concern random forest regression with feature importance was generated to determine which venues play a vital role in inching any property price upward.

## Random Forest Regression

Random Forest Regression was carried out to determine relationship between venue category and average property price. In this regard data frame venue_count_df was utilized which resulted in regression coefficient (R^2) of ~ -0.01 with mean squared error (mse) of ~ 6.04E+10. Grid Search CV was utilized to determine best parameters including max_depth, min_samples_leaf & n_estimators in order to optimize R^2 value. Below plot shows list price versus predicted price;


In addition to this, feature importance using random forest regression was performed to determine venues with respect to their importance on average property prices of neighbourhoods. Below is the plot showing venue with its importance with respect to average property price;




## Results

This results section provides an overview of the outcomes of the methodology and their relevance to the original problem of identifying price variation with respect to housing attributes for different Toronto neighbourhoods in conjunction with nearby venues to resolve the encounters faced by any individual during selection of a place.

### Histogram

Histogram for all property type was generated in order to provide an insight of the pricing frequency in the city of Toronto in general which lies between CAD 450,000 - 920,000. In addition to this, three more histograms were developed for condos, detached houses & semi/town houses for comparison purposes which illustrated that condos possess less pricing frequency in comparison to other property types making them more affordable for an individual.

### Boxplot

Box plot was generated to investigate the impact of property attributes such as number of bedrooms & bathrooms on the distribution of property price. For all the cases strong to moderate association was observed with the number of bathrooms & property price, whereas weak to moderate association was observed with the number of bedroom & property in the case of all property types & detached houses. In view to this, spearman's rank correlation was applied to the dataset which also showed similar results. Strength of correlation (Rs) was observed more in case of bathrooms in comparison to bedroom.

### Barplot

Bar plot was generated to provide a view of average property price of every neighbourhood which also illustrated that average property price across Toronto city is CAD 865,000 and most of neighbourhood lies above it. Rustic & Trinity Bellwoods are considered as the most expensive neighbourhoods of Toronto city.

### Data Visualization

Following insights were gathered as a result of generating Folium Chloropleth map for all the cases;

(i) Majority of neighbourhoods fall under the highest bracket of price for all property types leaving few that lies near to the boundary of Toronto city in North York Borough and in the central part of Scarborough District

(ii) Condos located in the downtown and midtown Toronto falls in the highest price bracket.

(iii) Detached houses in most of the neighbourhoods in Toronto falls under the highest price bracket, except few neighbourhoods which are located near to boundary of Toronto city in North York Borough and the ones which are in the central part of Scarborough district.

(iv) Most of town/semi houses falls under lower to mid-price bracket except few ones which are in midtown & west end district.

(v) Maximum number of condos are concentrated in downtown Toronto in Waterfront Communities-The island.

(vi) Maximum number of detached houses are concentrated in downtown Toronto in Waterfront Communities-The island

(vii) Neighbourhoods including Downsview Roding-CFB & Niagara have maximum number of town/semi houses

(viii) Maximum number of venues & unique venues category are concentrated in Midtown & Downtown Toronto

### Model Development & Evaluation: Machine Learning Approach

The regression machine learning technique provides additional insights in order to address the business problem. Regression types including Linear, Polynomial and K Nearest Neighbour was utilized for developing relation between property attributes, nearby venues with respect to property prices. The following table summarizes the results for all the regression techniques;

| | Linear Regression Results | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All Property Types | | Condos | | Detached Houses | | Town/Semi Houses | | Venue Count | Venue Category Count | Venue Category |
| | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | | | |
| Regression Coefficient (R^2) | -1.41E+24 | 0.41 | 0.15 | 0.26 | -1.23E+26 | 0.45 | 0.09 | 0.25 | -0.02 | -0.02 | -2.29 |
| Mean Squared Error (MSE) | 1.81E+35 | 7.59E+10 | 5.89E+10 | 5.15E+10 | 1.98E+37 | 8.96E+10 | 1.61E+11 | 1.34E+11 | 6.04.E+10 | 6.04E+10 | 1.96E+11 |

| | Polynomial Regression Results | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All Property Types | | Condos | | Detached Houses | | Town/Semi Houses | | Venue Count | Venue Category Count | Venue Category |
| | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | | | |
| Regression Coefficient (R^2) | -1.10E+23 | 0.41 | 0.15 | 0.26 | -1.14E+23 | 0.44 | -1.45E+26 | 0.25 | -0.01 | -0.01 | -1.31 |
| Mean Squared Error (MSE) | 1.42E+34 | 7.63E+10 | 5.89E+10 | 5.12E+10 | 1.85E+34 | 8.98E+10 | 2.58E+37 | 1.33E+11 | 5.99E+10 | 5.98E+10 | 1.37E+11 |

| | K Nearest Neighbour Regression Results | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | All Property Types | | Condos | | Detached Houses | | Town/Semi Houses | | Venue Count | Venue Category Count | Venue Category |
| | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | Bedroom | Bathroom | | | |
| Regression Coefficient (R^2) | 0.34 | 0.37 | 0.10 | 0.26 | 0.15 | 0.44 | 0.03 | 0.20 | -0.07 | -0.06 | -0.01 |
| Mean Squared Error (MSE) | 8.49E+10 | 8.16E+10 | 6.26E+10 | 5.14E+10 | 1.37E+11 | 9.04E+10 | 1.72E+11 | 1.42E+11 | 6.35E+10 | 6.33E+10 | 5.99E+10 |
| oot Mean Squared Error(RMSE | 2.91E+05 | 2.86E+05 | 2.50E+05 | 2.27E+05 | 3.71E+05 | 3.01E+05 | 4.15E+05 | 3.77E+05 | 2.52E+05 | 2.52E+05 | 2.45E+05 |

Based on above results, number of bathrooms shows moderate to strong association with property price whereas number of bedrooms shows poor association with property price. One key observation is that the KNN regression tends to improve R^2 value for bedrooms in the case of all property type and detached houses. Venue data shows poor association with property price for all the three regression techniques. In this regard, random forest regression with feature importance was performed which suggested cafes, dessert shops & fast food restaurants as the top three venues that play a vital role in inching any property price upward.

## Discussion & Recommendation

As a result of subject analysis, following trail can be followed in order to find suitable place to live;

(i) For any newcomer, the main district of interest for finding accommodation should be the neighbourhood located at the border of Toronto city in North York and in the central part of Scarborough.

(ii) Secondly, property type condo should be the focus due to its affordability in comparison to other property types.

(iii) Thirdly, number of bathrooms requirement to be optimized as per family need as it has a moderate to strong association with property price.

(iv) Lastly, consider those neighbourhoods which are without or at least have minimum number of venues such as cafes, dessert shops & fast food restaurants as they have a major impact on property price.

One more key fact which has always kept Toronto's real estate market in a bubble is due to less supply of accommodations in comparison to demand. This is due to continuous influx of locals from other provinces of country and immigrants from all around the globe which has led increase in Toronto population with a smaller number of places to live. Since Canada aims to increase invitations for more immigrants in coming years from all around the globe which will keep rising Toronto population leaving real estate market hype.

## Conclusion

Since the objective of the project was to investigate the relationship between property attributes and nearby venues with respect to property price. As a result of subject study and analysis done with integration of real estate and location data, following are the outcomes and observations;

- Exploratory data analysis in conjunction with modelling suggests bathroom count as key attribute for property price prediction.
- Rustic and Trinity Bellwoods seems to be the most expensive neighbourhoods in Toronto.
- Based on data visualization, neighbourhoods closer to the boundary of Toronto city in NorthYork and in the central part of Scarborough district are the least expensive ones.
- Based on data visualization, Neighbourhoods in Mid Town & Down Town Toronto including Waterfront Communities-The Island, Niagara, Kensington-ChinaTown, Trinity Bellwoods, Palmerston-Little Italy, University, Baystreet Corridor, Church-Young Corridor, Moss Park, Annex, Dovercourt-Wallace Emerson-Juncti leads with maximum property prices, venue counts & unique category venue with respect to all property types.
- With respect to modelling including Linear, Polynomial and KNN regression shows similar results in terms of regression coefficient and mean squared error, with exceptions of bedroom case where the all property types and detached case shows improved regression coefficient and mean squared error when KNN regression is applied on the data.
- Since regression techniques including linear, polynomial and knn regression didn't show any association of venue category data with neighbourhood average property price. In this regard, random forest regression with feature importance was performed which improved regression coefficient and showed feature importance with respect to average property price of neighbourhoods.
- Based on feature importance results; cafes, dessert shops & fast food restaurants are considered as top three venues that have major impact on average property price of neighbourhoods.

So, in a nutshell, locals moving to Toronto from other provinces of Canada and international immigrants planning to settle in Toronto should consider above mentioned factors as qualifiers for selection of an appropriate place for living.