CSDA1050 – Advanced Analytics Capstone

TRENDING HOUSING PRICE IN TORONTO AREA AT 2018

A Project Submitted By

SRIPRANAVAN SRITHARAN

(ID: 304378)

Research Background

Why is this important topic? Cities have always had pockets of wealth and poverty. Neighborhoods in the great cities of the industrialized world have undergone many transitions over the course of their history. However, the City of Toronto's neighborhood transition has been relatively sudden and dramatic, and the changes have serious consequences for Toronto residents.

It focuses on who lives where, based on the socio-economic status of the residents in each neighborhood, and how the average status of the residents in each neighborhood has changed certain year period.

Business Problems.

One of the largest investments that most individuals will make in their lifetime is the buying of a house. In these times of "hot" real-estate markets, the more meaningful information that individuals have on the predictors of price, the better.

1. Buyers want to get the most value for their money as well as getting all the points covered off on their "wishlist' for features they want to have in a home.
2. Sellers would like to have top dollar for their home without having their home sit on the market for an inordinate amount of time.
3. Real-estate agents want to help their clients in both groups reach their goals in the real-estate market.
4. Real-estate boards and municipalities are very interested in their area's real-estate market trends.

Our data consists of home attributes taken from Toronto, Canada an area that includes Seattle. It includes homes sold in 2018.

The Goal of this analysis is to answer the following question:

"Can the sale price of a house be predicted accurately based on physical characteristics such as the Population, House-hold, House-hold Income, Education Status, Age of living area (0 to 4 'Child' and 25 to 64)?"

Focuses on working:

Analysis the all each factor with Area ID (Area Code)

Compare all results with Area ID

Given best model factor to Trending Housing Price

.

Data Preparation and Cleaning

We are using two data sets in this project, such as Toronto2018 and Location_Toronto.

- data = pd.read_csv("Toronto2018.csv")

  print("Column headings:")

  print(df.columns)

```
  Column headings:
Index(['Name', 'Forward Sortation Area ID', 'Total Population',
       'Total Households', 'Total Household Population',
       'Male Population by Age | Males',
       'Households by Income| Median Household Income ($)',
       'Household Population 25 to 64 Years by Educational Attainment(%) |
Household Population 25 To 64 Years | No Certificate, Diploma Or Degree',
       'Household Population 25 to 64 Years by Educational Attainment (%)|
Household Population 25 To 64 Years | University Certificate, Diploma Or D
egree At Bachelor Level Or Above',
       'Household Population by 5-Year Mobility | Household Population For
5 Year Mobility Status',
       'Population 15 Years or Over by Marital Status | Total Population 1
5 Years Or Over',
       'Households by Income (Constant Year) | Total Households',
       'Households by Income (Constant Year) | Average Household Income (C
onstant Year 2005 $), 2018',
       'Household Population by Total Immigrants and Place of Birth | Tota
l Household Population',
       'Total Population by Age | Total Population | Total 0 To 4',
       'Detached Housing Prices| Median ($)',
       'Semi-Detached Housing Prices| Median ($)',
       'Condo Aprtment Prices| Median ($)'],
      dtype='object')
```

- df = pd.read_csv("Location_Toronto.csv")

  df = df_location

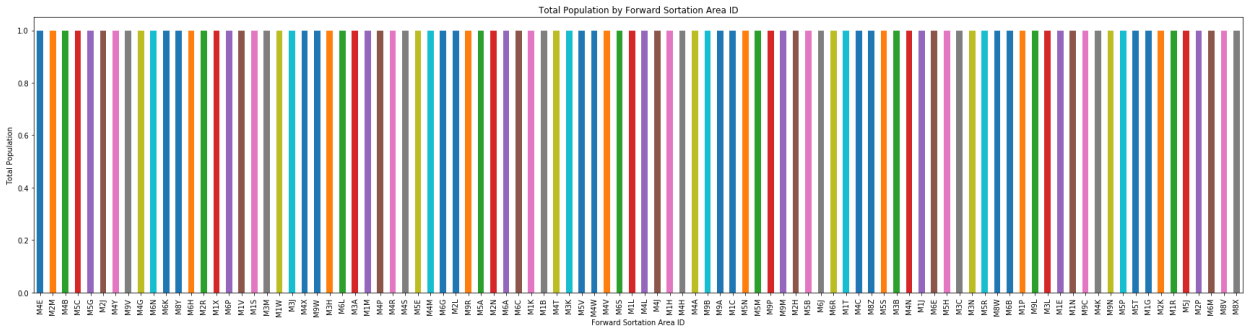
  print(df_location.columns)

```
Column headings:
Index(['FSA', 'Latitude', 'Longitude', 'Place Name'], dtype='object')
```

Cleaned the data and check the file .

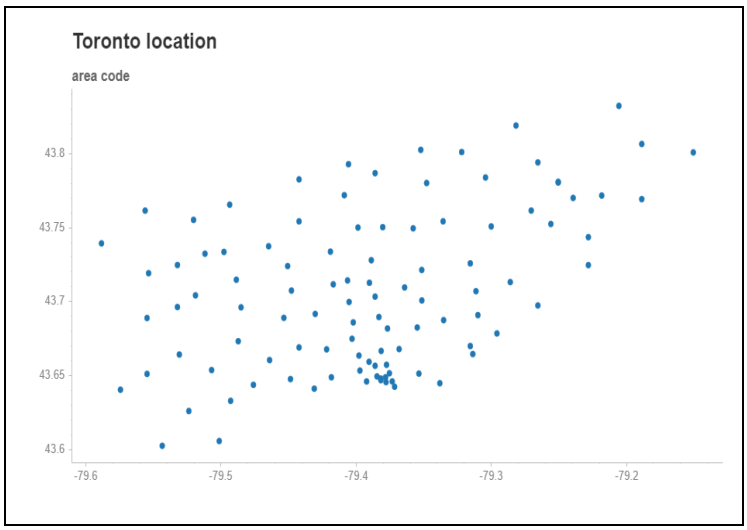
```
#### PACKAGE AND DATA IMPORTS
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
df = pd.read_csv("Toronto2018.csv")
df.head(5)
```

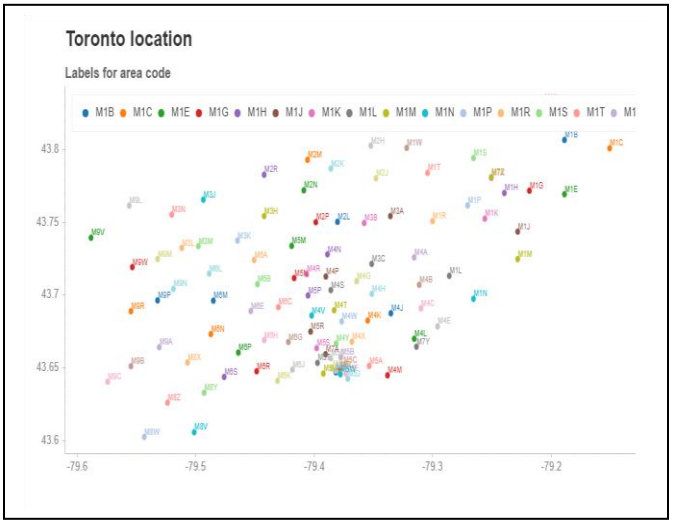
```
##### Out Put In .jpynb file#####
print(df['Name'])
# Look at the shape of the dataframe
df.shape
# Examine the fields
df.info()
#### CLEANING AND COMBINING
#########################
# creating 2018 data df using only one station
Toronto2018= data_df[Toronto2018_df['Forward Sortation Area ID'] == 1]
##check empty cells
df = pd.read_csv("Toronto2018.csv")
df.isnull().sum(axis=0)
# Look at the data summary
df_data = pd.read_csv("Toronto2018.csv")
df_data.describe()

Check the Area name in Toronto2018
```



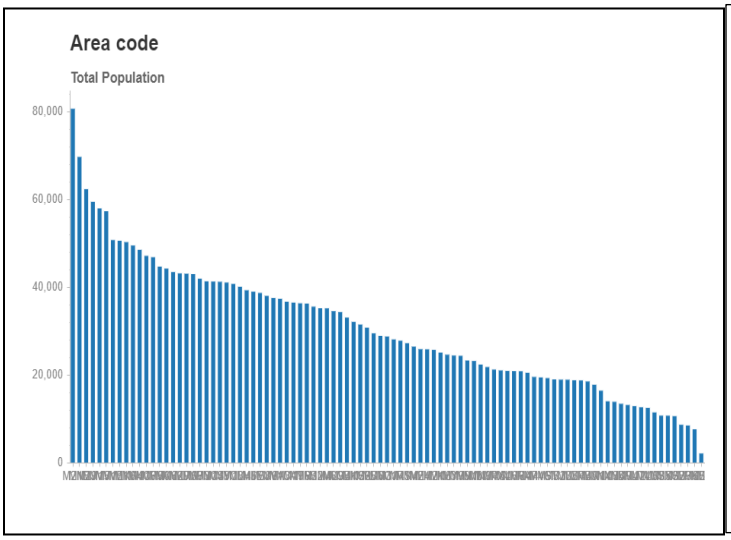
Modelling/Analysis

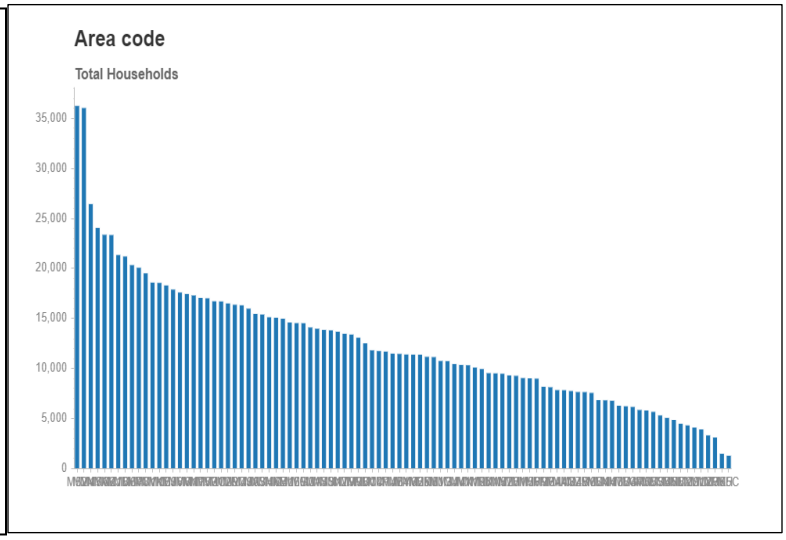
Location Mapping

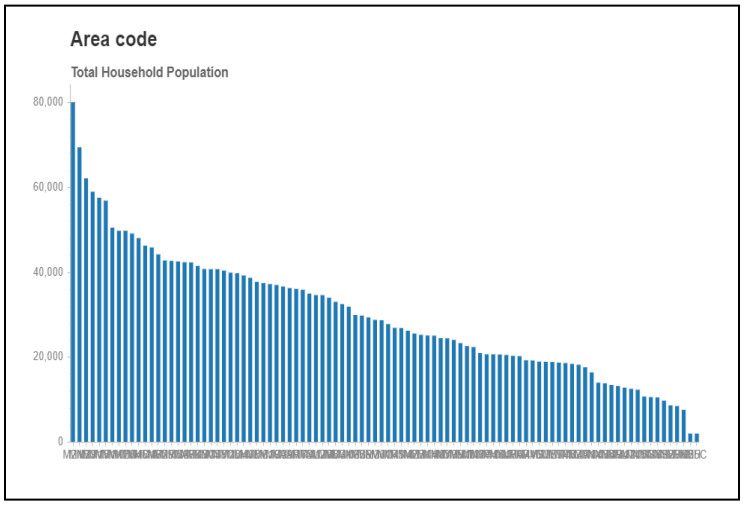


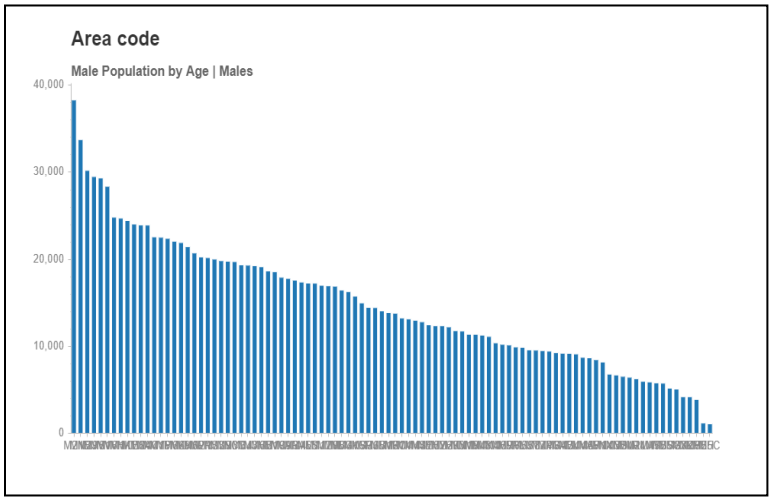


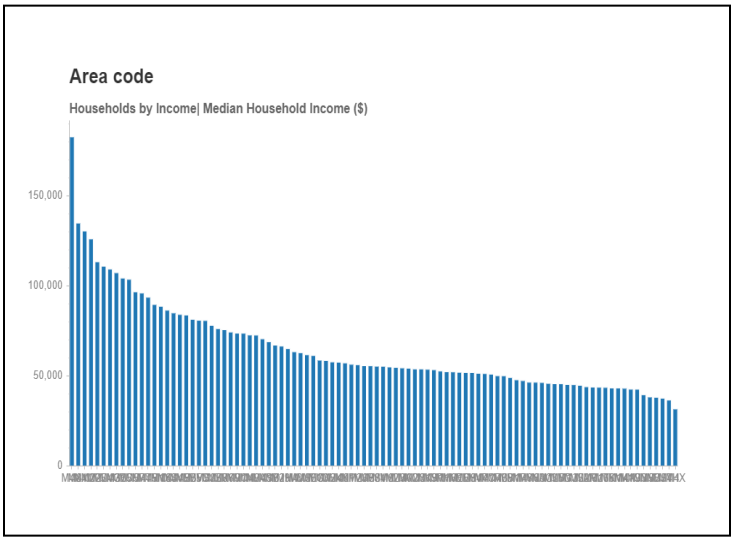
Plot the Bar chart for all factors with Toronto Area Name

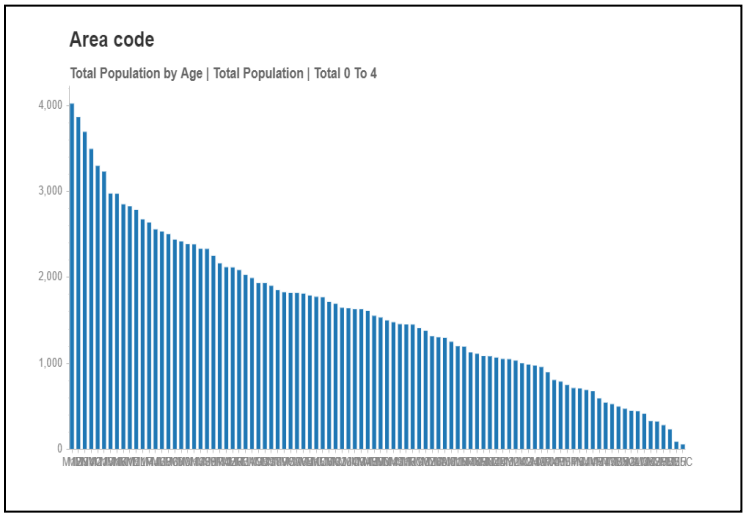




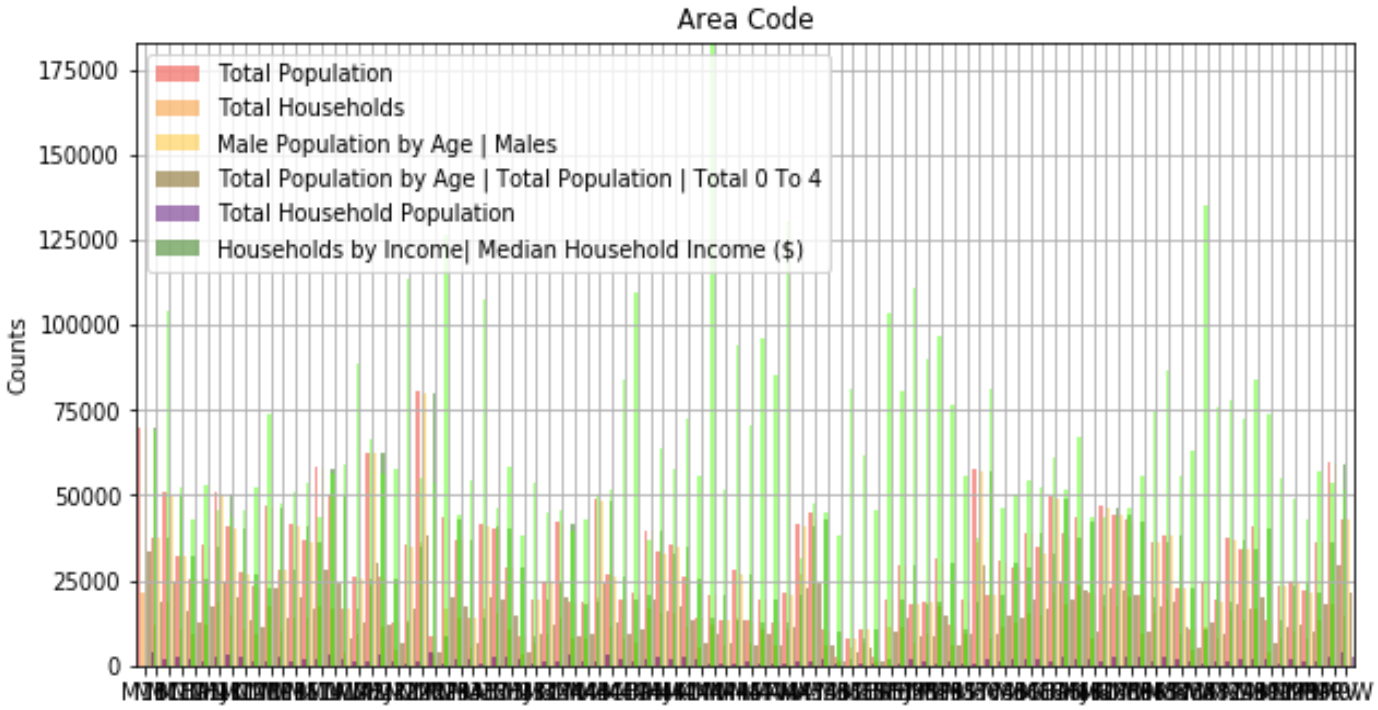




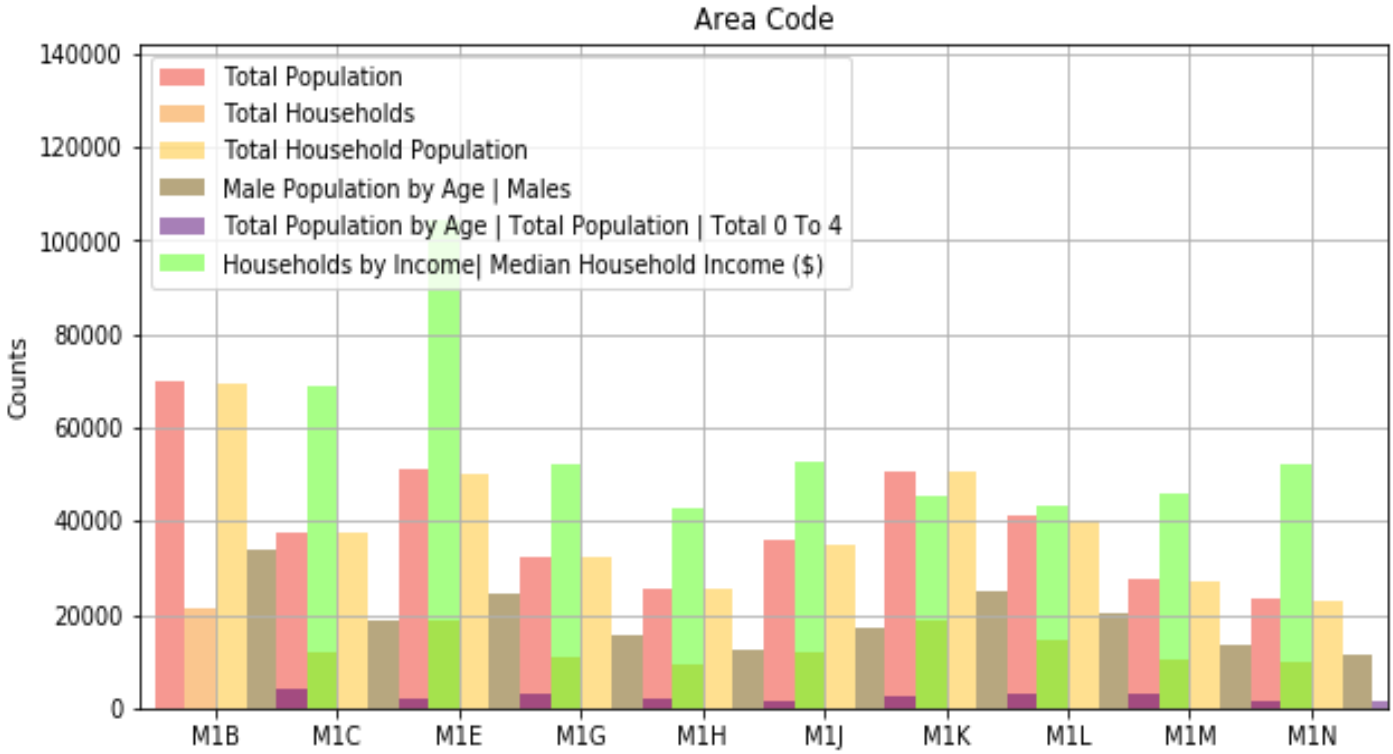




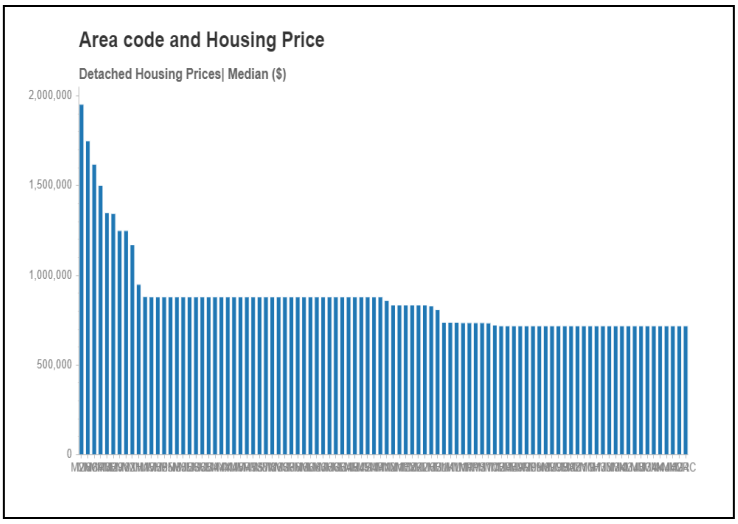
Group Bar chart for all factors in one picture

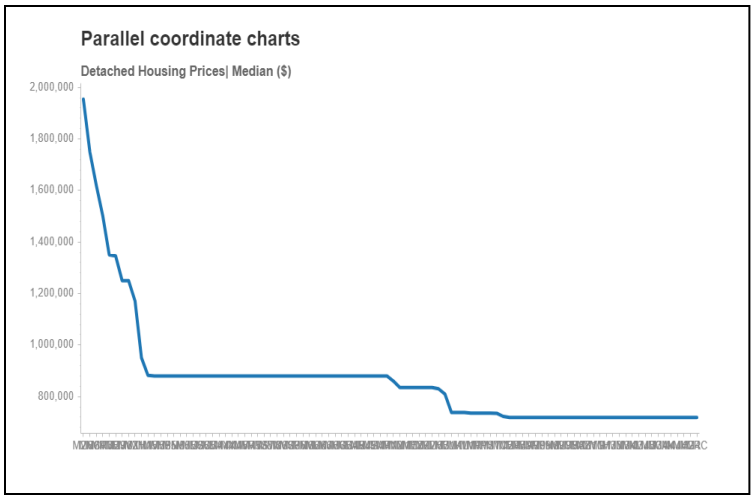


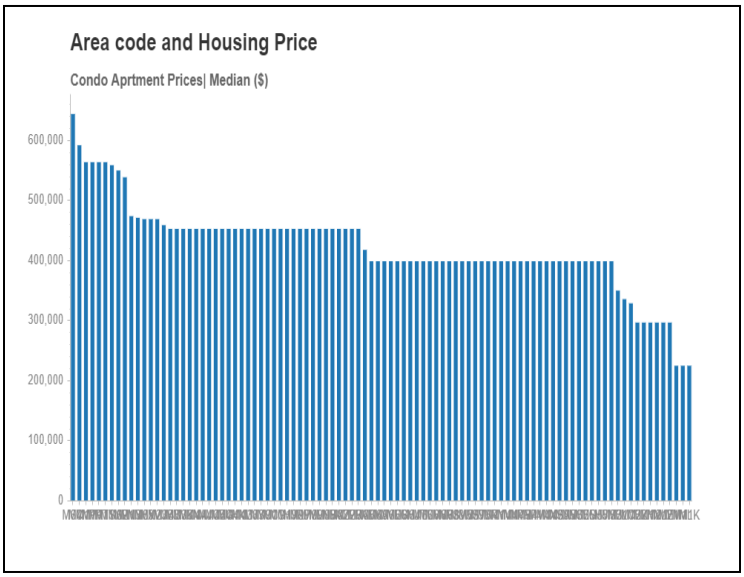
Group Bar for top 30

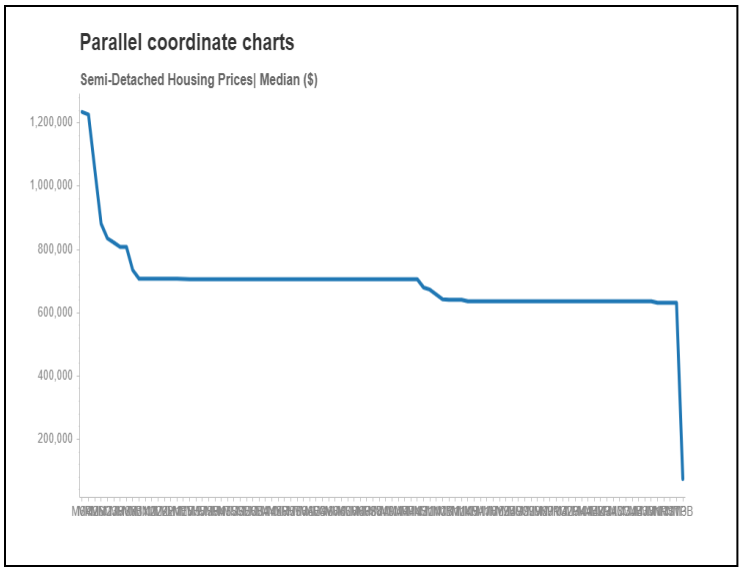


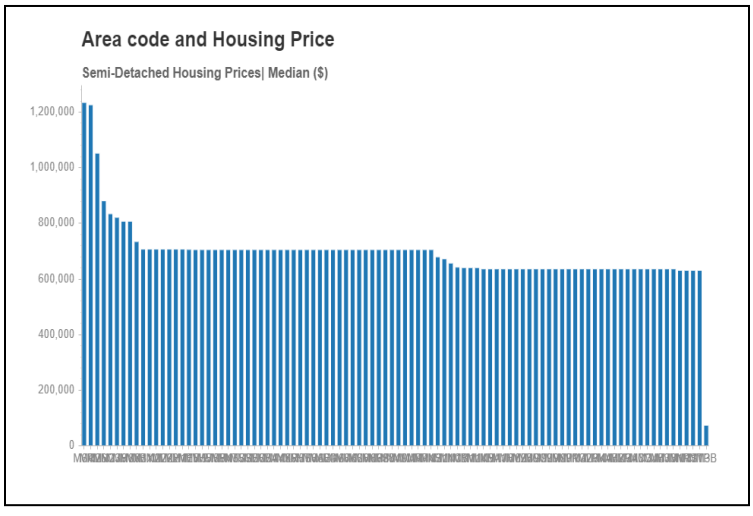
Plot the Housing price in Toronto Area

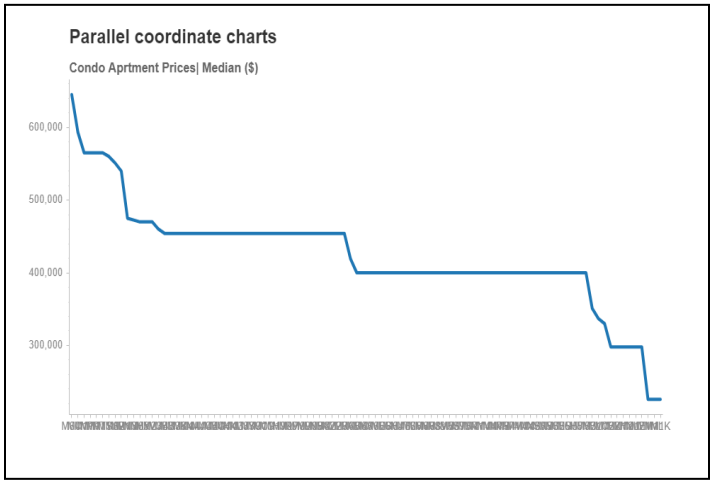






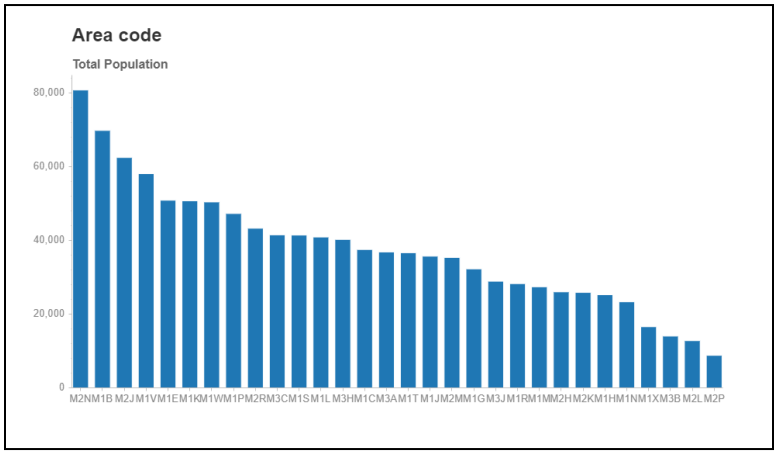


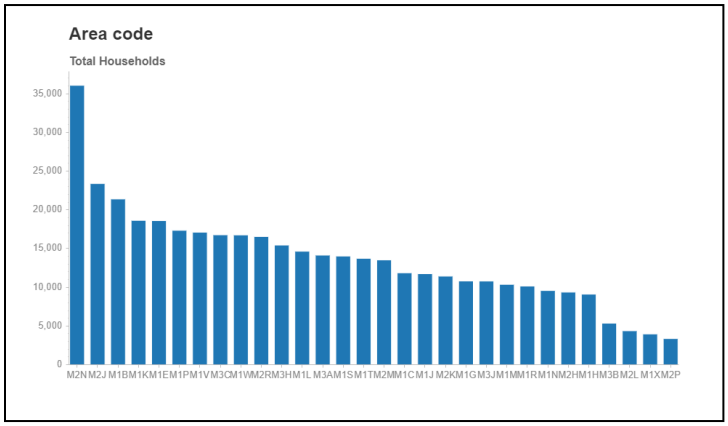


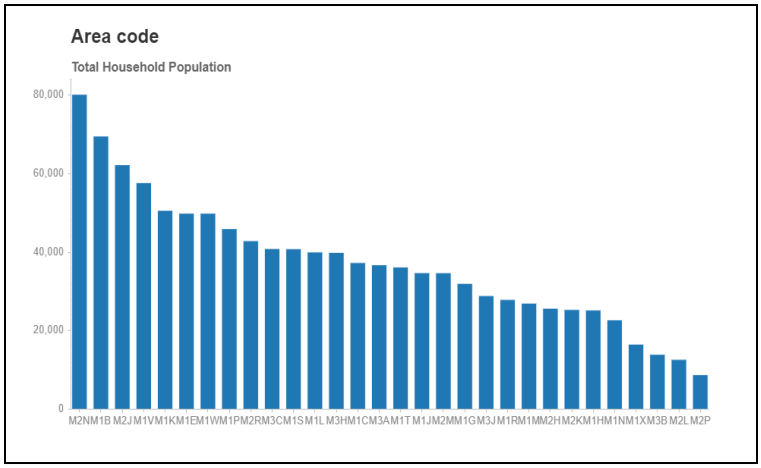


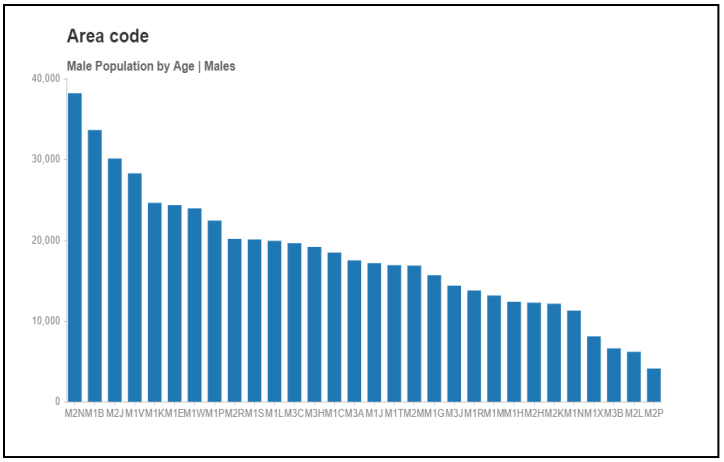
Results and Discussion

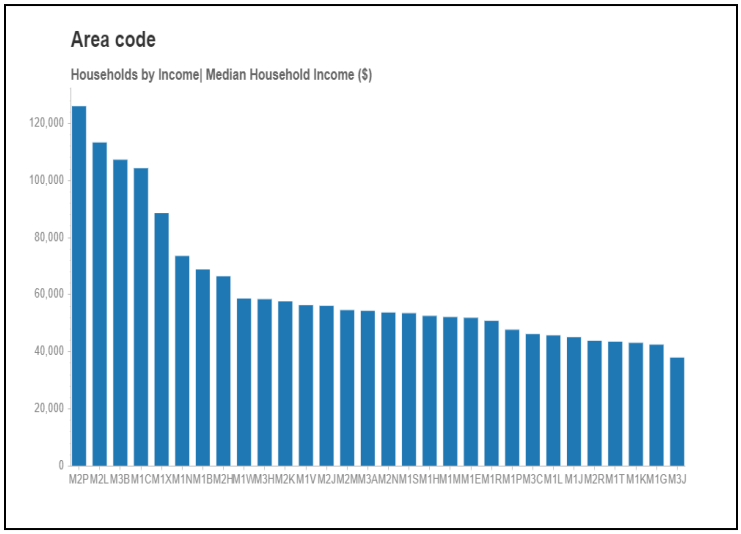
Plot the top 30 for all factors with Area Code

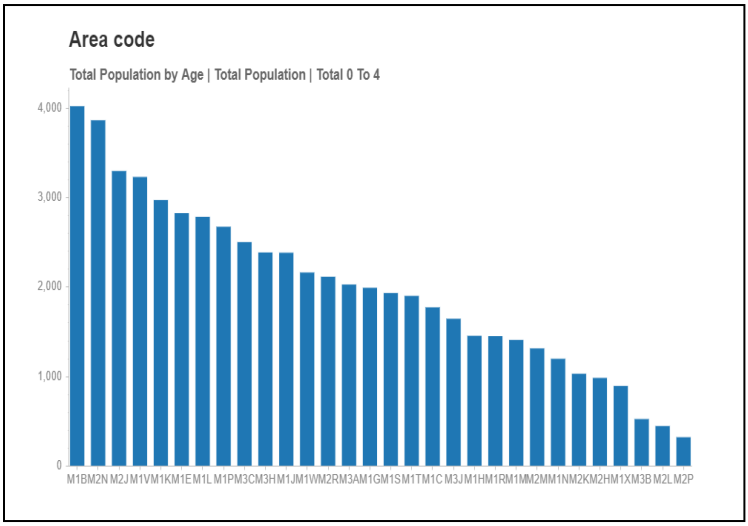






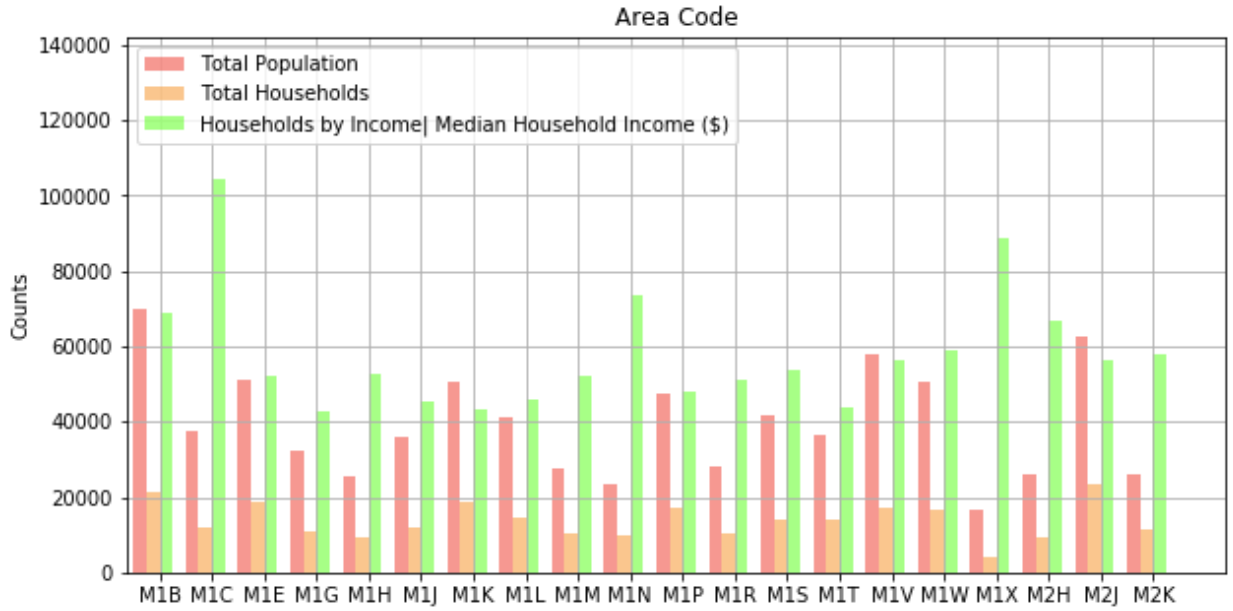






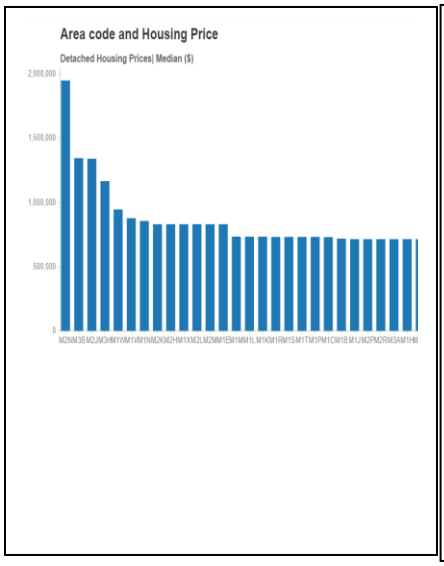
According to modelling, we found best three factors related to Housing price trend in Toronto area. Such as,

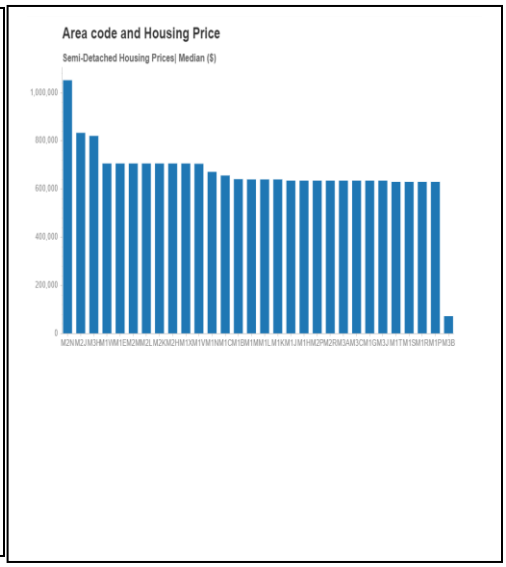
1. Total Population
2. Total Household
3. Household by Income

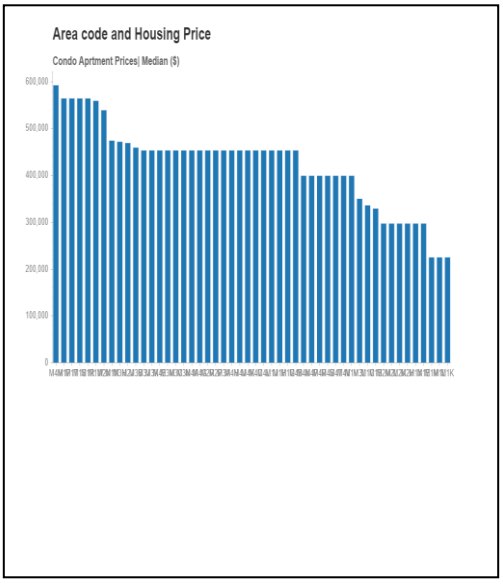


According to Modelling, we plot the best top 30 housing price trending area in Toronto City

(Detached Housing, Semi-Detached Housing, Condo Apartment)







Calculate some R-squared value using R programing,

>regmodel = lm(area name~Household Income,Toronto2018)

> summary(regmodel)

Call: lm(formula = zipcode ~ price, data = kc_house_data)

Residuals: Min 1Q Median 3Q Max -80.35 -45.56 -12.32 39.71 141.69

Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 9.808e+04 6.465e-01 151711.980 < 2e-16 *** price -7.754e-06 9.900e-07 -7.832 5.01e-15 *** ---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.43 on 21611 degrees of freedom Multiple R-squared: 0.002831, Adjusted R-squared: 0.002784 F-statistic: 61.34 on 1 and 21611 DF, p-value: 5.011e-15

Model Evaluation and Analysis Summary

- Highest impact on predicted price is Total Population, Household and Household income and total population and household are highest M2N, M1B and M2J and Highest household income M1C. Household income has reverse and unpredictable effect
- Highest housing price area M2N.
- Avg housing price increase is with Total Population and increase with Household.
- Linear regression with 2 features gives the most accurate price prediction (75%) and 1 feature is negative regression.