4/22/2017

NEIGHBOURHOODS IN TORONTO

DS8004 – FINAL PROJECT REPORT

Hira Fatima
RYERSON UNIVERSITY | MSC DATA SCIENCE AND ANALYTICS
|WINTER 2017

TABLE OF CONTENTS

INTRODUCTION

In this project, I start off by outlining the objective of this project, what data mining problem I am trying to solve, and what my dataset contains and where I got it from. I then move on to analyzing the data in more detail whereby learning about the variables (independent and dependent), and what is the proposed approach to solve my data mining problem in this project.

After discussing the cleaning and preprocessing steps, I jump right into a detailed descriptive analysis looking at various significant variables' values projected on Toronto's map. Following an in-depth analysis of correlations amongst variables, I then finally move onto the predictive analytics section where I try to answer the two main data mining questions. Are we able to predict house prices? And if so, which variables are the most significant predictors of house prices. In the end, I conclude the paper by discussing the problem of multicollinearity in the dataset and what is the best way to deal with it.

PROBLEM UNDERSTANDTING

DATA MINING PROBLEM?

As this is a data mining project, I wanted to mine the data related to the 140 Neighbourhoods in Toronto available on City of Toronto's website and use it to answer the following questions:

Predictive Analytics

- Can we predict average house price using some of the variables available?
- If so, which variables are the most significant predictors of average house prices in a given neighborhood?

Descriptive Analytics

- What more can I learn from the data using descriptive analytics and visualizations?

DATA SOURCE

Data for this project was collected from the City of Toronto website. City of Toronto has a Data Catalogue on their website with various dataset available containing different forms of information related to Toronto[1]. I picked 8 datasets from there under wellbeing Toronto section. The dataset names and their descriptions are listed below:

Dataset Name	Description
Demographics	This dataset contains demographics information for the 140 Neighbourhoods in Toronto
Housing	This dataset contains housing information for the 140 neighbourhoods in Toronto
Environment	This dataset contains Environment information for the 140 Neighbourhoods in Toronto
Safety	This dataset contains safety and crime information related to the 140 Neighbourhoods in Toronto
Economics	This dataset contains economical information for the 140 Neighbourhoods in Toronto
Civic	This dataset contains civic Information for the 140 Neighbourhoods in Toronto
Health	This dataset contains health information for the 140 Neighbourhoods in Toronto
Education	This dataset contains educational information for the 140 Neighbourhoods in Toronto
Neighbourhoods	This dataset contains the Neighbourhood IDs, their names and geospatial information which was used to create Toronto maps and project data onto it.

Each dataset contained 140 records for the 140 neighbourhoods in Toronto with about 10 to 15 variables on average for each dataset. After combining all the datasets by Neighbourhood ID, the master dataset contained 140 observations of 143 variables in Total. All variables contained continuous numerical data. Please see Appendix A, B and C for more details about the dataset.

DATA FILES

Data Files are embedded in this document. Please double click to open any file to view its contents.
The last file named DS8004Combined.xlsx is a consolidated version of the dataset where all individual datasets have been merged together using Neighbourhood ID.

Datasets

WB-Civics.xlsx WB-Demographics.xlsx WB-Economics.xlsx WB-Education.xlsx WB-Environment.xlsx WB-Health.xlsx

WB-Safety.xlsx WB-Housing.xlsx DS8004Combined.xlsx

Shape Files to Create Maps

neighbourhood_planning_areas_utm6.zip

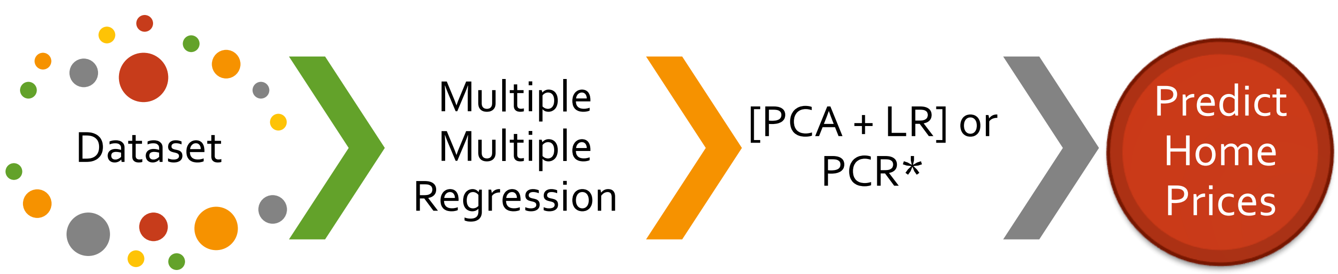
[1] http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=1a66e03bb8d1e310VgnVCM10000071d60f89RCRD

INDEPENDENT/DEPENDENT VARIABLES

For the prescriptive analytics part of analysis, the dependent variable is Average House Price and everything else is considered to be an independent variable.

PROPOSED APPROACH

Since all variables contain continuous numerical data and we want to predict the Average House Price given certain input parameters, it would make most sense to apply Regression Analysis. In this project, we will start of with Multiple Regression and check to see if there is multicollinearity amongst the variables or not. If there is multicollinearity, then before applying regression, PCA (Principle Component Analysis) would be required to eliminate multicollinearity problem amongst the input variables.



*PCA = Principle Component Analysis, LR = Multiple Regression, PCR = Principle Component Regression

DATA ACQUISITION

METHODOLOGY AND REPRODUCIBILITY

Since there were not many changes made to the dataset, it would be easy to reproduce the results with a new dataset with revised numbers. The only preprocessing steps included combining all the datasets in MS Excel using Neighbourhood ID, and removing spaces and special characters from the column names. As long as the column names remain the same, there should be no trouble reproducing the results with new and updated dataset.

CLEANING AND TRANSFORMATION

The following steps were applied to clean and transform the data:

- Removed special characters and spaces from column names
- When the dataset was imported into R from MS Excel, the values that were blank imported as NAs. These should have had zeros and therefore the NAs were replaced with zeros.
- Padded Neighbourhood ID with preceding zeros to match the shapefile. This was done so that the data could be projected on to maps using Neighbourhood ID as reference because in the shape file, it was given as a string value.

DESCRIPTIVE ANALYTICS

ANALYSIS OF RELATIONSHIP AMONG VARIABLES

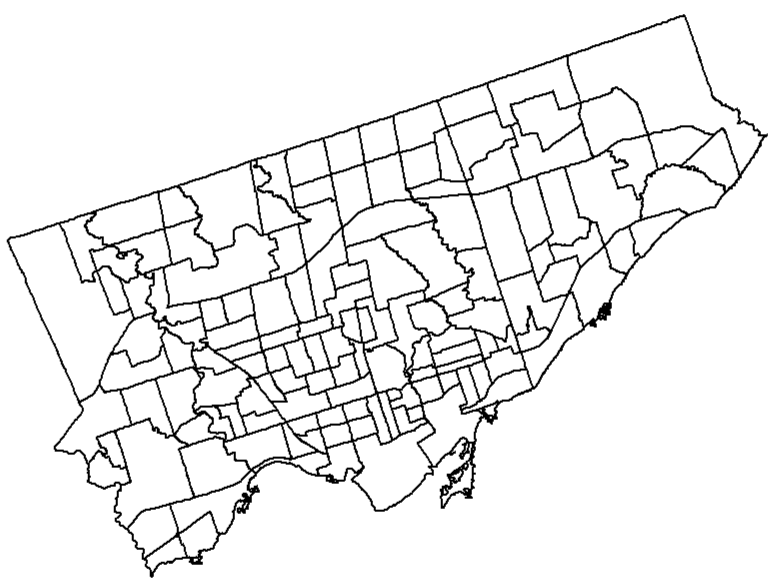
This file embedded here contains the correlation matrix for this dataset. It shows the relationship amongst all the variables contained in this dataset. Since the dataset was big with over 100 variables, it made more sense to extract the correlation matrix to MS Excel and analyze it there. The [embedded] document here has more details.

On a closer analysis of the variables, it s easy to tell that there are some obvious and expected correlations such as between language Chinese and population of Chinese people. However, there are some other interesting correlations too, such as Higher Income was highly positively correlated with House Prices. For more correlation details, please double click on the icon below to open the correlation matrix.

VISUALIZATIONS

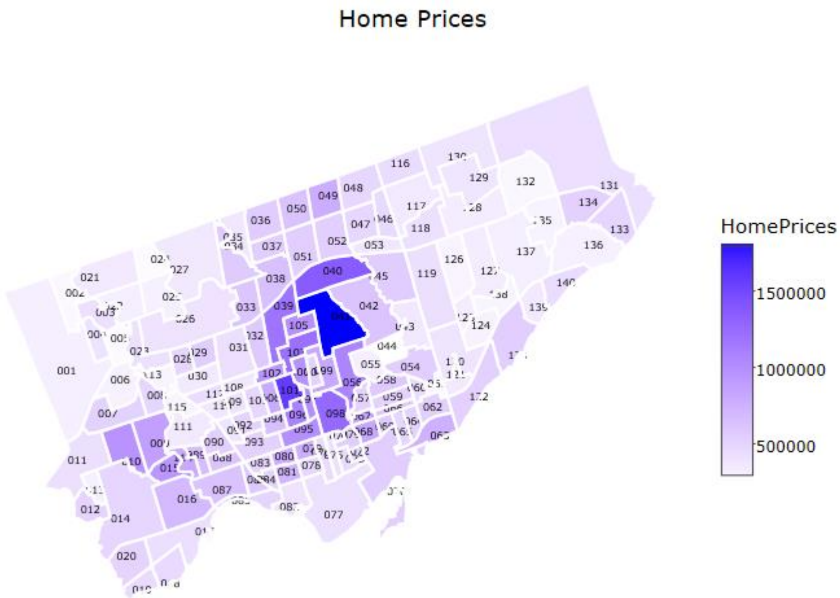
Data for various variables was projected on the map of Toronto to see the distribution of data spatially. It was not feasible to include over 100 charts for the 100 + variables, hence only the significant variables were plotted as shown below.

Let’s start off with a simple plot of Toronto’s neighbourhoods with boundaries as shown below.



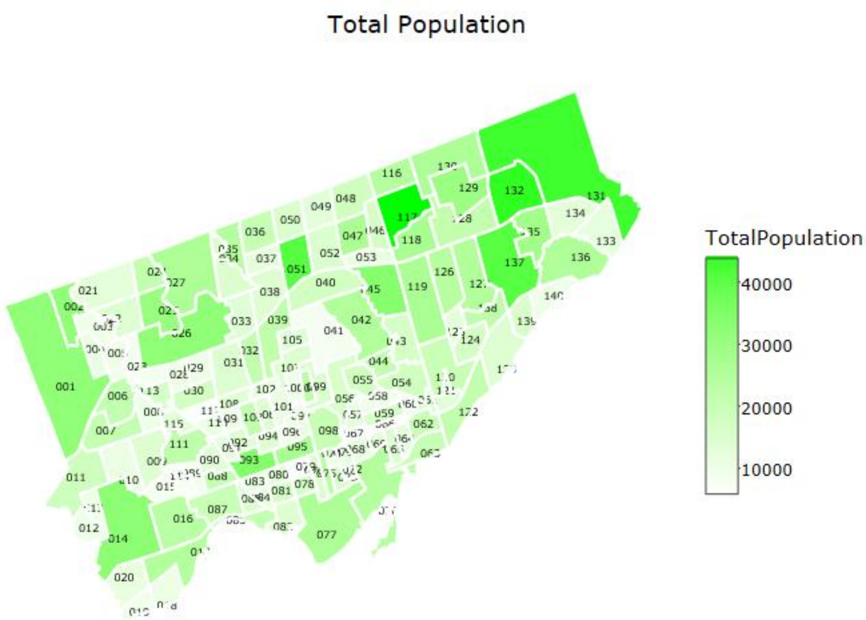
HOME PRICES

Neighbourhood Id: 041
Name: Bridle Path-Sunnybrook-York Mills - area around Bayview and York Mills Road
This neighbourhood has the highest average House Price value in the dataset.
Neighbourhood Id: 044
Name: Flemingdon Park - area around Eglinton and Don Mills
This neighbourhood has the lowest average House Price value in the dataset.



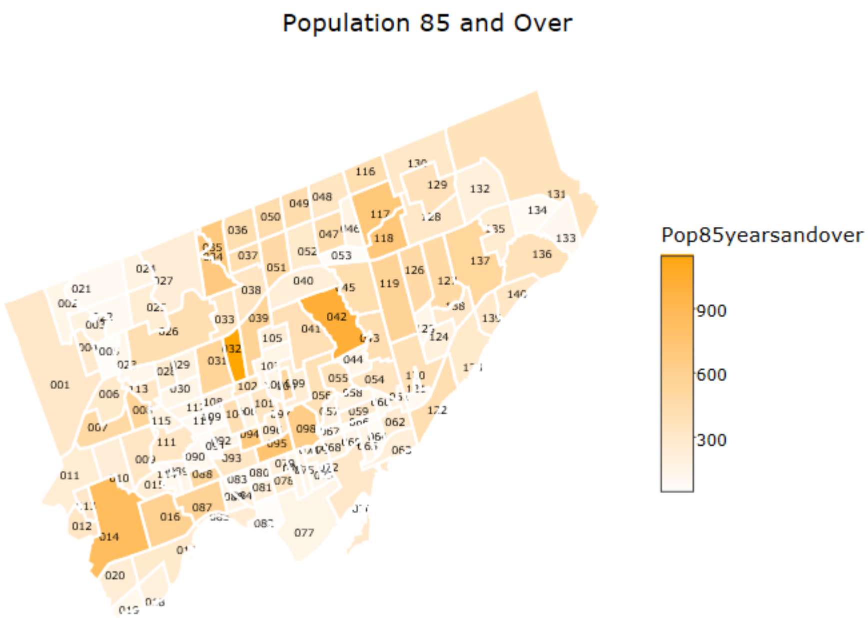
TOTAL POPULATION

Neighbourhood Id: 131
Name: Rouge - area around Shepperd and Meadowvale
This neighbourhood has the highest Total Population value in the dataset.



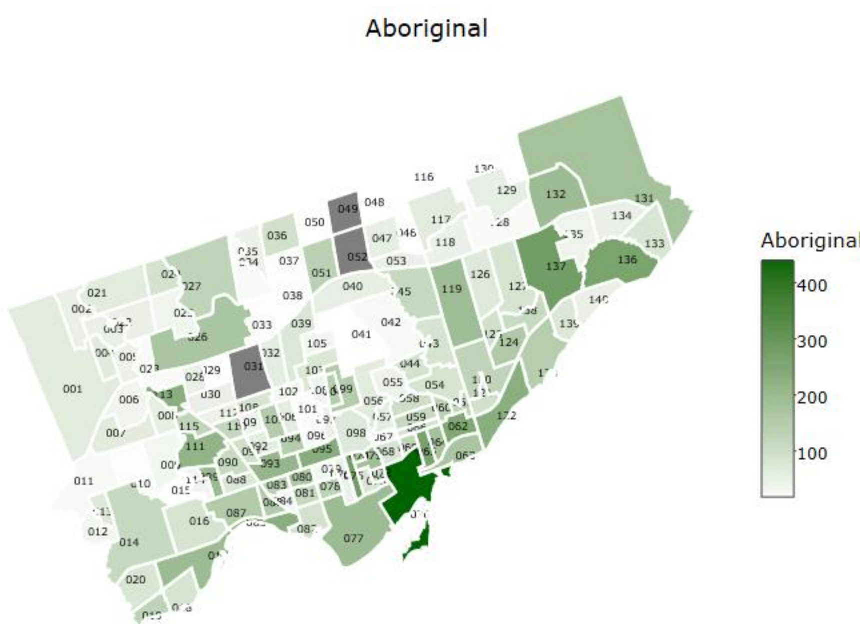
TOTAL POPULATION

Neighbourhood Id: 131
Name: Rouge - area around Shepperd and Meadowvale
This neighbourhood has the highest Total Population value in the dataset.



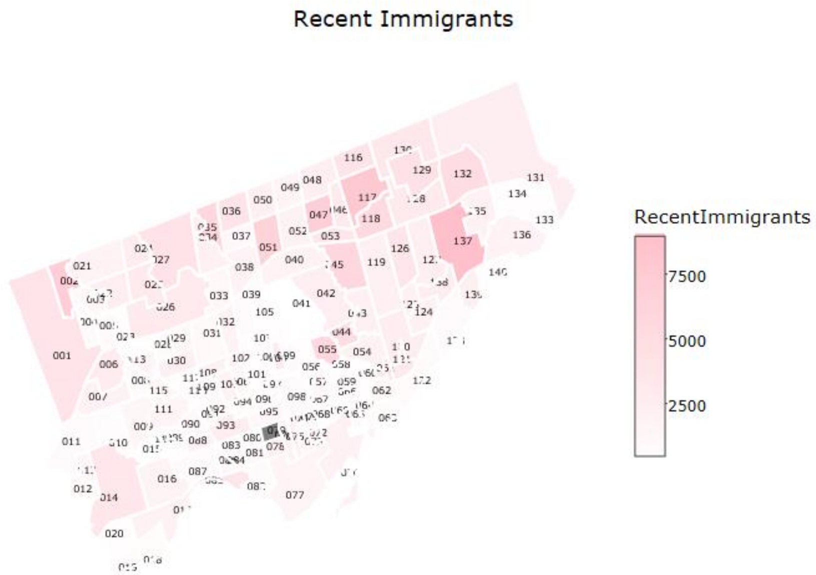
ABORIGINAL

Neighbourhood Id: 075
Name: South Riverdale - area around Danforth and Don Valley
This neighbourhood has the highest Aboriginal Population value in the dataset.



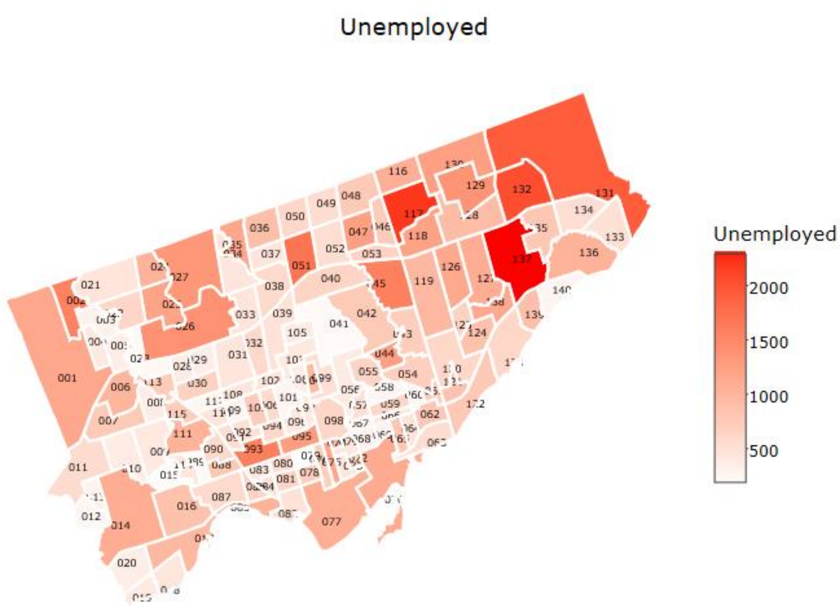
RECENT IMMIGRANTS

Neighbourhood Id: 137
Name: Woburn - area around Ellesmere and Markham
This neighbourhood has the highest Recent Immigration Population value in the dataset.



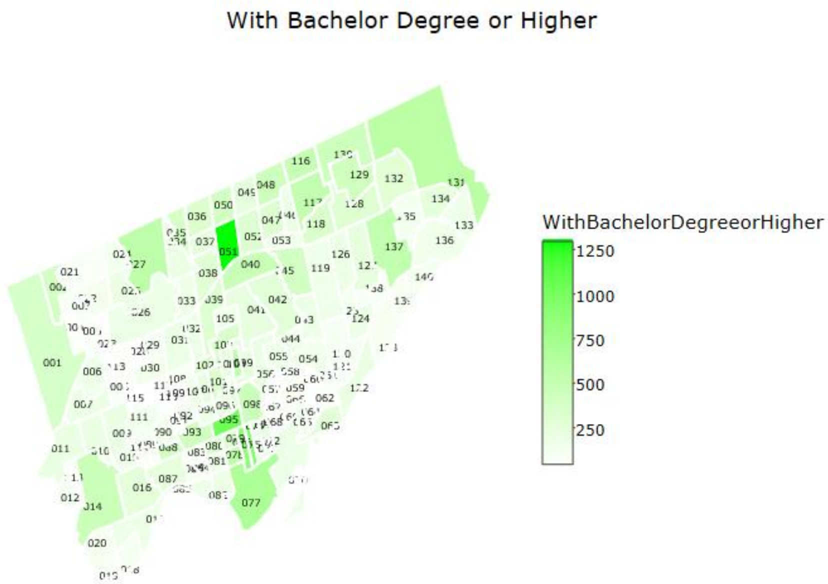
UNEMPLOYMENT

Neighbourhood Id: 137
Name: Woburn - area around Ellesmere and Markham
This neighbourhood has the highest unemployment value in the dataset.



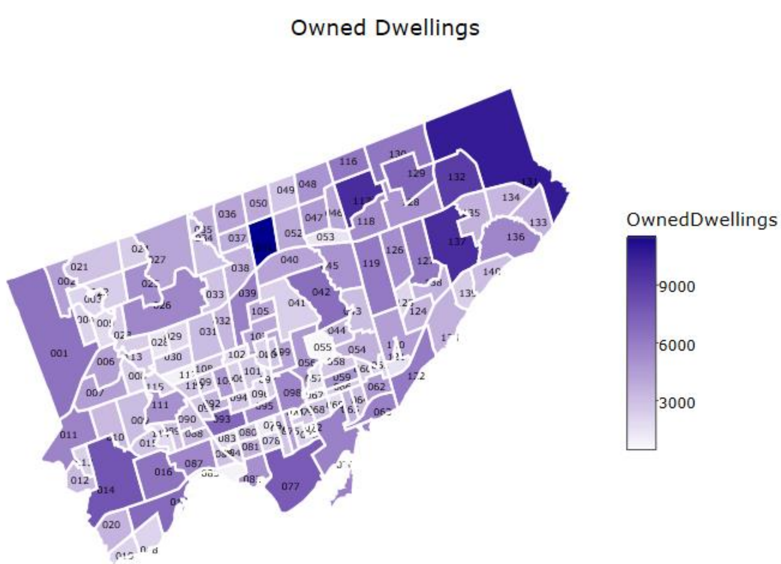
WITH BACHELOR OR HIGHER DEGREES

Neighbourhood Id: 051
Name: Willowdale East - area around Yonge and Finch
This neighbourhood has the highest Population of Individuals with Bachelor or Higher Degree value in the dataset.



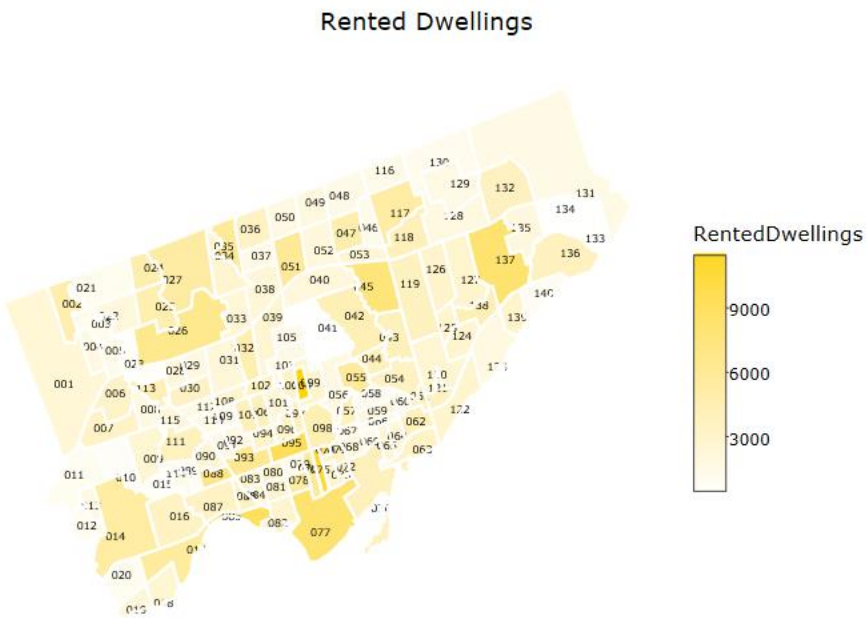
OWNED DWELLINGS

Neighbourhood Id: 131
Name: Rouge - area around Shepperd and Meadowvale
This neighbourhood has the highest value for Owned Dwellings the dataset.



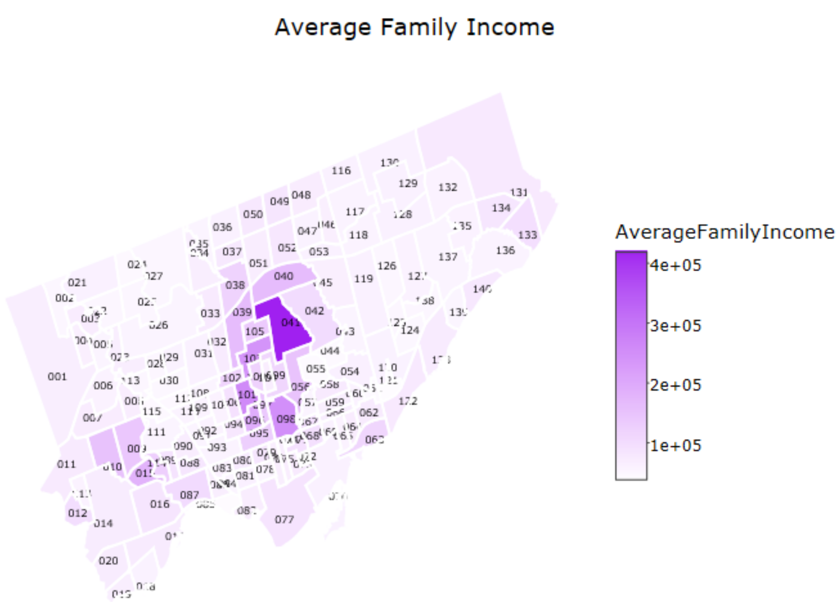
RENTED DWELLINGS

Neighbourhood Id: 137
Name: Woburn - area around Ellesmere and Markham
This neighbourhood has the highest value for Rented Dwellings in the dataset.



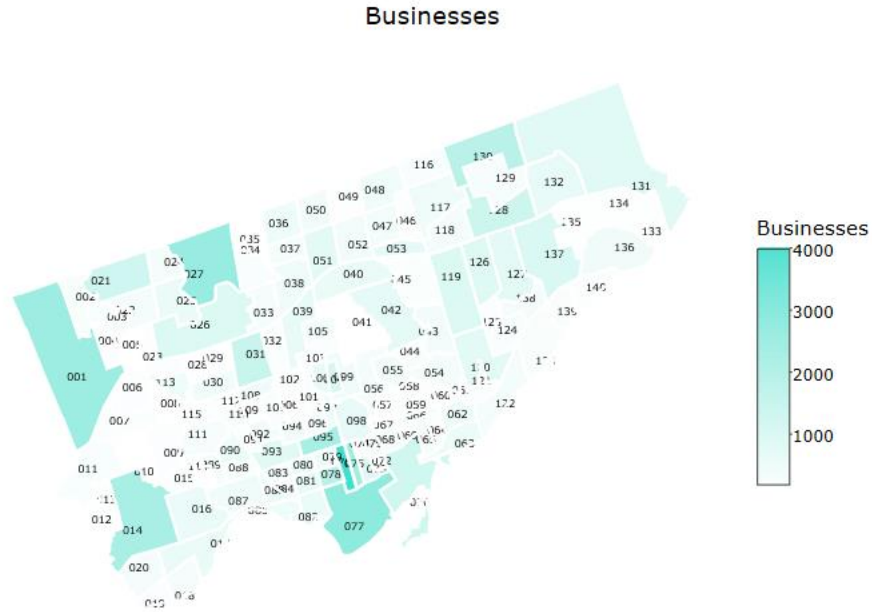
AVERAGE FAMILY INCOME

Neighbourhood Id: 041
Name: Bridle Path-Sunnybrook-York Mills - area around Bayview and York Mills Road
This neighbourhood has the highest average Family Income value in the dataset.



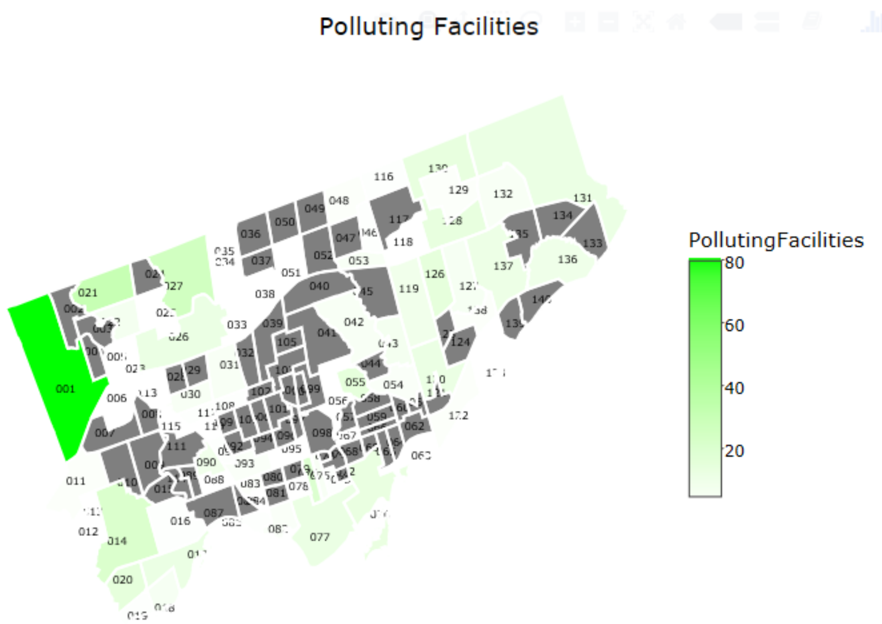
BUSINESS

Neighbourhood Id: 077
Name: Waterfront Communities-The Island - area around Downtown Toronto
This neighbourhood has the highest number of businesses in the dataset.



POLLUTING FACILITIES

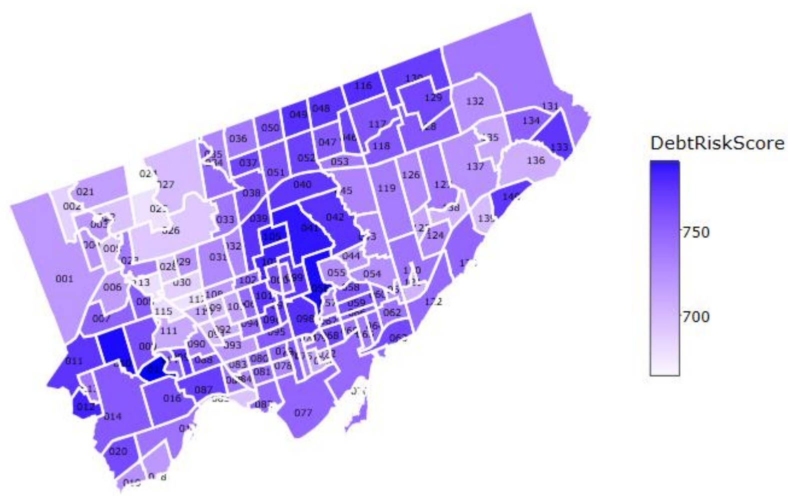
Neighbourhood Id: 001
Name: West Humber-Clairville - area around Highway 27 and Finch
This neighbourhood has the highest number of Polluting Facilities in the dataset.



DEBT RISK SCORE

Neighbourhood Id: 024
Name: Black Creek - area around Jane and Finch
This neighbourhood has the lowest value for Debt Risk Score in the dataset.
Neighbourhood Id: 015
Name: Kingsway South - area around Bloor and Royal York Rd
This neighbourhood has the highest value for Debt Risk Score in the dataset.

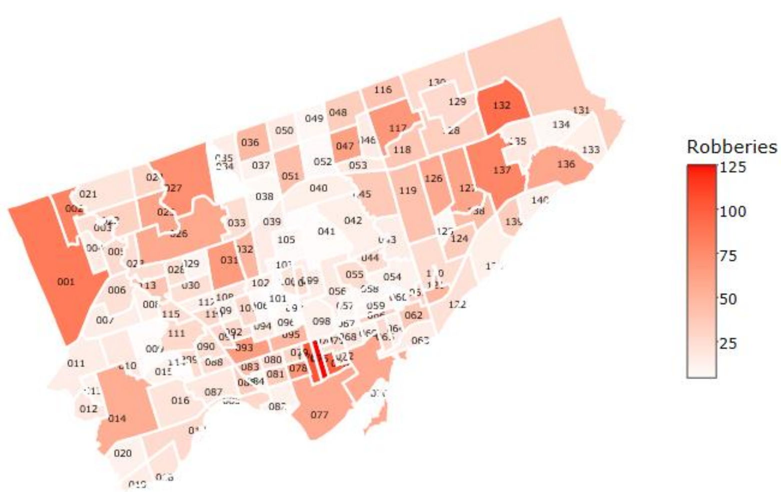
Debt Risk Score



ROBBERIES

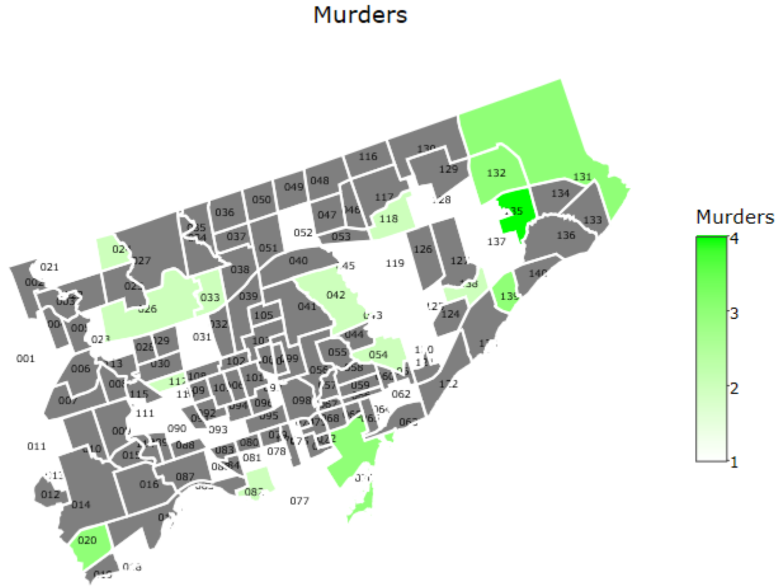
Neighbourhood Id: 075
Name: Church-Yonge Corridor - area around Yonge and Bloor
This neighbourhood has the highest value for Robberies in the dataset.

Robberies



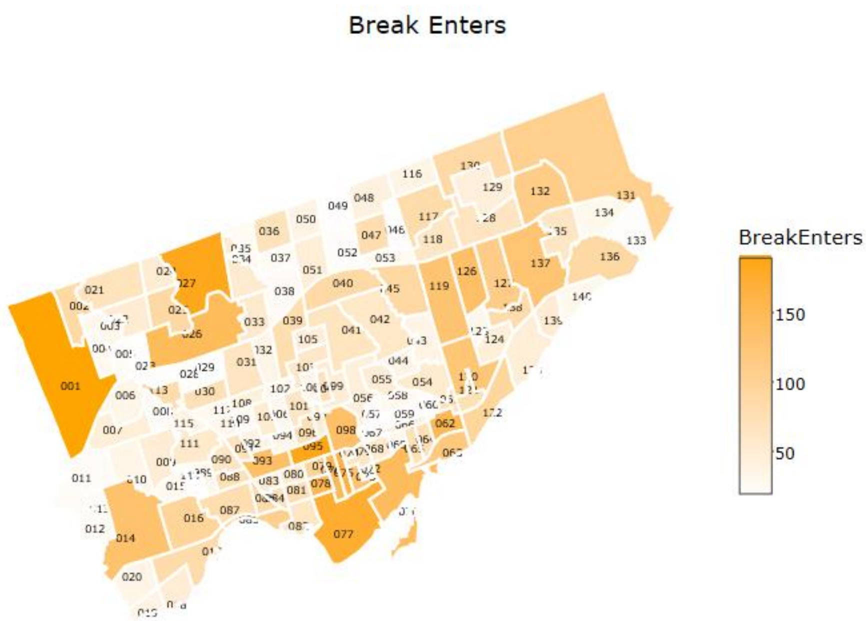
MURDERS

Neighbourhood Id: 135
Name: Morning Side - area around Ellesmere and Morningside
This neighbourhood has the highest value for Murders in the dataset.



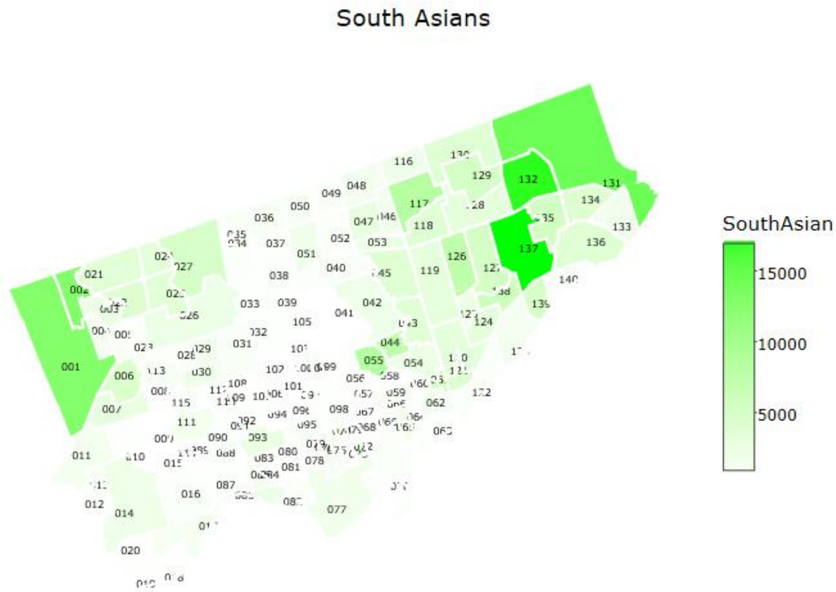
BREAK ENTERS

Neighbourhood Id: 001
Name: West Humber-Clairville - area around Highway 27 and Finch
This neighbourhood has the highest number of Break Enters in the dataset.



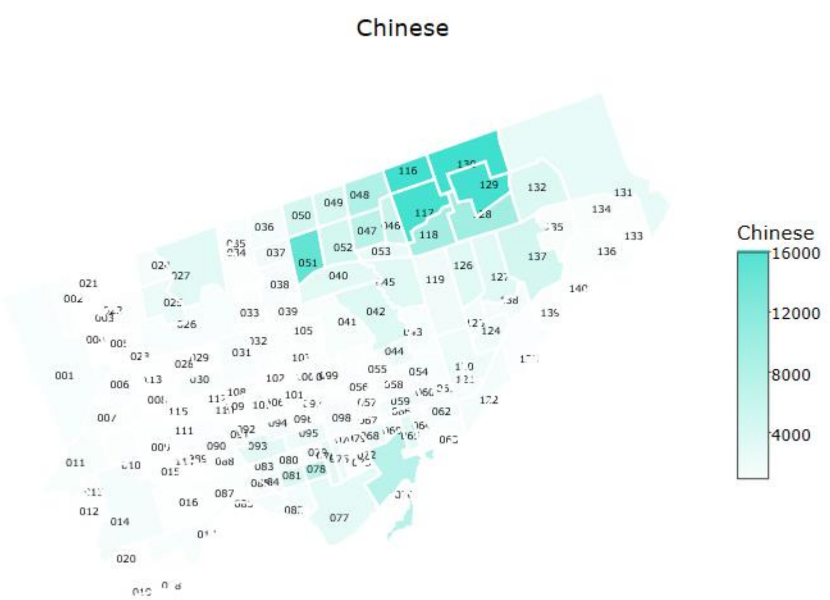
SOUTH ASIAN POPULATION

Neighbourhood Id: 137
Name: Woburn - area around Ellesmere and Markham
This neighbourhood has the highest South Asian Population value in the dataset.



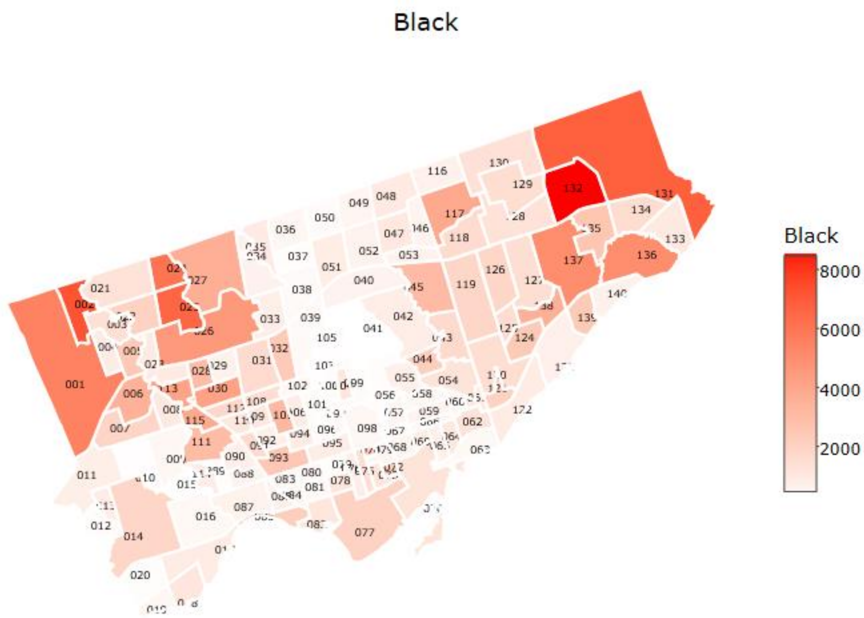
CHINESE POPULATION

Neighbourhood Id: 116, 117,129 and 130
Name: Neighbourhoods around Finch and Markham
This neighbourhood has the highest Chinese Population value in the dataset.



BLACK POPULATION

Neighbourhood Id: 132
Name: Malvern Neighbourhood around Shepperd and Neilson Road
This neighbourhood has the highest Black Population value in the dataset.



PREDICTIVE MODELING

FEATURE EVALUATION AND IMPACT ANALYSIS

The correlation matrix showed that a lot of the independent variables were correlated which meant that the dataset suffered with high multicollinearity. This problem was also evident when Multiple Regression was applied and it provided different results for different combinations of input variables as explain under the 'Model Design, Train and Validation' section. Furthermore, a check of VIF (Variance Inflation Factor) confirmed the multicollinearity as the values for VIF were much higher and so were the Kappa values.

FEATURE ENGINEERING AND EXTRACTION

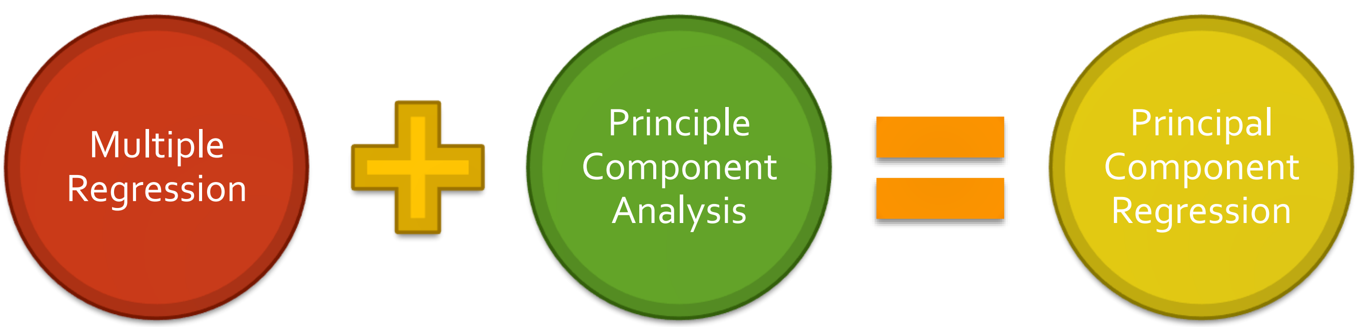
To find the right set of input variables that would correctly predict the Average House Price, I tried different techniques. First, I used simple Multiple Regression with K-Fold Cross Validation. Although it provided results with a high value for R squared, changing the input variables and trying out different combinations of input variables changed the significance of input variables vastly. For instance, in Model 1, CityGrantFunding was not a significant variable, however, in Model 2 where another variable called LowIncomeFamilies was added, CityGrantFunding's significance increased and it became significant in the model.

Then I tried to apply PCA before applying Multiple Regression. However, instead of applying these techniques separately, I decided to go with PCR (Principle Component Regression). Since the dataset suffered from high multicollinearity, at the end PCR, provided most satisfactory results even though R squared of predictions was lower than the simple Multiple Regression.

Before applying these models, irrelevant variables such as population, demographics and language variables were removed from the dataset as well. For example, there was no point to keep columns [Country: China] and [Language: Chinese].

Dataset was divided into test and train subsets randomly and K-Fold Cross Validation was also used to improve accuracy as the dataset was considerably small with only 140 observations.

For Multiple Regression, only the most significant variables [based on domain knowledge] were used.



MODEL DESIGN, TRAIN AND VALIDATION

In the first model, Multiple Regression was applied to evaluate the significance of four input variables. Based on the first model, Average Income and Debt Risk Score turned out to be the most significant variable.

MODEL 1

```
################## Model 1

predictionModel1 <- lm(HomePrices ~
AverageFamilyIncome+DebtRiskScore+CityGrantsFunding+WithBachelorDegreeorHigher, data = trainData)
summary(predictionModel1)
##
## Call:
## lm(formula = HomePrices ~ AverageFamilyIncome + DebtRiskScore +
##     CityGrantsFunding + WithBachelorDegreeorHigher, data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -325014  -80714  -13803   75357  372786
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.088e+06  3.799e+05  -2.865  0.00509 **
## AverageFamilyIncome        4.487e+00  2.880e-01  15.580  < 2e-16 ***
## DebtRiskScore              1.725e+03  5.365e+02   3.216  0.00175 **
## CityGrantsFunding         -1.539e-02  1.499e-02  -1.027  0.30712
## WithBachelorDegreeorHigher 1.203e+01  6.753e+01   0.178  0.85898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 116500 on 100 degrees of freedom
## Multiple R-squared:  0.8243, Adjusted R-squared:  0.8173
## F-statistic: 117.3 on 4 and 100 DF,  p-value: < 2.2e-16
prediction1 <- predict(predictionModel1, newdata = testData)
head(prediction1)
##        2        4       10       12       16       26
## 338354.6 448215.1 964702.6 705270.0 609697.1 354084.9
head(testData$HomePrices)
## [1] 251119 392271 971668 505350 690949 400486
#Calculate R-Squared for predicted values
SSE <- sum((testData$HomePrices - prediction1) ^ 2)
SST <- sum((testData$HomePrices - mean(testData$HomePrices)) ^ 2)
1 - SSE/SST
## [1] 0.8066456
```

The predictions had an R-squared value of .80 which is considerably high confirming that AverageIncome and DebtRiskScore are significant contributors to House Prices.

In the second model, another variable was added called LowIncomeFamilies. This was added to see how it would change the model and if it would affect/alter the significance of any of the previous variables. As shown in the results below, it did not change the significance of AverageIncome and DebtRiskScore, however, it did change the significance of other variables called CityGrantFundings and WhithBachelorDegree. In model 1 these two variables were not significant, but after adding LowIncomeFamilies, these variables became significant. The R-squared of the predictions also increased to .83

MODEL 2

```
################## Model 2

predictionModel2 <- lm(HomePrices ~
AverageFamilyIncome+DebtRiskScore+CityGrantsFunding+LowIncomeFamilies+WithBachelorDegreeorHigher,
data = trainData)
summary(predictionModel2)
##
## Call:
## lm(formula = HomePrices ~ AverageFamilyIncome + DebtRiskScore +
##     CityGrantsFunding + LowIncomeFamilies + WithBachelorDegreeorHigher,
##     data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -297653  -67039   -5123   62225  357332
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -8.312e+05  3.696e+05  -2.249  0.02672 *
## AverageFamilyIncome         4.273e+00  2.813e-01  15.190  < 2e-16 ***
## DebtRiskScore               1.486e+03  5.156e+02   2.881  0.00486 **
## CityGrantsFunding          -3.138e-02  1.504e-02  -2.087  0.03948 *
## LowIncomeFamilies          -2.065e+01  6.126e+00  -3.371  0.00107 **
## WithBachelorDegreeorHigher  1.836e+02  8.199e+01   2.240  0.02736 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 110900 on 99 degrees of freedom
## Multiple R-squared:  0.8424, Adjusted R-squared:  0.8344
## F-statistic: 105.8 on 5 and 99 DF,  p-value: < 2.2e-16
prediction2 <- predict(predictionModel2, newdata = testData)
head(prediction2)
##        2        4       10       12       16       26
## 328527.5 469670.2 969716.6 709290.9 565096.9 278293.7
head(testData$HomePrices)
## [1] 251119 392271 971668 505350 690949 400486
#Calculate R-Squared for predicted values
SSE <- sum((testData$HomePrices - prediction2) ^ 2)
SST <- sum((testData$HomePrices - mean(testData$HomePrices)) ^ 2)
1 - SSE/SST
## [1] 0.8338398
```

To further test what would happen if the input variables were changed, I ran one more iteration, this time with the addition of another variable called TeenPregnancy. With the addition of this variable, the significance of many other previous variables changed. For instance, DebtRiskScore which was significant in models 1 and 2 became insignificant in model 3. The R squared for predictions had a very minor decrease.

MODEL 3

```
################### Model 3

predictionModel3 <- lm(HomePrices ~
AverageFamilyIncome+DebtRiskScore+CityGrantsFunding+LowIncomeFamilies+WithBachelorDegreeorHigher+Tee
nPregnancy, data = trainData)
summary(predictionModel3)
##
## Call:
## lm(formula = HomePrices ~ AverageFamilyIncome + DebtRiskScore +
##     CityGrantsFunding + LowIncomeFamilies + WithBachelorDegreeorHigher +
##     TeenPregnancy, data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -291699  -66364   -8643   61682  359573
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -6.460e+05  5.078e+05  -1.272  0.20632
## AverageFamilyIncome         4.250e+00  2.857e-01  14.878  < 2e-16 ***
## DebtRiskScore               1.261e+03  6.665e+02   1.892  0.06145 .
## CityGrantsFunding          -2.933e-02  1.558e-02  -1.883  0.06268 .
## LowIncomeFamilies          -2.059e+01  6.149e+00  -3.348  0.00116 **
## WithBachelorDegreeorHigher  1.774e+02  8.311e+01   2.134  0.03534 *
## TeenPregnancy              -6.167e+02  1.154e+03  -0.534  0.59432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 111300 on 98 degrees of freedom
## Multiple R-squared:  0.8429, Adjusted R-squared:  0.8332
## F-statistic: 87.61 on 6 and 98 DF,  p-value: < 2.2e-16
prediction3 <- predict(predictionModel3, newdata = testData)
head(prediction3)
##        2        4       10       12       16       26
## 333495.0 470627.5 968243.1 716093.5 566843.0 277165.1
head(testData$HomePrices)
## [1] 251119 392271 971668 505350 690949 400486
#Calculate R-Squared for predicted values
SSE <- sum((testData$HomePrices - prediction3) ^ 2)
SST <- sum((testData$HomePrices - mean(testData$HomePrices)) ^ 2)
1 - SSE/SST
## [1] 0.8307024
```

With the unstable results retrieved from Models 1, 2, and 3 in mind, it only made sense to use PCA for dimensionality reduction and eliminating multicollinearity. In model 4, I applied PCR which is a combination of PCA and Multiple Regression. As mentioned before, I removed all the irrelevant and duplicate variables before feeding the dataset in model 4 to keep the list of variables at its minimum. A total of 112 variables were fed in and the model considered 93 components only. The graph shows that 60 components could provide optimal prediction results. Based on that, 60 components were used to predict the test dataset and it provided results with R squared value of .73. Although this value is slightly lower, I do feel more comfortable with Model 4 as it eliminated the problem of multicollinearity.

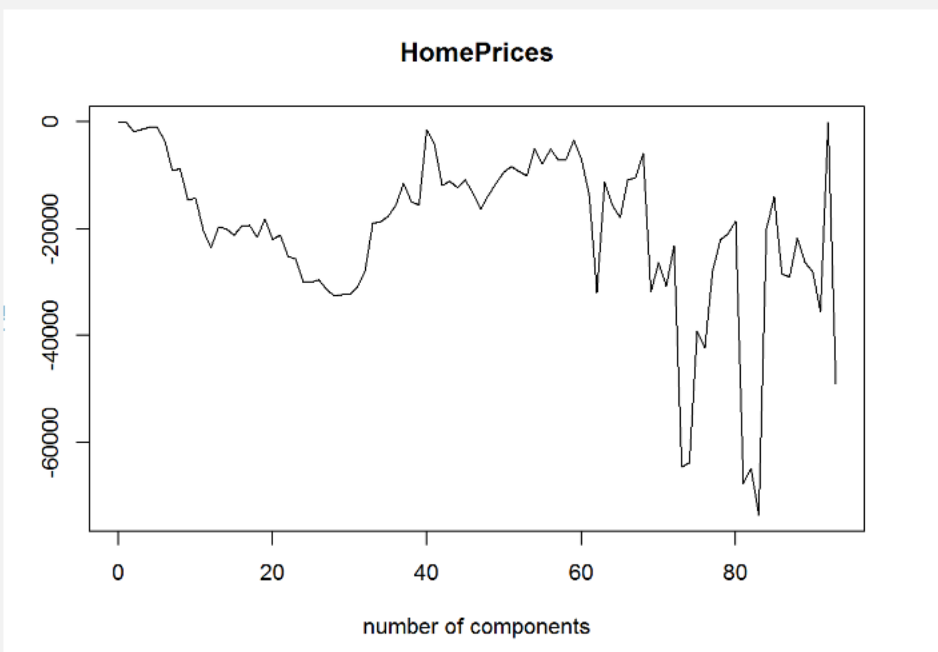
MODEL 4

```
################### Model 4

library(pls)
##
## Attaching package: 'pls'
## The following object is masked from 'package:corrplot':
##
##     corrplot
## The following object is masked from 'package:stats':
##
##     loadings
predictionModel4 <- pcr(HomePrices~., data = trainData, scale =TRUE, validation = "CV")
summary(predictionModel4)
## Data:    X dimension: 105 112
##  Y dimension: 105 1
## Fit method: svdpc
## Number of components considered: 93
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##       (Intercept)  1 comps   2 comps  3 comps  4 comps  5 comps   6 comps
## CV         273824  1457114  11509213  9938028  8580760  8810436  16137769
## adjCV      273824  1381017  10889881  9403288  8119126  8336388  15269182
##        7 comps   8 comps   9 comps  10 comps  11 comps  12 comps
## CV    25994142  25289736  32833511  32383445  38671390  41601139
## adjCV 24594928  23928462  31066118  30640287  36589749  39361802
##       13 comps  14 comps  15 comps  16 comps  17 comps  18 comps
## CV    37952110  38347899  39501255  37822275  37793429  39813436
## adjCV 35909194  36283661  37374928  35786327  35759034  37670306
##       19 comps  20 comps  21 comps  22 comps  23 comps  24 comps
## CV    36631565  40165370  39484825  42952284  43434430  46948653
## adjCV 34659717  38003297  37359387  40640191  41096384  44421434
##       25 comps  26 comps  27 comps  28 comps  29 comps  30 comps
## CV    46964427  46658482  48029583  48967851  48804921  48794673
## adjCV 44436356  44146879  45444175  46331937  46177776  46168081
##       31 comps  32 comps  33 comps  34 comps  35 comps  36 comps
## CV    47661933  45350343  37404760  37101669  36084317  33816275
## adjCV 45096317  42909160  35391287  35104513  34141925  31995974
##       37 comps  38 comps  39 comps  40 comps  41 comps  42 comps
## CV    29039697  33182367  33812164   9922305  17353708  29559393
```

```
## adjCV  27476529  31396194  31992097   9388271  16419598  27968244
##       43 comps  44 comps  45 comps  46 comps  47 comps  48 comps
## CV    28497320  30017432  28170016  31360162  34664233  31720551
## adjCV  26963339  28401616  26653645  29672065  32798280  30013060
##       49 comps  50 comps  51 comps  52 comps  53 comps  54 comps
## CV    29018867  26225649  24806199  26035265  27256452  18952860
## adjCV  27456810  24813944  23470902  24633809  25789260  17932658
##       55 comps  56 comps  57 comps  58 comps  59 comps  60 comps
## CV    23934559  19214673  22876225  22692368  15853031  22683050
## adjCV  22646189  18180380  21644831  21470871  14999714  21462046
##       61 comps  62 comps  63 comps  64 comps  65 comps  66 comps
## CV    31768948  48534696  28666246  33776175  36318397  28319649
## adjCV  30058848  45922093  27123164  31958024  34363397  26795223
##       67 comps  68 comps  69 comps  70 comps  71 comps  72 comps
## CV    27749871  20785266  48440235  43959503  47582496  41233062
## adjCV  26256117  19666423  45832721  41593185  45021153  39013508
##       73 comps  74 comps  75 comps  76 comps  77 comps  78 comps
## CV    68983686  68522526  53716860  55763229  45425859  40340124
## adjCV  65270317  64833982  50825302  52761515  42980606  38168636
##       79 comps  80 comps  81 comps  82 comps  83 comps  84 comps
## CV    39191571  36997916  70612526  69074953  73642153  38348767
## adjCV  37081910  35006335  66811475  65356669  69678017  36284472
##       85 comps  86 comps  87 comps  88 comps  89 comps  90 comps
## CV    32141438  45714013  46255075  39994954  43971539  45379543
## adjCV  30411285  43253246  43765181  37842042  41604569  42936780
##       91 comps  92 comps  93 comps
## CV    51070711   1679107  60024818
## adjCV  48321593   1588965  56793706
##
## TRAINING: % variance explained
##          1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X          32.74    44.71    53.45    59.48    63.69    67.65    70.40
## HomePrices    18.49    44.12    48.09    48.63    54.99    62.10    66.87
##          8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X          73.0    75.10    77.05    78.73    80.19    81.64
## HomePrices    66.9    73.75    73.75    74.72    75.12    78.04
##          14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
## X          82.95    84.08    85.16    86.09    86.96    87.77
## HomePrices    81.34    82.59    84.07    84.13    84.13    84.14
##          20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
## X          88.57    89.30    89.98    90.61    91.21    91.79
## HomePrices    84.36    84.53    84.83    85.20    85.90    86.72
##          26 comps  27 comps  28 comps  29 comps  30 comps  31 comps
## X          92.32    92.82    93.30    93.75    94.17    94.57
## HomePrices    87.03    87.33    87.37    87.68    87.68    87.80
##          32 comps  33 comps  34 comps  35 comps  36 comps  37 comps
## X          94.96    95.33    95.67    95.99    96.29    96.56
## HomePrices    88.34    88.71    88.80    88.90    89.04    89.04
##          38 comps  39 comps  40 comps  41 comps  42 comps  43 comps
## X          96.81    97.05    97.27    97.49    97.66    97.82
```

```
## HomePrices    89.69    89.70    91.05    91.83    91.91    92.56
##          44 comps  45 comps  46 comps  47 comps  48 comps  49 comps
## X           97.98    98.12    98.26    98.40    98.51    98.62
## HomePrices    93.34    93.64    93.88    93.89    93.92    93.94
##          50 comps  51 comps  52 comps  53 comps  54 comps  55 comps
## X           98.72    98.81    98.90    98.99    99.07    99.15
## HomePrices    94.67    94.88    94.91    94.91    94.91    94.94
##          56 comps  57 comps  58 comps  59 comps  60 comps  61 comps
## X           99.23    99.29    99.36    99.42    99.47    99.52
## HomePrices    94.95    94.95    95.12    95.13    96.12    96.12
##          62 comps  63 comps  64 comps  65 comps  66 comps  67 comps
## X           99.56    99.61    99.64    99.68    99.71    99.74
## HomePrices    96.36    96.52    96.56    96.71    96.83    96.83
##          68 comps  69 comps  70 comps  71 comps  72 comps  73 comps
## X           99.77    99.79    99.82    99.84    99.86    99.88
## HomePrices    96.83    96.85    97.08    97.17    97.39    97.58
##          74 comps  75 comps  76 comps  77 comps  78 comps  79 comps
## X           99.90    99.91    99.92    99.93    99.94    99.95
## HomePrices    97.59    97.93    98.04    98.04    98.07    98.08
##          80 comps  81 comps  82 comps  83 comps  84 comps  85 comps
## X           99.96    99.97    99.97    99.98    99.98    99.98
## HomePrices    98.57    98.58    98.58    98.65    98.65    98.74
##          86 comps  87 comps  88 comps  89 comps  90 comps  91 comps
## X           99.99    99.99    99.99    99.99    99.99   100.00
## HomePrices    98.89    99.11    99.12    99.15    99.20    99.27
##          92 comps  93 comps
## X          100.00   100.00
## HomePrices    99.28    99.28
validationplot(predictionModel4, val.type="R2")
```



```
prediction4 <- predict(predictionModel4, testData, ncomp = 60)

#Calculate R-Squared for predicted values
SSE <- sum((testData$HomePrices - prediction4) ^ 2)
SST <- sum((testData$HomePrices - mean(testData$HomePrices)) ^ 2)
1 - SSE/SST
## [1] 0.7346496
```

CONCLUSION

As part of this project, I was able to analyze different neighbourhoods in Toronto and learn where the highest or lowest concentration of various variables occurred while capitalizing on description analytics. For instance, Neighbourhood 137 showed that it has the highest recent Immigrant population along with the highest value for unemployment as well, which makes complete sense as it is challenging for newcomers to find jobs right away. Similarly, the graphs also showed many other interesting facts such as how neighbourhood 141 has the highest average family income and highest average house prices.

Using predictive analytics, I devised different models to predict the house prices and concluded that it made most sense to use Principle Component Regression as my dataset suffered from high multicollinearity. I was also able to figure out variables that were most significant in predicting average house prices including average income and debt risk score.

This project was a great exercise and exposed me to an important example of multicollinearity and how to deal with it.

APPENDICES

APPENDIX A - Column Names

```
##  [1] "Neighbourhood"             "NeighbourhoodId"
##  [3] "HomePrices"                "CityGrantsFunding"
##  [5] "DiversityIndex"            "VoterTurnout"
##  [7] "WatermainBreaks"           "TotalArea"
##  [9] "TotalPopulation"           "Pop-Males"
## [11] "Pop-Females"               "Pop0-4years"
## [13] "Pop5-9years"               "Pop10-14years"
## [15] "Pop15-19years"             "Pop20-24years"
## [17] "Pop25-29years"             "Pop30-34years"
## [19] "Pop35-39years"             "Pop40-44years"
## [21] "Pop45-49years"             "Pop50-54years"
## [23] "Pop55-59years"             "Pop60-64years"
## [25] "Pop65-69years"             "Pop70-74years"
## [27] "Pop75-79years"             "Pop80-84years"
## [29] "Pop85yearsandover"         "Pop6-12years"
## [31] "VisibleMinorityCategory"   "Chinese"
## [33] "SouthAsian"                "Black"
## [35] "Filipino"                  "LatinAmerican"
## [37] "SoutheastAsian"            "Arab"
## [39] "WestAsian"                 "Korean"
## [41] "Japanese"                  "OtherVisibleMinority"
## [43] "MultipleVisibleMinority"   "NotaVisibleMinority"
## [45] "Aboriginal"                "HomeLanguageCategory"
## [47] "Language-Chinese"          "Language-Italian"
## [49] "Language-Korean"           "Language-Persian(Farsi)"
## [51] "Language-Portuguese"       "Language-Russian"
## [53] "Language-Spanish"          "Language-Tagalog"
## [55] "Language-Tamil"            "Language-Urdu"
## [57] "MobilityCategory"          "Non-Movers"
## [59] "Movers"                    "RecentImmigrantsCategory"
## [61] "RecentImmigrants"          "SouthernAsia"
## [63] "SouthEastAsia"             "EasternAsia"
## [65] "WestAsia/MiddleEast"       "Africa"
## [67] "Europe"                    "Caribbean/Central/S.America"
## [69] "LabourForceCategory"       "InLabourForce"
## [71] "Unemployed"                "NotinLabourForce"
## [73] "Lessthangrade9"            "WithCollegeCertificate/Diploma"
## [75] "WithBachelorDegreeorHigher" "SeniorsLivingAlone"
## [77] "TotalTenants"              "HighShelterCosts"
## [79] "OwnedDwellings"            "RentedDwellings"
## [81] "HomeRepairsNeeded"         "TenantAverageRent"
## [83] "LowIncomeFamilies"         "LowIncomeSingles"
## [85] "LowIncomeChildren"         "FamilyIncomeCategory"
## [87] "AverageFamilyIncome"       "HouseholdIncomeCategory"
```

```
##  [89] "Pre-TaxHouseholdIncome"       "After-TaxHouseholdIncome"
##  [91] "AccesstoChildCare"            "BusinessLicensing"
##  [93] "Businesses"                   "ChildCareSpaces"
##  [95] "Inequality(Ginicoeff.)"       "LocalEmployment"
##  [97] "SocialAssistanceRecipients"   "CatholicSchoolGraduation"
##  [99] "CatholicSchoolLiteracy"       "CatholicUniversityApplicants"
## [101] "EarlyDevelopmentInstrument"   "LibraryActivity"
## [103] "LibraryOpenHours"             "LibraryProgramAttendance"
## [105] "LibraryPrograms"              "LibrarySpace"
## [107] "CityGreenRetrofits"           "GreenRebatePrograms"
## [109] "GreenSpaces"                  "PollutingFacilities"
## [111] "TreeCover"                    "BreastCancerScreenings"
## [113] "CervicalCancerScreenings"     "ColorectalCancerScreenings"
## [115] "CommunityFoodPrograms"        "DiabetesPrevalence"
## [117] "DineSafeInspections"          "FemaleFertility"
## [119] "HealthProviders"              "PrematureMortality"
## [121] "StudentNutrition"             "TeenPregnancy"
## [123] "HouseholdsAssisted"           "RentBankApplicants"
## [125] "SocialHousingTurnover"        "SocialHousingUnits"
## [127] "SocialHousingWaitingList"     "DebtRiskScore"
## [129] "AmbulanceCalls"               "AmbulanceReferrals"
## [131] "Arsons"                       "Assaults"
## [133] "BreakEnters"                  "DrugArrests"
## [135] "FireVehicleIncidents"         "FirearmsIncidents"
## [137] "FiresFireAlarms"              "HazardousIncidents"
## [139] "Murders"                      "Robberies"
## [141] "SexualAssaults"               "TCHCSafetyIncidents"
## [143] "Thefts"                       "VehicleThefts"
```

APPENDIX B – Structure of Dataset

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    140 obs. of  144 variables:
##  $ Neighbourhood            : chr  "West Humber-Clairville" "Mount Olive-Silverstone-Jamestown" "Thistletown-Beaumond Heights" "Rexdale-Kipling" ...
##  $ NeighbourhoodId          : chr  "001" "002" "003" "004" ...
##  $ HomePrices               : num  317508 251119 414216 392271 233832 ...
##  $ CityGrantsFunding        : num  520390 10040 158658 49210 42870 ...
##  $ DiversityIndex           : num  4.77 4.97 5.11 5.21 5.5 ...
##  $ VoterTurnout             : num  34.8 28.1 40.5 39.7 34 ...
##  $ WatermainBreaks          : num  30 8 9 5 10 21 5 31 22 19 ...
##  $ TotalArea                : num  30.1 4.6 3.4 2.5 2.9 ...
##  $ TotalPopulation          : num  32265 32130 9925 10725 9440 ...
##  $ Pop-Males                : num  16295 15900 4900 5205 4615 ...
##  $ Pop-Females              : num  15960 16230 5035 5525 4820 ...
##  $ Pop0-4years              : num  2005 2680 615 580 725 ...
##  $ Pop5-9years              : num  2135 2680 625 645 700 ...
##  $ Pop10-14years            : num  2325 2685 645 665 745 ...
##  $ Pop15-19years            : num  2180 2285 630 640 655 ...
##  $ Pop20-24years            : num  2565 2410 655 630 615 ...
##  $ Pop25-29years            : num  2465 2590 650 600 645 ...
##  $ Pop30-34years            : num  2400 2675 650 705 640 ...
##  $ Pop35-39years            : num  2440 2605 730 815 735 ...
##  $ Pop40-44years            : num  2595 2450 790 840 740 ...
##  $ Pop45-49years            : num  2375 2130 735 810 750 ...
##  $ Pop50-54years            : num  1955 1700 590 720 585 ...
##  $ Pop55-59years            : num  1800 1495 520 600 505 ...
##  $ Pop60-64years            : num  1415 1200 420 505 385 ...
##  $ Pop65-69years            : num  1150 910 400 435 310 ...
##  $ Pop70-74years            : num  1015 775 425 460 275 ...
##  $ Pop75-79years            : num  715 500 385 435 205 ...
##  $ Pop80-84years            : num  465 270 255 345 150 555 865 710 415 420 ...
##  $ Pop85yearsandover        : num  305 145 190 285 55 290 500 545 320 235 ...
##  $ Pop6-12years             : num  3126 3782 893 919 1020 ...
##  $ VisibleMinorityCategory  : num  31380 32105 9755 10445 9435 ...
##  $ Chinese                  : num  795 600 75 145 165 475 580 50 315 505 ...
##  $ SouthAsian               : num  12740 12920 2430 1515 965 ...
##  $ Black                    : num  5495 7225 1450 860 2475 ...
##  $ Filipino                 : num  1385 710 125 250 395 ...
##  $ LatinAmerican            : num  1340 1325 475 760 745 ...
##  $ SoutheastAsian           : num  465 625 75 200 185 135 210 50 60 15 ...
##  $ Arab                     : num  250 1180 110 75 140 300 105 40 10 NA ...
##  $ WestAsian                : num  155 1040 280 210 245 565 115 70 30 10 ...
##  $ Korean                   : num  130 65 65 25 75 500 275 75 335 310 ...
##  $ Japanese                 : num  30 75 NA 115 10 40 70 25 70 140 ...
##  $ OtherVisibleMinority     : num  1055 1250 100 100 150 ...
##  $ MultipleVisibleMinority  : num  585 365 60 100 145 185 215 35 30 30 ...
##  $ NotaVisibleMinority      : num  6930 4735 4505 6075 3730 ...
##  $ Aboriginal               : num  65 50 50 65 45 35 65 55 30 15 ...
##  $ HomeLanguageCategory     : num  31385 32100 9760 10440 9435 ...
```

```
## $ Language-Chinese          : num 425 475 25 100 90 220 295 45 200 190 ...
## $ Language-Italian          : num 535 455 360 155 350 725 940 510 305 150 ...
## $ Language-Korean           : num 20 55 40 20 40 425 205 45 195 250 ...
## $ Language-Persian(Farsi)   : num 55 365 130 100 130 355 70 10 NA NA ...
## $ Language-Portuguese       : num 20 70 70 55 40 80 160 190 65 40 ...
## $ Language-Russian          : num 35 45 20 40 50 115 105 200 455 75 ...
## $ Language-Spanish          : num 945 900 325 490 575 555 435 420 255 15 ...
## $ Language-Tagalog          : num 525 335 40 50 270 95 115 NA NA 10 ...
## $ Language-Tamil            : num 525 1275 290 70 80 ...
## $ Language-Urdu             : num 625 905 210 170 295 1350 600 NA NA NA ...
## $ MobilityCategory          : num 29355 29410 9150 9865 8715 ...
## $ Non-Movers                : num 16920 13965 6460 5865 5195 ...
## $ Movers                    : num 12430 15445 2695 4005 3515 ...
## $ RecentImmigrantsCategory  : num 3825 7125 950 865 925 ...
## $ RecentImmigrants          : num 3825 7125 950 860 925 ...
## $ SouthernAsia              : num 2355 4355 525 380 275 ...
## $ SouthEastAsia             : num 190 115 20 65 90 100 140 30 35 40 ...
## $ EasternAsia               : num 70 110 10 NA 40 230 140 10 145 130 ...
## $ WestAsia/MiddleEast       : num 35 890 145 105 75 250 35 130 20 10 ...
## $ Africa                    : num 425 755 130 60 285 500 140 65 20 20 ...
## $ Europe                    : num 105 45 40 100 35 305 420 355 975 160 ...
## $ Caribbean/Central/S.America  : num 580 805 60 125 125 245 200 210 135 10 ...
## $ LabourForceCategory       : num 24895 24070 7905 8560 7265 ...
## $ InLabourForce             : num 16535 15875 4895 5400 4610 ...
## $ Unemployed                : num 1165 1570 310 415 360 ...
## $ NotinLabourForce          : num 8385 8175 3005 3170 2655 ...
## $ Lessthangrade9            : num 1520 1705 500 530 560 ...
## $ WithCollegeCertificate/Diploma: num 3050 2975 785 735 705 ...
## $ WithBachelorDegreeorHigher  : num 370 400 90 55 60 160 190 60 200 205 ...
## $ SeniorsLivingAlone        : num 395 400 265 490 130 685 1130 805 585 385 ...
## $ TotalTenants              : num 2450 4815 1090 1750 1285 ...
## $ HighShelterCosts          : num 2955 3600 1040 1350 1040 ...
## $ OwnedDwellings            : num 6505 4440 2065 2125 1845 ...
## $ RentedDwellings           : num 2460 4820 1085 1750 1290 ...
## $ HomeRepairsNeeded         : num 365 980 185 300 320 845 495 265 400 135 ...
## $ TenantAverageRent         : num 850 875 875 835 895 ...
## $ LowIncomeFamilies         : num 7720 7715 2520 2780 2560 ...
## $ LowIncomeSingles          : num 725 1177 305 653 255 ...
## $ LowIncomeChildren         : num 643 1206 161 135 328 ...
## $ FamilyIncomeCategory      : num 7720 7720 2520 2775 2555 ...
## $ AverageFamilyIncome       : num 67240 52745 71300 65215 56515 ...
## $ HouseholdIncomeCategory   : num 8960 9265 3150 3880 3130 ...
## $ Pre-TaxHouseholdIncome    : num 63415 48145 55030 52430 53780 ...
## $ After-TaxHouseholdIncome  : num 63977 49601 54910 53779 55054 ...
## $ AccesstoChildCare         : num 0.384 0.246 0.325 0.24 0.322 ...
## $ BusinessLicensing         : num 695 106 116 49 31 59 117 49 67 40 ...
## $ Businesses                : num 2550 273 236 155 70 160 182 81 162 77 ...
## $ ChildCareSpaces           : num 180 45 25 60 60 129 131 60 30 311 ...
## $ Inequality(Ginicoeff.)    : num 0.359 0.42 0.382 0.388 0.388 ...
## $ LocalEmployment           : num 63385 3346 1350 1190 831 ...
```

```
## $ SocialAssistanceRecipients   : num  2702 6406 1082 1231 1759 ...
## $ CatholicSchoolGraduation     : num  0.81 0.793 0.73 0.864 0.653 ...
## $ CatholicSchoolLiteracy       : num  68.5 59.2 68.9 74.1 59.2 ...
##   [list output truncated]
```

APPENDIX C – Summary Stats

```
HomePrices     CityGrantsFunding DiversityIndex   VoterTurnout
## Min.   : 204104  Min.   :      0  Min.   :2.887  Min.   :21.05
## 1st Qu.: 374965  1st Qu.:  20349  1st Qu.:4.576  1st Qu.:33.47
## Median : 491210  Median :  95142  Median :4.876  Median :36.14
## Mean   : 548193  Mean   : 296155  Mean   :4.827  Mean   :35.99
## 3rd Qu.: 590216  3rd Qu.: 265704  3rd Qu.:5.085  3rd Qu.:38.46
## Max.   :1849084  Max.   :7312438  Max.   :5.659  Max.   :49.11
## WatermainBreaks   TotalArea     TotalPopulation   Pop-Males
## Min.   : 0.000  Min.   : 0.400  Min.   : 5450  Min.   : 2940
## 1st Qu.: 2.000  1st Qu.: 1.800  1st Qu.:11768  1st Qu.: 5621
## Median : 5.000  Median : 3.300  Median :15345  Median : 7372
## Mean   : 7.907  Mean   : 4.524  Mean   :17868  Mean   : 8603
## 3rd Qu.:12.000  3rd Qu.: 5.400  3rd Qu.:21776  3rd Qu.:10404
## Max.   :32.000  Max.   :37.600  Max.   :45865  Max.   :25555
##  Pop-Females    Pop0-4years     Pop5-9years     Pop10-14years
## Min.   : 2975  Min.   : 180.0  Min.   : 100.0  Min.   : 120.0
## 1st Qu.: 6018  1st Qu.: 555.0  1st Qu.: 577.5  1st Qu.: 623.8
## Median : 8040  Median : 805.0  Median : 777.5  Median : 850.0
## Mean   : 9265  Mean   : 963.7  Mean   : 954.1  Mean   :1007.2
## 3rd Qu.:11246  3rd Qu.:1178.8  3rd Qu.:1143.8  3rd Qu.:1256.2
## Max.   :26905  Max.   :3135.0  Max.   :3235.0  Max.   :3485.0
## Pop15-19years   Pop20-24years Pop25-29years   Pop30-34years
## Min.   : 230.0  Min.   : 325  Min.   : 260.0  Min.   : 165
## 1st Qu.: 647.5  1st Qu.: 775  1st Qu.: 828.8  1st Qu.: 825
## Median : 877.5  Median :1025  Median :1142.5  Median :1248
## Mean   :1042.9  Mean   :1229  Mean   :1357.2  Mean   :1396
## 3rd Qu.:1313.8  3rd Qu.:1508  3rd Qu.:1617.5  3rd Qu.:1682
## Max.   :3345.0  Max.   :3860  Max.   :4780.0  Max.   :4430
## Pop35-39years   Pop40-44years Pop45-49years   Pop50-54years
## Min.   : 315.0  Min.   : 410  Min.   : 345.0  Min.   : 320
## 1st Qu.: 976.2  1st Qu.:1002  1st Qu.: 928.8  1st Qu.: 805
## Median :1305.0  Median :1345  Median :1197.5  Median :1042
## Mean   :1449.2  Mean   :1518  Mean   :1384.6  Mean   :1202
## 3rd Qu.:1825.0  3rd Qu.:1922  3rd Qu.:1726.2  3rd Qu.:1511
## Max.   :3485.0  Max.   :3945  Max.   :3530.0  Max.   :3205
## Pop55-59years   Pop60-64years   Pop65-69years   Pop70-74years
## Min.   : 300.0  Min.   : 215.0  Min.   : 165.0  Min.   : 110.0
## 1st Qu.: 703.8  1st Qu.: 522.5  1st Qu.: 438.8  1st Qu.: 355.0
## Median : 932.5  Median : 682.5  Median : 580.0  Median : 530.0
## Mean   :1057.9  Mean   : 781.6  Mean   : 669.7  Mean   : 608.4
## 3rd Qu.:1327.5  3rd Qu.: 966.2  3rd Qu.: 871.2  3rd Qu.: 798.8
## Max.   :2810.0  Max.   :2145.0  Max.   :2050.0  Max.   :1800.0
## Pop75-79years   Pop80-84years   Pop85yearsandover  Pop6-12years
## Min.   :  95.0  Min.   :  60.0  Min.   :  30.0   Min.   : 224.0
## 1st Qu.: 315.0  1st Qu.: 243.8  1st Qu.: 163.8   1st Qu.: 887.5
## Median : 465.0  Median : 352.5  Median : 275.0   Median :1190.5
## Mean   : 534.9  Mean   : 403.5  Mean   : 307.5   Mean   :1445.7
## 3rd Qu.: 681.2  3rd Qu.: 535.0  3rd Qu.: 405.0   3rd Qu.:1731.5
```

```
## Max.   :1605.0  Max.   :1225.0  Max.   :1160.0   Max.   :5167.0
## VisibleMinorityCategory    Chinese        SouthAsian
## Min.   : 5440          Min.   :   50.0  Min.   :   65
## 1st Qu.:11395          1st Qu.:  423.8  1st Qu.:  390
## Median :15178          Median :  837.5  Median :  945
## Mean   :17638          Mean   : 2015.5  Mean   : 2125
## 3rd Qu.:21496          3rd Qu.: 1703.8  3rd Qu.: 2715
## Max.   :52310          Max.   :16790.0  Max.   :17920
##     Black           Filipino       LatinAmerican    SoutheastAsian
## Min.   :  10.0   Min.   :  25.0   Min.   :   0.0   Min.   :   0.0
## 1st Qu.: 458.8   1st Qu.: 198.8   1st Qu.: 140.0   1st Qu.:  60.0
## Median : 922.5   Median : 447.5   Median : 290.0   Median : 145.0
## Mean   :1479.1   Mean   : 729.8   Mean   : 460.0   Mean   : 265.1
## 3rd Qu.:1843.8   3rd Qu.:1001.2   3rd Qu.: 571.2   3rd Qu.: 278.8
## Max.   :8730.0   Max.   :4255.0   Max.   :3475.0   Max.   :3350.0
##      Arab           WestAsian          Korean           Japanese
## Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Min.   :  0.00
## 1st Qu.:  35.0   1st Qu.:  55.0   1st Qu.:  55.0   1st Qu.: 38.75
## Median :  75.0   Median : 127.5   Median : 107.5   Median : 70.00
## Mean   : 158.3   Mean   : 304.0   Mean   : 242.5   Mean   : 83.57
## 3rd Qu.: 190.0   3rd Qu.: 405.0   3rd Qu.: 217.5   3rd Qu.:115.00
## Max.   :1180.0   Max.   :3395.0   Max.   :4265.0   Max.   :300.00
## OtherVisibleMinority MultipleVisibleMinority NotaVisibleMinority
## Min.   :   0.00      Min.   :  10.0          Min.   : 1580
## 1st Qu.:  48.75      1st Qu.:  80.0          1st Qu.: 6045
## Median :  95.00      Median : 150.0          Median : 8458
## Mean   : 179.25      Mean   : 220.1          Mean   : 9361
## 3rd Qu.: 185.00      3rd Qu.: 283.8          3rd Qu.:11455
## Max.   :1485.00      Max.   :1210.0          Max.   :22250
##   Aboriginal      HomeLanguageCategory Language-Chinese  Language-Italian
## Min.   :  0.00   Min.   : 5440        Min.   :   10.0   Min.   :   0.0
## 1st Qu.: 40.00   1st Qu.:11391        1st Qu.:  197.5   1st Qu.:  35.0
## Median : 75.00   Median :15180        Median :  465.0   Median :  97.5
## Mean   : 94.57   Mean   :17638        Mean   : 1402.7   Mean   : 315.5
## 3rd Qu.:131.25   3rd Qu.:21498        3rd Qu.: 1116.2   3rd Qu.: 372.5
## Max.   :450.00   Max.   :52310        Max.   :13750.0   Max.   :3715.0
## Language-Korean  Language-Persian(Farsi) Language-Portuguese
## Min.   :   0.0   Min.   :   0.0          Min.   :   0.00
## 1st Qu.:  25.0   1st Qu.:  15.0          1st Qu.:  18.75
## Median :  55.0   Median :  62.5          Median :  47.50
## Mean   : 168.2   Mean   : 195.6          Mean   : 268.21
## 3rd Qu.: 160.0   3rd Qu.: 265.0          3rd Qu.: 168.75
## Max.   :3230.0   Max.   :2745.0          Max.   :5645.00
## Language-Russian Language-Spanish Language-Tagalog Language-Tamil
## Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
## 1st Qu.:  20.0   1st Qu.:  75.0   1st Qu.:  50.0   1st Qu.:   0.0
## Median :  62.5   Median : 180.0   Median : 117.5   Median :  45.0
## Mean   : 199.6   Mean   : 311.5   Mean   : 241.0   Mean   : 360.0
## 3rd Qu.: 150.0   3rd Qu.: 415.0   3rd Qu.: 331.2   3rd Qu.: 301.2
## Max.   :7070.0   Max.   :2725.0   Max.   :1440.0   Max.   :5425.0
```

```
##  Language-Urdu   MobilityCategory   Non-Movers       Movers
##  Min.   :   0.0   Min.   : 5265   Min.   : 2520   Min.   : 2260
##  1st Qu.:   7.5   1st Qu.:10820   1st Qu.: 6131   1st Qu.: 4520
##  Median :  50.0   Median :14288   Median : 7902   Median : 6438
##  Mean   : 219.5   Mean   :16670   Mean   : 9147   Mean   : 7522
##  3rd Qu.: 210.0   3rd Qu.:20329   3rd Qu.:11220   3rd Qu.: 9339
##  Max.   :3865.0   Max.   :48645   Max.   :24775   Max.   :25825
##  RecentImmigrantsCategory RecentImmigrants  SouthernAsia
##  Min.   : 100             Min.   :   0     Min.   :   0.0
##  1st Qu.: 630             1st Qu.: 635     1st Qu.:  45.0
##  Median :1532             Median :1532     Median : 147.5
##  Mean   :1906             Mean   :1910     Mean   : 494.8
##  3rd Qu.:2431             3rd Qu.:2428     3rd Qu.: 587.5
##  Max.   :9135             Max.   :9140     Max.   :5970.0
##  SouthEastAsia    EasternAsia     WestAsia/MiddleEast     Africa
##  Min.   :  0.0   Min.   :   0.00   Min.   :   0.0     Min.   :  0.0
##  1st Qu.: 40.0   1st Qu.:  58.75   1st Qu.:  25.0     1st Qu.: 30.0
##  Median :117.5   Median : 132.50   Median :  92.5     Median : 72.5
##  Mean   :181.4   Mean   : 422.61   Mean   : 203.0     Mean   :115.4
##  3rd Qu.:265.0   3rd Qu.: 420.00   3rd Qu.: 231.2     3rd Qu.:146.2
##  Max.   :845.0   Max.   :4660.00   Max.   :1830.0     Max.   :755.0
##      Europe       Caribbean/Central/S.America LabourForceCategory
##  Min.   :  10.0   Min.   :   0.0             Min.   : 5010
##  1st Qu.:  75.0   1st Qu.:  60.0             1st Qu.: 9848
##  Median : 145.0   Median : 135.0             Median :12658
##  Mean   : 257.9   Mean   : 185.9             Mean   :14720
##  3rd Qu.: 307.5   3rd Qu.: 225.0             3rd Qu.:17919
##  Max.   :2885.0   Max.   :1060.0             Max.   :41560
##  InLabourForce     Unemployed     NotinLabourForce Lessthangrade9
##  Min.   : 2995   Min.   : 155.0   Min.   : 1420   Min.   : 105.0
##  1st Qu.: 6632   1st Qu.: 438.8   1st Qu.: 3242   1st Qu.: 467.5
##  Median : 8360   Median : 617.5   Median : 4582   Median : 660.0
##  Mean   : 9578   Mean   : 728.2   Mean   : 5141   Mean   : 773.5
##  3rd Qu.:11685   3rd Qu.: 963.8   3rd Qu.: 6535   3rd Qu.: 945.0
##  Max.   :25160   Max.   :2390.0   Max.   :16410   Max.   :2565.0
##  WithCollegeCertificate/Diploma WithBachelorDegreeorHigher
##  Min.   : 440.0                Min.   :  25.0
##  1st Qu.: 893.8                1st Qu.: 100.0
##  Median :1222.5                Median : 180.0
##  Mean   :1480.2                Mean   : 233.4
##  3rd Qu.:1790.0                3rd Qu.: 305.0
##  Max.   :4920.0                Max.   :1300.0
##  SeniorsLivingAlone  TotalTenants   HighShelterCosts OwnedDwellings
##  Min.   : 115.0     Min.   :  130   Min.   : 300     Min.   :  300
##  1st Qu.: 385.0     1st Qu.: 1674   1st Qu.:1540     1st Qu.: 2416
##  Median : 535.0     Median : 2820   Median :2270     Median : 3325
##  Mean   : 637.7     Mean   : 3156   Mean   :2511     Mean   : 3800
##  3rd Qu.: 826.2     3rd Qu.: 3958   3rd Qu.:3190     3rd Qu.: 4924
##  Max.   :1810.0     Max.   :11900   Max.   :7705     Max.   :11745
##  RentedDwellings HomeRepairsNeeded TenantAverageRent LowIncomeFamilies
```

```
##  Min.   : 135   Min.   :  95.0   Min.   : 550.0   Min.   : 1205
##  1st Qu.: 1708   1st Qu.: 330.0   1st Qu.: 835.0   1st Qu.: 3082
##  Median : 2872   Median : 470.0   Median : 897.5   Median : 4002
##  Mean   : 3169   Mean   : 540.6   Mean   : 935.1   Mean   : 4645
##  3rd Qu.: 3958   3rd Qu.: 735.0   3rd Qu.:1030.0   3rd Qu.: 5761
##  Max.   :11900   Max.   :1465.0   Max.   :1405.0   Max.   :13860
##  LowIncomeSingles LowIncomeChildren FamilyIncomeCategory
##  Min.   : 108     Min.   :   0.0    Min.   :  975
##  1st Qu.: 656     1st Qu.: 134.5    1st Qu.: 3029
##  Median : 997     Median : 213.5    Median : 4005
##  Mean   :1144     Mean   : 339.1    Mean   : 4627
##  3rd Qu.:1398     3rd Qu.: 485.8    3rd Qu.: 5766
##  Max.   :4602     Max.   :1647.0    Max.   :13850
##  AverageFamilyIncome HouseholdIncomeCategory Pre-TaxHouseholdIncome
##  Min.   : 34825      Min.   : 2105           Min.   : 24775
##  1st Qu.: 58014      1st Qu.: 4615           1st Qu.: 46676
##  Median : 66728      Median : 6070           Median : 53663
##  Mean   : 80818      Mean   : 6928           Mean   : 58245
##  3rd Qu.: 82221      3rd Qu.: 8752           3rd Qu.: 62788
##  Max.   :423850      Max.   :18070           Max.   :208310
##  After-TaxHouseholdIncome AccesstoChildCare BusinessLicensing
##  Min.   : 25562           Min.   : 0.0800   Min.   :  30.00
##  1st Qu.: 48700           1st Qu.: 0.2431   1st Qu.:  93.75
##  Median : 55303           Median : 0.2930   Median : 135.00
##  Mean   : 59975           Mean   : 0.6024   Mean   : 189.62
##  3rd Qu.: 64681           3rd Qu.: 0.3487   3rd Qu.: 222.75
##  Max.   :211492           Max.   :41.0000   Max.   :1163.00
##    Businesses     ChildCareSpaces Inequality(Ginicoeff.) LocalEmployment
##  Min.   :  48.0   Min.   :  0.0   Min.   :  0.1956       Min.   :   300
##  1st Qu.: 171.2   1st Qu.: 55.0   1st Qu.:  0.3747       1st Qu.:  2050
##  Median : 343.0   Median : 93.5   Median :  0.3976       Median :  3828
##  Mean   : 534.8   Mean   :111.1   Mean   :  6.0443       Mean   :  9349
##  3rd Qu.: 593.8   3rd Qu.:159.5   3rd Qu.:  0.4145       3rd Qu.: 10067
##  Max.   :4194.0   Max.   :400.0   Max.   :792.0000       Max.   :175515
##  SocialAssistanceRecipients CatholicSchoolGraduation
##  Min.   :   0.0             Min.   :0.0000
##  1st Qu.: 656.5             1st Qu.:0.7835
##  Median :1312.0             Median :0.8283
##  Mean   :1765.9             Mean   :0.8140
##  3rd Qu.:2528.0             3rd Qu.:0.8753
##  Max.   :6786.0             Max.   :1.0000
##  CatholicSchoolLiteracy CatholicUniversityApplicants
##  Min.   : 40.00         Min.   : 0.00
##  1st Qu.: 68.50         1st Qu.:32.51
##  Median : 78.35         Median :39.21
##  Mean   : 77.01         Mean   :41.89
##  3rd Qu.: 85.71         3rd Qu.:52.08
##  Max.   :100.00         Max.   :80.00
##  EarlyDevelopmentInstrument LibraryActivity LibraryOpenHours
##  Min.   : 0.00              Min.   :1.000   Min.   : 3.000
```

```
##  1st Qu.: 50.00          1st Qu.:2.000   1st Qu.: 4.000
##  Median :100.00          Median :3.000   Median : 7.000
##  Mean   : 77.86          Mean   :3.271   Mean   : 6.571
##  3rd Qu.:100.00          3rd Qu.:5.000   3rd Qu.: 9.000
##  Max.   :100.00          Max.   :9.000   Max.   :10.000
##  LibraryProgramAttendance LibraryPrograms  LibrarySpace
##  Min.   :1.000         Min.   :1.000   Min.   :2.000
##  1st Qu.:2.000         1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000         Median :4.000   Median :5.000
##  Mean   :3.771         Mean   :4.136   Mean   :4.636
##  3rd Qu.:5.000         3rd Qu.:6.000   3rd Qu.:7.000
##  Max.   :9.000         Max.   :9.000   Max.   :9.000
##  CityGreenRetrofits GreenRebatePrograms  GreenSpaces
##  Min.   :0.000     Min.   : 19.0      Min.   : 0.00181
##  1st Qu.:1.000     1st Qu.:103.0      1st Qu.: 0.12482
##  Median :2.000     Median :153.0      Median : 0.26095
##  Mean   :2.329     Mean   :188.3      Mean   : 0.57499
##  3rd Qu.:3.000     3rd Qu.:247.0      3rd Qu.: 0.67761
##  Max.   :8.000     Max.   :763.0      Max.   :14.25833
##  PollutingFacilities   TreeCover       BreastCancerScreenings
##  Min.   : 0.000     Min.   :   61616   Min.   :47.50
##  1st Qu.: 0.000     1st Qu.:  523162   1st Qu.:56.52
##  Median : 1.000     Median : 1017744   Median :60.20
##  Mean   : 3.507     Mean   : 1281438   Mean   :60.06
##  3rd Qu.: 3.000     3rd Qu.: 1686216   3rd Qu.:63.15
##  Max.   :81.000     Max.   :12888044   Max.   :72.50
##  CervicalCancerScreenings ColorectalCancerScreenings CommunityFoodPrograms
##  Min.   :51.80         Min.   :28.30           Min.   : 0.000
##  1st Qu.:62.48         1st Qu.:37.08           1st Qu.: 1.000
##  Median :64.40         Median :38.65           Median : 2.000
##  Mean   :65.27         Mean   :39.14           Mean   : 2.929
##  3rd Qu.:68.62         3rd Qu.:41.30           3rd Qu.: 4.000
##  Max.   :77.60         Max.   :53.60           Max.   :18.000
##  DiabetesPrevalence DineSafeInspections FemaleFertility HealthProviders
##  Min.   : 4.60     Min.   :  0.000     Min.   :22.69   Min.   :  1.00
##  1st Qu.: 8.30     1st Qu.:  0.000     1st Qu.:35.87   1st Qu.:  9.75
##  Median :10.20     Median :  3.000     Median :45.06   Median : 23.00
##  Mean   :10.34     Mean   :  8.957     Mean   :44.70   Mean   : 34.85
##  3rd Qu.:12.53     3rd Qu.:  8.000     3rd Qu.:51.88   3rd Qu.: 50.50
##  Max.   :16.80     Max.   :127.000     Max.   :77.77   Max.   :192.00
##  PrematureMortality StudentNutrition  TeenPregnancy   HouseholdsAssisted
##  Min.   :129.3     Min.   :   0.00   Min.   : 0.00   Min.   :  0.000
##  1st Qu.:188.7     1st Qu.:  18.75   1st Qu.:16.00   1st Qu.:  0.000
##  Median :224.5     Median : 400.00   Median :25.40   Median :  1.000
##  Mean   :233.9     Mean   : 868.21   Mean   :27.47   Mean   :  8.743
##  3rd Qu.:266.2     3rd Qu.:1264.75   3rd Qu.:37.62   3rd Qu.:  3.000
##  Max.   :572.1     Max.   :6172.00   Max.   :77.33   Max.   :280.000
##  RentBankApplicants SocialHousingTurnover SocialHousingUnits
##  Min.   : 0.00     Min.   : 0.000       Min.   :   0.0
##  1st Qu.: 5.00     1st Qu.: 0.000       1st Qu.: 166.8
```

```
## Median :10.00     Median : 1.232      Median : 462.5
## Mean   :15.22     Mean   : 2.665      Mean   : 654.3
## 3rd Qu.:21.25     3rd Qu.: 3.560      3rd Qu.: 898.8
## Max.   :74.00     Max.   :18.000      Max.   :3702.0
## SocialHousingWaitingList DebtRiskScore  AmbulanceCalls
## Min.   :  11.0        Min.   :661.0  Min.   : 385.0
## 1st Qu.: 151.0        1st Qu.:720.5  1st Qu.: 854.8
## Median : 273.0        Median :741.0  Median :1245.0
## Mean   : 358.2        Mean   :739.2  Mean   :1536.2
## 3rd Qu.: 485.2        3rd Qu.:759.0  3rd Qu.:1932.5
## Max.   :1331.0        Max.   :793.0  Max.   :5733.0
## AmbulanceReferrals    Arsons       Assaults      BreakEnters
## Min.   : 0.000    Min.   :0.000  Min.   : 14.00  Min.   : 16.00
## 1st Qu.: 4.000    1st Qu.:0.000  1st Qu.: 55.75  1st Qu.: 32.75
## Median : 5.500    Median :1.000  Median :104.00  Median : 54.50
## Mean   : 6.843    Mean   :1.479  Mean   :121.29  Mean   : 64.76
## 3rd Qu.: 9.250    3rd Qu.:2.000  3rd Qu.:145.25  3rd Qu.: 83.00
## Max.   :29.000    Max.   :9.000  Max.   :609.00  Max.   :193.00
##  DrugArrests    FireVehicleIncidents FirearmsIncidents FiresFireAlarms
## Min.   :  0.00  Min.   :  4.00      Min.   : 0.000    Min.   :  7.00
## 1st Qu.: 18.00  1st Qu.: 37.00      1st Qu.: 0.000    1st Qu.: 38.00
## Median : 43.00  Median : 60.00      Median : 1.000    Median : 53.00
## Mean   : 64.35  Mean   : 89.44      Mean   : 2.364    Mean   : 60.39
## 3rd Qu.: 82.50  3rd Qu.:106.50      3rd Qu.: 4.000    3rd Qu.: 73.50
## Max.   :497.00  Max.   :674.00      Max.   :21.000    Max.   :150.00
## HazardousIncidents    Murders        Robberies      SexualAssaults
## Min.   : 35.00    Min.   :0.0000  Min.   :  3.00  Min.   : 0.00
## 1st Qu.: 68.75    1st Qu.:0.0000  1st Qu.: 13.00  1st Qu.: 4.00
## Median :100.00    Median :0.0000  Median : 23.00  Median : 8.00
## Mean   :118.21    Mean   :0.4571  Mean   : 28.65  Mean   :10.02
## 3rd Qu.:151.00    3rd Qu.:1.0000  3rd Qu.: 35.00  3rd Qu.:13.00
## Max.   :381.00    Max.   :4.0000  Max.   :126.00  Max.   :41.00
## TCHCSafetyIncidents    Thefts       VehicleThefts
## Min.   :  0.0      Min.   : 0.000  Min.   :  6.00
## 1st Qu.:  0.0      1st Qu.: 2.750  1st Qu.: 20.00
## Median : 60.5      Median : 4.000  Median : 32.50
## Mean   :111.6      Mean   : 6.657  Mean   : 45.32
## 3rd Qu.:150.0      3rd Qu.: 8.000  3rd Qu.: 55.00
## Max.   :721.0      Max.   :49.000  Max.   :341.00
```

APPENDIX D – Code in R Markdown File Format

R MARKDOWN OUTPUT

DS8004ProjectCODEi
nRMarkdownFormat.h

R MARKDOWN CODE

DS8004ProjectCODEi
nRMarkdownFormat.l