November 8, 2021

supervisor: Ceni babAoglu, phd

Ryerson University

Predicting rental rate growth in the canadian market

Rommel Agustin 500884909

Table of Contents

[RECAP OF OBJECTIVE 3](#_Toc87191623)

[Research Question 3](#_Toc87191624)

[DATA ANALYSES AND PREPARATION 4](#_Toc87191625)

[*Dataset 1* 4](#_Toc87191626)

[Initial Review 4](#_Toc87191627)

[Analyses and Preparation 5](#_Toc87191628)

[*Dataset 2* 8](#_Toc87191629)

[Initial Review 8](#_Toc87191630)

[Analyses and Preparation 9](#_Toc87191631)

[*Dataset 3* 12](#_Toc87191632)

[Initial Review 13](#_Toc87191633)

[Analyses and Preparation 13](#_Toc87191634)

[*Dataset 4* 16](#_Toc87191635)

[Initial Review 16](#_Toc87191636)

[Analyses and Preparation 17](#_Toc87191637)

[*4 Datasets Combined* 18](#_Toc87191638)

[Initial Review 18](#_Toc87191639)

[PREDICTIVE MODELING 23](#_Toc87191640)

[Combined Dataset 23](#_Toc87191641)

# INTRODUCTION

## Research Question

The focus of the research is on understanding better how to build an accurate model to predict rental rate change in a specific Canadian market and a specific rental segment.

## Github Link

You can also find this document at Github together with the raw datasets (except for Population 2016 which has 30MB of data over the 25MB allowed by Github) and the overview of the stages and processes for this capstone project.

<https://github.com/rommelagustin/CIND820.git>

# DATA ANALYSES AND PREPARATION

*Started with 4 datasets with the following number of observations and variables* Graphical user interface, text, application, email

Description automatically generated

## *Dataset 1*

Income.csv

*Median household total income and after-tax income by household type (total – household type including census family structure), Canada and census metropolitan areas, 2016 Census – 100% Data*

## Initial Review

Table

Description automatically generated

A picture containing text

Description automatically generated

## Analyses and Preparation

A screenshot of a computer

Description automatically generated with medium confidence

Useful attributes/columns:

* Geographic name
* Household type
* Median household total income, 2005
* Median household total income, 2015

Data types on above columns are appropriate

Null (empty) cells represent data that are not applicable or not available for a specific reference period.

The last 8 variables to the right of the dataset are numeric.

The “Household type” variable which is a character data type contains subsets such as:

* *Total Household Type*
* *Census-family households*
* *Households consisting of only on census family*
* *One couple, with or without children*
* *One couple, with children*
* *One lone-parent*
* *Other census family households*
* *Non-census family households*

This complicates our analysis since we need only one household type which is the “Census-family households” to line up with the numeric data on number of households and household total income.

Before we can conduct our analyses, we need to transform the dataset to a subset wherein said “Census-family households” is the only Household Type attribute to line up with the other identified useful attributes/columns mentioned above.

After transforming raw data to the relevant columns and picking up only the relevant attributes of Geographic name (changed to “Cities”), Median household total income (changed to “HH Income”) for 2005 and Median household total income (changed to “HH Income”) for 2015. We get the following 3 attributes in the revised dataset “**Income\_Final**” with 160 observations and 3 variables from 1,440 observations and 17 variables :

Table

Description automatically generated

This is where we can start doing analyses of the dataset:

A screenshot of a computer

Description automatically generated with medium confidence

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

2005 Boxplot

Chart, box and whisker chart

Description automatically generated

2015 Boxplot

Chart, box and whisker chart

Description automatically generated

We can conclude that the data is normal and outliers are minimal. There are no missing values.

## *Dataset 2*

Pop\_2016Census.csv

*Population - Census Profile - Age, Sex, Type of Dwelling, Families, Households, Marital Status, Language, Income, Immigration and Ethnocultural Diversity, Housing, Aboriginal Peoples, Education, Labour, Journey to Work, Mobility and Migration, and Language of Work for Census Metropolitan Areas and Census Agglomerations, 2016 Census / Catalogue number: 98-401-X2016041 (Statistics Canada)*

## Initial Review

Table

Description automatically generated with low confidence

Text

Description automatically generated

## Analyses and Preparation

A screenshot of a computer

Description automatically generated with medium confidence

Useful attributes/columns:

* Geo\_Name
* DIM: Profile of Census Metropolitan Areas / Census Agglomerations (2247)
  + Needs further analysis to determine which group to limit search with
* DIM Sex (3): Member ID [1]: Total - Sex
* DIM Sex (3): Member ID [2]: Total - Male
* DIM Sex (3): Member ID [3]: Total - Female

Data types on above columns are appropriate

The last 3 variables to the right of the dataset are numeric.

The “DIM: Profile of Census Metropolitan Areas / Census Agglomerations (2247)” variable which is a character data type contains subsets, among others the following:

* *Population, 2016*
* *Population, 2011*
* *Population percentage change, 2011 to 2016*
* *Total private dwellings*
* *Private dwellings occupied by usual residents*
* *Population density per square kilometre*
* *Land area in square kilometres*
* *Total - Age groups and average age of the population - 100% data*
* *There are* ***2,239*** *other lines for EACH of 160 Geographic Name (City)*

We only need the 2 population metrics:

* *Population, 2016*
* *Population, 2011*

We need to pull the 2 rows out from its current attribute to become 2 separate attributes/columns and delete all remaining subsets under “DIM: Profile of Census Metropolitan Areas / Census Agglomerations (2247)” variable.

The values shown under attribute “DIM Sex (3): Member ID [1]: Total - Sex“ should line up with these planned separate columns of Population, 2016 and Population, 2011.

After transforming raw data to the relevant columns and picking up only the relevant attributes of Geographic name (changed to “Cities”), Population, 2011 and Population, 2016. We get the following 3 attributes in the revised dataset “**Pop2016Final**” with 158 observations and 3 variables from 353,914 observations and 14 variables:

Graphical user interface, table

Description automatically generated with medium confidence

This is where we can start doing analyses of the dataset:

A screenshot of a computer

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generated

The 2 numerical data attributes are consistent with each other. Normal distribution and there are no missing values.

## *Dataset 3*

Population.csv

*Population - Census Profile - Age, Sex, Type of Dwelling, Families, Households, Marital Status, Language, Income, Immigration and Ethnocultural Diversity, Housing, Aboriginal Peoples, Education, Labour, Journey to Work, Mobility and Migration, and Language of Work for Census Metropolitan Areas and Census Agglomerations, 2011 Census / Catalogue number: 98-316-XWE (Statistics Canada)*

## Initial Review

A picture containing text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## Analyses and Preparation

A screenshot of a computer

Description automatically generated with medium confidence

Useful attributes/columns:

* CMACA\_Name
* Characteristics
* Total
* Male
* Female

Data types on above columns are appropriate

The last 3 variables to the right of the dataset are numeric.

The “Characteristics” variable which is a character data type contains subsets, among others the following:

* *Population, 2006*
* *Population, 2011*
* *Population percentage change, 2006 to 2011*
* *Total private dwellings*
* *Private dwellings occupied by usual residents*
* *Population density per square kilometre*
* *Land area in square kilometres*
* *Total - Age groups and average age of the population - 100% data*
* *There are 2,239 other lines for EACH of 160 Geographic Name (City)*

We only need the 2 population metrics:

* *Population, 2006*
* *Population, 2011*

We need to pull it out from its current attribute under “Characteristics” to 2 separate attributes/columns and delete all remaining subsets under “Characteristics” variable.

The values shown under attribute “Total “ shall line up with these planned separate columns of Population, 2006 and Population, 2011.

After transforming raw data to the relevant columns and picking up only the relevant attributes of CMACA\_Name (changed to “Cities”), Population, 2011 and Population, 2016. We get the following 3 attributes in the revised dataset “**Pop06\_11Final**” with 155 observations and 3 variables from 73,160 observations and 14 variables:

Graphical user interface, application

Description automatically generated

This is where we can start doing analyses of the dataset:

A screenshot of a computer

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

The 2 numerical data attributes are consistent with each other. Normal distribution and there are no missing values.

## *Dataset 4*

Rent Growth.csv

*Commercial Rent -* [*https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810025501*](https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810025501) *Commercial rents services price index, monthly. Statistics Canada Table: 1810025501-eng*

*(Statistics Canada)*

## Initial Review

Diagram

Description automatically generated with low confidence

Text

Description automatically generated

## Analyses and Preparation

A screenshot of a computer

Description automatically generated with medium confidence

Useful attributes/columns:

* REF\_DATE
* GEO
* Building Type
* VALUE

Data types on above columns are appropriate

The “VALUE” variable is numeric. “GEO” contains a combination of individual Cities, Provinces and total Canada. While “Building Type” variable contains Total Building Type, Office, Retail and Industrial. “REF\_DATE” covers monthly data from Jan 2006 to Jun 2021, However, Jan 2006 to Jun 2021 is for “Canada” only. Data for individual cities and provinces have data from Jan 2019 to Jun 2021 only.

We require data from individual cities therefore our period will be limited to Jan 2019 to Jun 2021 only which will be 2019, 2020 and 2021 (3 years only) for our purpose.

After transforming raw data to the relevant columns and picking up only the relevant attributes of REF\_DATE, Cities, Total, Building type (from “Building type”), and VALUE. We get the following 4 attributes in the revised dataset “**Rent\_Subset**” with 750 observations and 4 variables from 1,956 observations and 15 variables

Table

Description automatically generated

We need to further transform data to pick up only Dec 2020 in the REF\_DATE , Cities from GEO and Total, building type from the Building Type variable.

In the GEO variable, we need to pick up only individual Cities and not the provinces and total Canada. However, after picking only Cities, we were left only with 13 Cities. Comparing this with the earlier 3 datasets (i.e., Pop2016Final, Pop06\_11Final and Income\_Final) which has 160 Cities. Although in the 13 Cities, it includes the 5 key cities we have identified that we wanted to assess in this study.

## *4 Datasets Combined*

PopCombined.csv

Using the datasets, Pop2016Final, Pop06\_11Final,Income\_Final and Rent\_Final, we combined each file to PopCombined. Only Rent\_Final had lesser rows due to limited number of Cities and those with missing data compared to Population and Income were filled with “0”.

On the individual datasets, we segregated into separate columns Population 2016, Population 2006, Household Income 2015, Household Income 2005, Rent Growth 2020 and Rent Growth 2019. On the combined dataset, we will have one column for Population, one for Household Income and one for Rent Growth on a by city basis. This facilitates requirements for modeling and validation. The resulting re-formatting is under **PopCombined2** dataset.

## Initial Review

Graphical user interface, table

Description automatically generated with medium confidence

A picture containing diagram

Description automatically generated

. Text

Description automatically generated with medium confidence

We were able to transfer and combine the datasets we needed. The characteristics and data types we validated earlier were kept intact.

Correlation Test: Pop and Price; HHIncome and Price

Graphical user interface, text, application, email

Description automatically generated

Based on the results of the Pearson Correlation Test above, both population and household income have a linear relationship with rent but not a strong relationship.

Graphical user interface, application

Description automatically generated

Text

Description automatically generated

A picture containing graphical user interface

Description automatically generated

Correlation Test: Pop and Price; HHIncome and Price

Text

Description automatically generated

Text, letter

Description automatically generated

Based on the results of the Pearson Correlation Test above, both population and household income have a linear relationship with rent but not a strong relationship.

# PREDICTIVE MODELING

## Combined Dataset

We will do a *Cross-Validation* to make sure that every item in the original dataset has the same chance of appearing in the training and test set for our modeling.

Library(“caret”)

Folds<-createFolds(PopCombined2$Cities) for (f in folds) { train<-PopCombined2[-f,] test<-PopCombined2[f,] }

Our modeling will be more of supervised learning method being that the dataset is labeled. We will use classification algorithm specifically *k-Nearest Neighbours (k-NN) and Decision Tree*.

**k-Nearest Neighbours**

In doing the testing, we will use a 70 : 30 training and test sets split. We will measure performance via Confusion Matrix.

train\_index<-sample(1:nrow(PopCombined2), 0.7 \* nrow(PopCombined2)

train.set<-PopCombined2[train\_index]

test.set<-PopCombined2[-train\_index]

train.set\_new<-train.set[-1]

test.set\_new<-test.set[-5]

PopCombined2\_train\_labels<-train.set$Cities

PopCombined2\_test\_labels<-test.set$Cities

*Prediction*

PopCombined2\_knn\_prediction<-knn(train = train.set\_new, test = test.set\_new, cl = PopCombined2\_train\_labels, k = 3)

Interpretation of Results

*Confusion Matrix*

CrossTable(x = PopCombined2\_test\_labels, y = PopCombined2\_knn\_prediction, prop.chisq = FALSE)

Interpretation of Results