<Predicting Commercial Rent Growth in Canada>

<Rommel Agustin 500884909>

Table of Contents

[About This Research 3](#_Toc89202692)

[Other Past Research 3](#_Toc89202693)

[Methodology 4](#_Toc89202694)

[Data source and methods 4](#_Toc89202695)

[Extraction, Transformation and Loading 5](#_Toc89202696)

[Population 2016 and 2011 dataset 5](#_Toc89202697)

[Population 2011 and 2006 dataset 7](#_Toc89202698)

[Household income 2015 and 2005 dataset 8](#_Toc89202699)

[Commercial rents dataset 10](#_Toc89202700)

[Outliers 10](#_Toc89202701)

[Data Integration 12](#_Toc89202702)

[Analyses 13](#_Toc89202703)

[k-Nearest Neighbours 13](#_Toc89202704)

[Logistic Regression 14](#_Toc89202705)

[Model Evaluation 17](#_Toc89202706)

[k-Nearest Neighbours Prediction 17](#_Toc89202707)

[Logistic Regression Using the Test Set 17](#_Toc89202708)

[Logistic Regression Using Unsplit Dataset (greater # of observations than test) 18](#_Toc89202709)

[Findings and Results 19](#_Toc89202710)

[Limitations 19](#_Toc89202711)

[Next Steps 19](#_Toc89202712)

[Conclusion 19](#_Toc89202713)

[References 21](#_Toc89202714)

# About This Research

Back in 2017, when Whole Foods was acquired by Amazon, alarm bells were sounding off. Concerns of disruption in the retail landscape not just for grocery with the adoption of online shopping as Amazon did with department stores and other brick and mortar operations. A retail apocalypse scenario was being painted since then and investors were wary of investing in companies that have this exposure such as mall operations. The question of the need for physical space and the reasonability of current market rents have been revived again with this current pandemic. The Canadian commercial retail real estate industry is at a crossroad wherein the physical space for stores is starting to be challenged by the e-commerce alternative. With several lockdowns and limitations of access to space, a lot of businesses have been closed for months or had limited transactions (i.e., online order for curbside pick-up, delivery) especially for non-essential businesses. Tenants have accelerated their e-commerce businesses and their online shopping presence by establishing websites and doing tie ups with delivery companies (i.e., Instacart, Door Dash, Uber). Commercial retail landlords were also impacted by lockdowns and business slow down that resulted in unpaid rents or worse, permanent closures of some of its tenants’ businesses.

Our aim is to determine if population and household income growth impacts retail rent growth. Knowing this will help identify regions that might be poised for growth and may help maximize retail rent revenues at locations of existing commercial properties. We will be able to identify other locations which may have growth rate potential where there are development opportunities.

# Other Past Research

There was various past research on predicting rental price but very few on commercial rents and more on housing market rent. There was also past research on rental rate growth but not in relation to household income and population as main drivers. Our research intends to shed light on the direct impact of rent growth by population and household income growth.

Wheeler1 fine tuned use of traditional linear regression model in predicting house prices with the Bayesian spatially varying coefficient models and geographically weighted linear regression for estimation of hedonic prices.

The study of Des Rosiers, Theriault and Lavoie2 was on extent of impact on rent levels of retail concentration at regional and super-regional shopping centres as well as categories of goods in 2 Canadian cities, Montreal, and Quebec City. Among its findings are attributes to consider that have higher relationship to rent change, among which are:

* Percentage rent acts as a complement than a substitute for base rent
* Store size by far the most prominent determinant of shopping centre rents. A 10% increase in GLA negatively impact rent with a 4% rent discount
* Age of shopping centre impacts rents negatively stressing importance of renovation and enhancement strategies for the centre
* Corroboration of the study that landlords can charge higher rents to lease duration/long established tenants
* Higher agglomeration economies do not automatically translate to higher base rents. Other factors impact it
* Retail categories capture agglomeration economies where the higher retail category order the higher the base rent
* Base rent of various goods and services categories is negatively affected by a higher degree of retail concentration. However, not all categories are affected similarly by retail concentration

Zhang3 concluded that tenant mix and retail rent have a positive relationship. The study confirms the qualitative side of the relationship touched on by other research wherein shopping districts with high tenant mix can reduce consumer’s time and transportation costs by providing them with a wider variety of goods and services thus enabling them to complete their shopping with one single trip. High foot traffic is high turnover due to more consumers. With this, tenants will be willing to pay higher rents.

The study of Hiebert4 considered demographics but pertains to the housing markets both owned residence and apartment rentals for Montreal, Toronto, and Vancouver not commercial rents. It focused on immigrants and refugees if they were a factor in assessing growth estimation and prediction for housing prices and rent levels.

# Methodology

## Data source and methods

The data sources we would need are for population, household income and commercial rent. These were publicly available from Statistics Canada. For population, we got datasets from the census in 2011 that also held 2006 census data5. We also used a separate census dataset from the latest 2016 census6 that contains 2011 census data as well. The dataset for household total income covered 2015 and 2005 information7. Commercial rents services price index dataset represents the growth rate of retail rent using the 2019 base year8. It included growth rates from 2006 up to 2021.

Performance of data pre-processing that includes review for outliers, missing values, among others as well as transforming and data reduction of the datasets. The datasets are subjected to validation of distribution, finding correlations, and patterns.

In doing a predictive model for the transformed dataset, cross-validation used to make sure that every item in the original dataset has the same chance of appearing in the training and test sets for modeling. Using supervised learning method in modeling since the datasets are labeled.



*Figure 1. Stages for Data Mining to Build Commercial Rent Prediction Model*

## Extraction, Transformation and Loading

We downloaded all these datasets directly from *Statistics Canada* website. In exploring the datasets, we encountered the following:

### Population 2016 and 2011 dataset

On review of dataset, it showed 353,914 observations and 14 variables.

Table

Description automatically generated

*Figure 2. Raw Population 2016 dataset in R*

The relevant columns for this dataset are:

* 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

All other columns were removed as they were completely unrelated for the study. Data types on attributes chosen were appropriate.

However, “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 pulled the 2 rows out from its current attribute to become 2 separate stand-alone attributes/columns and deleted all remaining subsets under “DIM: Profile of Census Metropolitan Areas / Census Agglomerations (2247)” variable.

***Data reduction***: 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 original dataset of 353,914 observations and 14 variables.

Table

Description automatically generated

*Figure 3. Cleaned-up Population 2016 dataset in R*

### Population 2011 and 2006 dataset

On review of dataset, there were 73,160 observations and 14 variables.

Graphical user interface, application

Description automatically generated

*Figure 4. Raw Population 2011 and 2006 dataset in R*

We identified relevant columns as:

* CMACA\_Name
* Characteristics
  + Needs further analysis to determine which group to limit search with
* Total

All other columns were removed as they were completely unrelated for the study. Data types on the remaining relevant attributes were appropriate.

However, 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 pulled the 2 rows out of its current attribute under “Characteristics” to become 2 separate stand-alone attributes/columns and deleted all remaining subsets under “Characteristics” variable.

The values shown under attribute “Total “populated the separate columns of Population, 2006 and Population, 2011.

***Data reduction***: 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 the original dataset of 73,160 observations and 14 variables.

Table

Description automatically generated

*Figure 5. Cleaned-up Population 2011 and 2006 dataset in R*

### Household income 2015 and 2005 dataset

On review of the dataset, there were 1,440 observations and 17 variables

A screenshot of a computer

Description automatically generated with medium confidence

*Figure 6. Raw Household Income dataset in R*

We identified relevant columns for this dataset as:

* Geographic name
* Household type
  + Needs further analysis to determine which group to limit search with
* Median household total income, 2005
* Median household total income, 2015
* Median household total income % change

All other columns were removed as they were completely unrelated for the study. Data types on the remaining relevant attributes were appropriate.

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*

We needed only the “Census-family households” to line up with the numeric data on the variable median household total income.

Before we can conduct our analyses, we transformed 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.

***Data reduction***: 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 these attributes in the revised dataset “**Income\_Final**” with 160 observations and 3 variables from the original dataset of 1,440 observations and 17 variables.

Table

Description automatically generated

*Figure 6. Cleaned-up Household Income dataset in R*

### Commercial rents dataset

On review of the dataset, there were 1,956 observations and 15 variables.

Graphical user interface, application

Description automatically generated

*Figure 7. Raw Commercial Rent dataset in R*

Shown below are the relevant attributes/columns in commercial rent dataset:

* REF\_DATE
* GEO
* Building Type
* VALUE

All other columns were removed as they were completely unrelated for the study. Data types on the remaining relevant attributes were 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.

***Data reduction***: Transformed raw data to the relevant columns and picking up only the relevant attributes of REF\_DATE, GEO, Total, Building type (from “Building type”), and VALUE. It is further refined to show only Cities under the “GEO” attribute since it currently contains total Canada and Provinces too. For the REF\_DATE, only 2020 December is required eliminating other periods. This resulted to revised dataset of “**RentSubset3**” with 390 observations and 4 variables from raw dataset with 1,956 observations and 15 variables.

### Outliers

Using R Studio, we performed statistical analysis on each dataset on all the numeric columns including boxplots. No material exceptions noted.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

*Figure 8. Check distribution of numeric data*

2005 Boxplot

Chart, box and whisker chart

Description automatically generated

2015 Boxplot

Chart, box and whisker chart

Description automatically generated

*Figure 9. Boxplots - Check for outliers*

### Data Integration

To have a singular list showing metrics as population, household income and rents in one dataset, the 4 datasets must be consolidated using the merge function in R as follows:

* Left join on Population 2016 with Population 2011 to get Revised Population (both datasets with 158 observations)
* Left join on Household income (with 160 observations) with Revised population to get Combined Population and Household income (dataset with 160 observations)
* Left join on the Combined Population and Household income with Rent

However, before performing the above joins, the population and household income values had to be converted to growth values year over year since they will be compared to the dependent variable in rent.

Added growth columns for Population and Household income and revised dataset to PopCombined3 where all values of Population and Household income represent percentages (%s).

In doing the consolidation, we successfully combined Population and Household Income. Next will be Rent dataset joined with the combined Population and Household Income dataset.

#### Dataset Observation Issue

However, in preparing to consolidate the Rent dataset to the combined Population and Household Income, we noted that the 390 observations in Rent only included 12 cities for the pertinent year required and remaining were zeroes. This means that our final Combined dataset will be further reduced to how many rent observations there are as population and household income will use them as results for training and test. The Rent column is the dependent variable and will be converted to “Increase” (if value >100) or “Decrease” (if value <100).

The further reduction of observations to 12 observations with 4 variables now poses an issue as we are already too far into the process to find another public dataset to work on. We did not foresee this in our initial analyses as all datasets were rich in observations earlier. Investigating the raw datasets from Statistics Canada has been difficult and long due to how metrics were subset in columns/attributes. Unless there is clean-up and further transformation, this issue will not be easily detected by just a scan of data alone. This is specifically true with the dependent variable, rent, which resulted in an imbalance.

#### Addressing Dataset Issue

In real-world situations outside of this project, if after exhausting search for public rent dataset for additional cities required and nothing comes up, we would reach out for other sources such as companies like CBRE, JLL, Colliers, among others to get the missing rent information. However, for this project, what is required is publicly available datasets and so we will retain use of this resulting combined dataset. We proceeded with the resulting combined dataset and subjected it to the extent of modeling and validation that can be done.

# Analyses

Modeling and Validation

Supervised learning method was chosen since the dataset is labeled. We used classification algorithm specifically k-Nearest Neighbours (k-NN) and Logistic Regression.

## k-Nearest Neighbours

We wanted to determine if there is a pattern for rent growth relating to changes in population and household income.

We normalized the numeric variables as Population, Household Income and Rent Growth are in different scales.

normalize<-function(x) {

(x – min(x)) / (max(x) – min(x)) }

Apply the normalize function to the dataset:

PopCombined3\_n<- as.data.frame(lapply(PopCombined3[2,4], normalize))

In doing the testing, we will use a 70 : 30 training and test sets split:

Set.seed(123)

train\_index<-sample(1:nrow(PopCombined3\_n), 0.7 \* nrow(PopCombined3\_n))

train.set<-PopCombined3\_n[train\_index,]

test.set<-PopCombined3\_n[-train\_index,]

Remove “Cities” column from training and test datasets

train.set3\_new<-train.set[-1]

test.set3\_new<-test.set[-1]

Store the labels for our training and test sets

PopCombined3\_train\_labels<-train.set$Cities

PopCombined3\_test\_labels<-test.set$Cities

*Prediction*

library (class)

PopCombined3\_test\_prediction<-knn(train = train.set3\_new, test = test.set3\_new, cl = PopCombined3\_train\_labels, k = 3)

Calendar

Description automatically generated

When we ran the line for prediction, it returned an **error** most likely due to the small number of observations in the combined dataset that we earlier identified that may become an issue.

We went back to the raw dataset to extend exploration of data to see if there is still a way to expand the dataset. However, the rent information, which is the dependent variable was truly limited. If the rent data can be obtained from other sources to fill the missing cities, we may be able to increase observations to an acceptable level and re-run prediction line successfully since population and household income have values in those other cities not just key cities.

## Logistic Regression

To determine relationship between our dependent variable (rent) and the independent variables (population and household income), we used regression analysis. Logistic regression since our dependent variable is categorical/binary.

The binomial logistic regression predicts the observations to fall in either increase (“1”) or decrease (“0”) in rent, category of population and household income.

In the PopCombined3 dataset, outcome/dependent variable Rent Growth changed from text “Increase” and “Decrease” to binary response of “1” for increase and “0” for decrease as factors.

Graphical user interface, application, Word

Description automatically generated

*Figure 10. Replace Rent Growth with 0 and 1 from Decrease and Increase*

To do the testing, we created 70:30 training and test sets split

Set.seed(123)

train\_index<-sample(1:nrow(PopCombined3\_n), 0.7 \* nrow(PopCombined3\_n))

train.set<-PopCombined3\_n[train\_index,]

test.set<-PopCombined3\_n[-train\_index,]

Removed “Cities” column from training and test datasets

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

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

Creation of Logistic regression model

*Prediction Using the Test Set*

Text

Description automatically generated

*Figure 11. Logistic regression prediction with test set*

*Coefficients*:

For every unit change in population growth, the log odd of rent growth decreases by 0.089

For every unit change in household income growth, the log odd of rent growth decreases by 0.193

Population growth (p-value = 0.584) and Household income growth (p-value = 0.564) are **not** statistically significant

*Prediction Using the Unsplit Dataset (greater # of observations than test set)*

Text

Description automatically generated

*Figure 12. Logistic regression prediction using unsplit data (combined dataset)*

*Coefficients*:

For every unit change in population growth, the log odd of rent growth decreases by 0.123

For every unit change in household income growth, the log odd of rent growth decreases by 0.086

Population growth (p-value = 0.314) and Household income growth (p-value = 0.423) are **not** statistically significant

*Confidence Intervals*



## Model Evaluation

*Confusion Matrix* was used to measure the performance of the model

### k-Nearest Neighbours Prediction

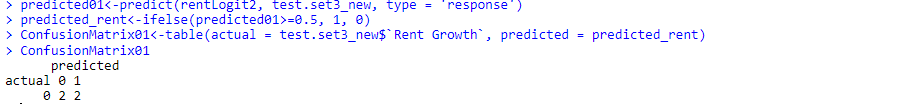
With a small number of observations in the dataset, the prediction model we built produced an error and we were not able to proceed in getting any results. Thus, we do not have any predictions to evaluate.

### Logistic Regression Using the Test Set

predicted01<-predict(rentLogit2, test.set3\_new, type = ‘response’)

predicted\_rent<-ifelse(predicted01>=0.5, 1, 0)

ConfusionMatrix01<-table(actual = test.set3\_new$`Rent Growth`, predicted = predicted\_rent)





*Figure 13. Confusion Matrix: Logistic Regression using Test Set*

The test set carried solely “0” or decrease in growth. Results of the prediction show 50 : 50 probability of between True Negative and False Positive

### Logistic Regression Using Unsplit Dataset (greater # of observations than test)

We wanted to see if there was an impact with the increase in data by using the unsplit dataset

predicted02<-predict(rentLogit, PopCombined4, type = ‘response’)

predicted\_rent02<-ifelse(predicted02>=0.5, 1, 0)

ConfusionMatrix02<-table(actual = PopCombined4$`Rent Growth`, predicted = predicted\_rent02)

Text

Description automatically generated



*Figure 14. Confusion Matrix: Logistic Regression using Unsplit Dataset*

# Findings and Results

We were not able to get any results for the k-Nearest Neighbours model since it returned an error due to the small number of observations in the dataset.

We still pursued using the small dataset and utilized the test set in the Logistic Regression model. It shows that population and household income are indirectly related to rent growth based on resulting coefficients. It also shows that both population and household income are not statistically significant. However, the Confusion Matrix we used to test its performance show it is 50% accurate. The test set contains only “0” or decrease which makes this result unreliable having no other instance of “1” or increase.

We made another run of the Logistic Regression model in order to see if additional data will improve the results by using the unsplit dataset (i.e., before training and test split). The result similarly shows through the coefficients that population and household income are indirectly related to rent growth and shows that both population and household income are not statistically significant. However, with the additional data including the “1” or increase in rent occurrence considered, accuracy of the model went up to 58%. It was able to predict increase when it is truly an increase (sensitivity) 60% of the time and it predicts decrease when it is actually a decrease (specificity) 57% of the time. When it predicts increase, it is correct 50% of the time (precision).

## Limitations

With the small number of data that resulted from our clean-up, not only did it become a problem in our model but also shows that any results we get will be inconclusive.

## Next Steps

Our next steps would be to obtain additional rent data for the missing cities. This will fill in the missing rent data and resolve the error in our k-Nearest Neighbours model enabling us to get the prediction and evaluate it after. This will also further improve results for the Logistic Regression model.

We can obtain additional rent data from other real estate companies or from research organizations such as CBRE, JLL or Colliers. We have checked the available reports from CBRE and Colliers. Most of them produce at least a quarterly report wherein commercial rents of Canadian cities are included. We will check if they have a csv file available, reach out to the research manager in that company and/or we can summarize all their reports to for our csv file to fill in our required missing data.

# Conclusion

Knowing that we have enough data for both the dependent and independent variables were our primary focus. However, the datasets from Statistics Canada posed a challenge. We needed time to clean-up, re-format (i.e., build columns from rows previously under a column), combine datasets(i.e., do joins), from as much as 353,914 observations and 14 variables on just one dataset from the 4 raw datasets became 12 observations and 4 variables for the consolidated dataset, in order to get to the point where we were able to present the relevant variables we require and at the same time discover that rent observations cover only the key cities and the rest were not provided resulting to data imbalance.

In both instances of Logistic Regression modeling, there is an indirect relationship with Population and Household Income to Rent which is for a unit increase in population or household income growth, there is a commensurate decrease in rent growth.

Although we were able to get some results on the Logistic Regression model, due to the limitation of data used to train and test, the **results were inconclusive**. However, our test shows that expanding/adding rent data further improves performance of the model and will provide more insights into predicting rent growth using population and household income growth.

Therefore, our next step is to fill in those missing rent data by reaching out to research organizations. Once we get the needed rent data, we will run our predictions models and are confident that we will get more meaningful insights and conclusions.

# References

1. Wheeler, David C. (2014). *A Bayesian approach to hedonic price analysis.* Papers in regional science. ISSN: 1056-8190. Volume: 93. Issue: 3 Page: 663 DOI: 10.1111/pirs.12003
2. Des Rosiers, F., Thériault, M., & Lavoie, C. (2009). *Retail Concentration and Shopping Center Rents - A Comparison of Two Cities*. The Journal of Real Estate Research, 31(2), 165-207.

<http://ezproxy.lib.ryerson.ca/login?url=https://www.proquest.com/scholarly-journals/retail-concentration-shopping-center-rents/docview/200319553/se-2?accountid=13631>

1. Zhang, Song (2020). *Tenant Mix and Retail Rents in High Street Shopping Districts.* The journal of real estate finance and economics. ISSN: 0895-5638. DOI: 10.1007/s11146-020-09768-3
2. Hiebert, D. (2017). *Immigrants and Refugees in the Housing Markets of Montreal, Toronto and Vancouver, 2011.* Canadian Journal of Urban Research, 26(2), 52. <http://ezproxy.lib.ryerson.ca/login?url=https://www.proquest.com/scholarly-journals/immigrants-refugees-housing-markets-montreal/docview/2159615860/se-2?accountid=13631>
3. *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).
4. *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).
5. *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, 2006 Census – 20% Sample data* / Catalogue no. 98-402-X2016006 (Statistics Canada).
6. *Commercial rents services price index, monthly.* Statistics Canada Table: 1810025501-eng <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810025501>