NOVEMBER 8, 2021

PREDICTING RENTAL RATE GROWTH IN THE CANADIAN MARKET

ROMMEL AGUSTIN 500884909

SUPERVISOR: CENI BABAOGLU, PHD
RYERSON UNIVERSITY

Table of Contents

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.

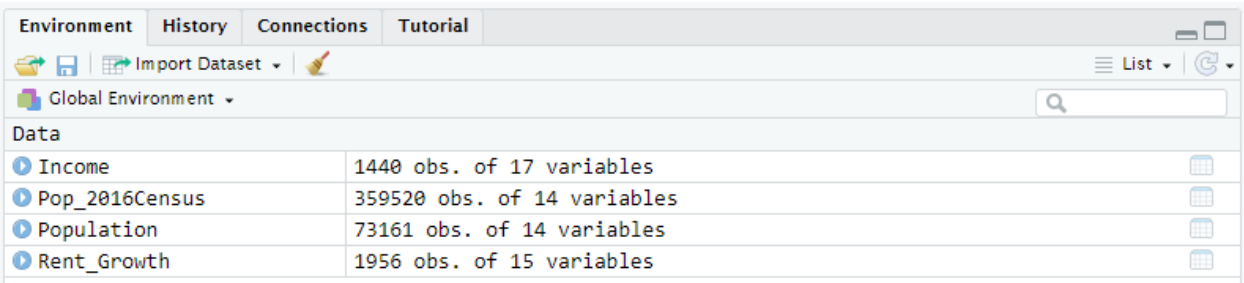
Github Link

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



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

```
> summary(Income)
 Geographic code    Geographic name    Geographic type    Geographic name, Province or territory
 Length:1440        Length:1440        Length:1440        Length:1440
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



 Geographic code, Province or territory Additional province code Global non-response rate Data quality flag  Household type
 Min.   :10.00                           Min.   :13.00            Min.   : 2.800           Length:1440        Length:1440
 1st Qu.:24.00                           1st Qu.:24.00            1st Qu.: 3.900           Class :character   Class :character
 Median :35.00                           Median :24.00            Median : 4.400           Mode  :character   Mode  :character
 Mean   :37.04                           Mean   :30.75            Mean   : 4.566
 3rd Qu.:48.00                           3rd Qu.:38.00            3rd Qu.: 4.900
 Max.   :61.00                           Max.   :48.00            Max.   :14.300
                                         NA's   :1332
 Number of households, 2006 Number of households, 2016 Median household total income (2015 constant dollars), 2005
 Min.   :     15            Min.   :     25            Min.   :     0
 1st Qu.:   1994            1st Qu.:   2256            1st Qu.: 53411
 Median :   4958            Median :   5570            Median : 71652
 Mean   :  28482            Mean   :  32124            Mean   : 70221
 3rd Qu.:  14470            3rd Qu.:  15720            3rd Qu.: 86555
 Max.   :1801255            Max.   :2135910            Max.   :184025

 Median household total income (2015 constant dollars), 2015 Median household total income (2015 constant dollars), % change
 Min.   : 19883                                              Min.   :-19.30
 1st Qu.: 59416                                              1st Qu.:  7.40
 Median : 81014                                              Median : 13.30
 Mean   : 80784                                              Mean   : 14.93
 3rd Qu.: 99328                                              3rd Qu.: 20.60
 Max.   :284706                                              Max.   :125.40
                                                             NA's   :1
 Median household after-tax income (2015 constant dollars), 2005
 Min.   :     0
 1st Qu.: 47723
 Median : 62130
 Mean   : 60969
 3rd Qu.: 74466
 Max.   :148095

 Median household after-tax income (2015 constant dollars), 2015
 Min.   : 19840
 1st Qu.: 53576
 Median : 70727
 Mean   : 70218
 3rd Qu.: 85802
 Max.   :222326
```

```
> str(Income)
tibble [1,440 × 17] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Geographic code                                                  : chr [1:1440] "001" "001" "001" "001" ...
 $ Geographic name                                                  : chr [1:1440] "St. John's" "St. John's" "St. John's" "St. John's" ...
 $ Geographic type                                                  : chr [1:1440] "CMA" "CMA" "CMA" "CMA" ...
 $ Geographic name, Province or territory                           : chr [1:1440] "Newfoundland and Labrador" "Newfoundland and Labrador" "Newfoundland and Labrador" "Newfc
dland and Labrador" ...
 $ Geographic code, Province or territory                           : num [1:1440] 10 10 10 10 10 10 10 10 10 10 ...
 $ Additional province code                                         : num [1:1440] NA NA NA NA NA NA NA NA NA NA ...
 $ Global non-response rate                                         : num [1:1440] 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 ...
 $ Data quality flag                                                : chr [1:1440] "00000" "00000" "00000" "00000" ...
 $ Household type                                                   : chr [1:1440] "Total \u0096 Household type including census family structure" "Census-family households'
"Households consisting of only one census family without additional persons" "One couple, with or without children in their census family" ...
 $ Number of households, 2006                                       : num [1:1440] 70660 51495 47415 39385 16450 ...
 $ Number of households, 2016                                       : num [1:1440] 85015 58665 54185 45960 22390 ...
 $ Median household total income (2015 constant dollars), 2005      : num [1:1440] 62554 77211 76247 84775 68864 ...
 $ Median household total income (2015 constant dollars), 2015      : num [1:1440] 79750 102864 101339 111972 88239 ...
 $ Median household total income (2015 constant dollars), % change  : num [1:1440] 27.5 33.2 32.9 32.1 28.1 39 31 31.2 28.3 33.1 ...
 $ Median household after-tax income (2015 constant dollars), 2005  : num [1:1440] 53516 65252 64154 70510 57914 ...
 $ Median household after-tax income (2015 constant dollars), 2015  : num [1:1440] 68121 86542 85047 92698 74938 ...
 $ Median household after-tax income (2015 constant dollars), % change: num [1:1440] 27.3 32.6 32.6 31.5 29.4 38.2 28.1 30.8 27.7 32.2 ...
 - attr(*, "spec")=
  .. cols(
  ..   `Geographic code` = col_character(),
  ..   `Geographic name` = col_character(),
  ..   `Geographic type` = col_character(),
  ..   `Geographic name, Province or territory` = col_character(),
  ..   `Geographic code, Province or territory` = col_double(),
  ..   `Additional province code` = col_double(),
  ..   `Global non-response rate` = col_double(),
  ..   `Data quality flag` = col_character(),
  ..   `Household type` = col_character(),
  ..   `Number of households, 2006` = col_double(),
  ..   `Number of households, 2016` = col_double(),
  ..   `Median household total income (2015 constant dollars), 2005` = col_double(),
  ..   `Median household total income (2015 constant dollars), 2015` = col_double(),
  ..   `Median household total income (2015 constant dollars), % change` = col_double(),
  ..   `Median household after-tax income (2015 constant dollars), 2005` = col_double(),
  ..   `Median household after-tax income (2015 constant dollars), 2015` = col_double(),
  ..   `Median household after-tax income (2015 constant dollars), % change` = col_double()
  .. )
```

Analyses and Preparation

Geographic code	Cities	Geographic type	Geographic name, Province or territory	Geographic code, Province or territory	Additional province code	Global non-response rate	Data quality flag	Household type	Number of households, 2006	Number of households, 2016	HH Income 2005	HH Income 2015	Median household total income (2015 constant dollars), % change	Median household after-tax income (2015 constant dollars), 2005	Median household after-tax income (2015 constant dollars), 2015
001	St. John's	CMA	Newfoundland and Labrador	10	NA	3.5	00000	Total ◆ Household type including census family structure	70660	85015	62554	79750	27.5	53516	68
001	St. John's	CMA	Newfoundland and Labrador	10	NA	3.5	00000	Census-family households	51495	58665	77211	102864	33.2	65252	86
001	St. John's	CMA	Newfoundland and Labrador	10	NA	3.5	00000	Households consisting of only one census family without ad...	47415	54185	76247	101339	32.9	64154	85
001	St. John's	CMA	Newfoundland and Labrador	10	NA	3.5	00000	One couple, with or without children in their census family	39385	45960	84775	111972	32.1	70510	92
001	St. John's	CMA	Newfoundland and Labrador	10	NA	3.5	00000	One couple, without children in their census family	16450	22390	68864	88239	28.1	57914	74
001	St. John's	CMA	Newfoundland and Labrador	10	NA	3.5	00000	One couple, with children in their census family	22935	23570	97097	134980	39.0	80161	110
001	St. John's	CMA	Newfoundland and Labrador	10	NA	3.5	00000	One lone-parent census family	8030	8225	38175	50005	31.0	36140	46
001	St. John's	CMA	Newfoundland and Labrador	10	NA	3.5	00000	Other census-family households	4080	4485	89790	117803	31.2	78341	102
001	St. John's	CMA	Newfoundland and Labrador	10	NA	3.5	00000	Non-census-family households	19165	26350	31016	39805	28.3	27685	35
005	Bay Roberts	CA	Newfoundland and Labrador	10	NA	3.5	00000	Total ◆ Household type including census family structure	3930	4505	48541	64594	33.1	43213	57
005	Bay Roberts	CA	Newfoundland and Labrador	10	NA	3.5	00000	Census-family households	3210	3440	56520	79714	41.0	50391	69
005	Bay Roberts	CA	Newfoundland and Labrador	10	NA	3.5	00000	Households consisting of only one census family without ad...	2945	3210	54974	78043	42.0	49074	67
005	Bay Roberts	CA	Newfoundland and Labrador	10	NA	3.5	00000	One couple, with or without children in their census family	2640	2805	59360	85362	43.8	52482	73
005	Bay Roberts	CA	Newfoundland and Labrador	10	NA	3.5	00000	One couple, without children in their census family	1185	1465	45645	62464	36.8	40268	55
005	Bay Roberts	CA	Newfoundland and Labrador	10	NA	3.5	00000	One couple, with children in their census family	1455	1340	73826	115627	56.6	63556	96
005	Bay Roberts	CA	Newfoundland and Labrador	10	NA	3.5	00000	One lone-parent census family	305	410	24025	39552	64.6	24025	37
005	Bay Roberts	CA	Newfoundland and Labrador	10	NA	3.5	00000	Other census-family households	260	230	65582	102656	56.5	60338	90
005	Bay Roberts	CA	Newfoundland and Labrador	10	NA	3.5	00000	Non-census-family households	720	1060	18794	24038	27.9	18428	23

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 :

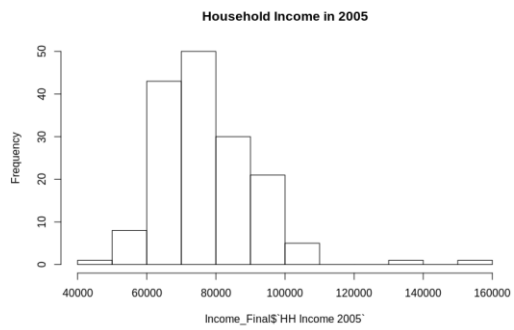
Filter

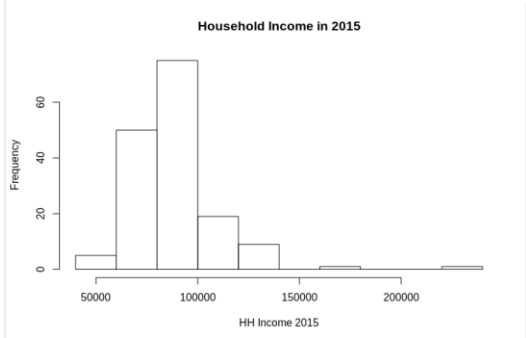
	Cities	HH Income 2005	HH Income 2015
1	St. John's	77211	102864
2	Bay Roberts	56520	79714
3	Grand Falls-Windsor	63266	77120
4	Gander	69454	88037
5	Corner Brook	65967	83095
6	Charlottetown	73680	82252
7	Summerside	62696	70290
8	Campbellton (New Brunswick part)	57497	67462
9	Halifax	82633	91252
10	Kentville	61302	70620
11	Truro	63053	72221
12	New Glasgow	63397	73062
13	Cape Breton	61087	71467
14	Campbellton (Quebec part)	48288	50688
15	Hawkesbury (Quebec part)	53340	59200
16	Ottawa - Gatineau (Quebec part)	85652	92828
17	Moncton	73868	80407
18	Saint John	72301	82988
19	Fredericton	75473	84630
20	Bathurst	63957	70979
21	Miramichi	61053	74069
22	Campbellton	56251	65106
23	Edmundston	64274	71399

Showing 1 to 27 of 160 entries, 3 total columns

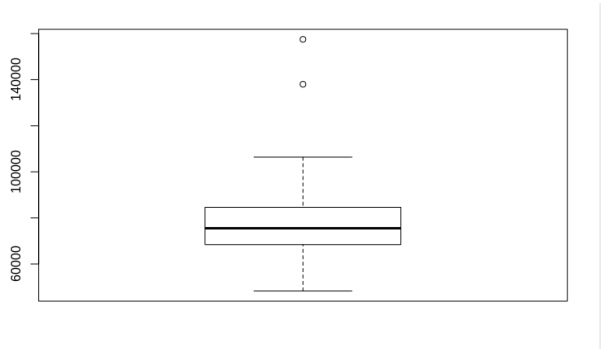
This is where we can start doing analyses of the dataset:

```
> summary(Income_Final)
    Cities          HH Income 2005   HH Income 2015
 Length:160         Min.   : 48288   Min.   : 50688
 Class :character   1st Qu.: 68460   1st Qu.: 75875
 Mode  :character   Median : 75509   Median : 86084
                    Mean   : 77614   Mean   : 88623
                    3rd Qu.: 84457   3rd Qu.: 95893
                    Max.   :157496   Max.   :220888
```

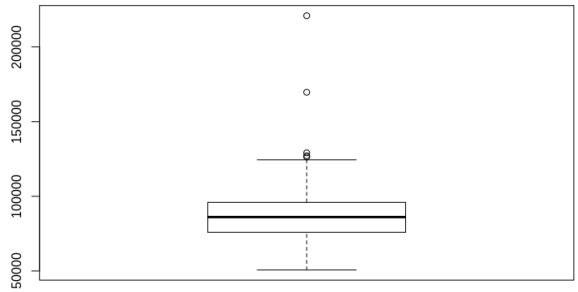




2005 Boxplot



2015 Boxplot



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

```
> View(Pop_2016Census)
> summary(Pop_2016Census)
  CENSUS_YEAR   GEO_CODE (POR)       GEO_LEVEL      GEO_NAME              GNR             GNR_LF       DATA_QUALITY_FLAG
 Min.   :2016   Length:359520      Min.   :2.00   Length:359520      Min.   : 2.800   Min.   : 2.600   Length:359520
 1st Qu.:2016   Class :character   1st Qu.:2.00   Class :character   1st Qu.: 3.900   1st Qu.: 4.400   Class :character
 Median :2016   Mode  :character   Median :2.00   Mode  :character   Median : 4.400   Median : 5.000   Mode  :character
 Mean   :2016                      Mean   :2.05                      Mean   : 4.566   Mean   : 5.437
 3rd Qu.:2016                      3rd Qu.:2.00                      3rd Qu.: 4.900   3rd Qu.: 5.900
 Max.   :2016                      Max.   :3.00                      Max.   :14.300   Max.   :20.700

  ALT_GEO_CODE     DIM: Profile of Census Metropolitan Areas/Census Agglomerations (2247)
 Min.   :  10001   Length:359520
 1st Qu.:  24456   Class :character
 Median :  35566   Mode  :character
 Mean   : 202445
 3rd Qu.:  48841
 Max.   :4884048

 Member ID: Profile of Census Metropolitan Areas/Census Agglomerations (2247)
 Min.   :   1
 1st Qu.: 562
 Median :1124
 Mean   :1124
 3rd Qu.:1686
 Max.   :2247

 Notes: Profile of Census Metropolitan Areas/Census Agglomerations (2247) Dim: Sex (3): Member ID: [1]: Total - Sex
 Min.   :  1.0                                                            Min.   :    -21
 1st Qu.: 66.0                                                            1st Qu.:      0
 Median :122.0                                                            Median :     10
 Mean   :121.4                                                            Mean   :  12226
 3rd Qu.:178.0                                                            3rd Qu.:    785
 Max.   :238.0                                                            Max.   :5928040
 NA's   :323520                                                           NA's   :20
 Dim: Sex (3): Member ID: [2]: Male Dim: Sex (3): Member ID: [3]: Female
 Length:359520                      Length:359520
 Class :character                   Class :character
 Mode  :character                   Mode  :character
```

```
> str(Pop_2016Census)
tibble [359,520 × 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ CENSUS_YEAR                                                          : num [1:359520] 2016 2016 2016 2016 2016 ...
 $ GEO_CODE (POR)                                                       : chr [1:359520] "001" "001" "001" "001" ...
 $ GEO_LEVEL                                                            : num [1:359520] 2 2 2 2 2 2 2 2 2 2 ...
 $ GEO_NAME                                                             : chr [1:359520] "St. John's" "St. John's" "St. Joh
n's" "St. John's" ...
 $ GNR                                                                  : num [1:359520] 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5
3.5 ...
 $ GNR_LF                                                               : num [1:359520] 5.3 5.3 5.3 5.3 5.3 5.3 5.3 5.3 5.3
5.3 ...
 $ DATA_QUALITY_FLAG                                                    : chr [1:359520] "00000" "00000" "00000" "00000" ...
 $ ALT_GEO_CODE                                                         : num [1:359520] 10001 10001 10001 10001 10001 ...
 $ DIM: Profile of Census Metropolitan Areas/Census Agglomerations (2247)      : chr [1:359520] "Population, 2016" "Population, 201
1" "Population percentage change, 2011 to 2016" "Total private dwellings" ...
 $ Member ID: Profile of Census Metropolitan Areas/Census Agglomerations (2247): num [1:359520] 1 2 3 4 5 6 7 8 9 10 ...
 $ Notes: Profile of Census Metropolitan Areas/Census Agglomerations (2247)    : num [1:359520] 1 2 NA 3 4 NA NA 5 NA NA ...
 $ Dim: Sex (3): Member ID: [1]: Total - Sex                                   : num [1:359520] 205955 196954 4.6 92353 85015 ...
 $ Dim: Sex (3): Member ID: [2]: Male                                          : chr [1:359520] "..." "..." "..." "..." "..." ...
 $ Dim: Sex (3): Member ID: [3]: Female                                        : chr [1:359520] "..." "..." "..." "..." "..." ...
 - attr(*, "problems")= tibble [20 x 5] (S3: tbl_df/tbl/data.frame)
  ..$ row     : int [1:20] 355883 355884 355885 355886 355887 355893 355894 355895 355896 355897 ...
  ..$ col     : chr [1:20] "Dim: Sex (3): Member ID: [1]: Total - Sex" "Dim: Sex (3): Member ID: [1]: Total - Sex" "Dim: Sex (3): Me
mber ID: [1]: Total - Sex" "Dim: Sex (3): Member ID: [1]: Total - Sex" ...
  ..$ expected: chr [1:20] "a double" "a double" "a double" "a double" ...
  ..$ actual  : chr [1:20] "..." "..." "..." "..." ...
  ..$ file    : chr [1:20] "'Pop_2016Census.csv'" "'Pop_2016Census.csv'" "'Pop_2016Census.csv'" "'Pop_2016Census.csv'" ...
 - attr(*, "spec")=
  .. cols(
  ..   CENSUS_YEAR = col_double(),
  ..   `GEO_CODE (POR)` = col_character(),
  ..   GEO_LEVEL = col_double(),
  ..   GEO_NAME = col_character(),
  ..   GNR = col_double(),
  ..   GNR_LF = col_double(),
  ..   DATA_QUALITY_FLAG = col_character(),
  ..   ALT_GEO_CODE = col_double(),
  ..   `DIM: Profile of Census Metropolitan Areas/Census Agglomerations (2247)` = col_character(),
  ..   `Member ID: Profile of Census Metropolitan Areas/Census Agglomerations (2247)` = col_double(),
  ..   `Notes: Profile of Census Metropolitan Areas/Census Agglomerations (2247)` = col_double(),
  ..   `Dim: Sex (3): Member ID: [1]: Total - Sex` = col_double(),
  ..   `Dim: Sex (3): Member ID: [2]: Male` = col_character(),
  ..   `Dim: Sex (3): Member ID: [3]: Female` = col_character()
  .. )
```

Analyses and Preparation

Income × | Population × | Pop_2016Census × | Rent_Growth ×

Filter

	CENSUS_YEAR	GEO_CODE (POR)	GEO_LEVEL	GEO_NAME	GNR	GNR_LF	DATA_QUALITY_FLAG	ALT_GEO_CODE	DIM: Profile of Census Metropolitan Areas/Census Agglomerations (2247)	Member ID: Profile of Census Metropolitan Areas/Census Agglomerations (2247)	Notes: Profile of Census Metropolitan Areas/Census Agglomerations (2247)	Dim: Sex (3): Member ID: [1]: Total - Sex	Dim: Sex (3): Member ID: [2]: Male	Dim: Sex (3): Member ID: [3]: Female
1	2016	001	2	St. John's	3.5	5.3	00000	10001	Population, 2016	1	1	205955.00	...	...
2	2016	001	2	St. John's	3.5	5.3	00000	10001	Population, 2011	2	2	196954.00	...	...
3	2016	001	2	St. John's	3.5	5.3	00000	10001	Population percentage change, 2011 to 2016	3	NA	4.60	...	...
4	2016	001	2	St. John's	3.5	5.3	00000	10001	Total private dwellings	4	3	92353.00	...	...
5	2016	001	2	St. John's	3.5	5.3	00000	10001	Private dwellings occupied by usual residents	5	4	85015.00	...	...
6	2016	001	2	St. John's	3.5	5.3	00000	10001	Population density per square kilometre	6	NA	255.90	...	...
7	2016	001	2	St. John's	3.5	5.3	00000	10001	Land area in square kilometres	7	NA	804.79	...	...
8	2016	001	2	St. John's	3.5	5.3	00000	10001	Total - Age groups and average age of the population - 100...	8	5	205955.00	99875	106080
9	2016	001	2	St. John's	3.5	5.3	00000	10001	0 to 14 years	9	NA	32465.00	16650	15815
10	2016	001	2	St. John's	3.5	5.3	00000	10001	0 to 4 years	10	NA	10280.00	5295	4985
11	2016	001	2	St. John's	3.5	5.3	00000	10001	5 to 9 years	11	NA	11385.00	5785	5600
12	2016	001	2	St. John's	3.5	5.3	00000	10001	10 to 14 years	12	NA	10795.00	5575	5225
13	2016	001	2	St. John's	3.5	5.3	00000	10001	15 to 64 years	13	NA	142940.00	69645	73290
14	2016	001	2	St. John's	3.5	5.3	00000	10001	15 to 19 years	14	NA	10955.00	5590	5360
15	2016	001	2	St. John's	3.5	5.3	00000	10001	20 to 24 years	15	NA	13965.00	6830	7135
16	2016	001	2	St. John's	3.5	5.3	00000	10001	25 to 29 years	16	NA	15165.00	7615	7555
17	2016	001	2	St. John's	3.5	5.3	00000	10001	30 to 34 years	17	NA	14895.00	7295	7595

Showing 1 to 20 of 359,520 entries, 14 total columns

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:

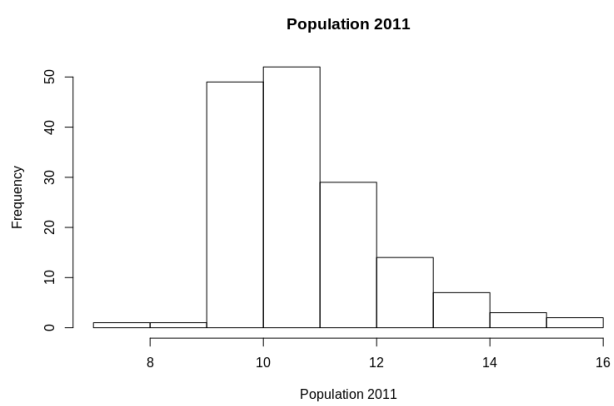
Filter

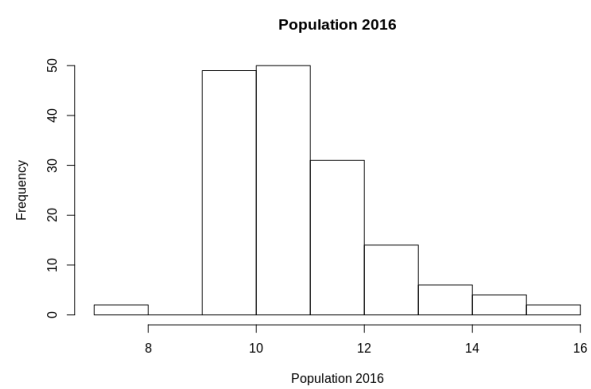
	Cities	Population, 2011	Population, 2016	LogPop2016	LogPop2011
1	Abbotsford - Mission	170191	180518	12.103586	12.044677
2	Alma	33018	32849	10.399677	10.404808
3	Arnprior	15485	15973	9.678655	9.647627
4	Baie-Comeau	28465	27692	10.228899	10.256431
5	Barrie	187013	197059	12.191258	12.138933
6	Bathurst	31936	31110	10.345285	10.371489
7	Bay Roberts	10871	11083	9.313168	9.293854
8	Belleville	101668	103472	11.547056	11.529468
9	Brandon	54847	58003	10.968250	10.912303
10	Brantford	135501	134203	11.807109	11.816734
11	Brockville	39024	38553	10.559789	10.571932
12	Brooks	23430	24662	10.113019	10.061773
13	Calgary	1214839	1392609	14.146690	14.010122
14	Campbell River	36096	37861	10.541677	10.493937
15	Campbellton	17361	15746	9.664342	9.761982
16	Campbellton (New Brunswick part)	14039	13114	9.481436	9.549594
17	Campbellton (Quebec part)	3322	2632	7.875499	8.108322
18	Camrose	17286	18742	9.838522	9.757652
19	Canmore	12288	13992	9.546241	9.416378
20	Cape Breton	101619	98722	11.500063	11.528986
21	Carleton Place	29180	31451	10.356186	10.281239
22	Centre Wellington	26693	28191	10.246758	10.192157
23	Charlottetown	65523	69325	11.146561	11.090157

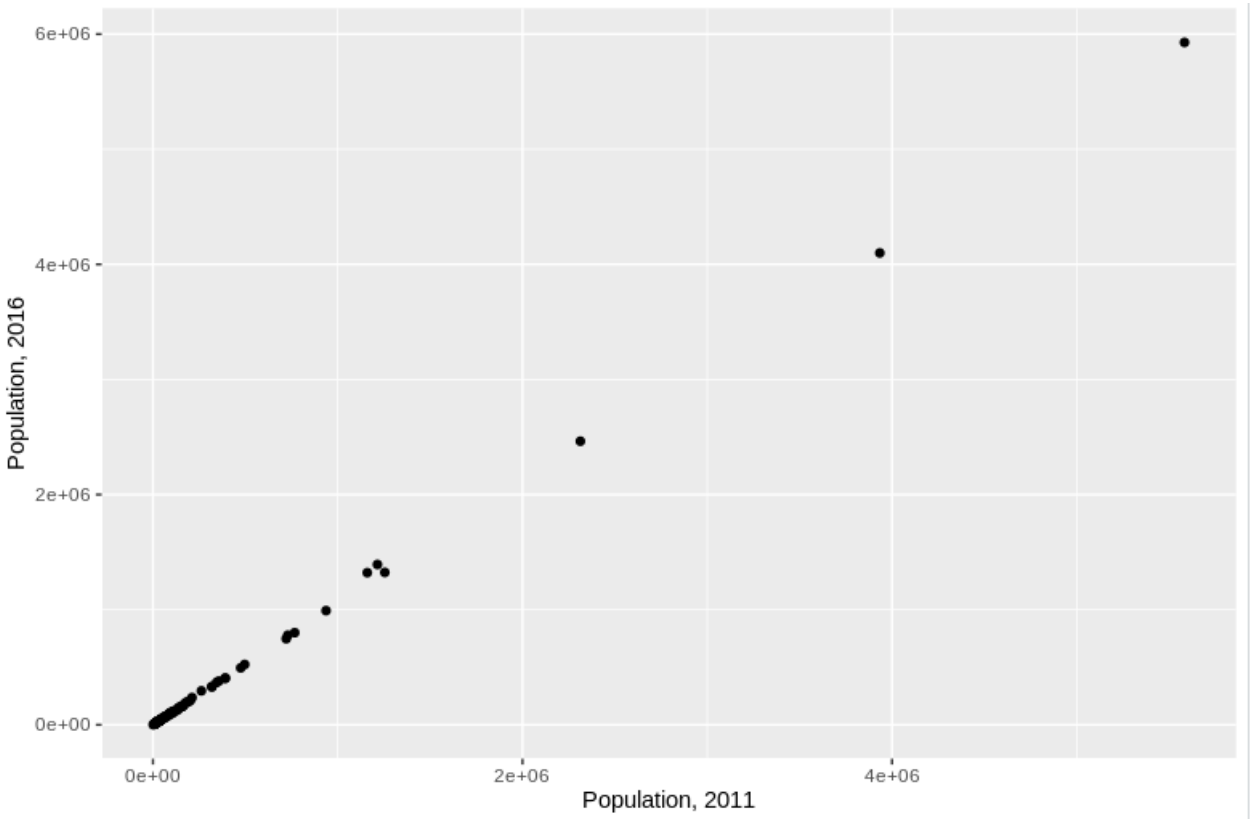
Showing 1 to 26 of 158 entries, 5 total columns

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

```
> summary(Pop2016Final)
   Cities          Population, 2011  Population, 2016
 Length:158        Min.   :   1577   Min.   :   1711
 Class :character  1st Qu.:  17493   1st Qu.:  18014
 Mode  :character  Median :  34920   Median :  34818
                   Mean   : 182967   Mean   : 193489
                   3rd Qu.:  98646   3rd Qu.: 101972
                   Max.   :5583064   Max.   :5928040
```







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

```
> View(Population)
> summary(Population)
 Census Profile - Census Metropolitan Areas and Census Agglomerations (CMAs/Cas)      X2                 X3
 Length:73161                                                                      Length:73161       Length:73161
 Class :character                                                                  Class :character   Class :character
 Mode  :character                                                                  Mode  :character   Mode  :character
      X4                 X5                 X6                 X7                 X8                 X9
 Length:73161       Length:73161       Length:73161       Length:73161       Length:73161       Length:73161
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
     X10                X11                X12                X13                X14
 Length:73161       Length:73161       Length:73161       Length:73161       Mode:logical
 Class :character   Class :character   Class :character   Class :character   NA's:73161
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
> |
```

```
Console   Terminal ×   Jobs ×
~/
tibble [73,160 × 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Geo_Code       : num [1:73160] 1 1 1 1 1 1 1 1 1 1 ...
 $ Prov_Name      : chr [1:73160] "Newfoundland and Labrador" "Newfoundland and Labrador" "Newfoundland and Labrador" "Newfoundland and Labrado
r" ...
 $ CMACA_Name     : chr [1:73160] "St. John's" "St. John's" "St. John's" "St. John's" ...
 $ Type           : chr [1:73160] "CMA" "CMA" "CMA" "CMA" ...
 $ Topic          : chr [1:73160] "Population and dwelling counts" "Population and dwelling counts" "Population and dwelling counts" "Populatio
n and dwelling counts" ...
 $ Characteristics: chr [1:73160] "Population in 2011" "Population in 2006" "2006 to 2011 population change (%)" "Total private dwellings" ...
 $ Note           : num [1:73160] 1 1 NA 2 3 NA NA 4 NA NA ...
 $ Total          : num [1:73160] 196966 181113 8.8 84542 78960 ...
 $ Flag_Total     : logi [1:73160] NA NA NA NA NA NA ...
 $ Male           : num [1:73160] NA NA NA NA NA ...
 $ Flag_Male      : chr [1:73160] "..." "..." "..." "..." ...
 $ Female         : num [1:73160] NA NA NA NA NA ...
 $ Flag_Female    : chr [1:73160] "..." "..." "..." "..." ...
 $ X14            : logi [1:73160] NA NA NA NA NA NA ...
 - attr(*, "problems")= tibble [68 × 5] (S3: tbl_df/tbl/data.frame)
  ..$ row     : int [1:68] 1418 1890 6138 6610 7082 7556 7558 8970 9442 9914 ...
  ..$ col     : chr [1:68] "Flag_Total" "Flag_Total" "Flag_Total" "Flag_Total" ...
  ..$ expected: chr [1:68] "1/0/T/F/TRUE/FALSE" "1/0/T/F/TRUE/FALSE" "1/0/T/F/TRUE/FALSE" "1/0/T/F/TRUE/FALSE" ...
  ..$ actual  : chr [1:68] "A" "A" "A" "A" ...
  ..$ file    : chr [1:68] "'POPULATION_2006_2011.csv'" "'POPULATION_2006_2011.csv'" "'POPULATION_2006_2011.csv'" "'POPULATION_2006_2011.csv'"
...
 - attr(*, "spec")=List of 3
  ..$ cols   :List of 14
  .. ..$ Geo_Code       : list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Prov_Name      : list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ CMACA_Name     : list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Type           : list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Topic          : list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Characteristics: list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_character" "collector"
  .. ..$ Note           : list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Total          : list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Flag_Total     : list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_logical" "collector"
  .. ..$ Male           : list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_double" "collector"
  .. ..$ Flag_Male      : list()
  .. .. .. - attr(*, "class")= chr [1:2] "collector_character" "collector"
```

Analyses and Preparation

Income × | Population × | Pop_2016Census × | Rent_Growth ×

Filter

	Census Profile - Census Metropolitan Areas and Census Agglomerations (CMAs/Cas)	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	Geo_Code	Prov_Name	CMACA_Name	Type	Topic	Characteristics	Note	Total	Flag_Total	Male	Flag_Male	Female	Flag_Female	NA
2	001	Newfoundland and Labrador	St. John's	CMA	Population and dwelling counts	Population in 2011	1	196966	NA	NA	...	NA	...	NA
3	001	Newfoundland and Labrador	St. John's	CMA	Population and dwelling counts	Population in 2006	1	181113	NA	NA	...	NA	...	NA
4	001	Newfoundland and Labrador	St. John's	CMA	Population and dwelling counts	2006 to 2011 population change (%)	NA	8.8	NA	NA	...	NA	...	NA
5	001	Newfoundland and Labrador	St. John's	CMA	Population and dwelling counts	Total private dwellings	2	84542	NA	NA	...	NA	...	NA
6	001	Newfoundland and Labrador	St. John's	CMA	Population and dwelling counts	Private dwellings occupied by usual residents	3	78960	NA	NA	...	NA	...	NA
7	001	Newfoundland and Labrador	St. John's	CMA	Population and dwelling counts	Population density per square kilometre	NA	244.8	NA	NA	...	NA	...	NA
8	001	Newfoundland and Labrador	St. John's	CMA	Population and dwelling counts	Land area (square km)	NA	804.65	NA	NA	...	NA	...	NA
9	001	Newfoundland and Labrador	St. John's	CMA	Age characteristics	Total population by age groups	4	196965	NA	94730	NA	102230	NA	NA
10	001	Newfoundland and Labrador	St. John's	CMA	Age characteristics	0 to 4 years	NA	10725	NA	5440	NA	5285	NA	NA
11	001	Newfoundland and Labrador	St. John's	CMA	Age characteristics	5 to 9 years	NA	10225	NA	5285	NA	4945	NA	NA
12	001	Newfoundland and Labrador	St. John's	CMA	Age characteristics	10 to 14 years	NA	10300	NA	5280	NA	5015	NA	NA
13	001	Newfoundland and Labrador	St. John's	CMA	Age characteristics	15 to 19 years	NA	11325	NA	5660	NA	5660	NA	NA
14	001	Newfoundland and Labrador	St. John's	CMA	Age characteristics	15 years	NA	2170	NA	1085	NA	1090	NA	NA
15	001	Newfoundland and Labrador	St. John's	CMA	Age characteristics	16 years	NA	2130	NA	1100	NA	1030	NA	NA
16	001	Newfoundland and Labrador	St. John's	CMA	Age characteristics	17 years	NA	2120	NA	1060	NA	1055	NA	NA
17	001	Newfoundland and Labrador	St. John's	CMA	Age characteristics	18 years	NA	2330	NA	1160	NA	1170	NA	NA

Showing 1 to 20 of 73,161 entries, 14 total columns

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*
- *<u>There are 2,239 other lines</u> 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:

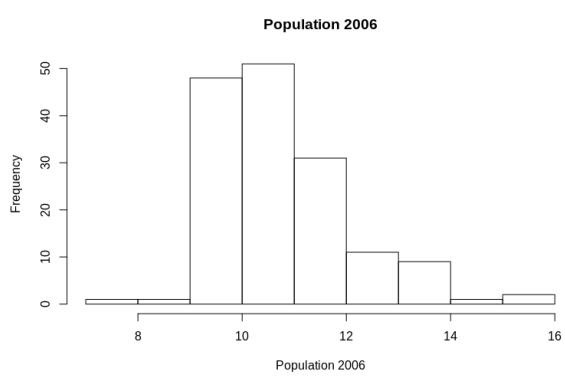
Filter

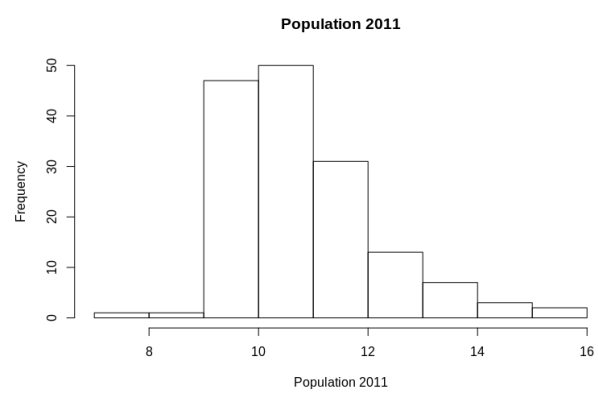
	Cities	Population in 2006	Population in 2011	LogPop2006	LogPop2011
1	Abbotsford - Mission	159020	170191	11.976785	12.044677
2	Alma	31864	33018	10.369232	10.404808
3	Amos	17176	17090	9.751268	9.746249
4	Baie-Comeau	29674	28789	10.298027	10.267749
5	Barrie	177061	187013	12.084250	12.138933
6	Bathurst	34106	33484	10.437229	10.418823
7	Bay Roberts	10507	10871	9.259797	9.293854
8	Belleville	91518	92540	11.424291	11.435396
9	Brandon	48256	53229	10.784275	10.882359
10	Brantford	124607	135501	11.732920	11.816734
11	Brockville	39668	39024	10.588300	10.571932
12	Brooks	22452	23430	10.019135	10.061773
13	Calgary	1079310	1214839	13.891833	14.010122
14	Campbell River	34707	36096	10.454697	10.493937
15	Campbellton	17878	17842	9.791326	9.789311
16	Campbellton (New Brunswick part)	14816	14520	9.603463	9.583282
17	Campbellton (Quebec part)	3062	3322	8.026824	8.108322
18	Camrose	15630	17286	9.656947	9.757652
19	Canmore	12039	12288	9.395907	9.416378
20	Cape Breton	105928	101619	11.570515	11.528986
21	Centre Wellington	26049	26693	10.167735	10.192157
22	Charlottetown	59325	64487	10.990786	11.074219
23	Chatham-Kent	108589	104075	11.595325	11.552867

Showing 1 to 26 of 155 entries, 5 total columns

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

```
> summary(Pop06_11Final)
   Cities           Population in 2006 Population in 2011
 Length:155         Min.   :   1398    Min.   :   1577
 Class :character   1st Qu.:  16844    1st Qu.:  17762
 Mode  :character   Median :  36288    Median :  37754
                    Mean   : 173298    Mean   : 185368
                    3rd Qu.:  92579    3rd Qu.:  98388
                    Max.   :5113149    Max.   :5583064
```





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 Commercial rents services price index, monthly. Statistics Canada Table: 1810025501-eng (Statistics Canada)

Initial Review

```
> summary(Rent_Growth)
   REF_DATE           GEO               DGUID           Building Type          UOM                UOM_ID
 Length:1956        Length:1956        Length:1956        Length:1956        Length:1956        Min.   :401
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:401
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :401
                                                                                                Mean   :401
                                                                                                3rd Qu.:401
                                                                                                Max.   :401
 SCALAR_FACTOR        SCALAR_ID    VECTOR            COORDINATE        VALUE          STATUS         SYMBOL
 Length:1956        Min.   :0   Length:1956        Min.   : 1.10   Min.   : 80.70   Mode:logical   Mode:logical
 Class :character   1st Qu.:0   Class :character   1st Qu.:10.10   1st Qu.: 98.20   NA's:1956      NA's:1956
 Mode  :character   Median :0   Mode  :character   Median :16.30   Median : 99.80
                    Mean   :0                      Mean   :16.98   Mean   : 98.79
                    3rd Qu.:0                      3rd Qu.:26.20   3rd Qu.:100.70
                    Max.   :0                      Max.   :31.10   Max.   :109.90
 TERMINATED        DECIMALS
 Mode:logical   Min.   :1
 NA's:1956      1st Qu.:1
                Median :1
                Mean   :1
                3rd Qu.:1
                Max.   :1
```

```
> str(Rent_Growth)
tibble [1,956 × 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ REF_DATE     : chr [1:1956] "2006-01" "2006-02" "2006-03" "2006-04" ...
 $ GEO          : chr [1:1956] "Canada" "Canada" "Canada" "Canada" ...
 $ DGUID        : chr [1:1956] "2016A000011124" "2016A000011124" "2016A000011124" "2016A000011124" ...
 $ Building Type: chr [1:1956] "Total, building type" "Total, building type" "Total, building type" "Total, building type" ...
 $ UOM          : chr [1:1956] "2019=100" "2019=100" "2019=100" "2019=100" ...
 $ UOM_ID       : num [1:1956] 401 401 401 401 401 401 401 401 401 401 ...
 $ SCALAR_FACTOR: chr [1:1956] "units" "units" "units" "units" ...
 $ SCALAR_ID    : num [1:1956] 0 0 0 0 0 0 0 0 0 0 ...
 $ VECTOR       : chr [1:1956] "v1210497010" "v1210497010" "v1210497010" "v1210497010" ...
 $ COORDINATE   : num [1:1956] 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
 $ VALUE        : num [1:1956] 80.9 80.7 81.2 81.7 81.5 81.8 81.1 81 81.2 81.2 ...
 $ STATUS       : logi [1:1956] NA NA NA NA NA NA ...
 $ SYMBOL       : logi [1:1956] NA NA NA NA NA NA ...
 $ TERMINATED   : logi [1:1956] NA NA NA NA NA NA ...
 $ DECIMALS     : num [1:1956] 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "spec")=
  .. cols(
  ..   REF_DATE = col_character(),
  ..   GEO = col_character(),
  ..   DGUID = col_character(),
  ..   `Building Type` = col_character(),
  ..   UOM = col_character(),
  ..   UOM_ID = col_double(),
  ..   SCALAR_FACTOR = col_character(),
  ..   SCALAR_ID = col_double(),
  ..   VECTOR = col_character(),
  ..   COORDINATE = col_double(),
  ..   VALUE = col_double(),
  ..   STATUS = col_logical(),
  ..   SYMBOL = col_logical(),
  ..   TERMINATED = col_logical(),
  ..   DECIMALS = col_double()
  .. )
```

Analyses and Preparation

Income × | Population × | Pop_2016Census × | Rent_Growth ×

Filter

	REF_DATE	GEO	DGUID	Building Type	UOM	UOM_ID	SCALAR_FACTOR	SCALAR_ID	VECTOR	COORDINATE	VALUE	STATUS	SYMBOL	TERMINATED	DECIMALS
1	2006-01	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	80.9	NA	NA	NA	1
2	2006-02	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	80.7	NA	NA	NA	1
3	2006-03	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.2	NA	NA	NA	1
4	2006-04	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.7	NA	NA	NA	1
5	2006-05	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.5	NA	NA	NA	1
6	2006-06	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.8	NA	NA	NA	1
7	2006-07	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.1	NA	NA	NA	1
8	2006-08	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.0	NA	NA	NA	1
9	2006-09	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.2	NA	NA	NA	1
10	2006-10	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.2	NA	NA	NA	1
11	2006-11	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.2	NA	NA	NA	1
12	2006-12	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	82.2	NA	NA	NA	1
13	2007-01	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.7	NA	NA	NA	1
14	2007-02	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.9	NA	NA	NA	1
15	2007-03	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	82.1	NA	NA	NA	1
16	2007-04	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.3	NA	NA	NA	1
17	2007-05	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.4	NA	NA	NA	1
18	2007-06	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	81.2	NA	NA	NA	1
19	2007-07	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	82.6	NA	NA	NA	1
20	2007-08	Canada	2016A000011124	Total, building type	2019=100	401	units	0	v1210497010	1.1	82.7	NA	NA	NA	1

Showing 1 to 24 of 1,956 entries, 15 total columns

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 <u>our period will be limited to Jan 2019 to Jun 2021 only</u> 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

Filter

	REF_DATE	GEO	Building Type	VALUE
1	2019-01	St. John's, Newfoundland and Labrador	Total, building type	100.3
2	2019-01	Charlottetown, Prince Edward Island	Total, building type	100.2
3	2019-01	Halifax, Nova Scotia	Total, building type	100.2
4	2019-01	Saint John, New Brunswick	Total, building type	99.7
5	2019-01	Montréal, Quebec	Total, building type	98.9
6	2019-01	Montréal, Quebec	Office buildings	100.3
7	2019-01	Montréal, Quebec	Retail buildings	97.7
8	2019-01	Montréal, Quebec	Industrial buildings and warehouses	98.5
9	2019-01	Québec, Quebec	Total, building type	99.4
10	2019-01	Ottawa-Gatineau, Ontario Part, Ontario/Quebec	Total, building type	99.6
11	2019-01	Toronto, Ontario	Total, building type	98.6
12	2019-01	Toronto, Ontario	Office buildings	99.4
13	2019-01	Toronto, Ontario	Retail buildings	98.0
14	2019-01	Toronto, Ontario	Industrial buildings and warehouses	98.4
15	2019-01	Winnipeg, Manitoba	Total, building type	99.1
16	2019-01	Saskatoon, Saskatchewan	Total, building type	100.0
17	2019-01	Calgary, Alberta	Total, building type	100.1
18	2019-01	Calgary, Alberta	Office buildings	101.3
19	2019-01	Calgary, Alberta	Retail buildings	99.6
20	2019-01	Calgary, Alberta	Industrial buildings and warehouses	99.3
21	2019-01	Edmonton, Alberta	Total, building type	99.7
22	2019-01	Vancouver, British Columbia	Total, building type	99.2

Showing 1 to 26 of 750 entries, 4 total columns

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

Filter

	Cities	Population	HH Income	Rent
1	St. John's	205955	102864	99.0
2	Bay Roberts	11083	79714	0.0
3	Grand Falls-Windsor	14171	77120	0.0
4	Gander	13234	88037	0.0
5	Corner Brook	31917	83095	0.0
6	Charlottetown	69325	82252	100.2
7	Summerside	16587	70290	0.0
8	Halifax	403390	91252	97.2
9	Kentville	26222	70620	0.0
10	Truro	45753	72221	0.0
11	New Glasgow	34487	73062	0.0
12	Cape Breton	98722	71467	0.0
13	Moncton	144810	80407	0.0
14	Saint John	126202	82988	100.3
15	Fredericton	101760	84630	0.0
16	Bathurst	31110	70979	0.0
17	Miramichi	27523	74069	0.0
18	Campbellton	15746	65106	0.0
19	Campbellton (New Brunswick part)	13114	67462	0.0
20	Campbellton (Quebec part)	2632	50688	0.0
21	Edmundston	23524	71399	0.0
22	Matane	17926	70434	0.0

Showing 1 to 26 of 320 entries, 4 total columns

```
> summary(PopCombined2)
    Cities            Population        HH Income           Rent
 Length:320         Min.   :      0   Min.   : 48288   Min.   :  0.00
 Class :character   1st Qu.:  16536   1st Qu.: 71827   1st Qu.:  0.00
 Mode  :character   Median :  34296   Median : 80564   Median :  0.00
                    Mean   : 178761   Mean   : 83118   Mean   :  7.47
                    3rd Qu.:  96794   3rd Qu.: 90999   3rd Qu.:  0.00
                    Max.   :5928040   Max.   :220888   Max.   :105.00
```

```
> str(PopCombined2)
tibble [320 × 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Cities    : chr [1:320] "St. John's" "Bay Roberts" "Grand Falls-Windsor" "Gander" ...
 $ Population: num [1:320] 205955 11083 14171 13234 31917 ...
 $ HH Income : num [1:320] 102864 79714 77120 88037 83095 ...
 $ Rent      : num [1:320] 99 0 0 0 0 ...
 - attr(*, "spec")=
  .. cols(
  ..   Cities = col_character(),
  ..   Population = col_double(),
  ..   `HH Income` = col_double(),
  ..   Rent = col_double()
  .. )
```

Correlation Test: Pop and Price; HHIncome and Price

```
> test1<-cor.test(PopCombined2$Population, PopCombined2$Rent,method = "pearson")
> test1

        Pearson's product-moment correlation

data:  PopCombined2$Population and PopCombined2$Rent
t = 13.032, df = 318, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5136216 0.6571620
sample estimates:
      cor
0.5900347
```

```
> test1<-cor.test(PopCombined2$`HH Income`, PopCombined2$Rent,method = "pearson")
> test1

        Pearson's product-moment correlation

data:  PopCombined2$`HH Income` and PopCombined2$Rent
t = 1.7142, df = 318, p-value = 0.08747
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0141013  0.2031949
sample estimates:
       cor
0.09568666
```

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

Cross-Validation

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)

```
> Folds<-createFolds(PopCombined2$Cities)
> str(Folds)
List of 10
 $ Fold01: int [1:33] 47 48 65 66 70 91 95 96 102 108 ...
 $ Fold02: int [1:33] 9 33 37 45 50 60 61 71 74 76 ...
 $ Fold03: int [1:40] 10 15 21 24 27 30 41 55 59 62 ...
 $ Fold04: int [1:29] 3 19 20 23 32 43 63 90 94 100 ...
 $ Fold05: int [1:21] 11 12 16 42 54 72 77 85 87 93 ...
 $ Fold06: int [1:25] 7 14 17 25 38 51 52 67 78 80 ...
 $ Fold07: int [1:31] 1 4 35 36 40 57 99 117 120 133 ...
 $ Fold08: int [1:37] 8 22 26 28 31 68 81 86 89 98 ...
 $ Fold09: int [1:36] 5 13 29 34 39 49 53 58 73 75 ...
 $ Fold10: int [1:35] 2 6 18 44 46 56 69 83 106 111 ...
```

For (f in Folds) {

train<-PopCombined2[-f,]

Filter

	Cities	Population	HH Income	Rent
1	St. John's	205955	102864	99.0
2	Grand Falls-Windsor	14171	77120	0.0
3	Gander	13234	88037	0.0
4	Corner Brook	31917	83095	0.0
5	Summerside	16587	70290	0.0
6	Halifax	403390	91252	97.2
7	Kentville	26222	70620	0.0
8	Truro	45753	72221	0.0
9	New Glasgow	34487	73062	0.0
10	Cape Breton	98722	71467	0.0
11	Moncton	144810	80407	0.0
12	Saint John	126202	82988	100.3
13	Fredericton	101760	84630	0.0
14	Bathurst	31110	70979	0.0
15	Miramichi	27523	74069	0.0
16	Campbellton (New Brunswick part)	13114	67462	0.0
17	Campbellton (Quebec part)	2632	50688	0.0
18	Edmundston	23524	71399	0.0
19	Matane	17926	70434	0.0
20	Rimouski	55349	79709	0.0
21	Rivière-du-Loup	28902	77979	0.0
22	Baie-Comeau	27692	88168	0.0

Showing 1 to 25 of 285 entries, 4 total columns

test<-PopCombined2[f,] }

Show Attributes

Name	Type	Value
test1	list [9] (S3: htest)	List of length 9
statistic	double [1]	1.714203
parameter	integer [1]	318
p.value	double [1]	0.08746579
estimate	double [1]	0.09568666
null.value	double [1]	0
alternative	character [1]	'two.sided'
method	character [1]	'Pearson\'s product-moment correlation'
data.name	character [1]	'PopCombined2$`HH Income` and PopCombined2$Rent'
conf.int	double [2]	-0.0141 0.2032

Modeling

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

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

normalize<-function(x) { return ((x – min(x)) / max(x) – min(x)))}

Apply the normalize function to the dataset:

PopCombined2_n<- as.data.frame(lapply(PopCombined2[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(PopCombined2_n), 0.7 * nrow(PopCombined2_n)

train.set<-PopCombined2_n[train_index]

test.set<-PopCombined2_n[-train_index]

Remove “Cities” column from training and test datasets
train.set_new<-train.set[-1]

test.set_new<-test.set[-1]

Store the labels for our training and test sets

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

Model Evaluation

We will measure performance via Confusion Matrix.

Confusion Matrix

```
CrossTable(x = PopCombined2_test_labels, y = PopCombined2_knn_prediction, prop.chisq =
FALSE)
```

Interpretation of Results