# MODULE 1 R PRACTICE

Week 1


## Abstract

In this assignment, I will use everything that I learned so far to analyze data to complete a data rich and visually appealing report



Mohammad Hossein Movahedi

Movahedi.m@northeastern.edu

## Introduction

For this assignment, I will use the data from Kaggle (Wong, 2021). The MLS.csv dataset and use the APA system for the bibliography. Also, the R code that I used to process the data is available on my GitHub account, which address is mentioned in the bibliography part. I also put my code in the appendix part.

The objective is to learn what the data is saying and its attributes. Secondly, to make visually appealing plots.

## Content and context of the dataset

- Location - Neighbourhood in the Greater Toronto Area
- Comp - "Comp" stands for composite and considers the various housing types into a single value.
- SFDetach - "SFDetach" stands for Single Family Detached Home, or commonly referred to as houses
- SFAttach - "SFAttach" stands for Single Family Attached Home
- THouse - "THouse" stands for Townhouses
- Apart - "Apart" is the abbreviation for Apartments or Condominiums
- All prices mentioned under "Benchmark" columns are depicted in Canadian Dollars
- All YoY Changes are in the context of "Percentages."

The MLS® Home Price Index (MLS® HPI) examines the levels and trends of home prices in a given neighbourhood. A sophisticated statistical model is used to calculate the index. It considers quantitative (e.g., the number of rooms) and qualitative (e.g., basement access and finishes) aspects of a property.

The MLS® HPI is calculated using the value house purchasers place on various dwelling characteristics, which change over time. It compares house values throughout the United States "apples to apples," allowing for a more accurate comparison. (Super User, 2022)

## Methodology

This assignment is divided into three parts, each dealing with one of the aspects of the project.

### Part 1: the variables of interest

This dataset has 17 variables, and it is a vast dataset; therefore, I choose only to use single-family detached and attached variables across locations.

```
 > print(names(means))

[1] "Location"          "SFDetachIndex"     "SFDetachBenchmark" "SFDetachYoYChange"

[5] "SFAttachIndex"     "SFAttachBenchmark" "SFAttachYoYChange"
```

### Part 2: data cleaning

For this part, I corrected data formats and also deleted duplicated rows.

```
 # importing database for this project aasignment I used Toronto Home Price Index dataset

thpi <- read.csv("MLS.csv")

# cheaking dataset structure

str(thpi)

#change date format from string to date and also location as factor

thpi$Date <- as.Date(thpi$Date,"%Y-%d-%m")

thpi$Location <- as.factor(thpi$Location)

thpi[c("CompYoYChange", "SFDetachYoYChange","SFAttachYoYChange","ApartYoYChange")] <-
as.numeric()

# also some clos are in presentages so I change them too

thpi[c("CompYoYChange", "SFDetachYoYChange","SFAttachYoYChange","ApartYoYChange")] <-
lapply(thpi[c("CompYoYChange", "SFDetachYoYChange","SFAttachYoYChange","ApartYoYChange")],


function(x){x/100})
```

```
#I am ok with names so I wont change them

#deleting duplicated datas

thpi[!duplicated(thpi), ]

summary(thpi)
```

## Part 3: initial analysis

For this part, I calculated the mean of every variable grouped by location and then cleaned outliers multiple times until there were no outliers in the dataset.

```
 # I group my data by locations so I can make better plots

th <-
thpi[c("Location","SFDetachIndex","SFDetachBenchmark","SFDetachYoYChange","SFAttachIndex",

             "SFAttachBenchmark","SFAttachYoYChange")]

th <- as_tibble(th)

means <- th %>%

  group_by(Location) %>%

  dplyr::summarise_all(mean,na.rm = TRUE)

# from now on I only use means as my dataset

attach(means)









# calculating IQR and Q

IQR(c(means[,2]),na.rm=T)

Q <- means %>%

 summarise_if(is.numeric,~quantile(.,probs=c(.25, .75), na.rm =T))



iqr <- means %>%

  summarise_if(is.numeric,IQR,na.rm=T)



t<- rbind(Q,iqr)
```

```
up<- c(t[2,] +1.5*t[3,])

low <- c(t[1,] -1.5*t[3,])

t <-rbind(t,up)

t <-rbind(t,low)

# starting to clean

means_cl <- means

t

means_cl <- rm.outlier(means[2:7], fill = TRUE, median = FALSE, opposite = FALSE)

boxplot(means_cl)$out

# the below process has be repeted until no outlier is detected

means_cl <- rm.outlier(means_cl[1:6], fill = TRUE, median = FALSE, opposite = FALSE)

boxplot(means_cl)$out

# adding the location column again

means_cl['Location']<- means$Location

mcl <- means_cl
```

## Part 4: the results

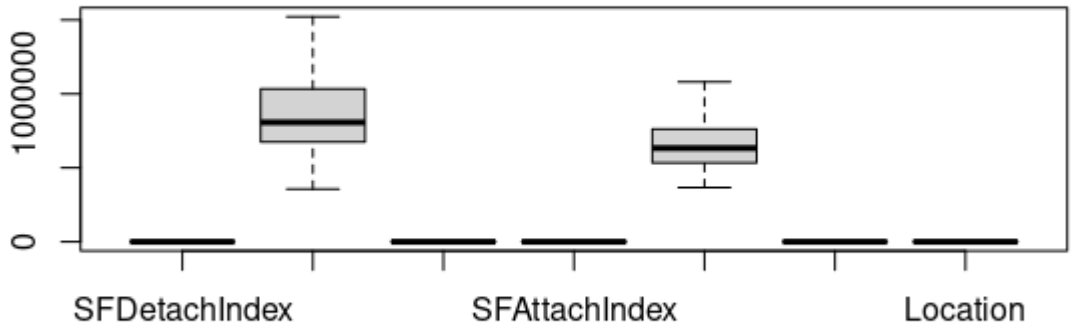
```
 > head(mcl,5)

  SFDetachIndex SFDetachBenchmark SFDetachYoYChange SFAttachIndex SFAttachBenchmark

1      213.1942          672089.9        0.10116957           NaN               NaN

2      239.8290          643097.1        0.10869130      248.4087          535508.7

3      256.8261          964318.8        0.09465362      260.4652          672268.1

4           NaN               NaN               NaN           NaN               NaN

5      224.7019          642750.0        0.10423889      247.6926          514855.6

  SFAttachYoYChange          Location

1               NaN Adjala-Tosorontio

2        0.11563188              Ajax

3        0.09392899            Aurora

4               NaN            Barrie
```

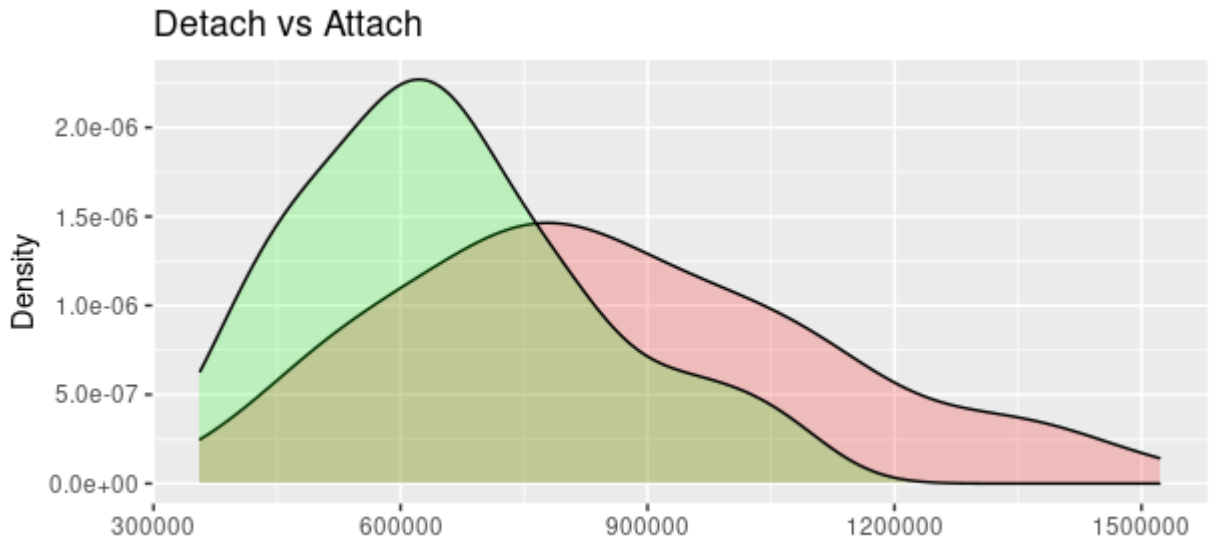
```
5         0.10812778      Bradford West
```

## Part 5: charts, tables and graphs created

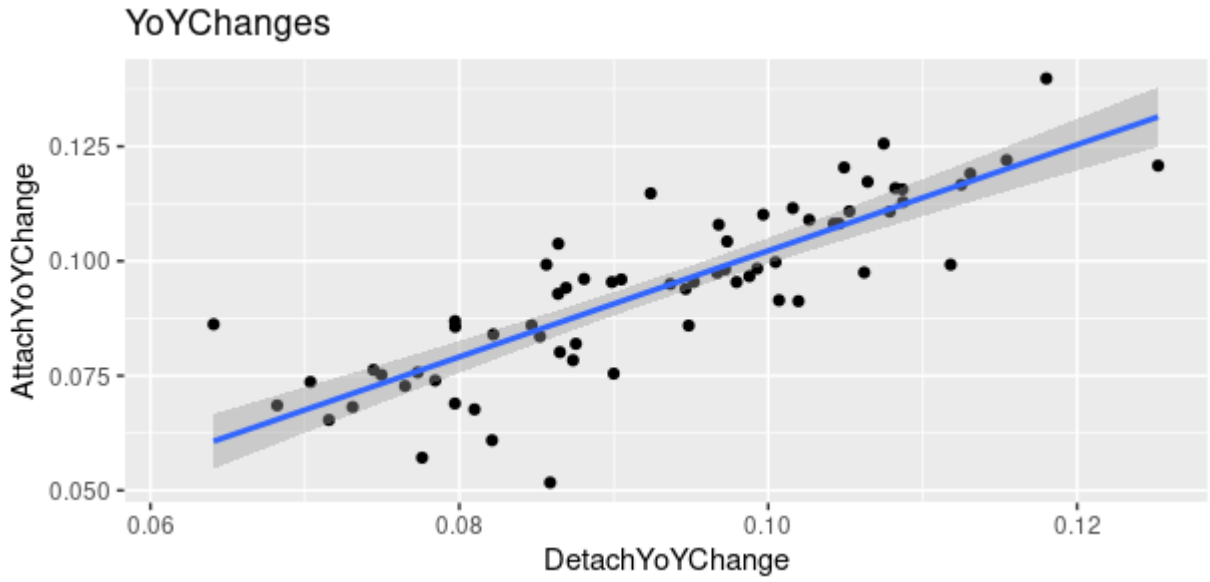
As can be seen in the plot box after data cleaning and Initial analysis, there are no outliers in the data set.

```
boxplot(means_cl)$out
```

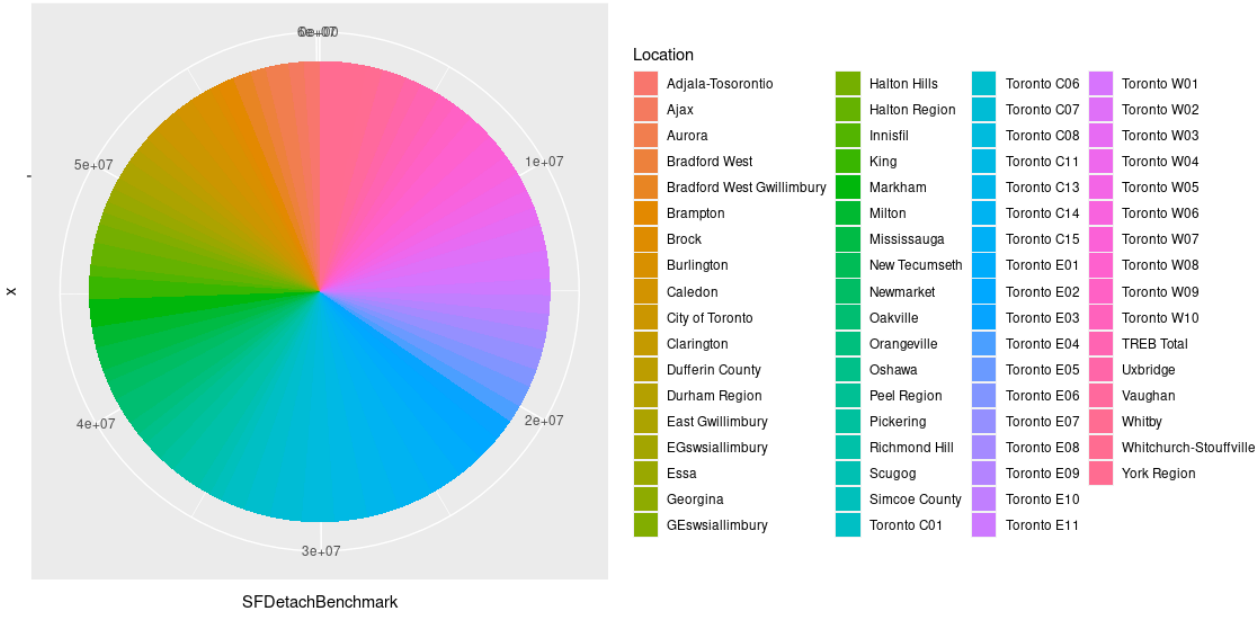



As seen in this density chart single-family detached benchmark (Red) has a much more comprehensive range than the single-family attached benchmark. Still, both means are about the same, which shows that houses with single-family attached are less dangerous than houses in general.




From the scatter plot of YoY changes, we can see that these two variables are related with reasonable confidence showing the price change in one can affect the other one.




As we can see from this database pie chart is not helpful.

Bibliography

Super User (2022). TRREB - MLS Home Price Index. [online] Trreb.ca. Available at: https://trreb.ca/index.php/market-news/mls-home-price-index [Accessed 2 Mar. 2022].

Wong, A. (2021). Toronto Home Price Index. [online] Kaggle.com. Available at: https://www.kaggle.com/alankmwong/toronto-home-price-index [Accessed 2 Mar. 2022].

Sthda.com. (2020). ggplot2 - Essentials - Easy Guides - Wiki - STHDA. [online] Available at: http://www.sthda.com/english/wiki/ggplot2-essentials [Accessed 2 Mar. 2022].

momova97 (2022). momova97/ALY6010_Movahedi: This is the place that I will keep my projects R code. [online] GitHub. Available at: https://github.com/momova97/ALY6010_Movahedi [Accessed 2 Mar. 2022].

Appendix

```
 #adding subheading and name

print("Mohammad Hossein Movahedi")

print("week 1 r practice")

#installing Important libraries

install.packages('FSA')

install.packages('FSAdata')

install.packages('magrittr')

install.packages('dplyr')

install.packages('tidyr')

install.packages('plyr')

install.packages('tidyverse')

install.packages('outliers')

install.packages('ggplot2')

library(ggplot2)

library(outliers)

library(FSA)

library(FSAdata)

library(magrittr)

library(dplyr)

library(tidyr)

library(plyr)

library(tidyverse)

library(scales)



# importing database for this project aasignment I used Toronto Home Price Index dataset

thpi <- read.csv("MLS.csv")

# cheaking dataset structure
```

```
str(thpi)

#change date format from string to date and also location as factor

thpi$Date <- as.Date(thpi$Date,"%Y-%d-%m")

thpi$Location <- as.factor(thpi$Location)

thpi[c("CompYoYChange", "SFDetachYoYChange","SFAttachYoYChange","ApartYoYChange")] <-
as.numeric()

# also some clos are in presentages so I change them too

thpi[c("CompYoYChange", "SFDetachYoYChange","SFAttachYoYChange","ApartYoYChange")] <-
lapply(thpi[c("CompYoYChange", "SFDetachYoYChange","SFAttachYoYChange","ApartYoYChange")],


function(x){x/100})

#I am ok with names so I wont change them

#deleting duplicated datas

thpi[!duplicated(thpi), ]

summary(thpi)



# I group my data by locations so I can make better plots

th <-
thpi[c("Location","SFDetachIndex","SFDetachBenchmark","SFDetachYoYChange","SFAttachIndex",

             "SFAttachBenchmark","SFAttachYoYChange")]

th <- as_tibble(th)

means <- th %>%

  group_by(Location) %>%

  dplyr::summarise_all(mean,na.rm = TRUE)

# from now on I only use means as my dataset

attach(means)







# calculating IQR and Q

IQR(c(means[,2]),na.rm=T)
```

```
Q <- means %>%

 summarise_if(is.numeric,~quantile(.,probs=c(.25, .75), na.rm =T))



iqr <- means %>%

  summarise_if(is.numeric,IQR,na.rm=T)



t<- rbind(Q,iqr)

up<- c(t[2,] +1.5*t[3,])

low <- c(t[1,] -1.5*t[3,])

t <-rbind(t,up)

t <-rbind(t,low)

# starting to clean

means_cl <- means

t

means_cl <- rm.outlier(means[2:7], fill = TRUE, median = FALSE, opposite = FALSE)

boxplot(means_cl)$out

# the below process has be repeted until no outlier is detected

means_cl <- rm.outlier(means_cl[1:6], fill = TRUE, median = FALSE, opposite = FALSE)

boxplot(means_cl)$out

# adding the location column again

means_cl['Location']<- means$Location

mcl <- means_cl

head(mcl,5)

#know we can get to the easy part

# histograms

AD <- data.frame(a = mcl$SFDetachBenchmark,b ="FD")

AD1 <-data.frame(a = mcl$SFAttachBenchmark,b = "FA")

AD <- rbind(AD,AD1)

ggplot(AD, aes(x=a))+
```

```
  geom_density(data=subset(AD,b == 'FD'),fill = "red", alpha = 0.2) +
  geom_density(data=subset(AD,b == 'FA'),fill = "green", alpha = 0.2)+
  labs(title ="Detach vs Attach", x="", y = "Density")
# scatter plot
ggplot(mcl, aes(x=SFDetachYoYChange, y=SFAttachYoYChange )) + geom_point()+
  labs(title ="YoYChanges", x="DetachYoYChange", y = "AttachYoYChange")+
  geom_smooth(method=lm)


#pie chart
bp<- ggplot(mcl, aes(x="", y=SFDetachBenchmark, fill=Location))+
  geom_bar(width = 1, stat = "identity")
bp + coord_polar("y", start=0)
```