CIND820 Markdown

R Markdown

#Install and load the appropriate packages/libraries for the project.

```
#install.packages("tidyverse")
#install.packages("ggplot2")
#install.packages("TSstudio")
#install.packages("dplyr")
#install.packages("corrplot")
#install.packages("janitor")
#install.packages("tidyr")
library(corrplot)

## Warning: package 'corrplot' was built under R version 4.0.2

## corrplot 0.84 loaded

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.2

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.2

## ── Attaching packages
─────────────────────────────────────────────────────────────────────────────
──────── tidyverse 1.3.0 ──

## ✓ tibble  3.0.1     ✓ dplyr   1.0.5
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.1
## ✓ purrr   0.3.4

## Warning: package 'tidyr' was built under R version 4.0.2

## Warning: package 'dplyr' was built under R version 4.0.2

## Warning: package 'forcats' was built under R version 4.0.2

## ── Conflicts
─────────────────────────────────────────────────────────────────────────────
─────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(TSstudio)
```

```
## Warning: package 'TSstudio' was built under R version 4.0.2
```

```
library(dplyr)
library(tidyr)
library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.0.2
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

Reading Data

```
mls<-read.csv ("/Users/tomkalantzis/Desktop/CIND820 FINAL PROJECT/Final
Project/MLS.csv",header=TRUE)
```

```
head(mls)
```

```
##            Location CompIndex CompBenchmark CompYoYChange SFDetachIndex
## 1 Adjala-Tosorontio     143.7        510000         12.00         143.7
## 2 Adjala-Tosorontio     140.8        499700         13.82         140.8
## 3 Adjala-Tosorontio     142.7        506400         15.83         142.7
## 4 Adjala-Tosorontio     138.4        491100         11.08         138.4
## 5 Adjala-Tosorontio     145.4        516000         16.51         145.4
## 6 Adjala-Tosorontio     150.3        533400         17.51         150.3
##   SFDetachBenchmark SFDetachYoYChange SFAttachIndex SFAttachBenchmark
## 1            510000             12.00            NA                NA
## 2            499700             13.82            NA                NA
## 3            506400             15.83            NA                NA
## 4            491100             11.08            NA                NA
## 5            516000             16.51            NA                NA
## 6            533400             17.51            NA                NA
##   SFAttachYoYChange THouseIndex THouseBenchmark THouseYoYChange ApartIndex
## 1                NA          NA              NA              NA         NA
## 2                NA          NA              NA              NA         NA
## 3                NA          NA              NA              NA         NA
## 4                NA          NA              NA              NA         NA
## 5                NA          NA              NA              NA         NA
## 6                NA          NA              NA              NA         NA
##   ApartBenchmark ApartYoYChange       Date
## 1             NA             NA 2015-07-01
## 2             NA             NA 2015-08-01
## 3             NA             NA 2015-09-01
## 4             NA             NA 2015-10-01
## 5             NA             NA 2015-11-01
## 6             NA             NA 2015-12-01
```

```
str(mls)
```

```
## 'data.frame':    4872 obs. of  17 variables:
##  $ Location         : chr  "Adjala-Tosorontio" "Adjala-Tosorontio"
"Adjala-Tosorontio" "Adjala-Tosorontio" ...
##  $ CompIndex        : num  144 141 143 138 145 ...
##  $ CompBenchmark    : num  510000 499700 506400 491100 516000 ...
##  $ CompYoYChange    : num  12 13.8 15.8 11.1 16.5 ...
##  $ SFDetachIndex    : num  144 141 143 138 145 ...
##  $ SFDetachBenchmark: num  510000 499700 506400 491100 516000 ...
##  $ SFDetachYoYChange: num  12 13.8 15.8 11.1 16.5 ...
##  $ SFAttachIndex    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ SFAttachBenchmark: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ SFAttachYoYChange: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ THouseIndex      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ THouseBenchmark  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ THouseYoYChange  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ ApartIndex       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ ApartBenchmark   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ ApartYoYChange   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Date             : chr  "2015-07-01" "2015-08-01" "2015-09-01" "2015-
10-01" ...

sum(is.na(mls))

## [1] 7297

summary(mls)

##    Location           CompIndex     CompBenchmark     CompYoYChange
##  Length:4872        Min.   :135.0   Min.   : 257000   Min.   :-19.580
##  Class :character   1st Qu.:215.7   1st Qu.: 582075   1st Qu.:  4.157
##  Mode  :character   Median :245.3   Median : 717600   Median :  9.220
##                     Mean   :241.8   Mean   : 762734   Mean   :  9.955
##                     3rd Qu.:266.6   3rd Qu.: 893225   3rd Qu.: 15.053
##                     Max.   :388.5   Max.   :2110000   Max.   : 46.780
##                     NA's   :12      NA's   :12        NA's   :12
##  SFDetachIndex   SFDetachBenchmark SFDetachYoYChange SFAttachIndex
##  Min.   :126.0   Min.   : 259100   Min.   :-21.320   Min.   :137.2
##  1st Qu.:218.7   1st Qu.: 676075   1st Qu.:  1.970   1st Qu.:219.1
##  Median :245.1   Median : 861100   Median :  8.130   Median :245.0
##  Mean   :242.9   Mean   : 936875   Mean   :  8.885   Mean   :244.1
##  3rd Qu.:267.9   3rd Qu.:1123600   3rd Qu.: 14.905   3rd Qu.:266.3
##  Max.   :385.4   Max.   :2536900   Max.   : 46.450   Max.   :709.0
##  NA's   :12      NA's   :12        NA's   :13        NA's   :136
##  SFAttachBenchmark SFAttachYoYChange   THouseIndex    THouseBenchmark
##  Min.   : 273300   Min.   :-100.000   Min.   :  0.0   Min.   :      0
##  1st Qu.: 537850   1st Qu.:   2.465   1st Qu.:208.5   1st Qu.: 435100
##  Median : 664400   Median :   8.345   Median :244.1   Median : 549500
##  Mean   : 709687   Mean   :   8.854   Mean   :242.0   Mean   : 575980
##  3rd Qu.: 824275   3rd Qu.:  14.810   3rd Qu.:275.2   3rd Qu.: 662600
##  Max.   :1561300   Max.   :  39.740   Max.   :437.6   Max.   :1750500
##  NA's   :136       NA's   :134        NA's   :1231    NA's   :1231
```

```
##  THouseYoYChange      ApartIndex    ApartBenchmark   ApartYoYChange
##  Min.   :-100.000   Min.   :106.0   Min.   :171400   Min.   :-10.600
##  1st Qu.:   3.973   1st Qu.:193.4   1st Qu.:375125   1st Qu.:  5.862
##  Median :   9.495   Median :236.8   Median :470750   Median :  9.960
##  Mean   :   9.303   Mean   :233.8   Mean   :478506   Mean   : 11.791
##  3rd Qu.:  15.515   3rd Qu.:268.9   3rd Qu.:562500   3rd Qu.: 15.620
##  Max.   :  56.490   Max.   :389.1   Max.   :973900   Max.   : 64.320
##  NA's   :1230       NA's   :1042    NA's   :1042     NA's   :1042
##      Date
##  Length:4872
##  Class :character
##  Mode  :character
##
##
##
##
```

Standard deviation for all the numerical attributes

```
sd(mls$THouseYoYChange ,na.rm=TRUE)
```

```
## [1] 14.67352
```

```
sd(mls$CompYoYChange ,na.rm=TRUE)
```

```
## [1] 9.052718
```

```
sd(mls$SFDetachBenchmark , na.rm=TRUE)
```

```
## [1] 372625.4
```

```
sd(mls$SFdetachYoYChange , na.rm=TRUE)
```

```
## [1] NA
```

```
sd(mls$SFAttachIndex , na.rm=TRUE)
```

```
## [1] 40.28068
```

```
sd(mls$SFAttachBenchmark, na.rm=TRUE)
```

```
## [1] 239959.3
```

```
sd(mls$THouseIndex, na.rm=TRUE)
```

```
## [1] 53.68013
```

```
sd(mls$THouseBenchmark, na.rm=TRUE)
```

```
## [1] 219344.5
```

```
sd(mls$ApartIndex, na.rm=TRUE)
```

```
## [1] 51.46811
```

```
sd(mls$ApartBenchmark, na.rm=TRUE)
```

```
## [1] 144789.1
```

```
sd(mls$ApartYoYChange, na.rm=TRUE)
```

```
## [1] 9.105863
```

```
sd(mls$ThouseYoYChange, na.rm=TRUE)
```

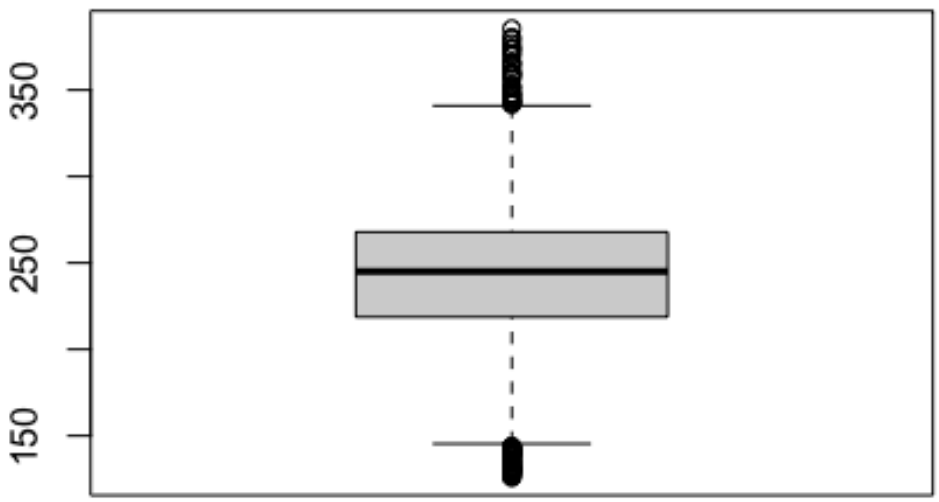
```
## [1] NA
```

```
sd(mls$SFAttachYoYChange, na.rm=TRUE)
```

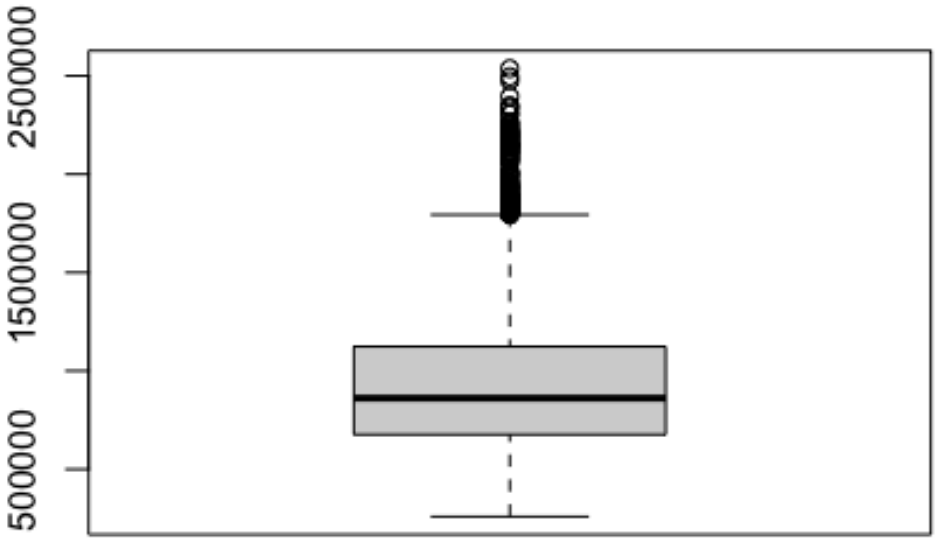
```
## [1] 9.921748
```

#Boxplots showing the outliers

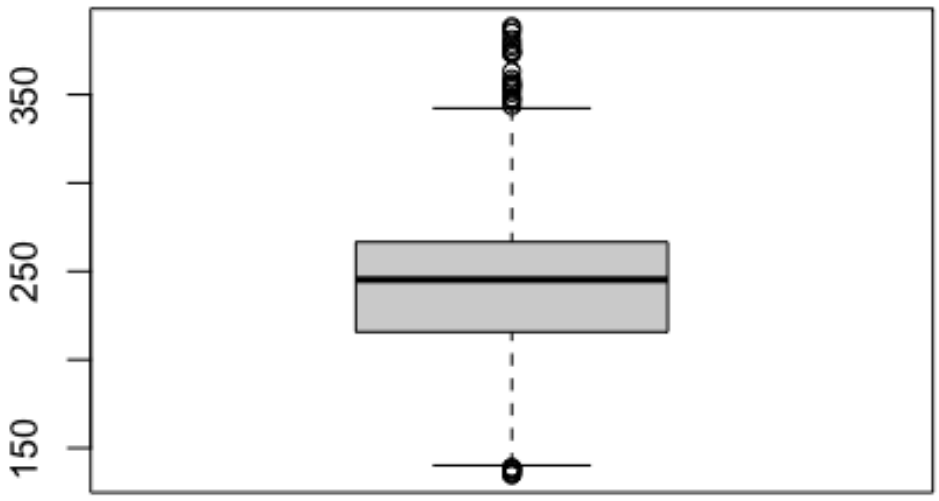
```
boxplot(mls$SFDetachIndex, na.rm=TRUE)
```



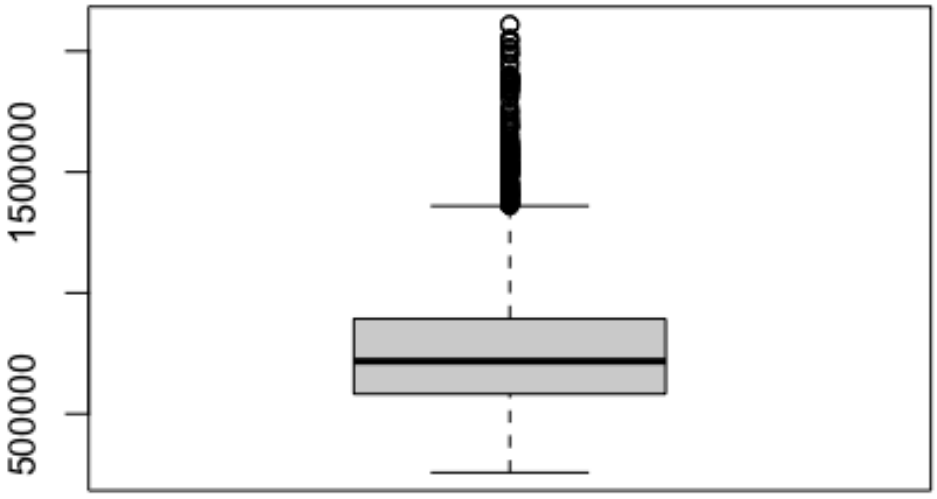

```
#boxplot(mls$CompyYoYChange, na.rm=TRUE)
boxplot(mls$SFDetachBenchmark, na.rm=TRUE)
```



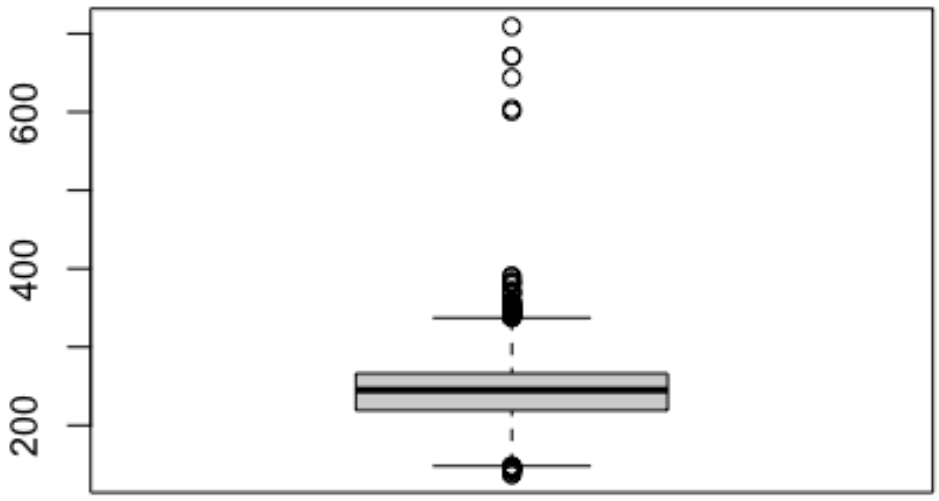

```
boxplot(mls$CompIndex, na.rm=TRUE)
```



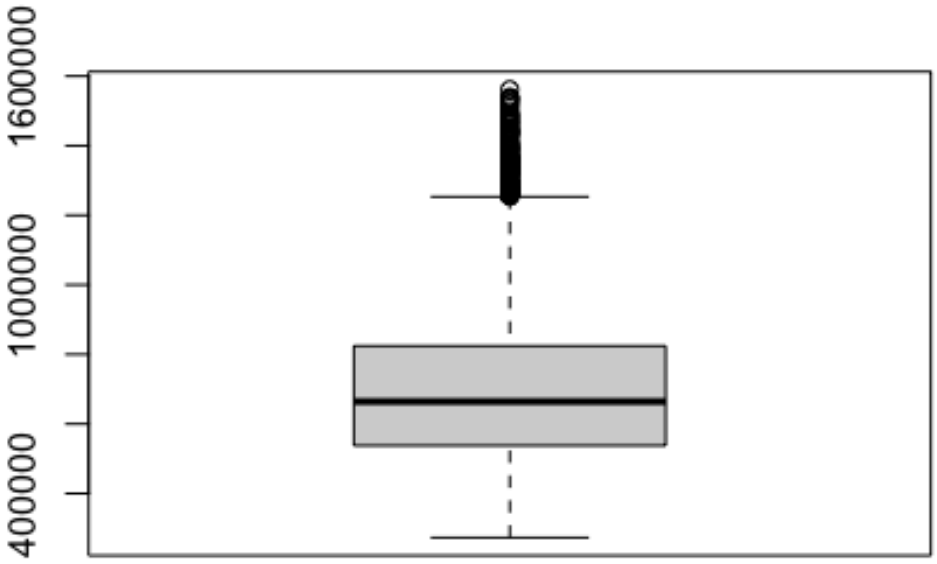

```
boxplot(mls$CompBenchmark, na.rm=TRUE)
```



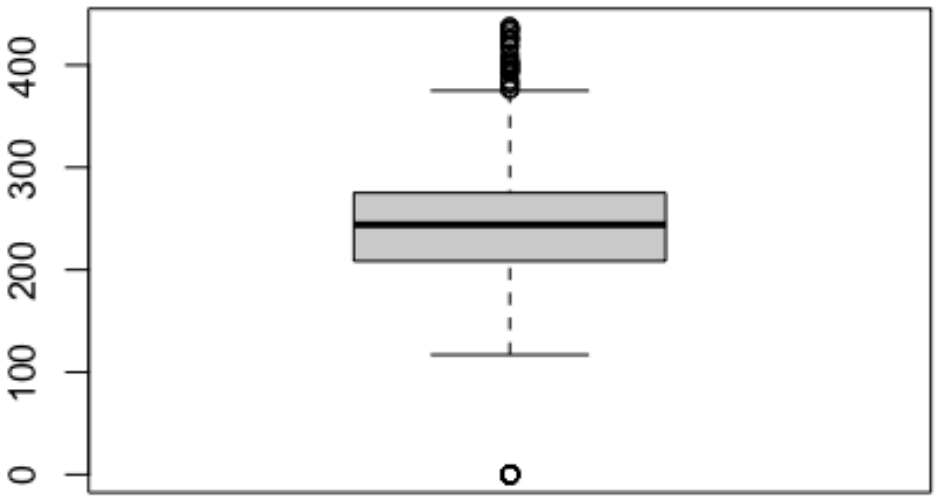

```
#boxplot(mls$CompYoYChange, na.rm=TRUE)
#boxplot(mls$SFDetachYoYChange, na.rm=TRUE)
boxplot(mls$SFAttachIndex, na.rm=TRUE)
```



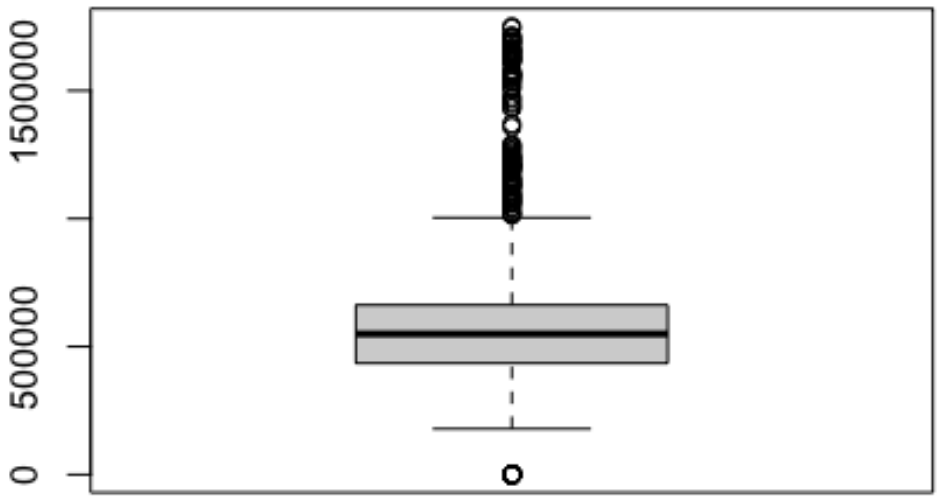

```
boxplot(mls$SFAttachBenchmark, na.rm=TRUE)
```



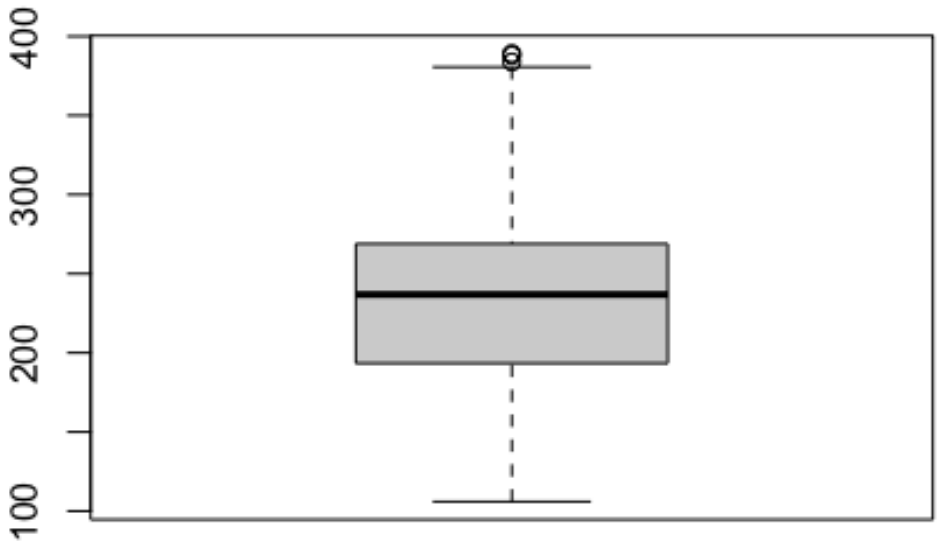

```
#boxplot(mls$SFAttachYoYChange, na.rm=TRUE)
boxplot(mls$THouseIndex, na.rm=TRUE)
```



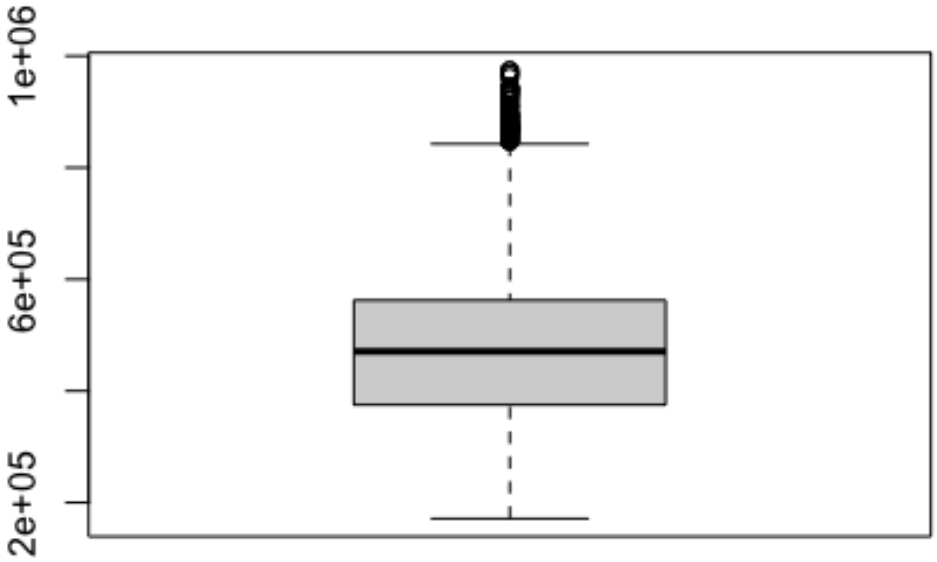

```r
boxplot(mls$THouseBenchmark, na.rm=TRUE)
```




```r
#boxplot(mls$THouseYoYChange, na.rm=TRUE)
boxplot(mls$ApartIndex, na.rm=TRUE)
```



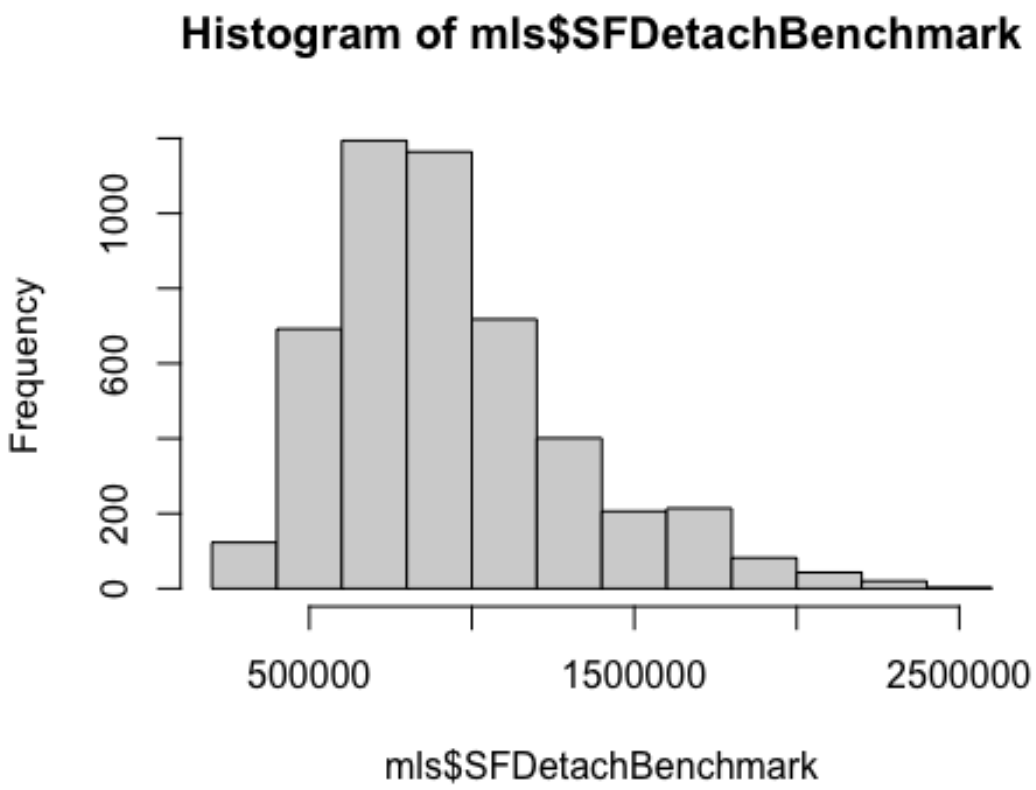

```
boxplot(mls$ApartBenchmark, na.rm=TRUE)
```



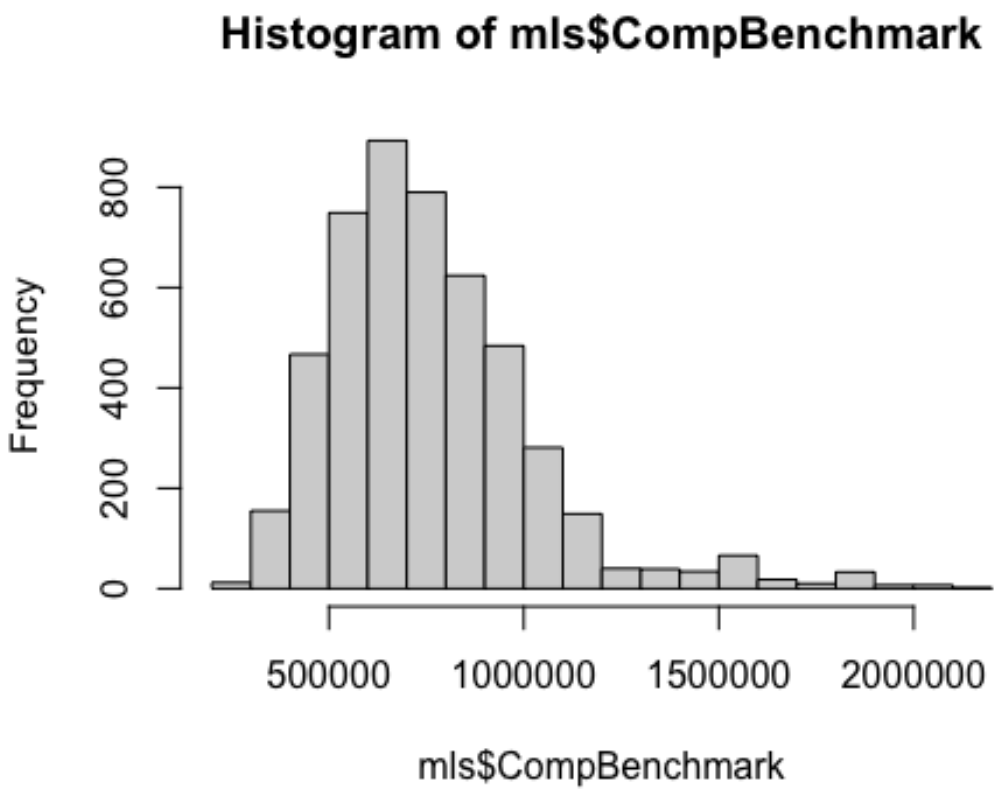

```
#boxplot(mls$ApartYoYChange, na.rm=TRUE)
```

#Histograms showing where the values are skewed

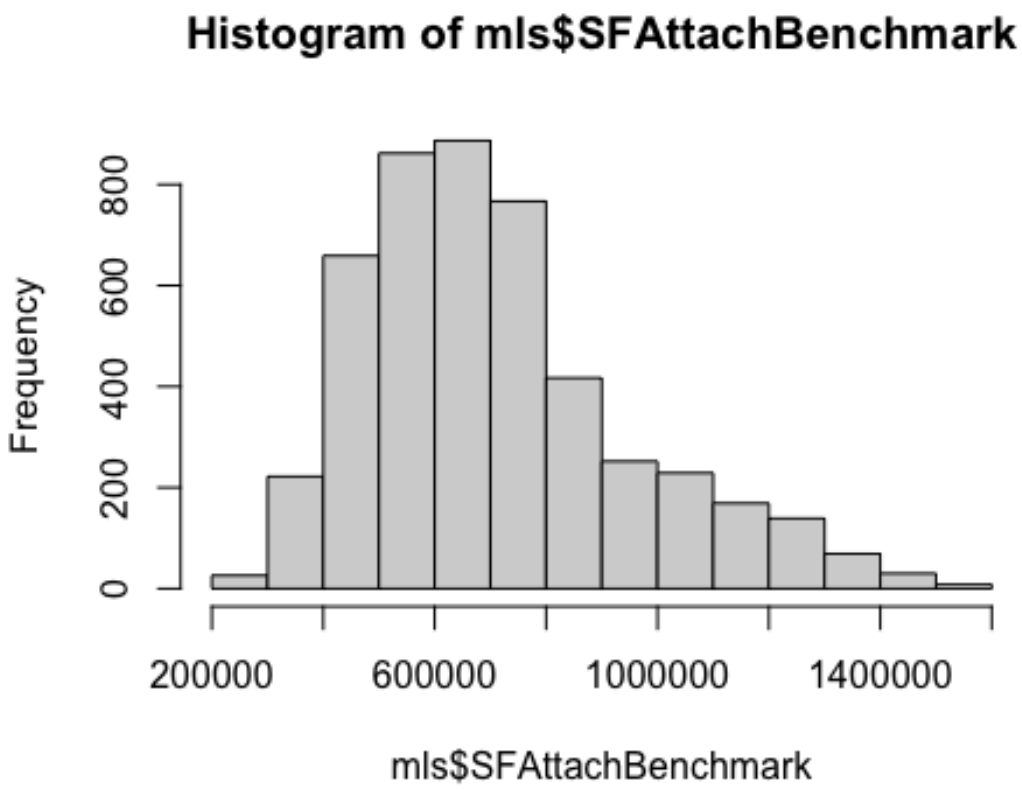
```
hist(mls$SFDetachBenchmark)
```



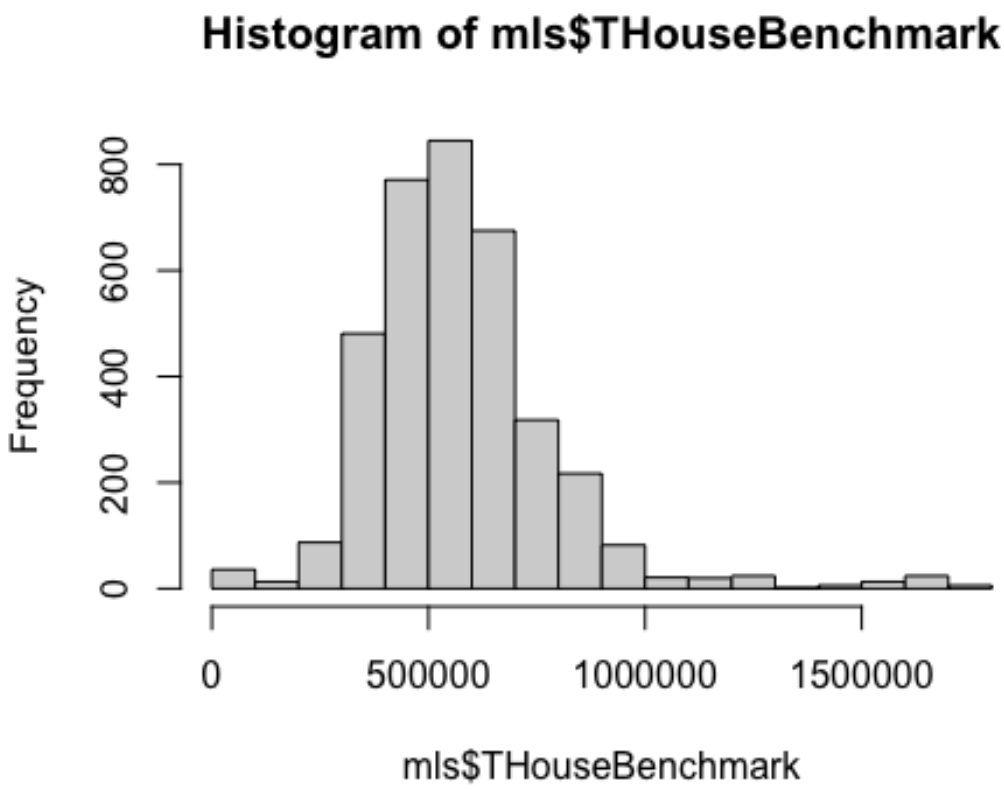

```
hist(mls$CompBenchmark)
```



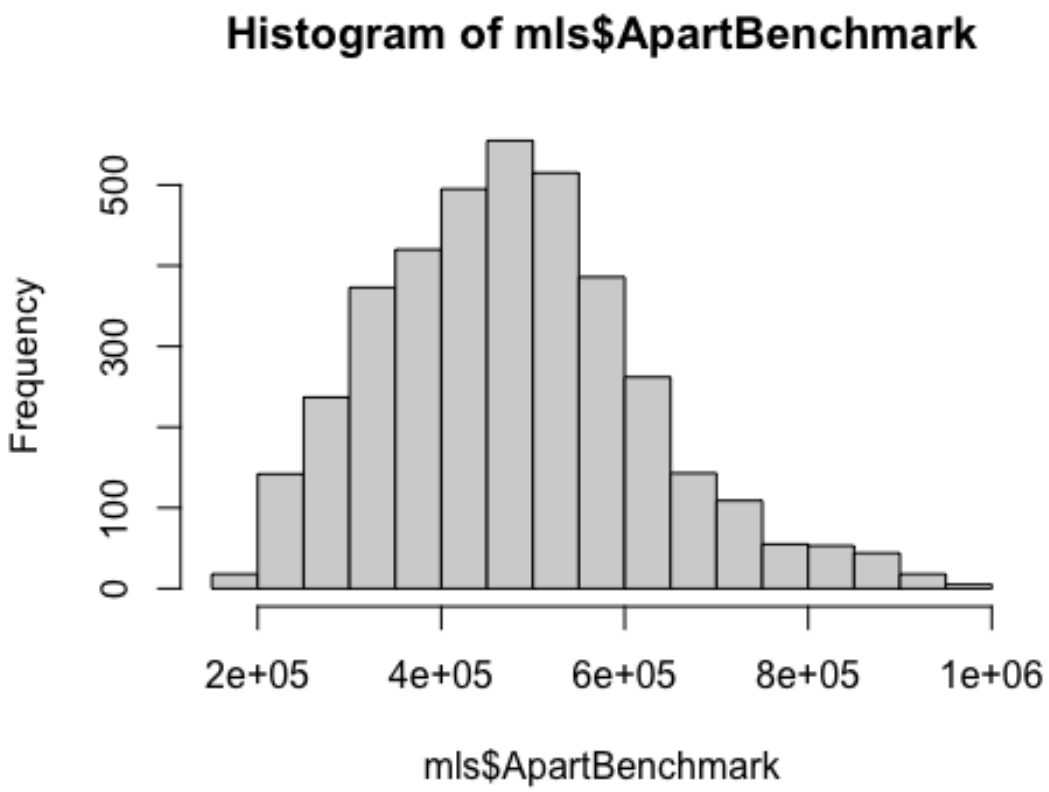

```
hist(mls$SFAttachBenchmark)
```




```r
hist(mls$THouseBenchmark)
```




```r
hist(mls$ApartBenchmark)
```



#Corellation between attributes

```
cor(mls$SFDetachBenchmark, mls$SFAttachBenchmark, use="complete.obs")
## [1] 0.9126085
cor(mls$SFDetachBenchmark, mls$THouseBenchmark, use="complete.obs")
## [1] 0.6922385
cor(mls$SFDetachBenchmark, mls$ApartBenchmark,  use="complete.obs")
## [1] 0.7092716
cor(mls$SFAttachBenchmark,mls$SFDetachBenchmark , use="complete.obs")
## [1] 0.9126085
cor(mls$SFAttachBenchmark,mls$THouseBenchmark , use="complete.obs")
## [1] 0.7126003
cor(mls$SFAttachBenchmark, mls$ApartBenchmark, use="complete.obs")
## [1] 0.7067582
cor(mls$THouseBenchmark,mls$SFDetachBenchmark , use="complete.obs")
```

```
## [1] 0.6922385
```

```
cor(mls$THouseBenchmark,mls$SFAttachBenchmark , use="complete.obs")
```

```
## [1] 0.7126003
```

```
cor(mls$THouseBenchmark,mls$ApartBenchmark , use="complete.obs")
```

```
## [1] 0.6714359
```

```
cor(mls$ApartBenchmark,mls$SFDetachBenchmark , use="complete.obs")
```

```
## [1] 0.7092716
```

```
cor(mls$ApartBenchmark, mls$SFAttachBenchmark, use="complete.obs")
```

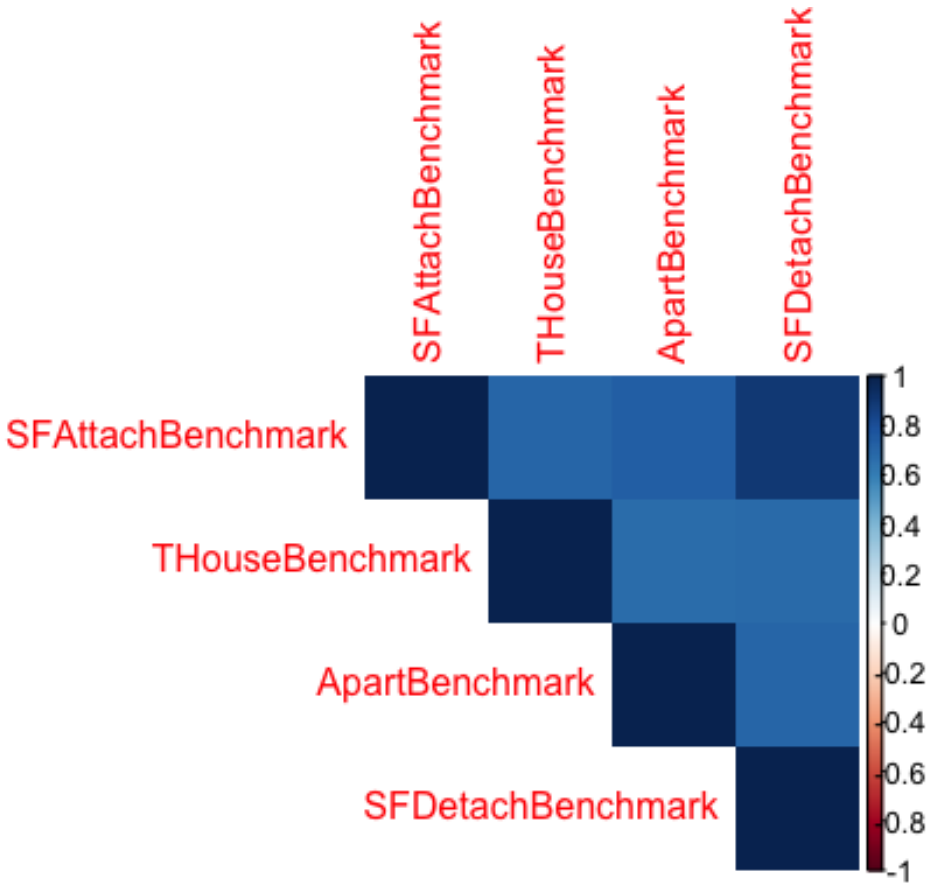
```
## [1] 0.7067582
```

```
cor(mls$ApartBenchmark,mls$THouseBenchmark , use="complete.obs")
```

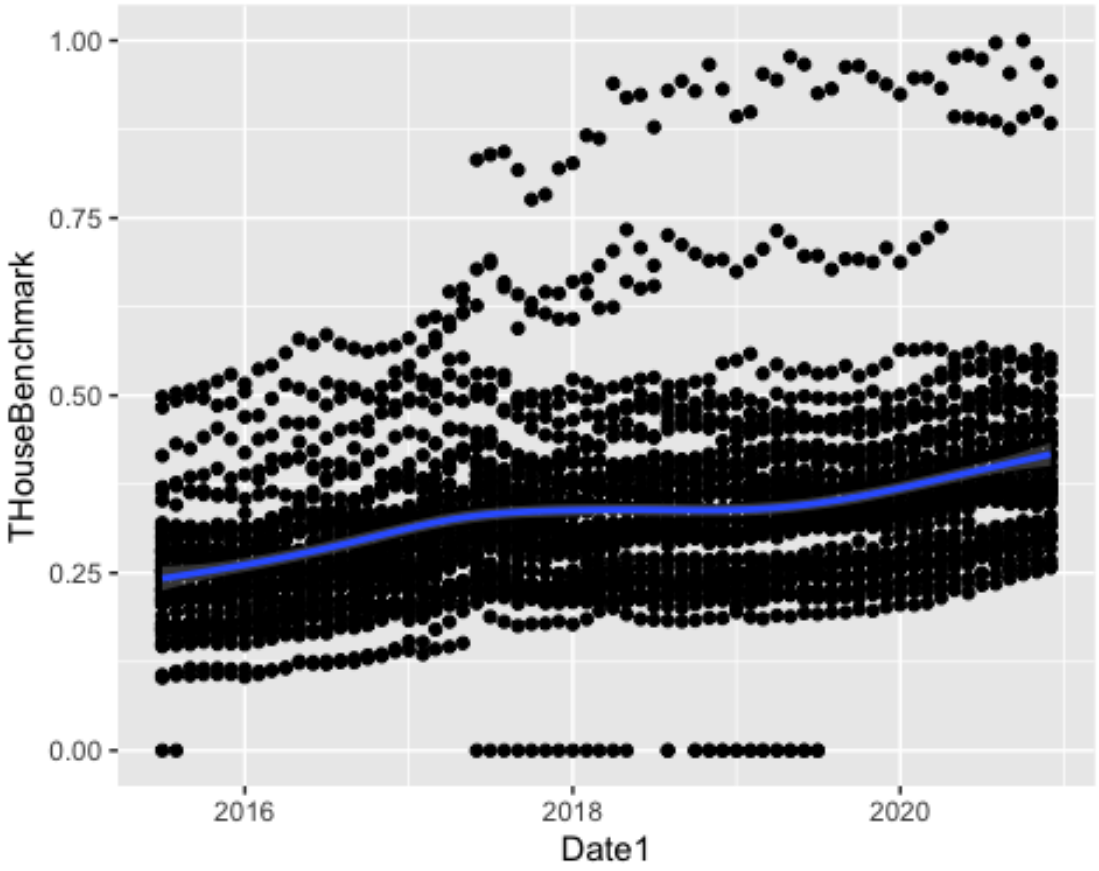
```
## [1] 0.6714359
```

#EDA

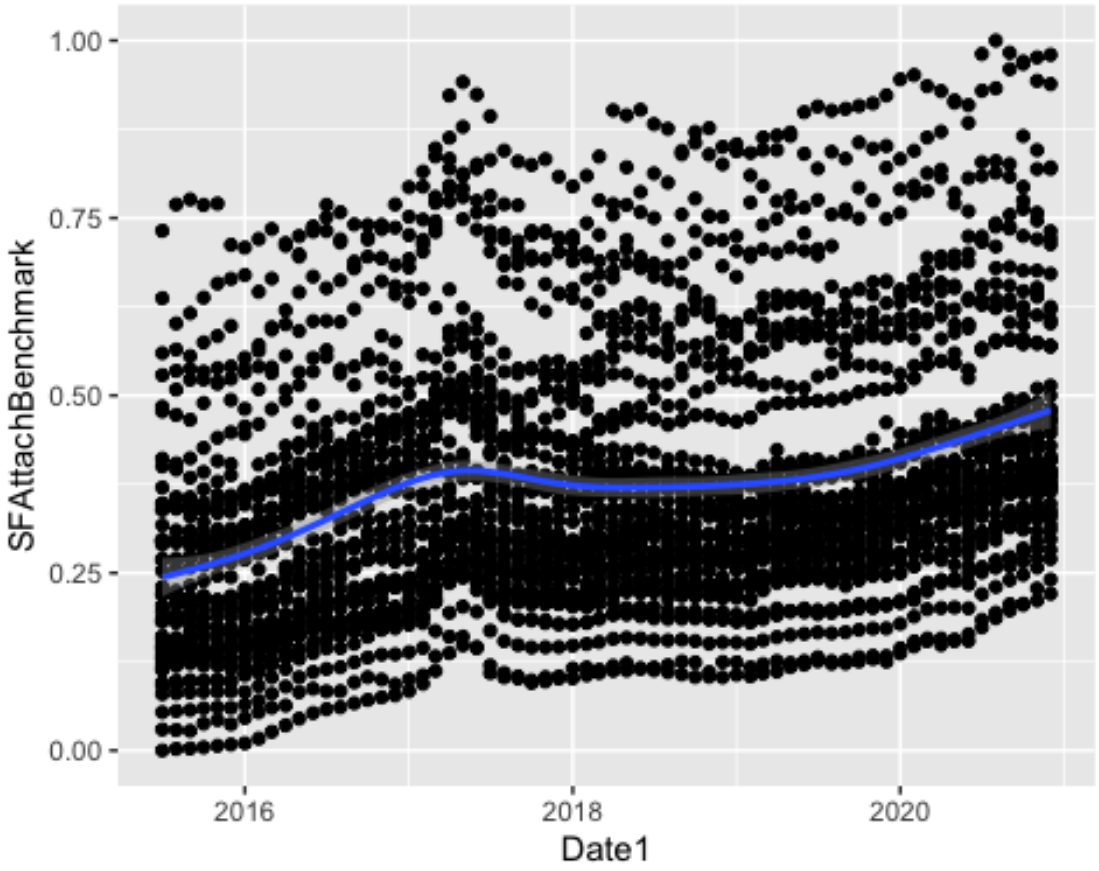
```
newmls<-
c("SFAttachBenchmark","THouseBenchmark","ApartBenchmark","SFDetachBenchmark",
"Date")
mls1<-mls[newmls]
#mls1

mls2<-na.omit(mls1)
#mls2

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

mls2$SFAttachBenchmark<-normalize(mls2$SFAttachBenchmark)
mls2$THouseBenchmark <- normalize(mls2$THouseBenchmark)
mls2$ApartBenchmark <- normalize(mls2$ApartBenchmark)
mls2$SFDetachBenchmark <- normalize (mls2$SFDetachBenchmark)

Date1<- as.Date(mls2$Date)
#Date1

mls3<-
cor(mls2[,c("SFAttachBenchmark","THouseBenchmark","ApartBenchmark","SFDetachB
enchmark")])
corrplot(mls3, method="shade",type="upper")
```



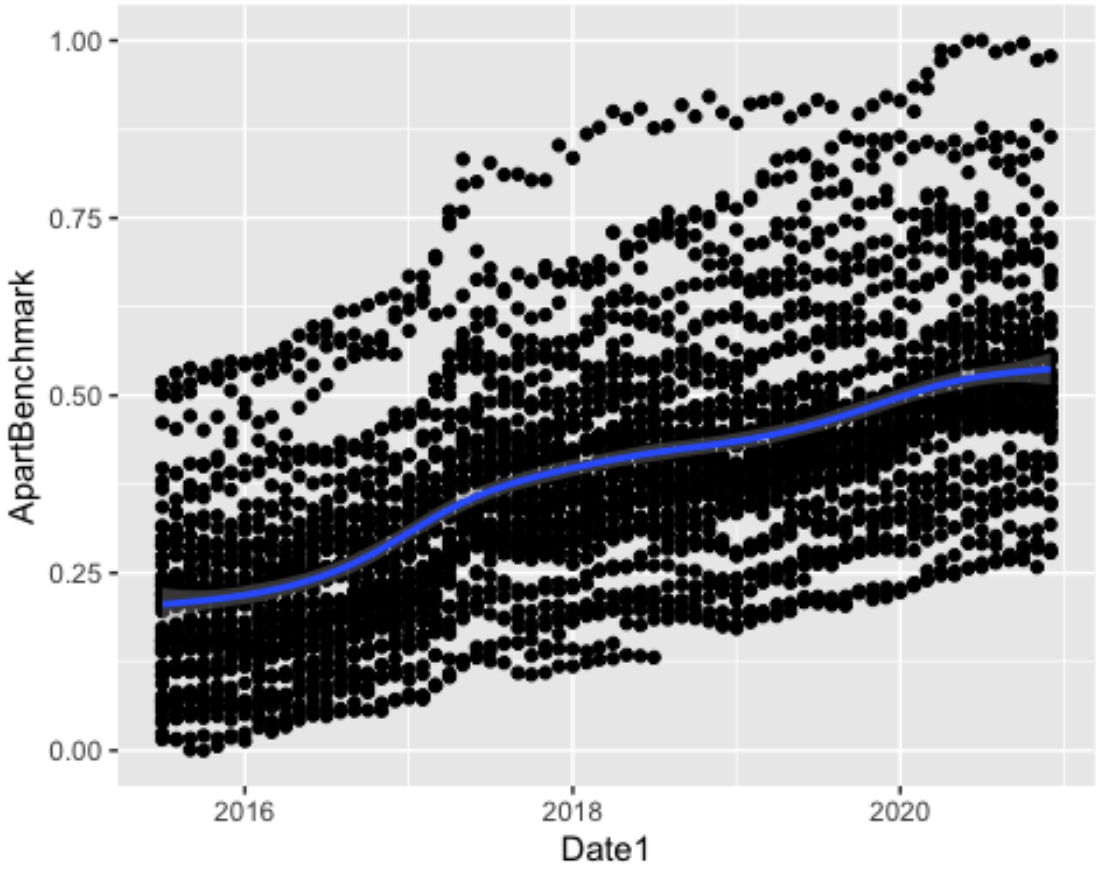

```
ggplot(mls2, aes(x=Date1, y=THouseBenchmark  )) +
        geom_point() +
        geom_smooth()

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



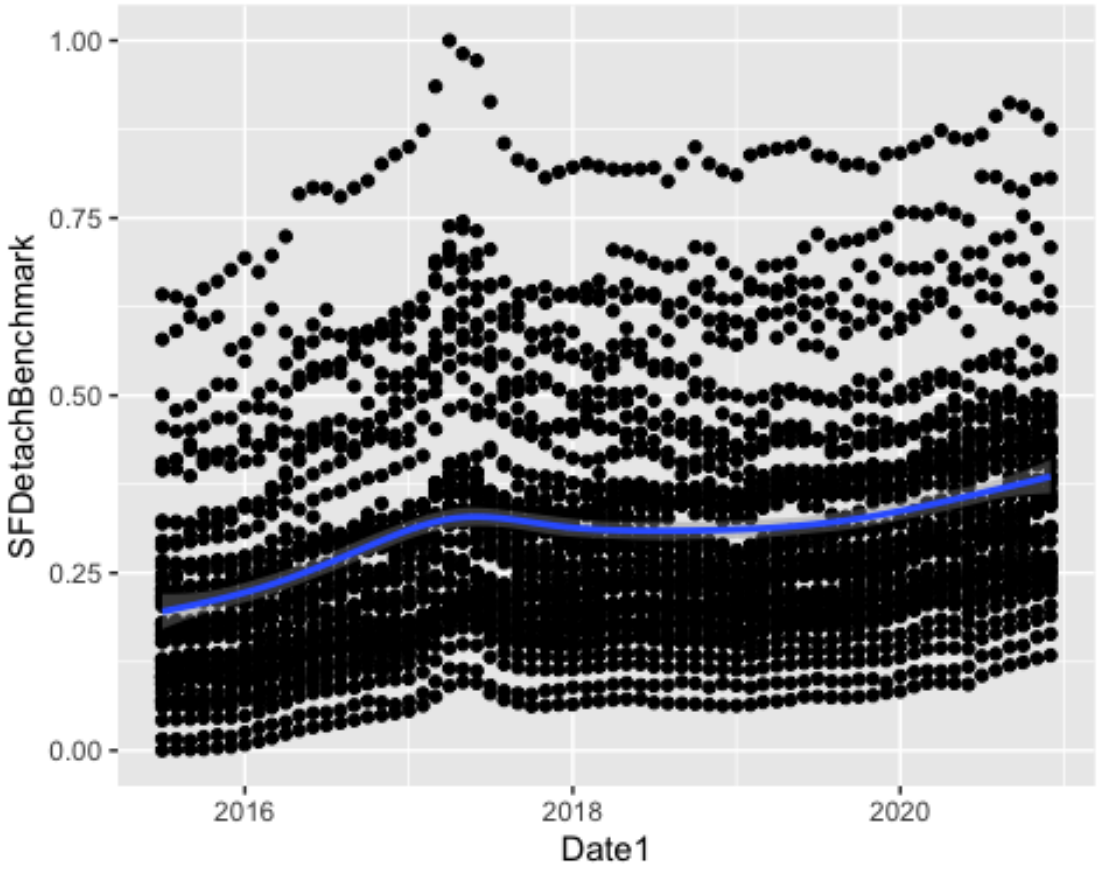

```
ggplot(mls2, aes(x=Date1, y=SFAttachBenchmark )) +
        geom_point() +
        geom_smooth()

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(mls2, aes(x=Date1, y=ApartBenchmark)) +
        geom_point() +
        geom_smooth()

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(mls2, aes(x=Date1, y=SFDetachBenchmark)) +
        geom_point() +
        geom_smooth()

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```


1.00
0.75
0.50
0.25
0.00
SFDetachBenchmark
2016
2018
2020
Date1