Retrieving a Dataset from GitHub with R

Assignment 5

Amanda

09-13-2020

I downloaded my file from “ajpiter R ProTips West Roxbury”

The link to the GitHub repository where I downloaded the file is [GitHub Repository] https://github.com/ajpiter/RProTips/blob/master/Projects/WestRoxburyHomes/WestRoxbury.csv

To save the dataset as an object, I used the following code:

```
# Save the dataset as an object
WestRoxbury.df <- read.csv("https://raw.githubusercontent.com/ajpiter/RProTips/master/Projects/WestRoxbu
```

##Exploring Data

I started exploring the dataset by running the following code:

```
#View(WestRoxbury.df)
dim(WestRoxbury.df)
```

```
## [1] 5802   14
```

```
names(WestRoxbury.df)
```

```
##  [1] "TOTAL.VALUE" "TAX"         "LOT.SQFT"    "YR.BUILT"    "GROSS.AREA"
##  [6] "LIVING.AREA" "FLOORS"      "ROOMS"       "BEDROOMS"    "FULL.BATH"
## [11] "HALF.BATH"   "KITCHEN"     "FIREPLACE"   "REMODEL"
```

```
head(WestRoxbury.df)
```

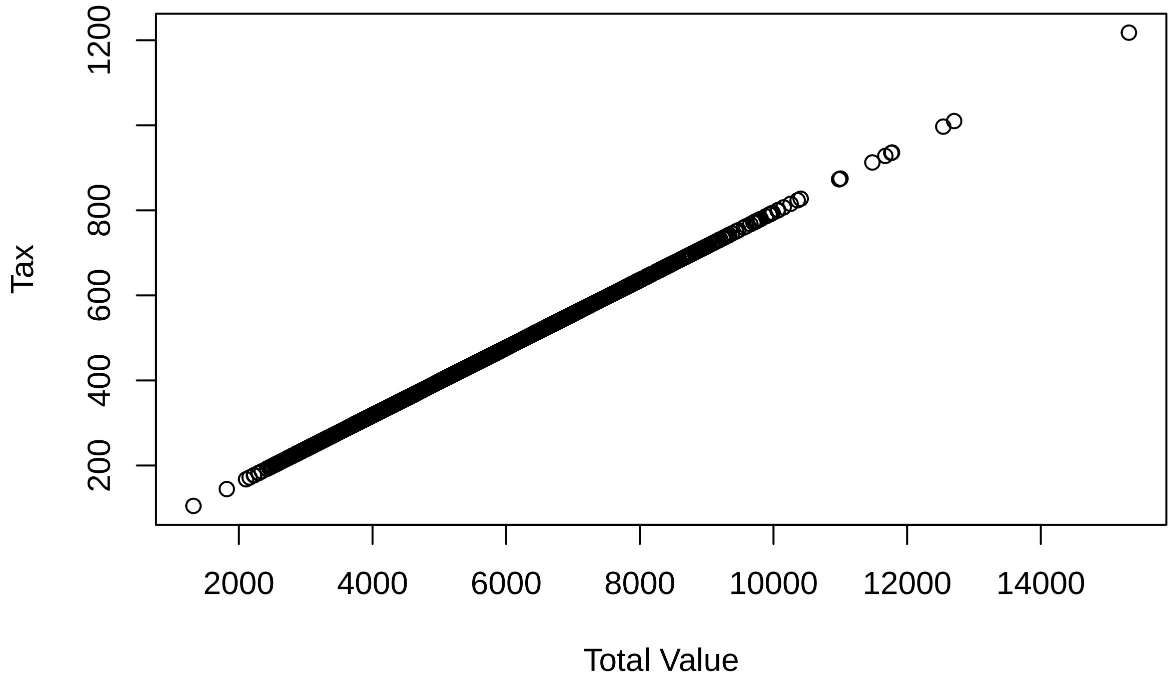
```
##   TOTAL.VALUE  TAX LOT.SQFT YR.BUILT GROSS.AREA LIVING.AREA FLOORS ROOMS
## 1       344.2 4330     9965     1880       2436        1352      2     6
## 2       412.6 5190     6590     1945       3108        1976      2    10
## 3       330.1 4152     7500     1890       2294        1371      2     8
## 4       498.6 6272    13773     1957       5032        2608      1     9
## 5       331.5 4170     5000     1910       2370        1438      2     7
## 6       337.4 4244     5142     1950       2124        1060      1     6
##   BEDROOMS FULL.BATH HALF.BATH KITCHEN FIREPLACE REMODEL
## 1        3         1         1       1         0    None
## 2        4         2         1       1         0  Recent
## 3        4         1         1       1         0    None
## 4        5         1         1       1         1    None
## 5        3         2         0       1         0    None
## 6        3         1         0       1         1     Old
```

```
summary(WestRoxbury.df)
```

```
##   TOTAL.VALUE          TAX          LOT.SQFT        YR.BUILT      GROSS.AREA
##  Min.   : 105.0   Min.   : 1320   Min.   :  997   Min.   :   0   Min.   : 821
##   1st Qu.: 325.1   1st Qu.: 4090   1st Qu.: 4772   1st Qu.:1920   1st Qu.:2347
##   Median : 375.9   Median : 4728   Median : 5683   Median :1935   Median :2700
##   Mean   : 392.7   Mean   : 4939   Mean   : 6278   Mean   :1937   Mean   :2925
##   3rd Qu.: 438.8   3rd Qu.: 5520   3rd Qu.: 7022   3rd Qu.:1955   3rd Qu.:3239
##   Max.   :1217.8   Max.   :15319   Max.   :46411   Max.   :2011   Max.   :8154
##    LIVING.AREA       FLOORS          ROOMS           BEDROOMS       FULL.BATH
##   Min.   : 504    Min.   :1.000   Min.   : 3.000   Min.   :1.00   Min.   :1.000
##   1st Qu.:1308    1st Qu.:1.000   1st Qu.: 6.000   1st Qu.:3.00   1st Qu.:1.000
##   Median :1548    Median :2.000   Median : 7.000   Median :3.00   Median :1.000
##   Mean   :1657    Mean   :1.684   Mean   : 6.995   Mean   :3.23   Mean   :1.297
##   3rd Qu.:1874    3rd Qu.:2.000   3rd Qu.: 8.000   3rd Qu.:4.00   3rd Qu.:2.000
##   Max.   :5289    Max.   :3.000   Max.   :14.000   Max.   :9.00   Max.   :5.000
##     HALF.BATH          KITCHEN         FIREPLACE          REMODEL
##   Min.   :0.0000   Min.   :1.000   Min.   :0.0000   Length:5802
##   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.0000   Class :character
##   Median :1.0000   Median :1.000   Median :1.0000   Mode  :character
##   Mean   :0.6139   Mean   :1.015   Mean   :0.7399
##   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##   Max.   :3.0000   Max.   :2.000   Max.   :4.0000
```

To further explore the datasets I created basic visualizations. The first was a scatterplot of home value compared to the amount of taxes for the home.

```r
#Scatterplot of Total Value versus Tax Amount
plot(WestRoxbury.df$TOTAL.VALUE ~ WestRoxbury.df$TAX, xlab= "Total Value", ylab = "Tax")
```

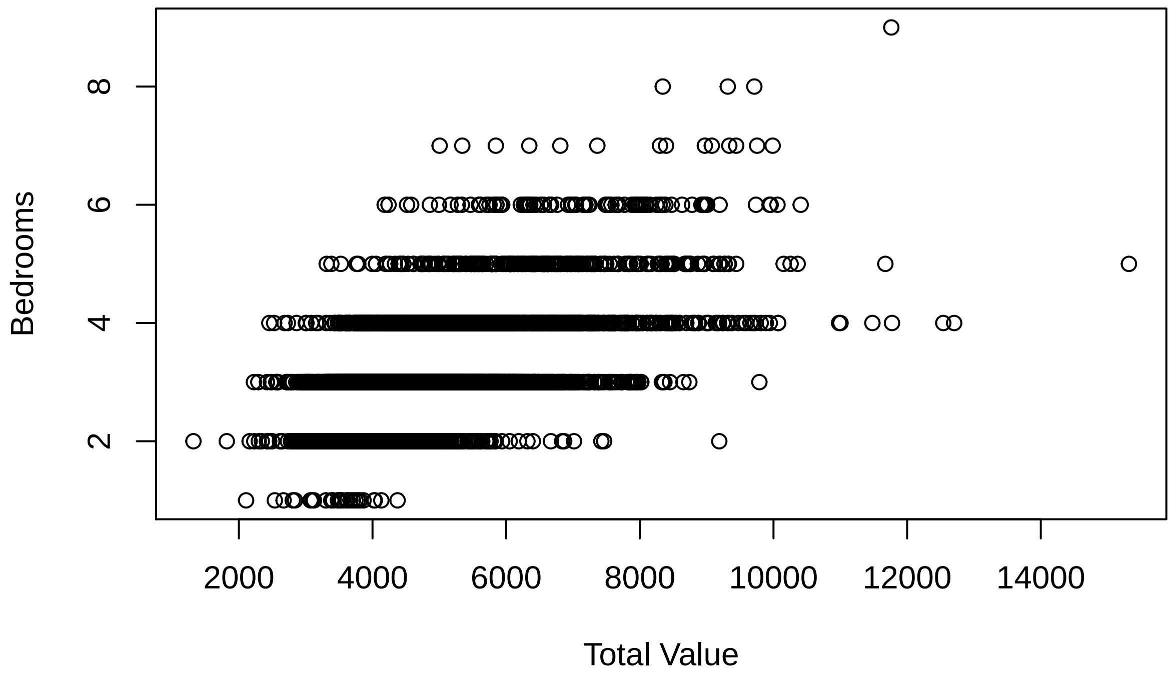


```r
#We can see from the plot that the tax amount is positively correlated to the home value
```

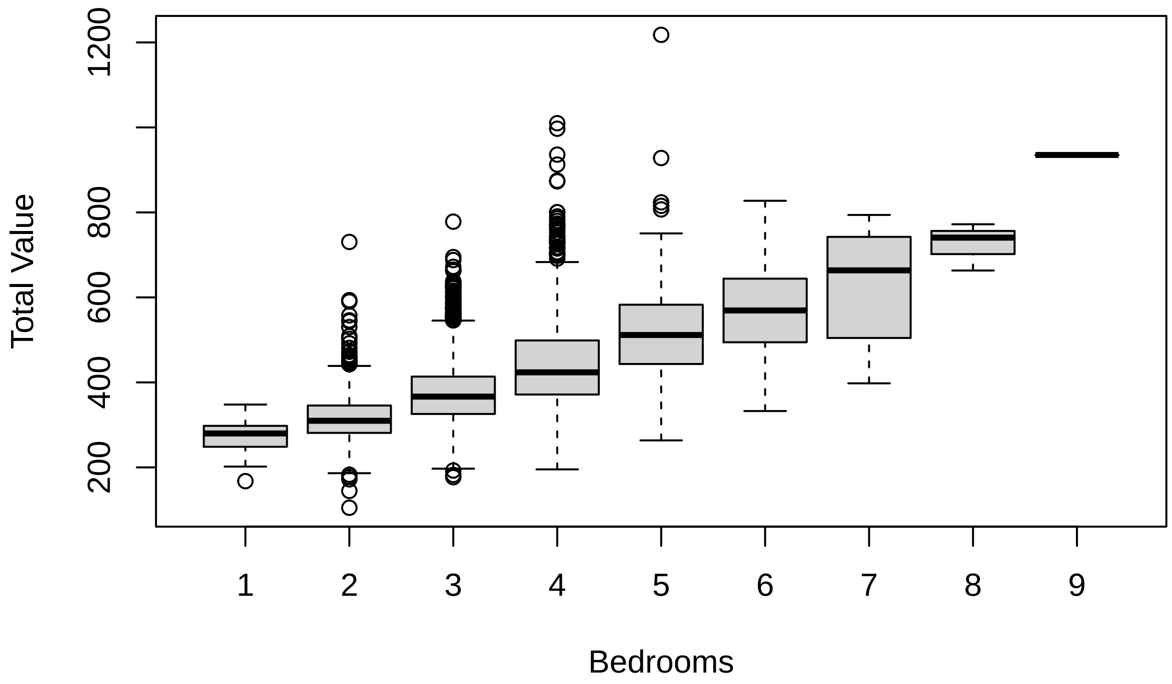
Since home value and tax value were strongly correlated, I decided to explore if the relationship between the number of bedrooms and the total home value.

##Simple Visualizations: Comparing Home Value & Bedrooms

```
plot(WestRoxbury.df$BEDROOMS ~ WestRoxbury.df$TAX, xlab= "Total Value", ylab = "Bedrooms")
```



```
#Homes with an equal number of Bedrooms have a wider range in total value
#A boxplot better represents the range of home values by bedroom, but dose not illustrate much
boxplot(WestRoxbury.df$TOTAL.VALUE ~ WestRoxbury.df$BEDROOMS, xlab = "Bedrooms", ylab = "Total Value")
```



A boxplot comparing the number of bedrooms with the home value was able to show that home values generally increased. There were outliers in 2, 3, and 4 bedroom homes that made them more valuable.

To dig deeper into the data, I subseted the home in the 4th interquartile range of Home Values. The High End subset would allow me to see if the number of bedrooms had the same impact.

The code I used to subset the data was:

```r
#To create better visualization
#Manipulate the Dataset to study high-end home values
#select all homes in the 4th interquartile range for home value
HomeValue4Q <- WestRoxbury.df$TOTAL.VALUE > 438
#HomeValue4Q
#create a new column
WestRoxbury.df$HomeValue4Q <- c(HomeValue4Q)
#View(WestRoxbury.df)

#check for NAs on Total Value and Tax before graph
#is.na(HomeValue4Q)
#is.na(WestRoxbury.df$TAX)

#subset the data based on homes with values in the 4th quartile range
WestRoxburyHighEnd.df <- subset(WestRoxbury.df, WestRoxbury.df$HomeValue4Q ==TRUE,
                  select=c(TOTAL.VALUE, TAX, LOT.SQFT, YR.BUILT,
                           GROSS.AREA, LIVING.AREA, FLOORS, ROOMS,
                           BEDROOMS, FULL.BATH, HALF.BATH, KITCHEN,
                           FIREPLACE, REMODEL, HomeValue4Q))
#View(WestRoxburyHighEnd.df)
```

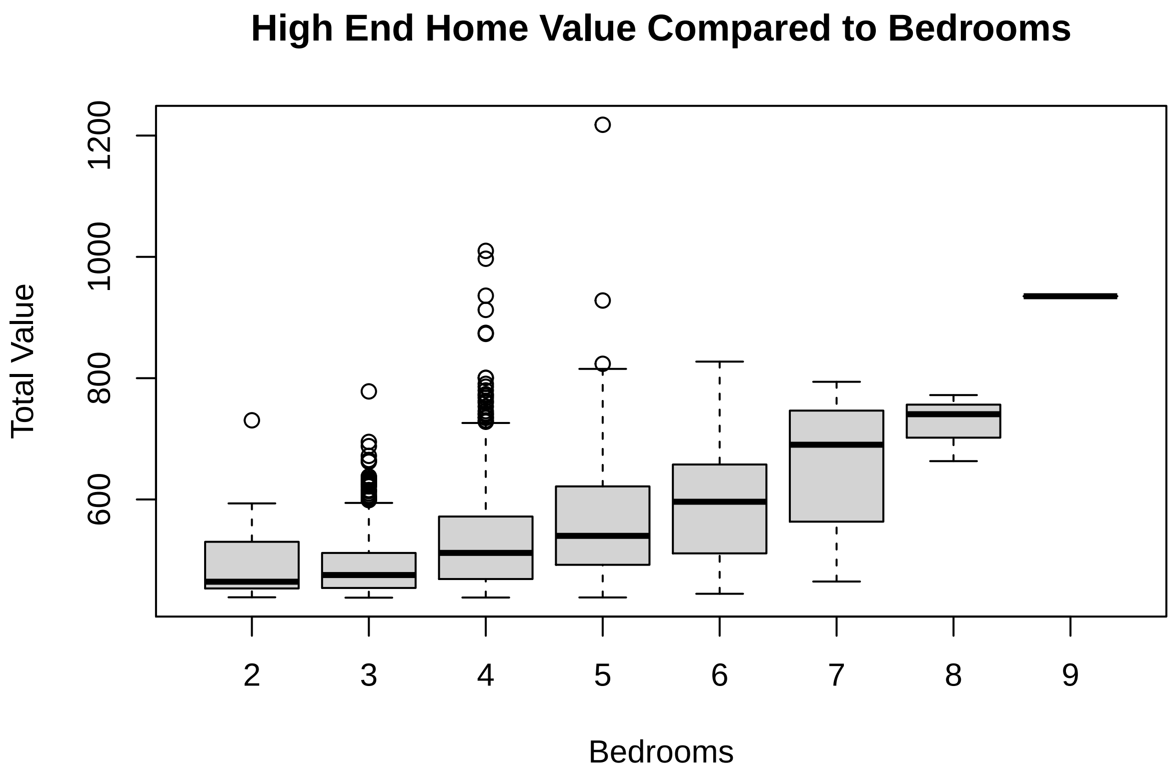
```r
head(WestRoxburyHighEnd.df)
```

```
##    TOTAL.VALUE  TAX LOT.SQFT YR.BUILT GROSS.AREA LIVING.AREA FLOORS ROOMS
## 4        498.6 6272    13773     1957       5032        2608      1     9
## 14       575.0 7233    12288     2004       4616        2378      2     9
## 46       490.7 6173     5683     1995       4100        2640      2     6
## 66       566.3 7124     8249     2007       4390        2708      2     8
## 87       479.1 6027     9642     1999       2952        1872      2     7
## 97       466.1 5863     8970     1999       2952        1872      2     7
##    BEDROOMS FULL.BATH HALF.BATH KITCHEN FIREPLACE REMODEL HomeValue4Q
## 4         5         1         1       1         1    None        TRUE
## 14        4         2         1       1         1    None        TRUE
## 46        3         1         1       1         1  Recent        TRUE
## 66        4         2         1       1         1    None        TRUE
## 87        4         2         1       1         1    None        TRUE
## 97        4         2         1       1         1    None        TRUE
```

Then I built a boxplot data visualization. I used the WestRoxburyHighEnd dataframe to look at homes with values in the 4th quartile range. Then I used a boxplot to illustrate the relationship between Home Value and the number of Bedrooms.

The code and the plot visualization are below.

```r
boxplot(WestRoxburyHighEnd.df$TOTAL.VALUE ~ WestRoxburyHighEnd.df$BEDROOMS, main = "High End Home Value
```



```
To compare the results of the High End Homes, I created additional dataframes for homes with values in
```

```
Home Value Interquartile Ranges
```

```
1st Quartile: WestRoxburyLowEnd.df
2nd Quartile: WestRoxburyLowMid.df
3rd Quartile: WestRoxburyHighMid.df
4th Quartile: WestRoxburyHighEnd.df

##Complex Visualization: Home Value & Bedrooms by Interquartile Range

‘‘‘r
#To understand if bedrooms impacts homes in all price ranges equally
#create side by side box plots

summary(WestRoxbury.df$TOTAL.VALUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   105.0   325.1   375.9   392.7   438.8  1217.8
```

```r
#select all homes in the 3rd interquartile range for home value
HomeValue3Q <- WestRoxbury.df$TOTAL.VALUE > 329 & WestRoxbury.df$TOTAL.VALUE < 438
#HomeValue3Q
#create a new column
WestRoxbury.df$HomeValue3Q <- c(HomeValue3Q)

#select all homes in the 2nd interquartile range for home value
HomeValue2Q <- WestRoxbury.df$TOTAL.VALUE > 325 & WestRoxbury.df$TOTAL.VALUE < 392
#HomeValue2Q
#create a new column
WestRoxbury.df$HomeValue2Q <- c(HomeValue2Q)

#select all homes in the 1st interquartile range for home value
HomeValue1Q <- WestRoxbury.df$TOTAL.VALUE > 0 & WestRoxbury.df$TOTAL.VALUE < 325
#HomeValue1Q
#create a new column
WestRoxbury.df$HomeValue1Q <- c(HomeValue1Q)

#View the Dataframe with 4 columns for home value
#View(WestRoxbury.df)
head(WestRoxbury.df)
```

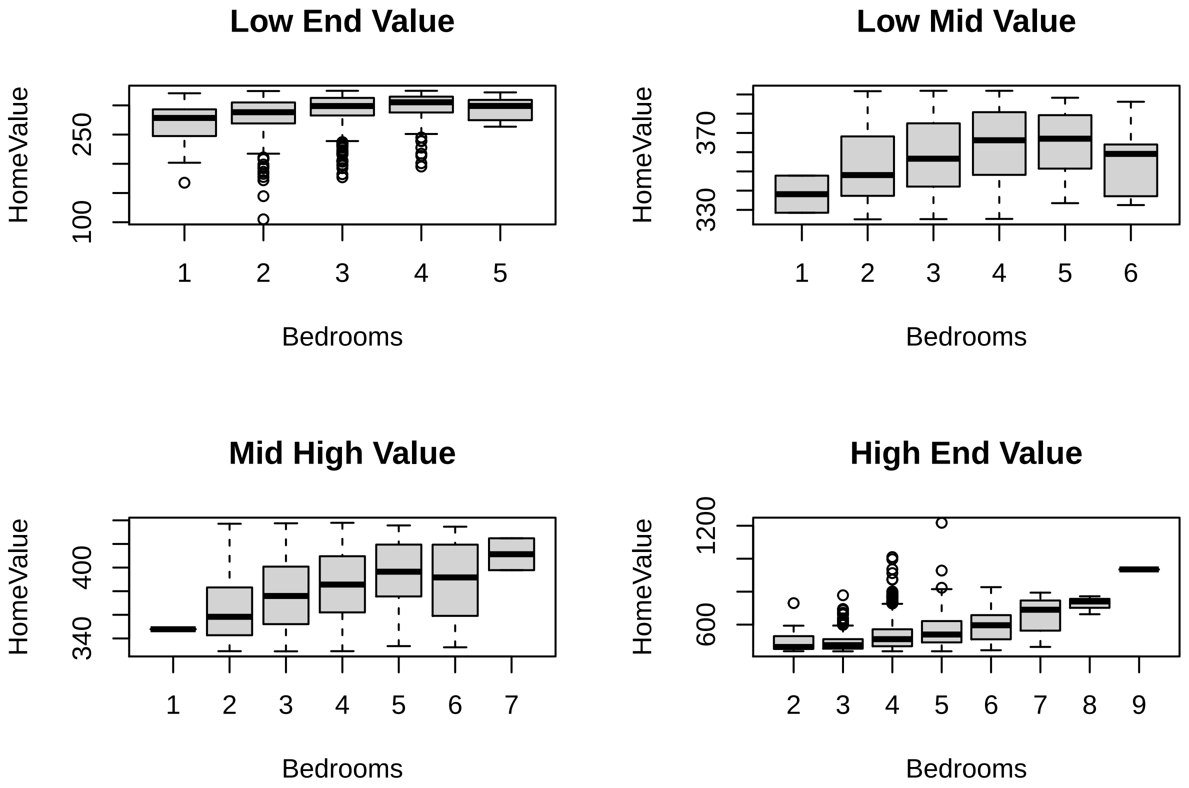
```
##   TOTAL.VALUE  TAX LOT.SQFT YR.BUILT GROSS.AREA LIVING.AREA FLOORS ROOMS
## 1       344.2 4330     9965     1880       2436        1352      2     6
## 2       412.6 5190     6590     1945       3108        1976      2    10
## 3       330.1 4152     7500     1890       2294        1371      2     8
## 4       498.6 6272    13773     1957       5032        2608      1     9
## 5       331.5 4170     5000     1910       2370        1438      2     7
## 6       337.4 4244     5142     1950       2124        1060      1     6
##   BEDROOMS FULL.BATH HALF.BATH KITCHEN FIREPLACE REMODEL HomeValue4Q
## 1        3         1         1       1         0    None       FALSE
## 2        4         2         1       1         0  Recent       FALSE
## 3        4         1         1       1         0    None       FALSE
## 4        5         1         1       1         1    None        TRUE
## 5        3         2         0       1         0    None       FALSE
## 6        3         1         0       1         1     Old       FALSE
##   HomeValue3Q HomeValue2Q HomeValue1Q
## 1        TRUE        TRUE       FALSE
## 2        TRUE       FALSE       FALSE
## 3        TRUE        TRUE       FALSE
## 4       FALSE       FALSE       FALSE
## 5        TRUE        TRUE       FALSE
## 6        TRUE        TRUE       FALSE
```

```
#subset the data based on homes with values in the 1st quartile range
WestRoxburyLowEnd.df <- subset(WestRoxbury.df, WestRoxbury.df$HomeValue1Q ==TRUE,
                                select=c(TOTAL.VALUE, TAX, LOT.SQFT, YR.BUILT,
                                         GROSS.AREA, LIVING.AREA, FLOORS, ROOMS,
                                         BEDROOMS, FULL.BATH, HALF.BATH, KITCHEN,
                                         FIREPLACE, REMODEL, HomeValue4Q))
#subset the data based on homes with values in the 1st quartile range
WestRoxburyLowMid.df <- subset(WestRoxbury.df, WestRoxbury.df$HomeValue2Q ==TRUE,
                               select=c(TOTAL.VALUE, TAX, LOT.SQFT, YR.BUILT,
                                        GROSS.AREA, LIVING.AREA, FLOORS, ROOMS,
                                        BEDROOMS, FULL.BATH, HALF.BATH, KITCHEN,
                                        FIREPLACE, REMODEL, HomeValue4Q))
#subset the data based on homes with values in the 1st quartile range
WestRoxburyHighMid.df <- subset(WestRoxbury.df, WestRoxbury.df$HomeValue3Q ==TRUE,
                               select=c(TOTAL.VALUE, TAX, LOT.SQFT, YR.BUILT,
                                        GROSS.AREA, LIVING.AREA, FLOORS, ROOMS,
                                        BEDROOMS, FULL.BATH, HALF.BATH, KITCHEN,
                                        FIREPLACE, REMODEL, HomeValue4Q))
```

Finally, I built a more complex visualization with 4 boxplots side by side. I used Home Values and Number of Bedrooms for the 2 variables in each plot.

The code and the example are below.

```
#side by side box plots
#use par() to split the plots into panels
help("par")
par(mfrow = c(2,2))
boxplot(WestRoxburyLowEnd.df$TOTAL.VALUE ~WestRoxburyLowEnd.df$BEDROOMS, main = "Low End Value",
        xlab="Bedrooms", ylab="HomeValue")
boxplot(WestRoxburyLowMid.df$TOTAL.VALUE ~WestRoxburyLowMid.df$BEDROOMS, main = "Low Mid Value",
        xlab="Bedrooms", ylab="HomeValue")
boxplot(WestRoxburyHighMid.df$TOTAL.VALUE ~WestRoxburyHighMid.df$BEDROOMS, main = "Mid High Value",
        xlab="Bedrooms", ylab="HomeValue")
boxplot(WestRoxburyHighEnd.df$TOTAL.VALUE ~ WestRoxburyHighEnd.df$BEDROOMS, main = "High End Value",
        xlab="Bedrooms", ylab="HomeValue")
```



By printing the box plots for each of the quartile ranges together we can see the variations on how the number of bedrooms impacts the honevalue.

In both the 2nd quartile and 3rd quartile home value ranges, an increase in bedrooms usually correlates with an increase in home value unless there is more than 5 bedrooms.

For homes with values in the lower quartile values, we see more outliers on the lower end of the scale. Bedrooms appear to be less likely to be a predictive factor.

For homes with vaules in the upper quartile range, we see more outliers on the higher end of the scale. Additionally the range in the upper quartile is larger, from 448 to 1200. However homes with values about 800 are displayed as outliers (excluding the 9 bedroom home). as many homes