Using dplyr and the Tidyverse to manipulate data

Assignment 7

Amanda

09-20-2020

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
```

```
#make a backup dataset just in case
WestRoxburyBackup <- WestRoxbury.df
```

Using Rename to Standarize Column Names rename() is not one of the six verbs, but helps with my sanity

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#rename(data, New Name = Old Name)
#rename Total Value to reflect it is in thousands
WestRoxbury.df<- rename(WestRoxbury.df, HomeValueThousands = TOTAL.VALUE)
#view(WestRoxbury.df)
#rename columns to follow CamelBack my preferred naming convention
head(WestRoxbury.df)
```

```
##   HomeValueThousands  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
```

```
WestRoxbury.df <- rename(WestRoxbury.df, HomeTax = TAX)
WestRoxbury.df <- rename(WestRoxbury.df, LotSqft = LOT.SQFT)
#you can rename more than one column at once
WestRoxbury.df <- rename(WestRoxbury.df, YrBuilt = YR.BUILT, GrossArea = GROSS.AREA, LivingArea = LIVIN
                         Floors = FLOORS, TotalRooms = ROOMS, Bedrooms = BEDROOMS, FullBaths = FULL.BAT
                         HalfBaths = HALF.BATH)
#view(WestRoxbury.df) #check your work
#missed a few
WestRoxbury.df <- rename(WestRoxbury.df, Kitchen = KITCHEN, Fireplace = FIREPLACE, Remodel =REMODEL)
head(WestRoxbury.df)
```

```
##   HomeValueThousands HomeTax LotSqft YrBuilt GrossArea LivingArea Floors
## 1              344.2    4330    9965    1880      2436       1352      2
## 2              412.6    5190    6590    1945      3108       1976      2
## 3              330.1    4152    7500    1890      2294       1371      2
## 4              498.6    6272   13773    1957      5032       2608      1
## 5              331.5    4170    5000    1910      2370       1438      2
## 6              337.4    4244    5142    1950      2124       1060      1
##   TotalRooms Bedrooms FullBaths HalfBaths Kitchen Fireplace Remodel
## 1          6        3         1         1       1         0    None
## 2         10        4         2         1       1         0  Recent
## 3          8        4         1         1       1         0    None
## 4          9        5         1         1       1         1    None
## 5          7        3         2         0       1         0    None
## 6          6        3         1         0       1         1     Old
```

#There are 6 dplyr "verbs" to help keep datasets tidy

##dplyr Verb #1 Mutate() #Mutate() adds or alters columns through a calculation on existing columuns

```
#Use Mutate to calculate the home value in single dollars vs thousands
WestRoxbury.df <- mutate(WestRoxbury.df, HomeValue = WestRoxbury.df$HomeValueThousands * 1000)

#Use Mutate to calculate the price per sq foot
WestRoxbury.df <- mutate(WestRoxbury.df, PricePerSqFt = WestRoxbury.df$HomeValue / WestRoxbury.df$Living
WestRoxbury.df <- mutate(WestRoxbury.df, TaxPerSqFt = WestRoxbury.df$HomeTax / WestRoxbury.df$LivingArea
WestRoxbury.df <- mutate(WestRoxbury.df, PricePerLotSqFt = WestRoxbury.df$HomeValue / WestRoxbury.df$Lo
WestRoxbury.df <- mutate(WestRoxbury.df, TaxPerLotSqFt = WestRoxbury.df$HomeTax / WestRoxbury.df$LotSqf
#view(WestRoxbury.df) #Check your work
head(WestRoxbury.df) #check your work
```

```
##   HomeValueThousands HomeTax LotSqft YrBuilt GrossArea LivingArea Floors
## 1              344.2    4330    9965    1880      2436       1352      2
## 2              412.6    5190    6590    1945      3108       1976      2
## 3              330.1    4152    7500    1890      2294       1371      2
## 4              498.6    6272   13773    1957      5032       2608      1
## 5              331.5    4170    5000    1910      2370       1438      2
## 6              337.4    4244    5142    1950      2124       1060      1
##   TotalRooms Bedrooms FullBaths HalfBaths Kitchen Fireplace Remodel HomeValue
## 1          6        3         1         1       1         0    None    344200
## 2         10        4         2         1       1         0  Recent    412600
## 3          8        4         1         1       1         0    None    330100
## 4          9        5         1         1       1         1    None    498600
```

```
## 5          7        3         2         0       1         0    None    331500
## 6          6        3         1         0       1         1     Old    337400
##   PricePerSqFt TaxPerSqFt PricePerLotSqFt TaxPerLotSqFt
## 1     254.5858   3.202663        34.54089     0.4345208
## 2     208.8057   2.626518        62.61002     0.7875569
## 3     240.7732   3.028446        44.01333     0.5536000
## 4     191.1810   2.404908        36.20126     0.4553837
## 5     230.5285   2.899861        66.30000     0.8340000
## 6     318.3019   4.003774        65.61649     0.8253598
```

#dplyr Verb #2 select() #select() helps keep only the columns you need and allows you to change the order of the columns

```
#we already know we don't need TotalValueThousands since HomeValue represents the same number converted
summary(WestRoxbury.df) #use summary to see if there are other values we can eliminate
```

```
##  HomeValueThousands    HomeTax         LotSqft         YrBuilt
##  Min.   : 105.0     Min.   : 1320   Min.   :  997   Min.   :   0
##  1st Qu.: 325.1     1st Qu.: 4090   1st Qu.: 4772   1st Qu.:1920
##  Median : 375.9     Median : 4728   Median : 5683   Median :1935
##  Mean   : 392.7     Mean   : 4939   Mean   : 6278   Mean   :1937
##  3rd Qu.: 438.8     3rd Qu.: 5520   3rd Qu.: 7022   3rd Qu.:1955
##  Max.   :1217.8     Max.   :15319   Max.   :46411   Max.   :2011
##    GrossArea      LivingArea       Floors        TotalRooms        Bedrooms
##  Min.   : 821   Min.   : 504   Min.   :1.000   Min.   : 3.000   Min.   :1.00
##  1st Qu.:2347   1st Qu.:1308   1st Qu.:1.000   1st Qu.: 6.000   1st Qu.:3.00
##  Median :2700   Median :1548   Median :2.000   Median : 7.000   Median :3.00
##  Mean   :2925   Mean   :1657   Mean   :1.684   Mean   : 6.995   Mean   :3.23
##  3rd Qu.:3239   3rd Qu.:1874   3rd Qu.:2.000   3rd Qu.: 8.000   3rd Qu.:4.00
##  Max.   :8154   Max.   :5289   Max.   :3.000   Max.   :14.000   Max.   :9.00
##    FullBaths       HalfBaths         Kitchen        Fireplace
##  Min.   :1.000   Min.   :0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :1.000   Median :1.0000   Median :1.000   Median :1.0000
##  Mean   :1.297   Mean   :0.6139   Mean   :1.015   Mean   :0.7399
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   :3.0000   Max.   :2.000   Max.   :4.0000
##    Remodel            HomeValue        PricePerSqFt     TaxPerSqFt
##  Length:5802        Min.   : 105000   Min.   :105.3   Min.   :1.324
##  Class :character   1st Qu.: 325125   1st Qu.:216.3   1st Qu.:2.721
##  Mode  :character   Median : 375900   Median :243.0   Median :3.056
##                     Mean   : 392686   Mean   :245.1   Mean   :3.083
##                     3rd Qu.: 438775   3rd Qu.:270.1   3rd Qu.:3.398
##                     Max.   :1217800   Max.   :489.0   Max.   :6.150
##  PricePerLotSqFt  TaxPerLotSqFt
##  Min.   : 13.13   Min.   :0.1651
##  1st Qu.: 54.87   1st Qu.:0.6902
##  Median : 66.48   Median :0.8362
##  Mean   : 67.54   Mean   :0.8496
##  3rd Qu.: 78.11   3rd Qu.:0.9825
##  Max.   :262.07   Max.   :3.2965
```

```
#Every home has at least one kitchen, and only a very few homes have 2 kitchens
#Since we are going to look at home values in the lower quartile we won't use kitchen
#We also removed GrossAres
#Then reordered the columns
```

```
WestRoxbury.df <- select(WestRoxbury.df, HomeValue, HomeTax, YrBuilt, LivingArea, PricePerSqFt, TaxPerSq
       TaxPerLotSqFt, Floors, TotalRooms, Bedrooms, FullBaths, HalfBaths, Fireplace, Remodel)
#View(WestRoxbury.df) #check your work
head(WestRoxbury.df)
```

```
##   HomeValue HomeTax YrBuilt LivingArea PricePerSqFt TaxPerSqFt LotSqft
## 1    344200    4330    1880       1352     254.5858   3.202663    9965
## 2    412600    5190    1945       1976     208.8057   2.626518    6590
## 3    330100    4152    1890       1371     240.7732   3.028446    7500
## 4    498600    6272    1957       2608     191.1810   2.404908   13773
## 5    331500    4170    1910       1438     230.5285   2.899861    5000
## 6    337400    4244    1950       1060     318.3019   4.003774    5142
##   PricePerLotSqFt TaxPerLotSqFt Floors TotalRooms Bedrooms FullBaths HalfBaths
## 1        34.54089     0.4345208      2          6        3         1         1
## 2        62.61002     0.7875569      2         10        4         2         1
## 3        44.01333     0.5536000      2          8        4         1         1
## 4        36.20126     0.4553837      1          9        5         1         1
## 5        66.30000     0.8340000      2          7        3         2         0
## 6        65.61649     0.8253598      1          6        3         1         0
##   Fireplace Remodel
## 1         0    None
## 2         0  Recent
## 3         0    None
## 4         1    None
## 5         0    None
## 6         1     Old
```

#Creating Numeric Variables #Before we start with ggplot2 it is helpful to convert data to numeric variables

```
class(WestRoxbury.df)
```

```
## [1] "data.frame"
```

```
summary(WestRoxbury.df) #Remodel is the only a character
```

```
##    HomeValue          HomeTax         YrBuilt       LivingArea
##  Min.   : 105000   Min.   : 1320   Min.   :   0   Min.   : 504
##  1st Qu.: 325125   1st Qu.: 4090   1st Qu.:1920   1st Qu.:1308
##  Median : 375900   Median : 4728   Median :1935   Median :1548
##  Mean   : 392686   Mean   : 4939   Mean   :1937   Mean   :1657
##  3rd Qu.: 438775   3rd Qu.: 5520   3rd Qu.:1955   3rd Qu.:1874
##  Max.   :1217800   Max.   :15319   Max.   :2011   Max.   :5289
##   PricePerSqFt     TaxPerSqFt       LotSqft      PricePerLotSqFt
##  Min.   :105.3   Min.   :1.324   Min.   :  997   Min.   : 13.13
##  1st Qu.:216.3   1st Qu.:2.721   1st Qu.: 4772   1st Qu.: 54.87
##  Median :243.0   Median :3.056   Median : 5683   Median : 66.48
##  Mean   :245.1   Mean   :3.083   Mean   : 6278   Mean   : 67.54
##  3rd Qu.:270.1   3rd Qu.:3.398   3rd Qu.: 7022   3rd Qu.: 78.11
##  Max.   :489.0   Max.   :6.150   Max.   :46411   Max.   :262.07
##  TaxPerLotSqFt        Floors        TotalRooms        Bedrooms
##  Min.   :0.1651   Min.   :1.000   Min.   : 3.000   Min.   :1.00
##  1st Qu.:0.6902   1st Qu.:1.000   1st Qu.: 6.000   1st Qu.:3.00
##  Median :0.8362   Median :2.000   Median : 7.000   Median :3.00
##  Mean   :0.8496   Mean   :1.684   Mean   : 6.995   Mean   :3.23
##  3rd Qu.:0.9825   3rd Qu.:2.000   3rd Qu.: 8.000   3rd Qu.:4.00
##  Max.   :3.2965   Max.   :3.000   Max.   :14.000   Max.   :9.00
```

```
##    FullBaths       HalfBaths        Fireplace        Remodel
##  Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Length:5802
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :1.000   Median :1.0000   Median :1.0000   Mode  :character
##  Mean   :1.297   Mean   :0.6139   Mean   :0.7399
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   :3.0000   Max.   :4.0000
```

```
class(WestRoxbury.df$REMODEL) #another way to see remodel is a character
```

```
## [1] "NULL"
```

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

```
##   HomeValue HomeTax YrBuilt LivingArea PricePerSqFt TaxPerSqFt LotSqft
## 1    344200    4330    1880       1352     254.5858   3.202663    9965
## 2    412600    5190    1945       1976     208.8057   2.626518    6590
## 3    330100    4152    1890       1371     240.7732   3.028446    7500
## 4    498600    6272    1957       2608     191.1810   2.404908   13773
## 5    331500    4170    1910       1438     230.5285   2.899861    5000
## 6    337400    4244    1950       1060     318.3019   4.003774    5142
##   PricePerLotSqFt TaxPerLotSqFt Floors TotalRooms Bedrooms FullBaths HalfBaths
## 1        34.54089     0.4345208      2          6        3         1         1
## 2        62.61002     0.7875569      2         10        4         2         1
## 3        44.01333     0.5536000      2          8        4         1         1
## 4        36.20126     0.4553837      1          9        5         1         1
## 5        66.30000     0.8340000      2          7        3         2         0
## 6        65.61649     0.8253598      1          6        3         1         0
##   Fireplace Remodel
## 1         0    None
## 2         0  Recent
## 3         0    None
## 4         1    None
## 5         0    None
## 6         1     Old
```

```
#We could either convert Remodel or delete it.
#Example of Converting Remodel in WestRoxbury.df
WestRoxbury.df$Remodel <- as.factor(WestRoxbury.df$Remodel) #converts to factor
class(WestRoxbury.df$REMODEL) #facor
```

```
## [1] "NULL"
```

```
levels(WestRoxbury.df$REMODEL) # None, Old, Recent
```

```
## NULL
```

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

```
##  [1] "HomeValue"       "HomeTax"         "YrBuilt"         "LivingArea"
##  [5] "PricePerSqFt"    "TaxPerSqFt"      "LotSqft"         "PricePerLotSqFt"
##  [9] "TaxPerLotSqFt"   "Floors"          "TotalRooms"      "Bedrooms"
## [13] "FullBaths"       "HalfBaths"       "Fireplace"       "Remodel"
```

```
WestRoxbury.df <- model.matrix(~ 0 + HomeValue + HomeTax + +YrBuilt + LivingArea + PricePerSqFt + TaxPe
                                 LotSqft + PricePerLotSqFt + TaxPerLotSqFt + Floors + TotalRooms +
                                 + Bedrooms + FullBaths + HalfBaths + Fireplace + Remodel,
                               data = WestRoxbury.df)
```

```
#view(WestRoxbury.df) #REMODEL has been split into 3 columns
head(WestRoxbury.df)
```

```
##   HomeValue HomeTax YrBuilt LivingArea PricePerSqFt TaxPerSqFt LotSqft
## 1    344200    4330    1880       1352     254.5858   3.202663    9965
## 2    412600    5190    1945       1976     208.8057   2.626518    6590
## 3    330100    4152    1890       1371     240.7732   3.028446    7500
## 4    498600    6272    1957       2608     191.1810   2.404908   13773
## 5    331500    4170    1910       1438     230.5285   2.899861    5000
## 6    337400    4244    1950       1060     318.3019   4.003774    5142
##   PricePerLotSqFt TaxPerLotSqFt Floors TotalRooms Bedrooms FullBaths HalfBaths
## 1        34.54089     0.4345208      2          6        3         1         1
## 2        62.61002     0.7875569      2         10        4         2         1
## 3        44.01333     0.5536000      2          8        4         1         1
## 4        36.20126     0.4553837      1          9        5         1         1
## 5        66.30000     0.8340000      2          7        3         2         0
## 6        65.61649     0.8253598      1          6        3         1         0
##   Fireplace RemodelNone RemodelOld RemodelRecent
## 1         0           1          0             0
## 2         0           0          0             1
## 3         0           1          0             0
## 4         1           1          0             0
## 5         0           1          0             0
## 6         1           0          1             0
```

```
class(WestRoxbury.df) #matrix
```

```
## [1] "matrix" "array"
```

```
WestRoxbury.df <- as.data.frame(WestRoxbury.df)
```

#deplyr Verb 3 #filter() is used to keep only rows matching some logical criteria

```
#For the housing dataset I want to keep the home values in the bottom quartile
summary(WestRoxbury.df) #The first quartile home values is up to 325125
```

```
##    HomeValue          HomeTax         YrBuilt       LivingArea
##  Min.   : 105000   Min.   : 1320   Min.   :   0   Min.   : 504
##  1st Qu.: 325125   1st Qu.: 4090   1st Qu.:1920   1st Qu.:1308
##  Median : 375900   Median : 4728   Median :1935   Median :1548
##  Mean   : 392686   Mean   : 4939   Mean   :1937   Mean   :1657
##  3rd Qu.: 438775   3rd Qu.: 5520   3rd Qu.:1955   3rd Qu.:1874
##  Max.   :1217800   Max.   :15319   Max.   :2011   Max.   :5289
##   PricePerSqFt     TaxPerSqFt       LotSqft      PricePerLotSqFt
##  Min.   :105.3   Min.   :1.324   Min.   :  997   Min.   : 13.13
##  1st Qu.:216.3   1st Qu.:2.721   1st Qu.: 4772   1st Qu.: 54.87
##  Median :243.0   Median :3.056   Median : 5683   Median : 66.48
##  Mean   :245.1   Mean   :3.083   Mean   : 6278   Mean   : 67.54
##  3rd Qu.:270.1   3rd Qu.:3.398   3rd Qu.: 7022   3rd Qu.: 78.11
##  Max.   :489.0   Max.   :6.150   Max.   :46411   Max.   :262.07
##  TaxPerLotSqFt        Floors        TotalRooms       Bedrooms
##  Min.   :0.1651   Min.   :1.000   Min.   : 3.000   Min.   :1.00
##  1st Qu.:0.6902   1st Qu.:1.000   1st Qu.: 6.000   1st Qu.:3.00
##  Median :0.8362   Median :2.000   Median : 7.000   Median :3.00
##  Mean   :0.8496   Mean   :1.684   Mean   : 6.995   Mean   :3.23
##  3rd Qu.:0.9825   3rd Qu.:2.000   3rd Qu.: 8.000   3rd Qu.:4.00
```

```
##  Max.   :3.2965   Max.   :3.000   Max.   :14.000   Max.   :9.00
##    FullBaths       HalfBaths       Fireplace       RemodelNone
##  Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :1.297   Mean   :0.6139   Mean   :0.7399   Mean   :0.7491
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   :3.0000   Max.   :4.0000   Max.   :1.0000
##    RemodelOld     RemodelRecent
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.1001   Mean   :0.1508
##  3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
```

```
WestRoxbury1Q.df <- filter(WestRoxbury.df, HomeValue < 325125)
summary(WestRoxbury1Q.df) #All Home Values are between 105000 and 325100
```

```
##    HomeValue         HomeTax       YrBuilt       LivingArea    PricePerSqFt
##  Min.   :105000   Min.   :1320   Min.   :1800   Min.   : 504   Min.   :105.3
##  1st Qu.:276950   1st Qu.:3484   1st Qu.:1925   1st Qu.:1020   1st Qu.:210.0
##  Median :295800   Median :3721   Median :1950   Median :1223   Median :239.1
##  Mean   :289970   Mean   :3647   Mean   :1942   Mean   :1234   Mean   :244.6
##  3rd Qu.:310350   3rd Qu.:3904   3rd Qu.:1959   3rd Qu.:1400   3rd Qu.:277.8
##  Max.   :325100   Max.   :4089   Max.   :1993   Max.   :2428   Max.   :488.3
##    TaxPerSqFt       LotSqft      PricePerLotSqFt  TaxPerLotSqFt
##  Min.   :1.324   Min.   :  997   Min.   : 19.71   Min.   :0.2479
##  1st Qu.:2.641   1st Qu.: 4000   1st Qu.: 50.69   1st Qu.:0.6377
##  Median :3.008   Median : 4728   Median : 61.52   Median :0.7738
##  Mean   :3.077   Mean   : 4986   Mean   : 62.88   Mean   :0.7909
##  3rd Qu.:3.494   3rd Qu.: 5827   3rd Qu.: 71.32   3rd Qu.:0.8971
##  Max.   :6.143   Max.   :13977   Max.   :238.97   Max.   :3.0057
##      Floors        TotalRooms       Bedrooms       FullBaths
##  Min.   :1.000   Min.   : 3.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.: 5.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :1.000   Median : 6.000   Median :3.000   Median :1.000
##  Mean   :1.317   Mean   : 5.966   Mean   :2.708   Mean   :1.132
##  3rd Qu.:1.500   3rd Qu.: 7.000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :2.500   Max.   :10.000   Max.   :5.000   Max.   :3.000
##    HalfBaths       Fireplace     RemodelNone      RemodelOld
##  Min.   :0.0000   Min.   :0.00   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00   Median :1.0000   Median :0.00000
##  Mean   :0.3611   Mean   :0.51   Mean   :0.8635   Mean   :0.08615
##  3rd Qu.:1.0000   3rd Qu.:1.00   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :2.0000   Max.   :2.00   Max.   :1.0000   Max.   :1.00000
##  RemodelRecent
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.05031
##  3rd Qu.:0.00000
##  Max.   :1.00000
```

```
#I might need the other quartiles as comparisons later, so I create them now
summary(WestRoxbury.df)
```

```
##    HomeValue          HomeTax         YrBuilt       LivingArea
##  Min.   : 105000   Min.   : 1320   Min.   :   0   Min.   : 504
##  1st Qu.: 325125   1st Qu.: 4090   1st Qu.:1920   1st Qu.:1308
##  Median : 375900   Median : 4728   Median :1935   Median :1548
##  Mean   : 392686   Mean   : 4939   Mean   :1937   Mean   :1657
##  3rd Qu.: 438775   3rd Qu.: 5520   3rd Qu.:1955   3rd Qu.:1874
##  Max.   :1217800   Max.   :15319   Max.   :2011   Max.   :5289
##   PricePerSqFt     TaxPerSqFt       LotSqft      PricePerLotSqFt
##  Min.   :105.3   Min.   :1.324   Min.   :  997   Min.   : 13.13
##  1st Qu.:216.3   1st Qu.:2.721   1st Qu.: 4772   1st Qu.: 54.87
##  Median :243.0   Median :3.056   Median : 5683   Median : 66.48
##  Mean   :245.1   Mean   :3.083   Mean   : 6278   Mean   : 67.54
##  3rd Qu.:270.1   3rd Qu.:3.398   3rd Qu.: 7022   3rd Qu.: 78.11
##  Max.   :489.0   Max.   :6.150   Max.   :46411   Max.   :262.07
##  TaxPerLotSqFt        Floors        TotalRooms       Bedrooms
##  Min.   :0.1651   Min.   :1.000   Min.   : 3.000   Min.   :1.00
##  1st Qu.:0.6902   1st Qu.:1.000   1st Qu.: 6.000   1st Qu.:3.00
##  Median :0.8362   Median :2.000   Median : 7.000   Median :3.00
##  Mean   :0.8496   Mean   :1.684   Mean   : 6.995   Mean   :3.23
##  3rd Qu.:0.9825   3rd Qu.:2.000   3rd Qu.: 8.000   3rd Qu.:4.00
##  Max.   :3.2965   Max.   :3.000   Max.   :14.000   Max.   :9.00
##    FullBaths       HalfBaths       Fireplace       RemodelNone
##  Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :1.297   Mean   :0.6139   Mean   :0.7399   Mean   :0.7491
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   :3.0000   Max.   :4.0000   Max.   :1.0000
##    RemodelOld     RemodelRecent
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.1001   Mean   :0.1508
##  3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
```

```
#There are multiple ways to filter() based on multiple criteria
#passed in one at a time
WestRoxbury2Q.df <- filter(WestRoxbury.df, HomeValue > 325125, HomeValue < 375900)
summary(WestRoxbury2Q.df) #Check your work
```

```
##    HomeValue         HomeTax        YrBuilt       LivingArea    PricePerSqFt
##  Min.   :325200   Min.   :4091   Min.   :1820   Min.   : 752   Min.   :128.4
##  1st Qu.:338375   1st Qu.:4256   1st Qu.:1925   1st Qu.:1288   1st Qu.:222.4
##  Median :350000   Median :4403   Median :1939   Median :1414   Median :248.3
##  Mean   :350057   Mean   :4403   Mean   :1938   Mean   :1442   Mean   :249.4
##  3rd Qu.:361800   3rd Qu.:4551   3rd Qu.:1954   3rd Qu.:1582   3rd Qu.:270.3
##  Max.   :375800   Max.   :4727   Max.   :2002   Max.   :2612   Max.   :489.0
##    TaxPerSqFt       LotSqft      PricePerLotSqFt  TaxPerLotSqFt
##  Min.   :1.615   Min.   : 1830   Min.   : 16.94   Min.   :0.2130
##  1st Qu.:2.797   1st Qu.: 4668   1st Qu.: 55.28   1st Qu.:0.6953
```

```
##  Median :3.123   Median : 5285   Median : 66.48   Median :0.8362
##  Mean   :3.137   Mean   : 5707   Mean   : 66.34   Mean   :0.8345
##  3rd Qu.:3.399   3rd Qu.: 6298   3rd Qu.: 75.67   3rd Qu.:0.9517
##  Max.   :6.150   Max.   :20000   Max.   :200.93   Max.   :2.5273
##      Floors       TotalRooms       Bedrooms       FullBaths
##  Min.   :1.00   Min.   : 4.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.00   1st Qu.: 6.000   1st Qu.:3.000   1st Qu.:1.000
##  Median :2.00   Median : 7.000   Median :3.000   Median :1.000
##  Mean   :1.63   Mean   : 6.628   Mean   :3.047   Mean   :1.199
##  3rd Qu.:2.00   3rd Qu.: 7.000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :2.50   Max.   :12.000   Max.   :6.000   Max.   :3.000
##    HalfBaths        Fireplace       RemodelNone       RemodelOld
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :0.00000
##  Mean   :0.5138   Mean   :0.6685   Mean   :0.8046   Mean   :0.08978
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :2.0000   Max.   :2.0000   Max.   :1.0000   Max.   :1.00000
##  RemodelRecent
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1057
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

```
#Using logical operators & and, or |
WestRoxbury3Q.df <- filter(WestRoxbury.df, HomeValue > 375900 & HomeValue < 438775)
WestRoxbury4Q.df <- filter(WestRoxbury.df, HomeValue > 438775)
```

#deplyr Verb 4 #summarise() is used to aggregate multiple values together and provide them to a function as a single result

```
summarise(WestRoxbury1Q.df, HomeValueMean = mean(HomeValue))
```

```
##   HomeValueMean
## 1      289970.1
```

```
#Many summarise() math functions can be seen using summary()
#summarise() is useful after preforming group_by() so we will return later
```

#deeplyr Verb 5 #group_by() applies a grouping for which operations will be preformed

```
WR1QRemodelGroup <- group_by(WestRoxbury1Q.df, RemodelNone)
head(WR1QRemodelGroup) #don't freak out that it dosen't look like anything
```

```
## # A tibble: 6 x 18
## # Groups:   RemodelNone [2]
##   HomeValue HomeTax YrBuilt LivingArea PricePerSqFt TaxPerSqFt LotSqft
##       <dbl>   <dbl>   <dbl>      <dbl>        <dbl>      <dbl>   <dbl>
## 1    320400    4030    1950       1200          267       3.36   10000
## 2    313000    3937    1960       1485          211.      2.65    5000
## 3    315500    3968    1889       1290          245.      3.08    5000
## 4    298200    3751    1940        864          345.      4.34    5000
## 5    313100    3938    1880       1438          218.      2.74    6949
## 6    317500    3994    1920       1232          258.      3.24    4450
## # ... with 11 more variables: PricePerLotSqFt <dbl>, TaxPerLotSqFt <dbl>,
```

```
## #   Floors <dbl>, TotalRooms <dbl>, Bedrooms <dbl>, FullBaths <dbl>,
## #   HalfBaths <dbl>, Fireplace <dbl>, RemodelNone <dbl>, RemodelOld <dbl>,
## #   RemodelRecent <dbl>
```

```
#a group_by() is like a placeholder to preform other operations
WR1QBedroomsGroup <- group_by(WestRoxbury1Q.df, Bedrooms)
head(WR1QBedroomsGroup)
```

```
## # A tibble: 6 x 18
## # Groups:   Bedrooms [1]
##   HomeValue HomeTax YrBuilt LivingArea PricePerSqFt TaxPerSqFt LotSqft
##       <dbl>   <dbl>   <dbl>      <dbl>        <dbl>      <dbl>   <dbl>
## 1    320400    4030    1950       1200          267       3.36   10000
## 2    313000    3937    1960       1485          211.      2.65    5000
## 3    315500    3968    1889       1290          245.      3.08    5000
## 4    298200    3751    1940        864          345.      4.34    5000
## 5    313100    3938    1880       1438          218.      2.74    6949
## 6    317500    3994    1920       1232          258.      3.24    4450
## # ... with 11 more variables: PricePerLotSqFt <dbl>, TaxPerLotSqFt <dbl>,
## #   Floors <dbl>, TotalRooms <dbl>, Bedrooms <dbl>, FullBaths <dbl>,
## #   HalfBaths <dbl>, Fireplace <dbl>, RemodelNone <dbl>, RemodelOld <dbl>,
## #   RemodelRecent <dbl>
```

```
#the most common being summarise()
```

#deplyr Verb 4 summarise() #to summarise(groupbyfunction, tibllecolumnname = mathfunction(columnname)) #summarise() code works in .r file is not working in .rmd file

```
#To show the mean home value via the remodel status in the lower quartile
summarise(WR1QBedroomsGroup, HomeValueMean = mean(HomeValue))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 2
##   Bedrooms HomeValueMean
##      <dbl>         <dbl>
## 1        1       269193.
## 2        2       282958.
## 3        3       293910.
## 4        4       297680.
## 5        5       293243.
```

```
#Within summarise() you can show multiple math functions
summarise(WR1QBedroomsGroup, HomeValueMean = mean(HomeValue), HomeValueMedian = median(HomeValue),
          HomeValueSd = sd(HomeValue), HomeValueInnerQuartile = IQR(HomeValue), HomeValueMin = min(Home
          HomeValueMax = max(HomeValue))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 7
##   Bedrooms HomeValueMean HomeValueMedian HomeValueSd HomeValueInnerQ~
##      <dbl>         <dbl>           <dbl>       <dbl>            <dbl>
## 1        1       269193.          278450      36474.            44525
## 2        2       282958.          288300      29224.            35850
## 3        3       293910.          298900      24802.            30200
## 4        4       297680.          305450      26563.            27225
## 5        5       293243.          299000      22973.            34700
## # ... with 2 more variables: HomeValueMin <dbl>, HomeValueMax <dbl>
```

```r
#Not always best to put all math functions in the same tibble
#interquartile range alone
summarise(WR1QBedroomsGroup, HomeValueQuartile = quantile(HomeValue))
```

```
## `summarise()` regrouping output by 'Bedrooms' (override with `.groups` argument)
```

```
## # A tibble: 25 x 2
## # Groups:   Bedrooms [5]
##    Bedrooms HomeValueQuartile
##       <dbl>             <dbl>
##  1        1            167600
##  2        1            247925
##  3        1            278450
##  4        1            292450
##  5        1            320700
##  6        2            105000
##  7        2            269150
##  8        2            288300
##  9        2            305000
## 10        2            325100
## # ... with 15 more rows
```

```r
#interquartile range added to above makes it diffcult to quickly read
summarise(WR1QBedroomsGroup, HomeValueMean = mean(HomeValue), HomeValueMedian = median(HomeValue),
          HomeValueSd = sd(HomeValue), HomeValueInnerQuartile = IQR(HomeValue), HomeValueMin = min(Home
          HomeValueMax = max(HomeValue), HomeValueQuartile = quantile(HomeValue))
```

```
## `summarise()` regrouping output by 'Bedrooms' (override with `.groups` argument)
```

```
## # A tibble: 25 x 8
## # Groups:   Bedrooms [5]
##    Bedrooms HomeValueMean HomeValueMedian HomeValueSd HomeValueInnerQ~
##       <dbl>         <dbl>           <dbl>       <dbl>            <dbl>
##  1        1       269193.          278450      36474.            44525
##  2        1       269193.          278450      36474.            44525
##  3        1       269193.          278450      36474.            44525
##  4        1       269193.          278450      36474.            44525
##  5        1       269193.          278450      36474.            44525
##  6        2       282958.          288300      29224.            35850
##  7        2       282958.          288300      29224.            35850
##  8        2       282958.          288300      29224.            35850
##  9        2       282958.          288300      29224.            35850
## 10        2       282958.          288300      29224.            35850
## # ... with 15 more rows, and 3 more variables: HomeValueMin <dbl>,
## #   HomeValueMax <dbl>, HomeValueQuartile <dbl>
```

```r
#Save the easier to read tibble
WR1QSummaryBedroom.tb <- summarise(WR1QBedroomsGroup, HomeValueMean = mean(HomeValue), HomeValueMedian =
                                   median(HomeValue), HomeValueSd = sd(HomeValue), HomeValueInnerQua
                                   IQR(HomeValue), HomeValueMin = min(HomeValue), HomeValueMax =
                                   max(HomeValue))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
#Not all functions allow you to list the column
#n() is used for the count of bedrooms
summarise(WR1QBedroomsGroup, BedroomsCount = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 2
##   Bedrooms BedroomsCount
##      <dbl>         <int>
## 1        1            28
## 2        2           499
## 3        3           799
## 4        4           118
## 5        5             7
```

```
#n_distinct() shows the number of district HomeValues per Bedrooms
summarise(WR1QBedroomsGroup, BedroomsDistinct = n_distinct(HomeValue))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 2
##   Bedrooms BedroomsDistinct
##      <dbl>            <int>
## 1        1               28
## 2        2              396
## 3        3              475
## 4        4              107
## 5        5                7
```

```
#By adding them to the same tibble you can see there is a unique home value for every 1 bedroom home
summarise(WR1QBedroomsGroup, BedroomsCount = n(), BedroomsDistinct = n_distinct(HomeValue))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 3
##   Bedrooms BedroomsCount BedroomsDistinct
##      <dbl>         <int>            <int>
## 1        1            28               28
## 2        2           499              396
## 3        3           799              475
## 4        4           118              107
## 5        5             7                7
```

```
#You can also preform math within a summarise()
WR1QBedroomsUniqueValue.tb <- summarise(WR1QBedroomsGroup, BedroomsCount = n(), BedroomsDistinct = n_dis
          BedroomsSharedValues = BedroomsCount - BedroomsDistinct)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

#Dplyr Verb 6 arrange() #arrange() is similar to sort command in Excel

```
summary(WestRoxbury1Q.df) #check column names and ranges
```

```
##    HomeValue         HomeTax        YrBuilt       LivingArea    PricePerSqFt
##  Min.   :105000   Min.   :1320   Min.   :1800   Min.   : 504   Min.   :105.3
##  1st Qu.:276950   1st Qu.:3484   1st Qu.:1925   1st Qu.:1020   1st Qu.:210.0
##  Median :295800   Median :3721   Median :1950   Median :1223   Median :239.1
##  Mean   :289970   Mean   :3647   Mean   :1942   Mean   :1234   Mean   :244.6
##  3rd Qu.:310350   3rd Qu.:3904   3rd Qu.:1959   3rd Qu.:1400   3rd Qu.:277.8
##  Max.   :325100   Max.   :4089   Max.   :1993   Max.   :2428   Max.   :488.3
##    TaxPerSqFt       LotSqft      PricePerLotSqFt  TaxPerLotSqFt
##  Min.   :1.324   Min.   :  997   Min.   : 19.71   Min.   :0.2479
##  1st Qu.:2.641   1st Qu.: 4000   1st Qu.: 50.69   1st Qu.:0.6377
```

```
##  Median :3.008   Median : 4728   Median : 61.52   Median :0.7738
##  Mean   :3.077   Mean   : 4986   Mean   : 62.88   Mean   :0.7909
##  3rd Qu.:3.494   3rd Qu.: 5827   3rd Qu.: 71.32   3rd Qu.:0.8971
##  Max.   :6.143   Max.   :13977   Max.   :238.97   Max.   :3.0057
##      Floors        TotalRooms        Bedrooms       FullBaths
##  Min.   :1.000   Min.   : 3.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.: 5.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :1.000   Median : 6.000   Median :3.000   Median :1.000
##  Mean   :1.317   Mean   : 5.966   Mean   :2.708   Mean   :1.132
##  3rd Qu.:1.500   3rd Qu.: 7.000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :2.500   Max.   :10.000   Max.   :5.000   Max.   :3.000
##    HalfBaths        Fireplace     RemodelNone       RemodelOld
##  Min.   :0.0000   Min.   :0.00   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00   Median :1.0000   Median :0.00000
##  Mean   :0.3611   Mean   :0.51   Mean   :0.8635   Mean   :0.08615
##  3rd Qu.:1.0000   3rd Qu.:1.00   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :2.0000   Max.   :2.00   Max.   :1.0000   Max.   :1.00000
##  RemodelRecent
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.05031
##  3rd Qu.:0.00000
##  Max.   :1.00000
```

```
head(arrange(WestRoxbury.df, HomeValue)) #head() lets you check your work
```

```
##     HomeValue HomeTax YrBuilt LivingArea PricePerSqFt TaxPerSqFt LotSqft
## 326    105000    1320    1910        504     208.3333   2.619048     997
## 259    144600    1819    1920        797     181.4304   2.282309    1017
## 245    167600    2108    1900        690     242.8986   3.055072    1980
## 609    171800    2161    1912        754     227.8515   2.866048    2313
## 578    176900    2225    1920       1680     105.2976   1.324405    1037
## 311    177400    2231    1910        600     295.6667   3.718333    3624
##     PricePerLotSqFt TaxPerLotSqFt Floors TotalRooms Bedrooms FullBaths
## 326       105.31595     1.3239719    1.0          4        2         1
## 259       142.18289     1.7885939    1.5          5        2         1
## 245        84.64646     1.0646465    1.0          3        1         1
## 609        74.27583     0.9342845    1.0          5        2         1
## 578       170.58824     2.1456123    2.0          6        3         1
## 311        48.95143     0.6156181    1.0          5        2         1
##     HalfBaths Fireplace RemodelNone RemodelOld RemodelRecent
## 326         0         0           1          0             0
## 259         0         0           1          0             0
## 245         0         0           1          0             0
## 609         0         0           1          0             0
## 578         0         0           1          0             0
## 311         0         0           1          0             0
```

```
#Use desc() to arrange from highest to lowest
head(arrange(WestRoxbury1Q.df, desc(HomeValue)))
```

```
##      HomeValue HomeTax YrBuilt LivingArea PricePerSqFt TaxPerSqFt LotSqft
## 1606    325100    4089    1935        840     387.0238   4.867857    7220
```

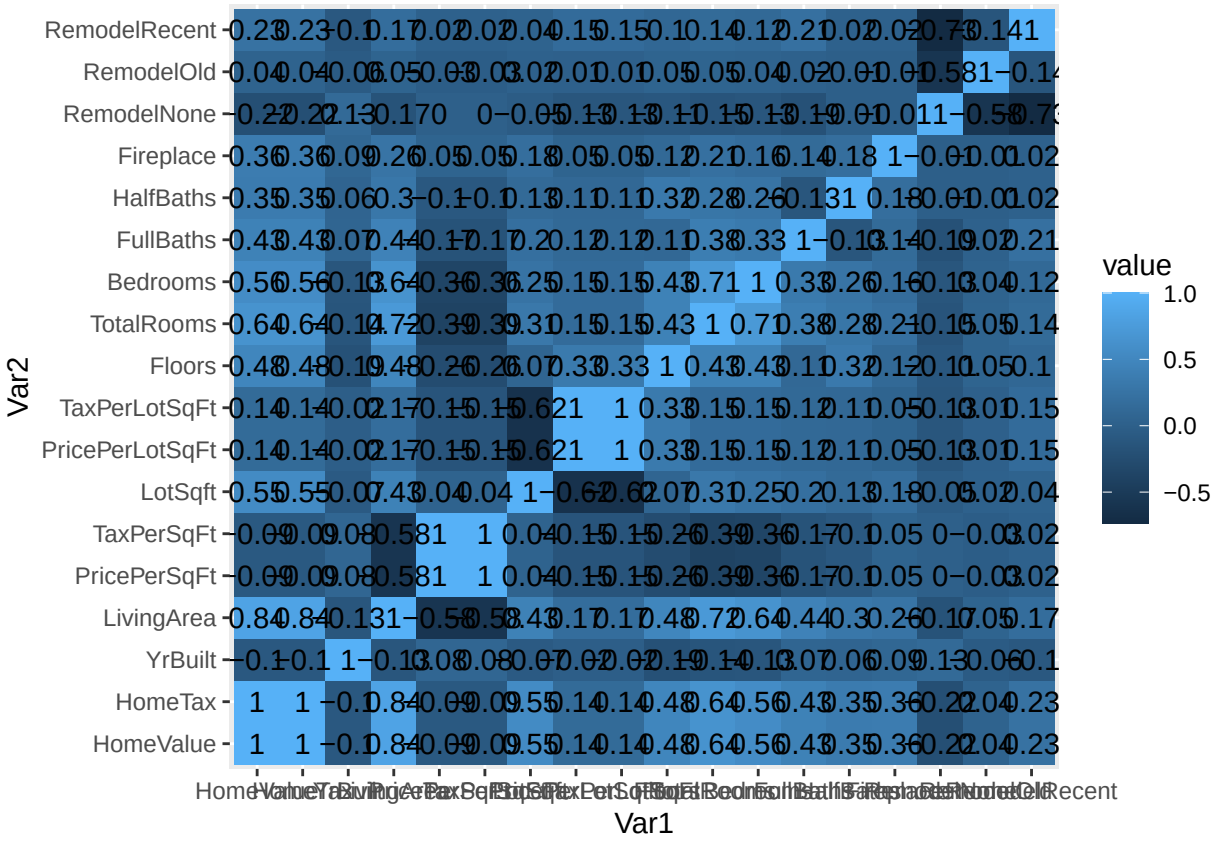
```
## 236     325000    4088    1910       1276     254.7022   3.203762    3850
## 1154    325000    4088    1940       1248     260.4167   3.275641    3168
## 1251    325000    4088    1890       1932     168.2195   2.115942    9714
## 4251    325000    4088    1930       1330     244.3609   3.073684    4026
## 582     324900    4087    1960       1487     218.4936   2.748487    5668
##      PricePerLotSqFt TaxPerLotSqFt Floors TotalRooms Bedrooms FullBaths
## 1606        45.02770     0.5663435      1          5        2         1
## 236         84.41558     1.0618182      2          8        3         1
## 1154       102.58838     1.2904040      2          7        3         1
## 1251        33.45687     0.4208359      2          8        4         2
## 4251        80.72529     1.0153999      1          7        3         2
## 582         57.32181     0.7210656      1          7        4         2
##      HalfBaths Fireplace RemodelNone RemodelOld RemodelRecent
## 1606         0         0           0          1             0
## 236          1         0           0          1             0
## 1154         1         1           0          0             1
## 1251         0         0           1          0             0
## 4251         0         0           0          1             0
## 582          0         0           1          0             0
```

```
#order matters when arranging by multiple columns
#Homes are ordered by price, if there are 2 homes with the same price then ordered by bedroom
#view(arrange(WestRoxbury1Qt.df, HomeValue, Bedrooms))
#Homes are ordered so all 1 bedrooms are arranged in assending value, followed by all 2 bedrooms
#view(arrange(WestRoxbury1Qt.df, Bedrooms, HomeValue))
```

#To Visualize the Data using ggplot2

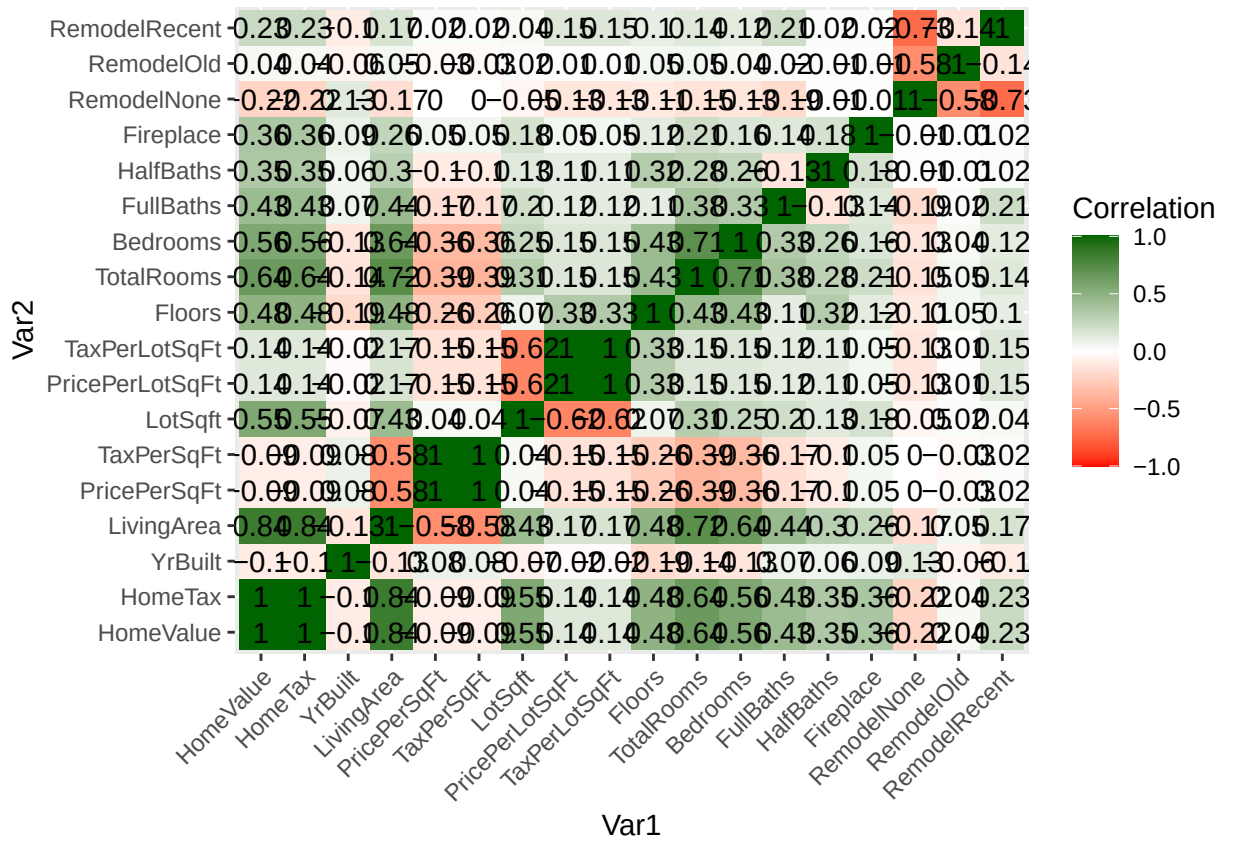
```
library(ggplot2)
library(reshape2)

#To determine if any of the Remodel Factors have an impact on our dataset we
#Create a Matrix using ggplot
cor.mat <- round(cor(WestRoxbury.df), 2)
melted.cor.mat <- melt(cor.mat)
ggplot(melted.cor.mat, aes(x=Var1, y=Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(x=Var1, y=Var2, label = value))
```



#Improving Correlation Matrix #The above Matrix is really hard to read #To improve legiability we can change the orientation of the labels on the X axis and update the colors.

```
ggplot(data = melted.cor.mat, aes(x=Var1, y=Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(x=Var1, y=Var2, label = value)) +
  scale_fill_gradient2(low = "red", high = "darkgreen", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Correlation") +
  theme(axis.text.x = element_text(angle = 45, size = 9, hjust = 1))
```



#To Make it More legible #Narrow Variables Using select() to help narrow down our plot

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

```
##    HomeValue          HomeTax         YrBuilt       LivingArea
##  Min.   : 105000   Min.   : 1320   Min.   :   0   Min.   : 504
##  1st Qu.: 325125   1st Qu.: 4090   1st Qu.:1920   1st Qu.:1308
##  Median : 375900   Median : 4728   Median :1935   Median :1548
##  Mean   : 392686   Mean   : 4939   Mean   :1937   Mean   :1657
##  3rd Qu.: 438775   3rd Qu.: 5520   3rd Qu.:1955   3rd Qu.:1874
##  Max.   :1217800   Max.   :15319   Max.   :2011   Max.   :5289
##   PricePerSqFt     TaxPerSqFt       LotSqft      PricePerLotSqFt
##  Min.   :105.3   Min.   :1.324   Min.   :  997   Min.   : 13.13
##  1st Qu.:216.3   1st Qu.:2.721   1st Qu.: 4772   1st Qu.: 54.87
##  Median :243.0   Median :3.056   Median : 5683   Median : 66.48
##  Mean   :245.1   Mean   :3.083   Mean   : 6278   Mean   : 67.54
##  3rd Qu.:270.1   3rd Qu.:3.398   3rd Qu.: 7022   3rd Qu.: 78.11
##  Max.   :489.0   Max.   :6.150   Max.   :46411   Max.   :262.07
##  TaxPerLotSqFt       Floors        TotalRooms       Bedrooms
##  Min.   :0.1651   Min.   :1.000   Min.   : 3.000   Min.   :1.00
##  1st Qu.:0.6902   1st Qu.:1.000   1st Qu.: 6.000   1st Qu.:3.00
##  Median :0.8362   Median :2.000   Median : 7.000   Median :3.00
##  Mean   :0.8496   Mean   :1.684   Mean   : 6.995   Mean   :3.23
##  3rd Qu.:0.9825   3rd Qu.:2.000   3rd Qu.: 8.000   3rd Qu.:4.00
##  Max.   :3.2965   Max.   :3.000   Max.   :14.000   Max.   :9.00
##    FullBaths       HalfBaths        Fireplace       RemodelNone
##  Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
```

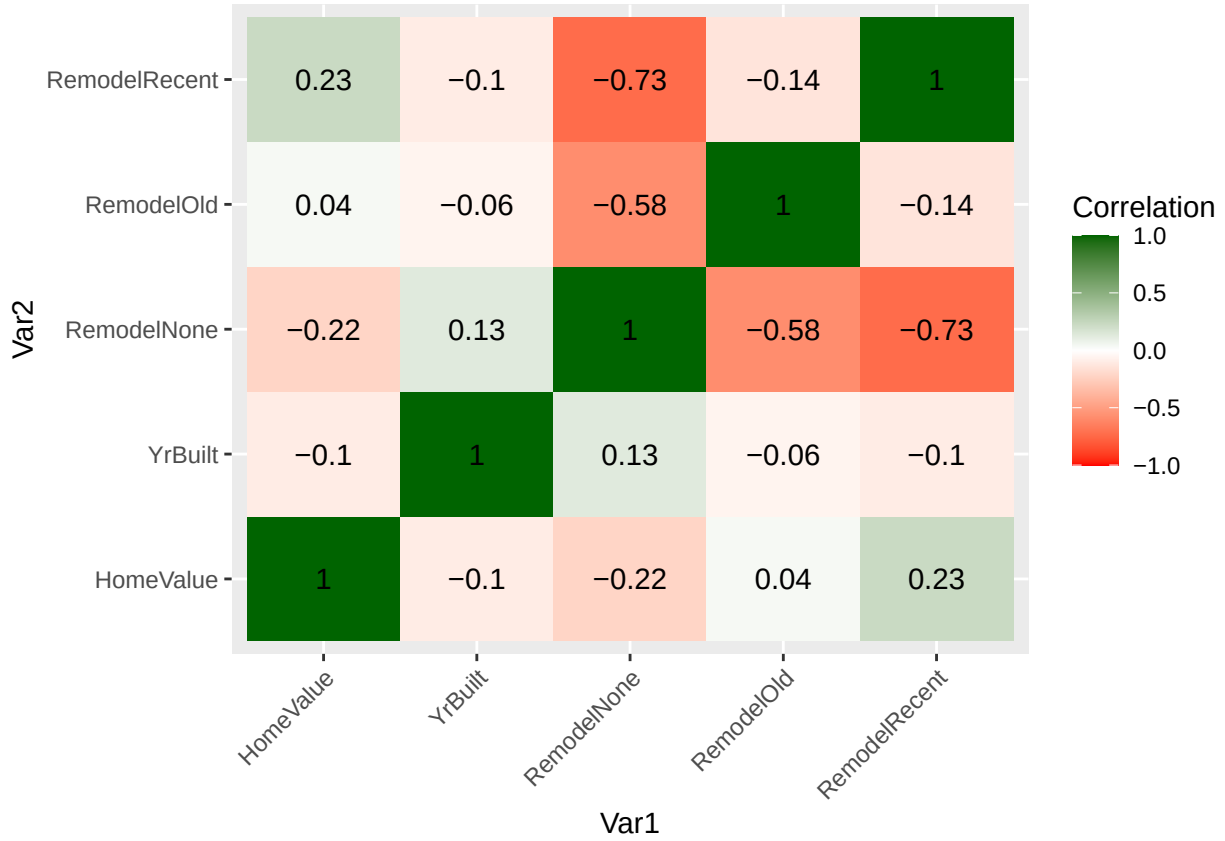
```
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :1.297   Mean   :0.6139   Mean   :0.7399   Mean   :0.7491
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   :3.0000   Max.   :4.0000   Max.   :1.0000
##    RemodelOld     RemodelRecent
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.1001   Mean   :0.1508
##  3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
```

```
WestRoxburyRemodel.df <- select(WestRoxbury.df, HomeValue, YrBuilt,
                                RemodelNone, RemodelOld, RemodelRecent)
head(WestRoxburyRemodel.df)
```

```
##   HomeValue YrBuilt RemodelNone RemodelOld RemodelRecent
## 1    344200    1880           1          0             0
## 2    412600    1945           0          0             1
## 3    330100    1890           1          0             0
## 4    498600    1957           1          0             0
## 5    331500    1910           1          0             0
## 6    337400    1950           0          1             0
```

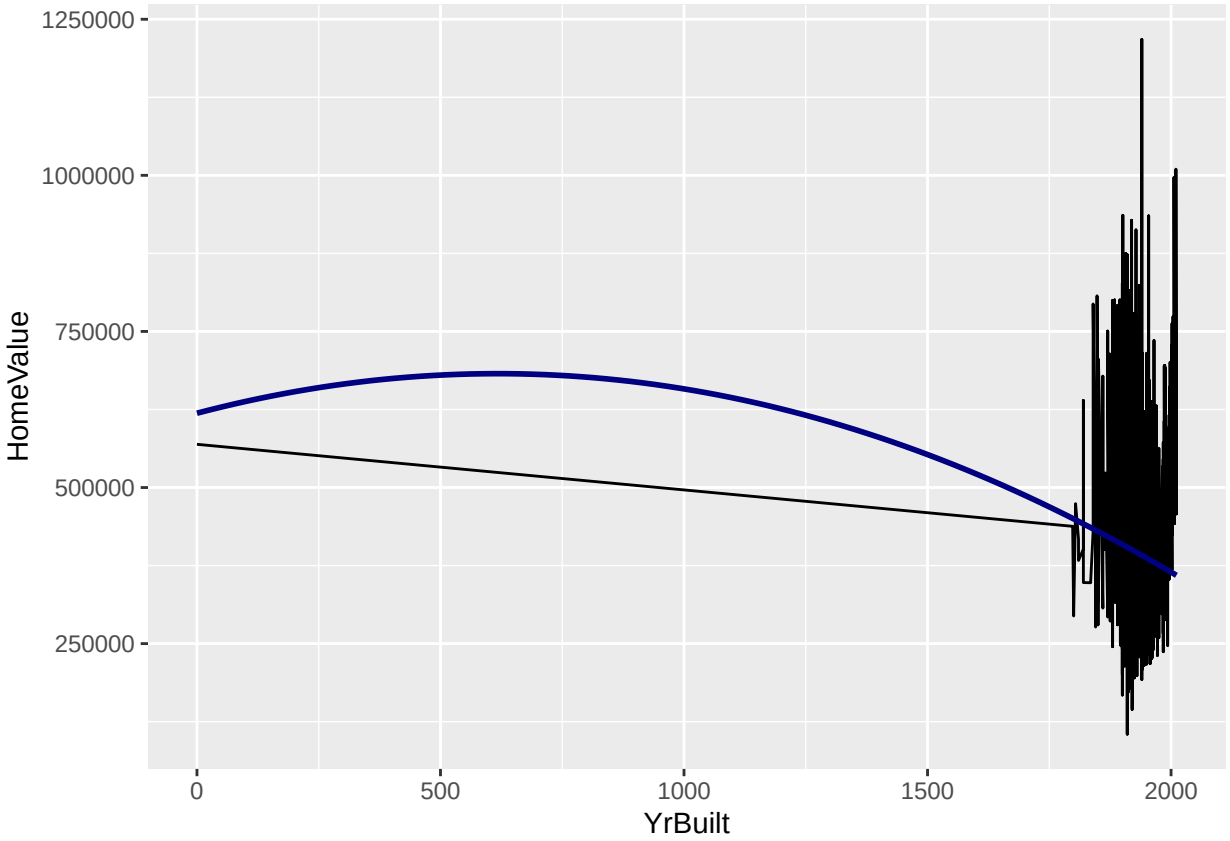
#Then we create a correlation matrix for the remodel factors

```
cor.mat.remodel <- round(cor(WestRoxburyRemodel.df), 2)
melted.cor.mat.remodel <- melt(cor.mat.remodel)
ggplot(data = melted.cor.mat.remodel, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  geom_text(aes(x=Var1, y=Var2, label = value)) +
  scale_fill_gradient2(low = "red", high = "darkgreen", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Correlation") +
  theme(axis.text.x = element_text(angle = 45, size = 9, hjust = 1))
```




```
#Scatterplot of Home Value versus Year Built

ggplot(WestRoxbury.df, aes(YrBuilt, HomeValue)) +
  geom_line() +
  geom_smooth(formula = y ~ poly(x, 2), method = "lm", colour = "navy", se = FALSE, na.rm = TRUE)
```



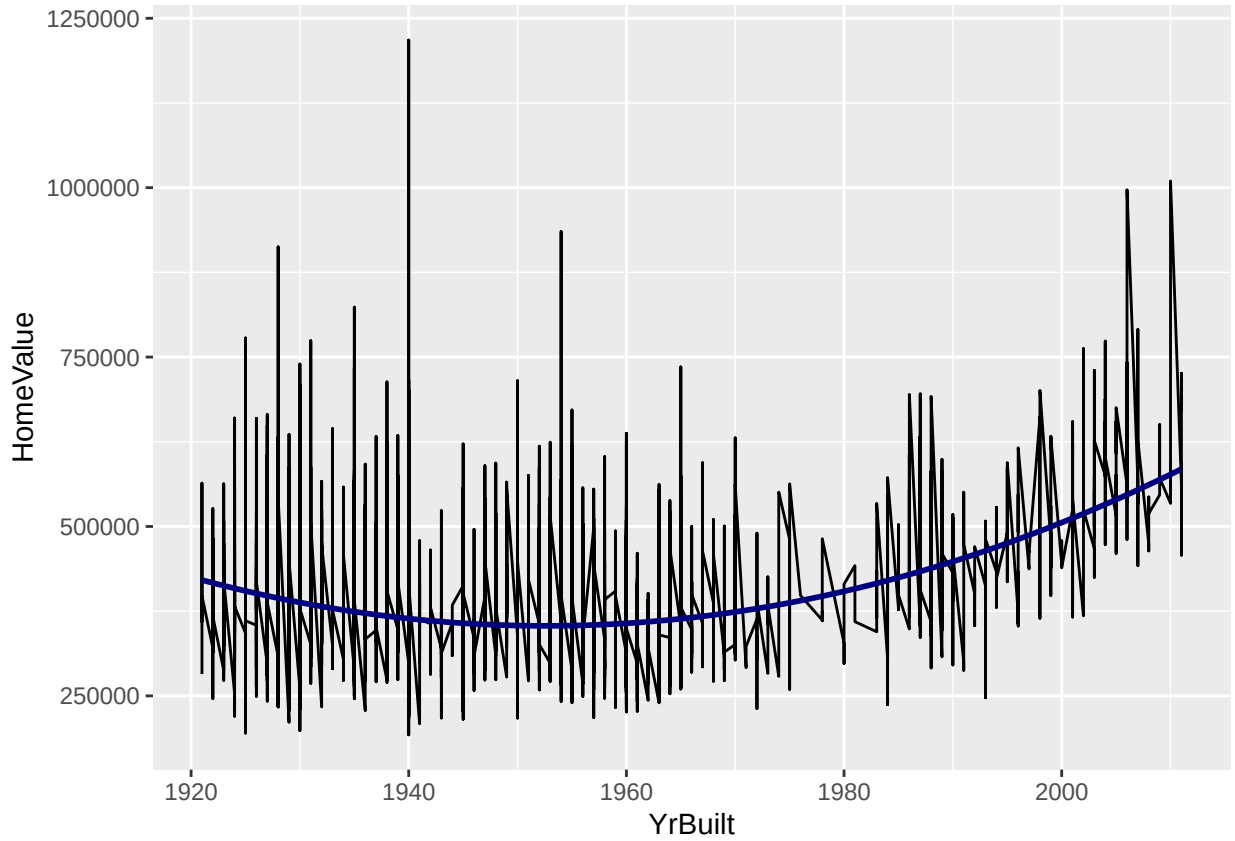
#Narrow to years in the dataset to make it more useful

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

```
##    HomeValue          HomeTax         YrBuilt       LivingArea
##  Min.   : 105000   Min.   : 1320   Min.   :   0   Min.   : 504
##  1st Qu.: 325125   1st Qu.: 4090   1st Qu.:1920   1st Qu.:1308
##  Median : 375900   Median : 4728   Median :1935   Median :1548
##  Mean   : 392686   Mean   : 4939   Mean   :1937   Mean   :1657
##  3rd Qu.: 438775   3rd Qu.: 5520   3rd Qu.:1955   3rd Qu.:1874
##  Max.   :1217800   Max.   :15319   Max.   :2011   Max.   :5289
##   PricePerSqFt     TaxPerSqFt       LotSqft      PricePerLotSqFt
##  Min.   :105.3   Min.   :1.324   Min.   :  997   Min.   : 13.13
##  1st Qu.:216.3   1st Qu.:2.721   1st Qu.: 4772   1st Qu.: 54.87
##  Median :243.0   Median :3.056   Median : 5683   Median : 66.48
##  Mean   :245.1   Mean   :3.083   Mean   : 6278   Mean   : 67.54
##  3rd Qu.:270.1   3rd Qu.:3.398   3rd Qu.: 7022   3rd Qu.: 78.11
##  Max.   :489.0   Max.   :6.150   Max.   :46411   Max.   :262.07
##  TaxPerLotSqFt        Floors        TotalRooms       Bedrooms
##  Min.   :0.1651   Min.   :1.000   Min.   : 3.000   Min.   :1.00
##  1st Qu.:0.6902   1st Qu.:1.000   1st Qu.: 6.000   1st Qu.:3.00
##  Median :0.8362   Median :2.000   Median : 7.000   Median :3.00
##  Mean   :0.8496   Mean   :1.684   Mean   : 6.995   Mean   :3.23
##  3rd Qu.:0.9825   3rd Qu.:2.000   3rd Qu.: 8.000   3rd Qu.:4.00
##  Max.   :3.2965   Max.   :3.000   Max.   :14.000   Max.   :9.00
##    FullBaths       HalfBaths       Fireplace       RemodelNone
##  Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
```

```
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :1.297   Mean   :0.6139   Mean   :0.7399   Mean   :0.7491
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   :3.0000   Max.   :4.0000   Max.   :1.0000
##    RemodelOld     RemodelRecent
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.1001   Mean   :0.1508
##  3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
```

```
WestRoxburyYears.df <- filter(WestRoxbury.df, YrBuilt > 1920)
ggplot(WestRoxburyYears.df, aes(YrBuilt, HomeValue)) +
  geom_line() +
  geom_smooth(formula = y ~ poly(x, 2), method = "lm", colour = "navy", se = FALSE, na.rm = TRUE)
```



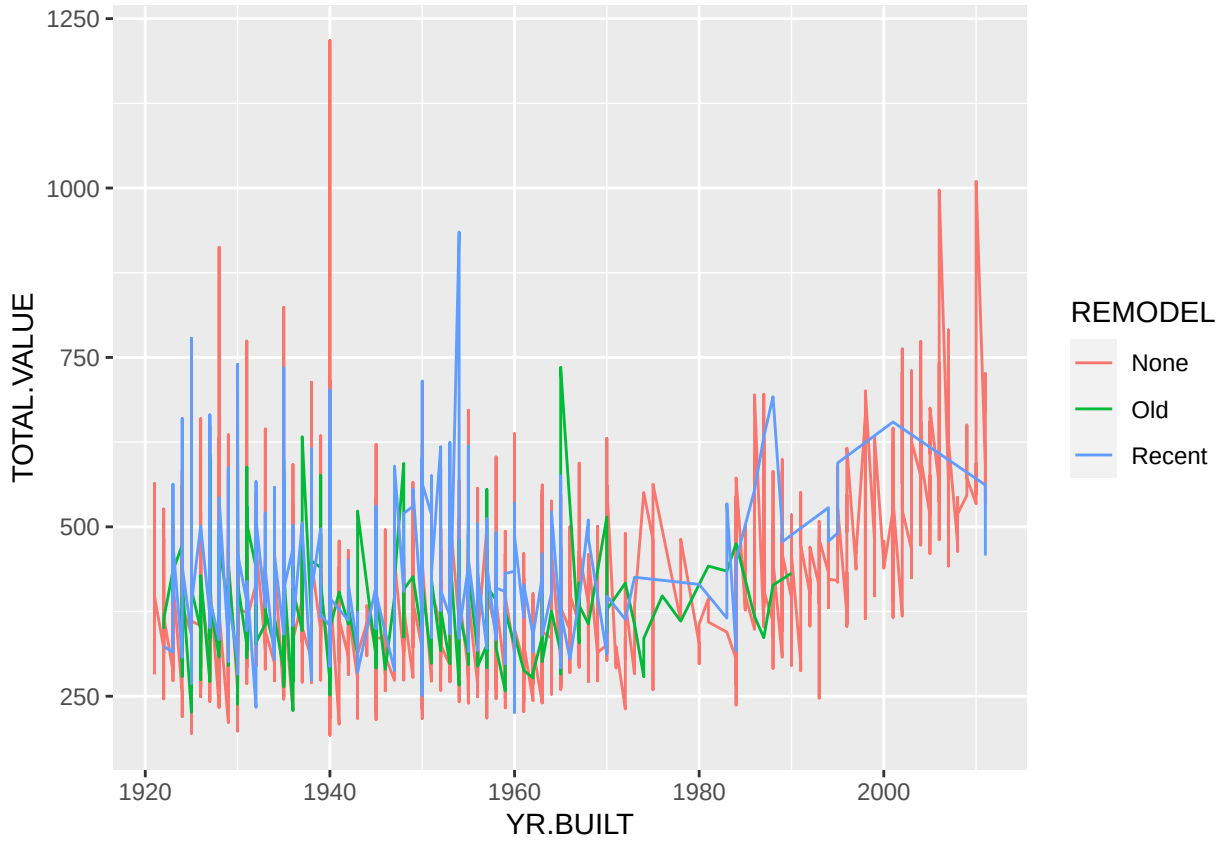

```
#We can see from the plot that the tax amount is positively correlated to the home value
#Simple Visualizations: Comparing Home Value & Bedrooms
```

```
summary(WestRoxburyBackup)
```

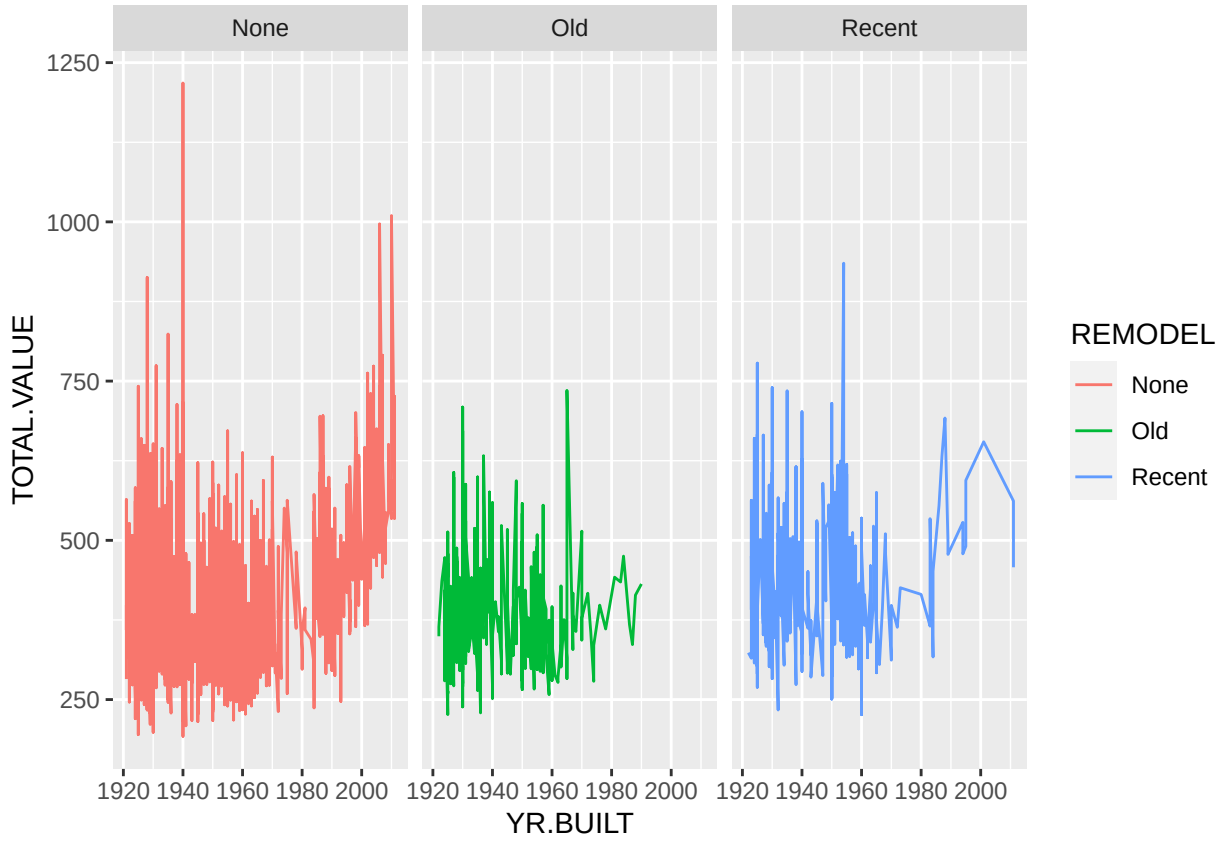
```
##   TOTAL.VALUE          TAX           LOT.SQFT        YR.BUILT      GROSS.AREA
##  Min.   : 105.0   Min.   : 1320   Min.   :  997   Min.   :   0   Min.   : 821
##  1st Qu.: 325.1   1st Qu.: 4090   1st Qu.: 4772   1st Qu.:1920   1st Qu.:2347
##  Median : 375.9   Median : 4728   Median : 5683   Median :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
```

```
WestRoxburyBackupYears.df <- filter(WestRoxburyBackup, YR.BUILT > 1920)
ggplot(WestRoxburyBackupYears.df, aes(YR.BUILT, TOTAL.VALUE, color = REMODEL)) +
  geom_line()
```




```
WestRoxburyBackupYears.df <- filter(WestRoxburyBackup, YR.BUILT > 1920)
ggplot(WestRoxburyBackupYears.df, aes(YR.BUILT, TOTAL.VALUE, color = REMODEL)) +
  geom_line() +
  facet_wrap(facets = vars(REMODEL))
```




```
summary(WestRoxburyBackup)
```

```
##   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
```

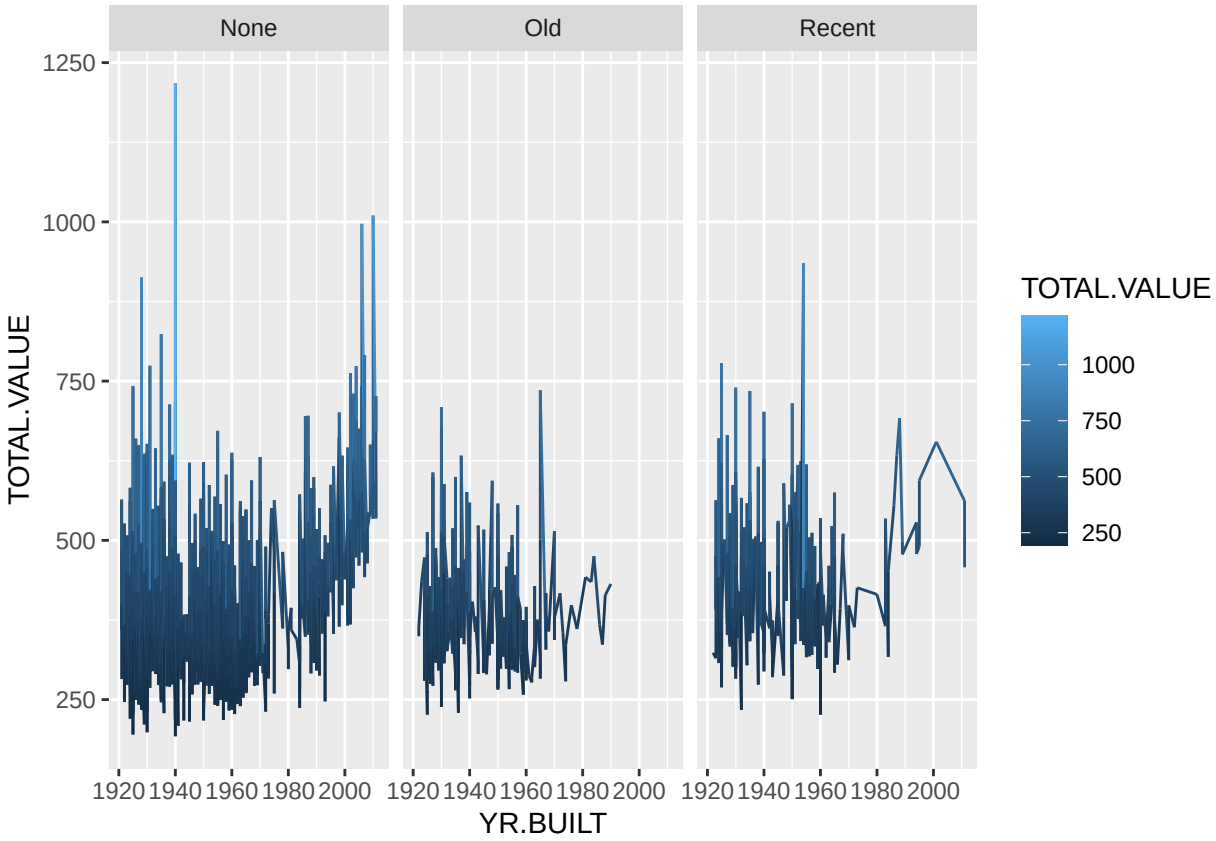
```
#Create New Columns within the WestRoxburyBackupdataset to split by quartile
TotalValue1Q <- WestRoxburyBackup$TOTAL.VALUE < 325
WestRoxburyBackup$TotalValue1Q <- c(TotalValue1Q)
```

```r
#2nd Quartile
TotalValue2Q <- WestRoxburyBackup$TOTAL.VALUE > 325 & WestRoxburyBackup$TOTAL.VALUE < 392
WestRoxburyBackup$TotalValue2Q <- c(TotalValue2Q)
#3rd Quartile
TotalValue3Q <- WestRoxburyBackup$TOTAL.VALUE > 392 & WestRoxburyBackup$TOTAL.VALUE < 438
WestRoxburyBackup$TotalValue3Q <- c(TotalValue3Q)
#4th Quartile
TotalValue4Q <- WestRoxburyBackup$TOTAL.VALUE > 438
WestRoxburyBackup$TotalValue4Q <- c(TotalValue4Q)

head(WestRoxburyBackup)
```

```
##   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 TotalValue1Q
## 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        FALSE
## 5        3         2         0       1         0    None        FALSE
## 6        3         1         0       1         1     Old        FALSE
##   TotalValue2Q TotalValue3Q TotalValue4Q
## 1         TRUE        FALSE        FALSE
## 2        FALSE         TRUE        FALSE
## 3         TRUE        FALSE        FALSE
## 4        FALSE        FALSE         TRUE
## 5         TRUE        FALSE        FALSE
## 6         TRUE        FALSE        FALSE
```

```r
WestRoxburyBackupYears.df <- filter(WestRoxburyBackup, YR.BUILT > 1920)
ggplot(WestRoxburyBackupYears.df, aes(YR.BUILT, TOTAL.VALUE, color = TOTAL.VALUE)) +
  geom_line() +
  facet_wrap(facets = vars(REMODEL))
```



#I have to stop improving the graphs or I'll fail the midterm because I didn't study. #I want to put the above with a line for each quartile