Housing

Include libraries:

```
library(tidyverse)
library(ggplot2)
library(readr)
```

Links and ideas:

VERY HELPFUL LINK: https://rstudio-pubs-static.s3.amazonaws.com/150743_fbe2be64165349798440e35351653b16.html

https://www.datacamp.com/community/tutorials/linear-regression-R#coefficients 1) look at p-value –> H0:b=0, H1:b!=0. accept H0 if p>a (i.e.) for large p. We want a small p!, b should be different from 0! In simple terms, a p-value indicates whether or not you can reject or accept a hypothesis. The hypothesis, in this case, is that the predictor is not meaningful for your model. 2) look at R2 –> the better the closer to 1. R2=(explained variation of the model)/(Total variation of the model). R2 = 0.99 –> explains 99% of the variability. In the blue rectangle, notice that there's two different R^2, one multiple and one adjusted. The multiple is the R^2 that you saw previously. One problem with this R^2 is that it cannot decrease as you add more independent variables to your model, it will continue increasing as you make the model more complex, even if these variables don't add anything to your predictions (like the example of the number of siblings). For this reason, the adjusted R^2 is probably better to look at if you are adding more than one variable to the model, since it only increases if it reduces the overall error of the predictions. 3) plot the residuals –> plot(lm_model$residuals). Should be random, no pattern! If pattern -> If you have more data, your simple linear model will not be able to generalize well.

Multicolinearity: http://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis Multicollinearity increases the standard errors of the coefficients. Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from 0. In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant. Without multicollinearity (and thus, with lower standard errors), those coefficients might be significant.

idea: 1)visualize one linear regression with 1 variable 2)add them all 3)analyse and remove, come up with new ones (e.g. muliply # of bedrooms with . . .)

Introduction

The goal of this thesis is to study and predict housing prices (response variable) based on predictor variables that provide data on the physical characteristics of properties, such as their size, age, location, etc.

Exploratory data analysis:

Lets have a short look at our data set:

```
#read data in and have a look at it:
house <- read_csv("data/realestate.csv")
head(house)
```

```
## # A tibble: 6 x 12
##      ID  Price  Sqft Bedroom Bathroom Airconditioning Garage  Pool
##   <int>  <int> <int>   <int>    <int>           <int>  <int> <int>
## 1     1 360000  3032       4        4               1      2     0
## 2     2 340000  2058       4        2               1      2     0
## 3     3 250000  1780       4        3               1      2     0
## 4     4 205500  1638       4        2               1      2     0
## 5     5 275500  2196       4        3               1      2     0
## 6     6 248000  1966       4        3               1      5     1
## # ... with 4 more variables: YearBuild <int>, Quality <int>, Lot <int>,
## #   AdjHighway <int>
```

```
#we only work with "house":
attach(house)

#delete column "ID":
house <- house[,-1]

#dim of house:
dim(house)
```

```
## [1] 522  11
```

```
names(house)
```

```
##  [1] "Price"           "Sqft"            "Bedroom"
##  [4] "Bathroom"        "Airconditioning" "Garage"
##  [7] "Pool"            "YearBuild"       "Quality"
## [10] "Lot"             "AdjHighway"
```

```
#is "Pool" binary or does t count the number of pools?
house %>% select(Pool) %>% distinct()
```

```
## # A tibble: 2 x 1
##    Pool
##   <int>
## 1     0
## 2     1
```

```
#which values does "Quality" take?:
house %>% select(Quality) %>% distinct()
```

```
## # A tibble: 3 x 1
##   Quality
##     <int>
## 1       2
## 2       3
## 3       1
```

```
#which values does "AdjHighway" take?:
house %>% select(AdjHighway) %>% distinct()
```

```
## # A tibble: 2 x 1
##   AdjHighway
##        <int>
## 1          0
## 2          1
```

```
View(house)
```

We store our data set "realestat.csv" into a variable called house. The first insight is that the column "ID" is not needed, as it only stores the index of the houses. As a result, we can delete this column. The data set now contains 11 columns and 522 rows. The variables are as follows: "Price" (quantitative; the price of the property in question), "Sqft" (quantitative; size), "Bedroom" (quantitative; number of bedrooms), "Bathroom" (quantitative; number of bathrooms), "Airconditioning" (qualitative; yes: 1, no: 0), "Garage" (quantitative; number of garages (or parking spots)), "Pool" (qualitative; yes: 1, no: 0), "YearBuild" (qualitative; year in which the house was built), "Quality" (qualitative; the build-quality of the house on a scale of 1 to 3), "Lot" (quantitative, the surface of the land), and finally "AdjHighway"(qualitative; 1: next to a highway, 0: not close or next to a highway).

EDA

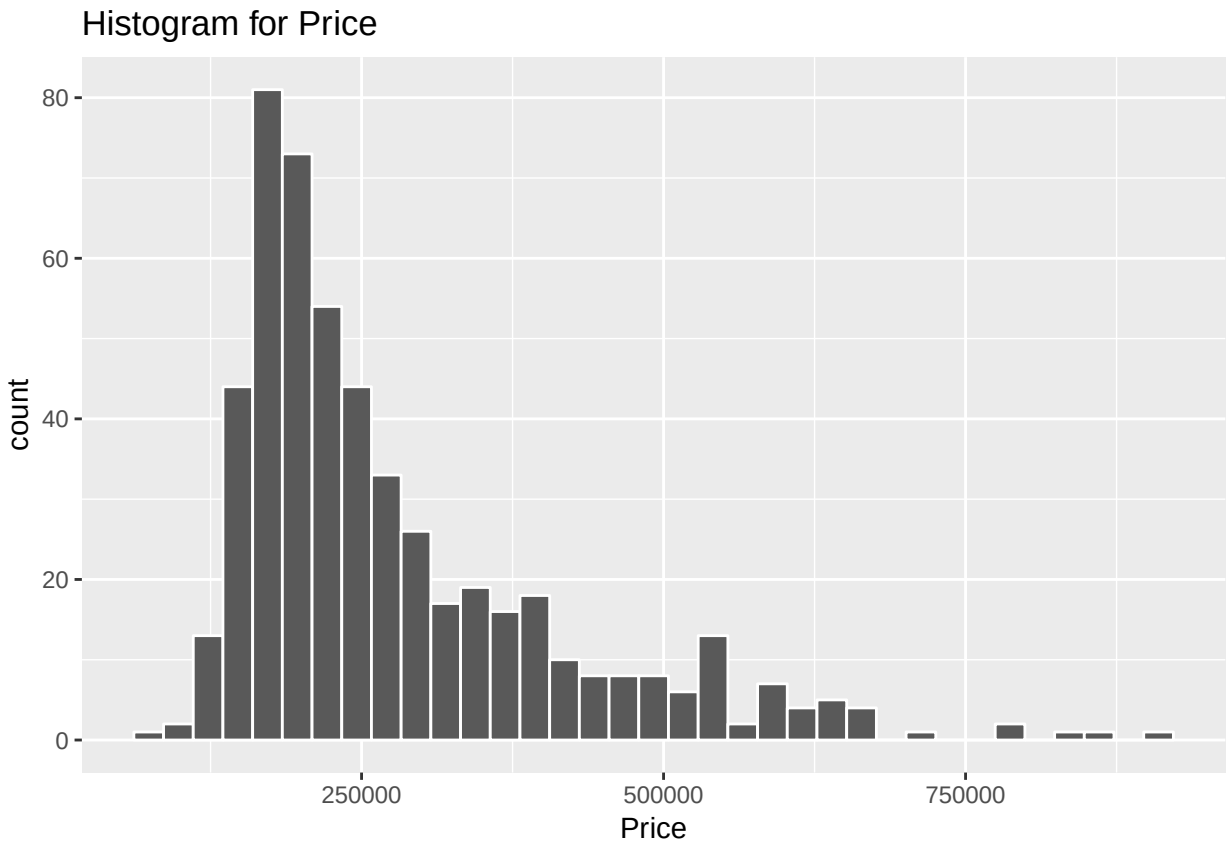
First, we are going to study the variables themselves and their impact on the price of the house. Next, we are going to generate to simple regression models based on the results of the analysis. Finally, we will look at the relations that might exist amongst te variables in order to remove a few of them from oure model (to improve it by reducing the effects of multicollinearity) and to create new variables that might improve our model.

Lets start by analysing the "Price" variable:

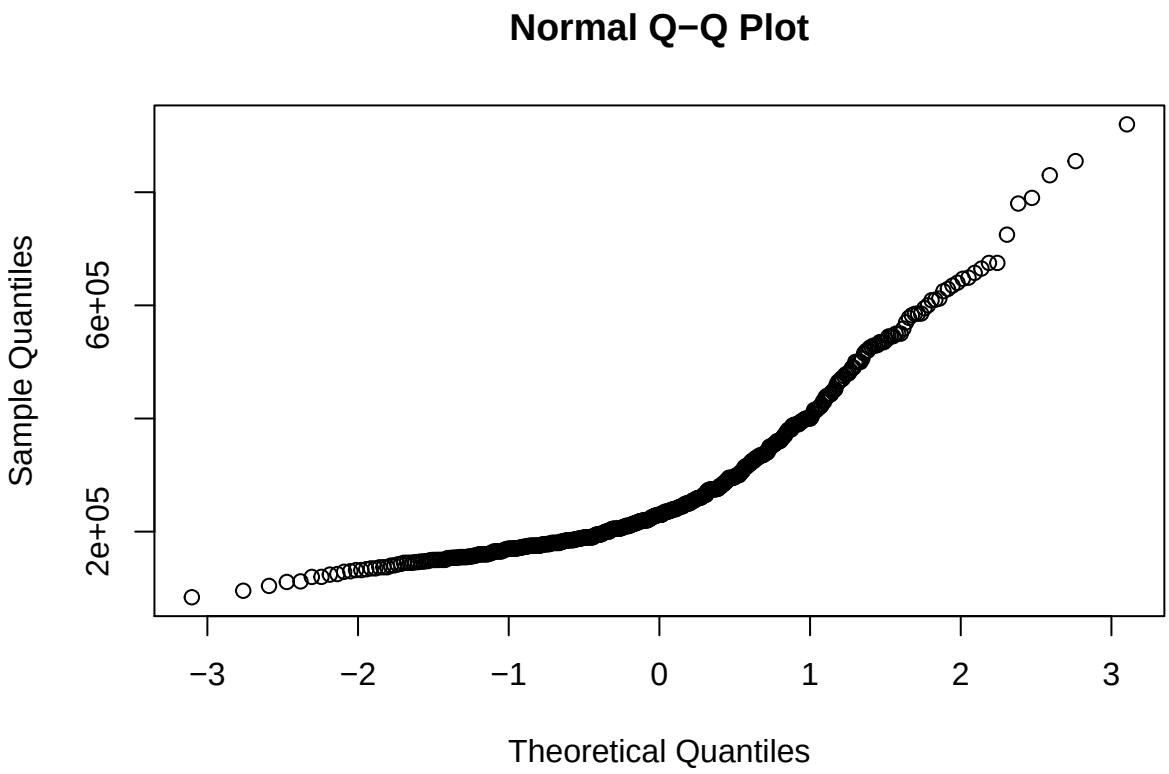
```
#summary statistics:
summary(Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   84000  180000  229900  277894  335000  920000
```

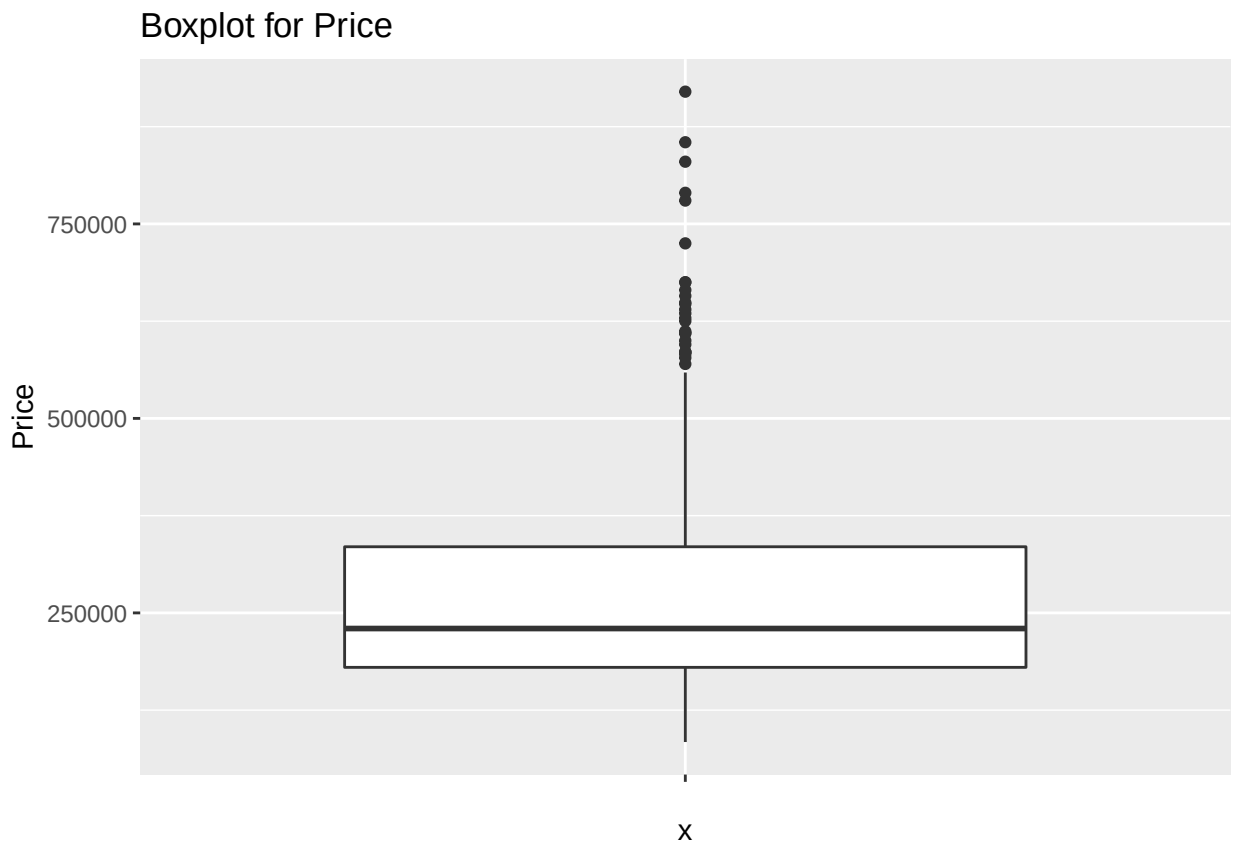
```
#histogram:
ggplot(data=house, aes(Price)) +
  geom_histogram(col="white",
                 bins=35) +
  labs(title="Histogram for Price")
```




```
#is it normal?
qqnorm(Price)
```




```
#boxplot:
ggplot(data=house, aes(x="", y=Price)) +
  geom_boxplot()+
  labs(title="Boxplot for Price")
```




```
#Check for impossible values:
sum(is.na(Price))
```

```
## [1] 0
```

```
sum(Price[Price < 0])
```

```
## [1] 0
```

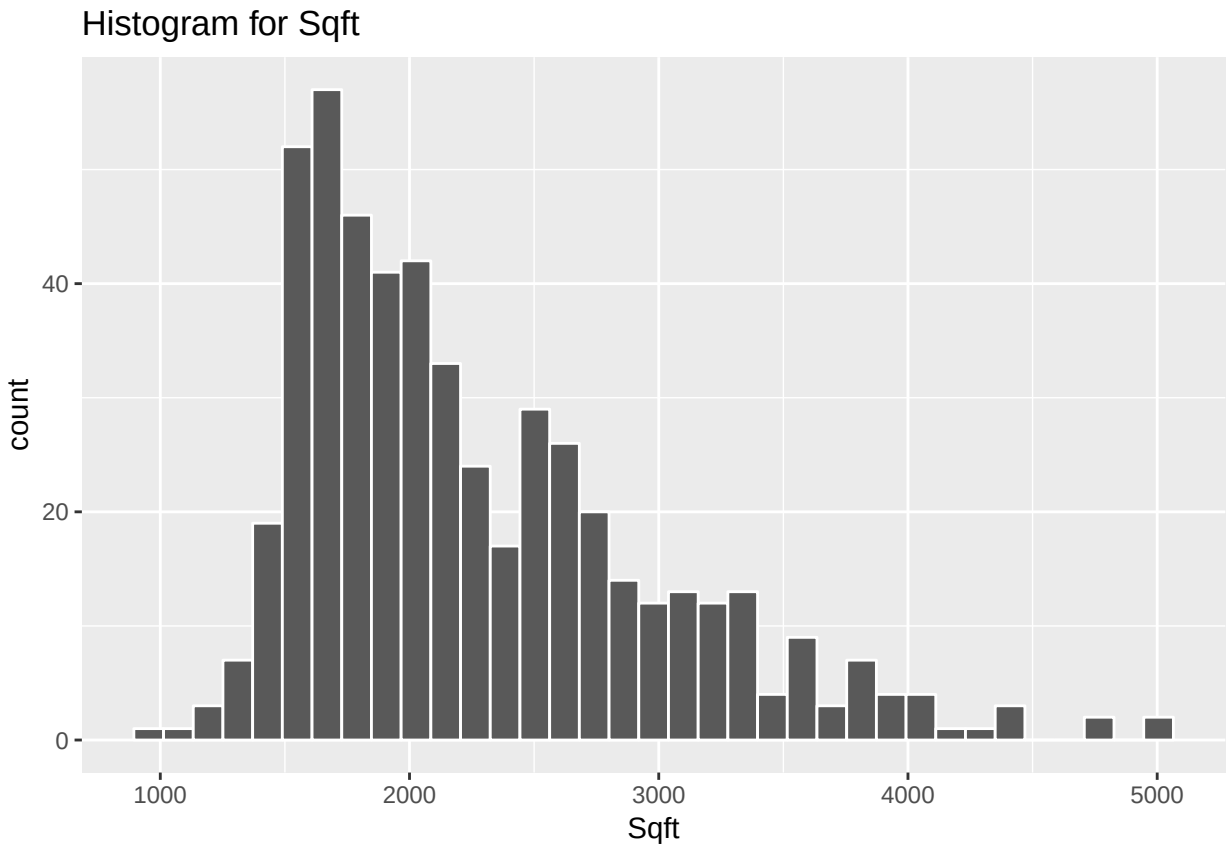
Looking at the histogram, the price follows approximately a normal distribution. However, it is right skewed, meaning that it has a long right tail. However, the qq-plot shows that the normal assumptions is very weak. The median price of a house is 229900, $while the mean price is 277894$. This is due to the fact that there are a few outliers, reaching a maximum price of 920000$. The boxplot of the price-variable verifies our assumption with the outliers. The data is indeed right skewed with a lot of ouliers. Furthermore, there are no missing values and no negative values, which shows that the data seem to be complete and not faulty.

"Sqft":

```
#summary:
summary(Sqft)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     980    1701    2061    2261    2636    5032
```

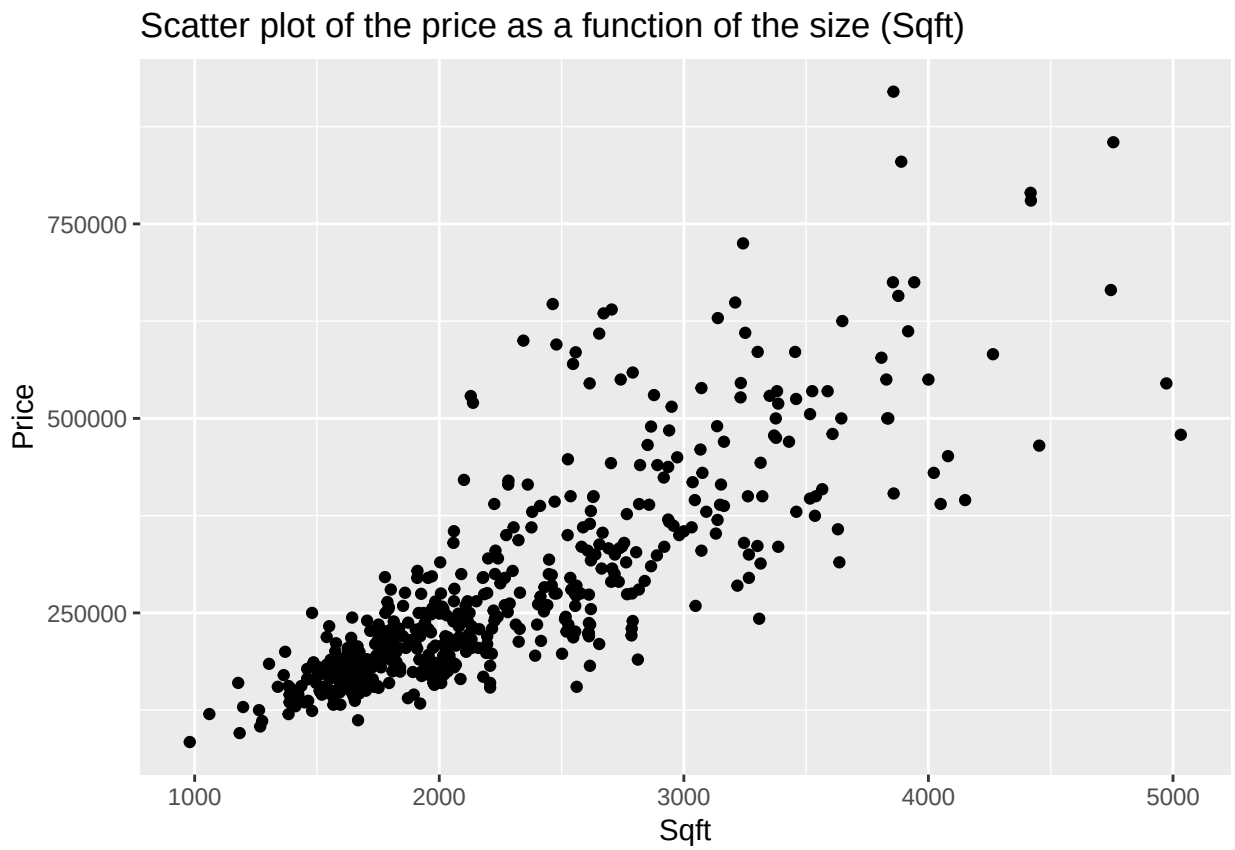
```
#histogram:
ggplot(data=house, aes(Sqft)) +
  geom_histogram(col="white",
                 bins=35) +
  labs(title="Histogram for Sqft")
```




```
#null values?
sum(is.na(Sqft))
```

```
## [1] 0
```

```
#scatter plot:
ggplot(data = house, aes(x=Sqft,y=Price))+
  geom_point()+
  labs(title="Scatter plot of the price as a function of the size (Sqft)")
```



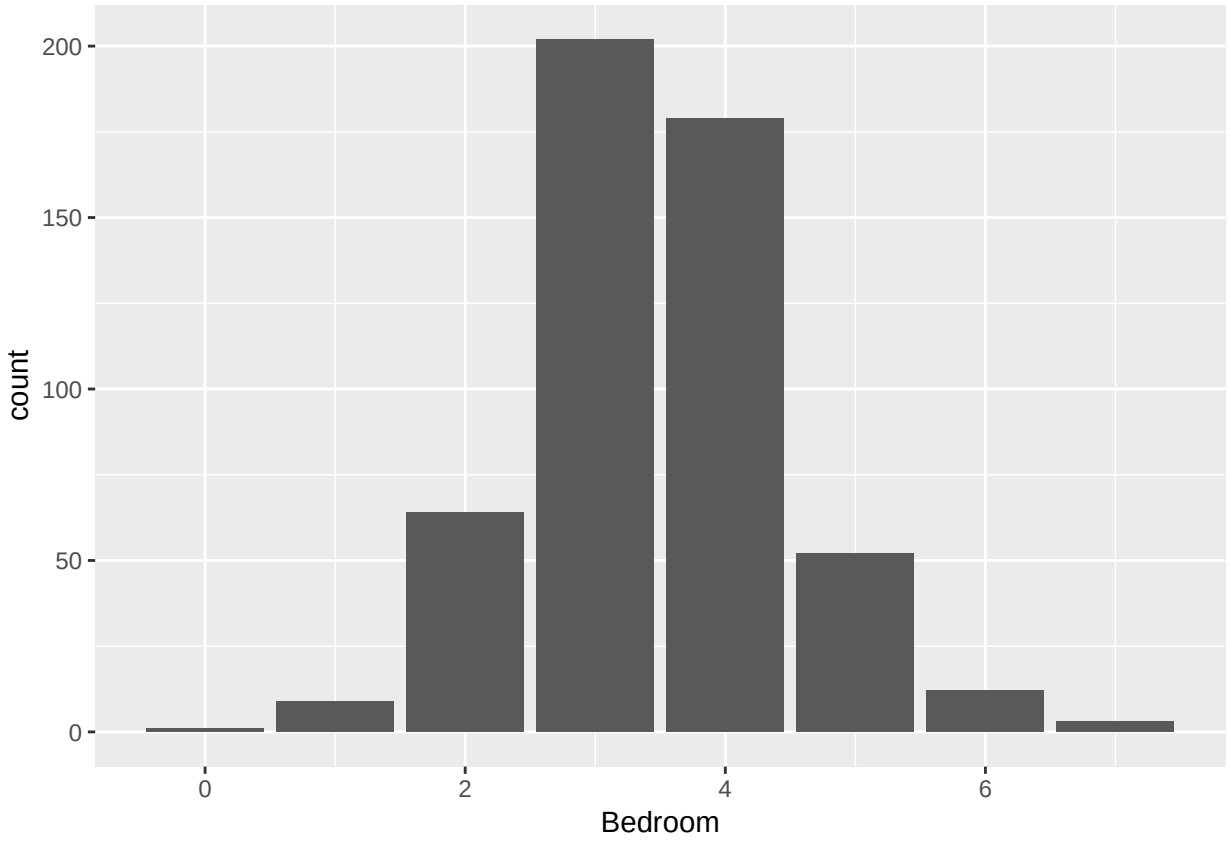

```
#correlation:
cor(Price,Sqft)
```

```
## [1] 0.8194701
```

"Sqft" is similarly distributed to "Price". The mean size of a house is 2261, the median is 2061. The distribution is right-skewed as well. There are no null values and all values are larger than 0. Looking at the scatter plot between the "Price" variable and the "Sqft" variable, there seems to be a linear relationship between the 2 variables. The high positive correlation of 0.819 seems to reinforce our initial assumption of a linear relationship between the two variables.

"Bedroom":

```
#barplot:
ggplot(data=house, aes(Bedroom)) +
  geom_bar()
```




```
#table of the different values:
table(Bedroom)
```

```
## Bedroom
##   0   1   2   3   4   5   6   7
##   1   9  64 202 179  52  12   3
```

```
#null values?
sum(is.na(Bedroom))
```

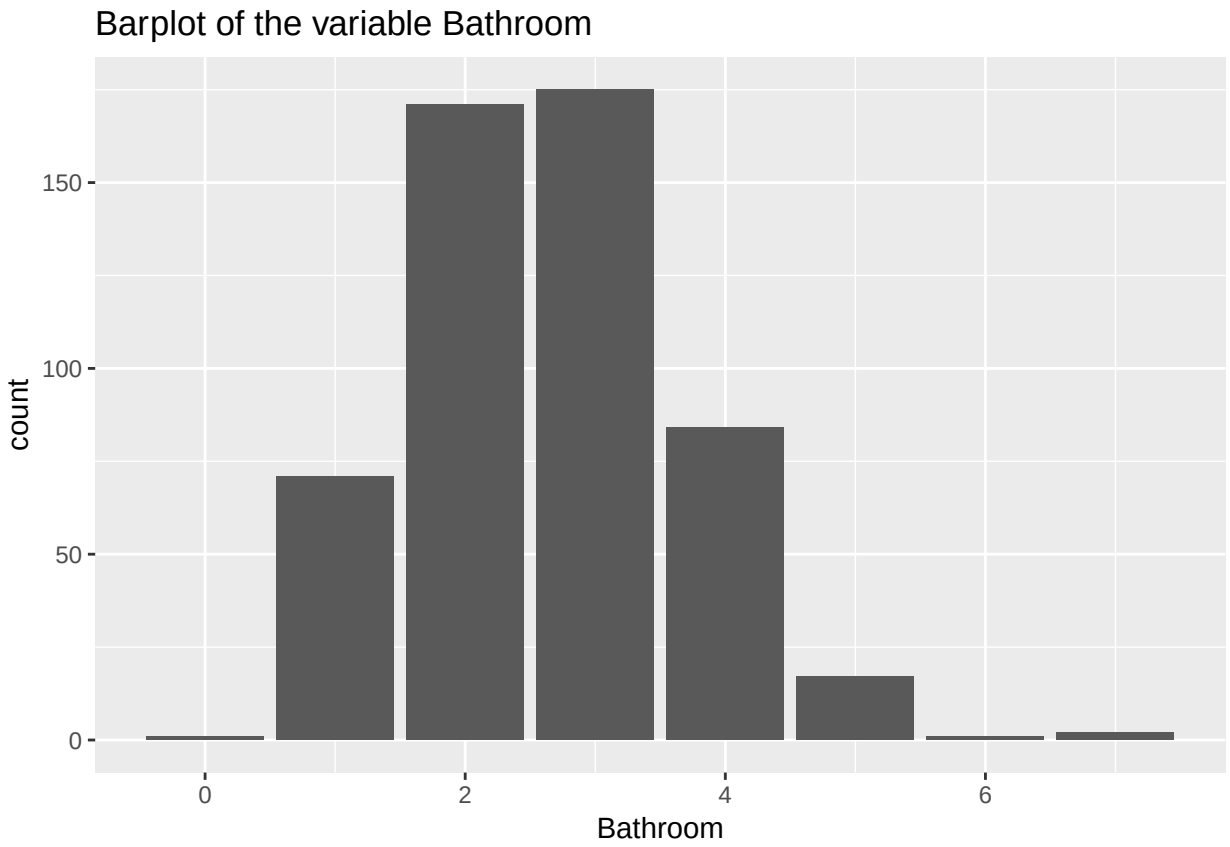
```
## [1] 0
```

```
#mean for the different values of the categorical variable:
mean_cat <- house %>%
              group_by(Bedroom) %>%
              summarise(m=mean(Price))
#plot the mean values for the diffrent bedrooms:
ggplot(data=mean_cat, aes(x=Bedroom,y=m)) +
  geom_point()+
  labs(title="Plot of the mean values of the price against the number of bedrooms")
```

Most of the houses have 3 bedrooms (202 bedrooms), followed by 4 bedrooms (179), then 2 (64). There are no missing values and the data lies within $\{0,\ldots,6\}$. There seems to be an interesting relationship between the number of bedrooms and the price of the house. We plotted the mean values for the price having a certain amount of bedrooms. We can observe a linear trend. More bedrooms seem to increase the price of the property.

"Bathroom":

```
ggplot(data=house, aes(Bathroom)) +
  geom_bar()+
  labs(title="Barplot of the variable Bathroom")
```




```
table(Bathroom)
```

```
## Bathroom
##   0   1   2   3   4   5   6   7
##   1  71 171 175  84  17   1   2
```

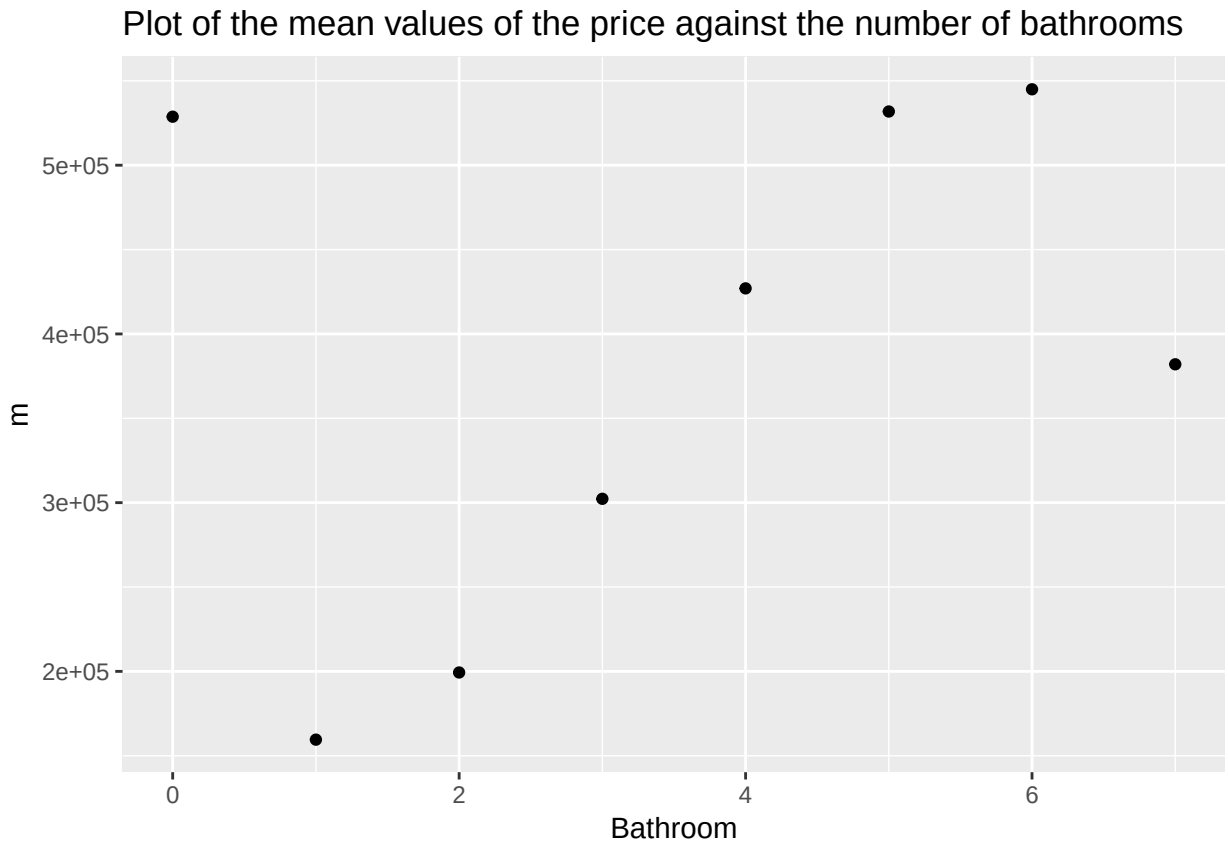
```
#null values?
sum(is.na(Bathroom))
```

```
## [1] 0
```

```
#mean for the different values of the categorical variable:
mean_cat <- house %>%
              group_by(Bathroom) %>%
              summarise(m=mean(Price))
mean_cat
```

```
## # A tibble: 8 x 2
##   Bathroom       m
##      <int>   <dbl>
## 1        0 528750
## 2        1 159533.
## 3        2 199341.
## 4        3 302281.
## 5        4 427004.
## 6        5 531844.
## 7        6 545000
## 8        7 382000
```

```
#plot the mean values for the diffrent bedrooms:
ggplot(data=mean_cat, aes(x=Bathroom,y=m)) +
  geom_point()+
  labs(title="Plot of the mean values of the price against the number of bathrooms")
```



Most houses have 3 bathrooms, followed by houses with 2 and 4 bathrooms. There are no missing values and the data lies within {0,1,...,7}. As a result, there seems to be no invalid data points. Similar to the bedrooms, the number of bathrooms seems to have an influence n the price. More bathrooms means that the property will be sold a higher price (general tendency).

"Airconditioning":

```
#table + calculate percentage:
ac_table<-table(Airconditioning)
ac_table
```

```
## Airconditioning
##   0   1
##  88 434
```

```
100*ac_table[2]/(ac_table[1]+ac_table[2])
```

```
##        1
## 83.14176
```

```
#null values?
sum(is.na(Airconditioning))
```

```
## [1] 0
```

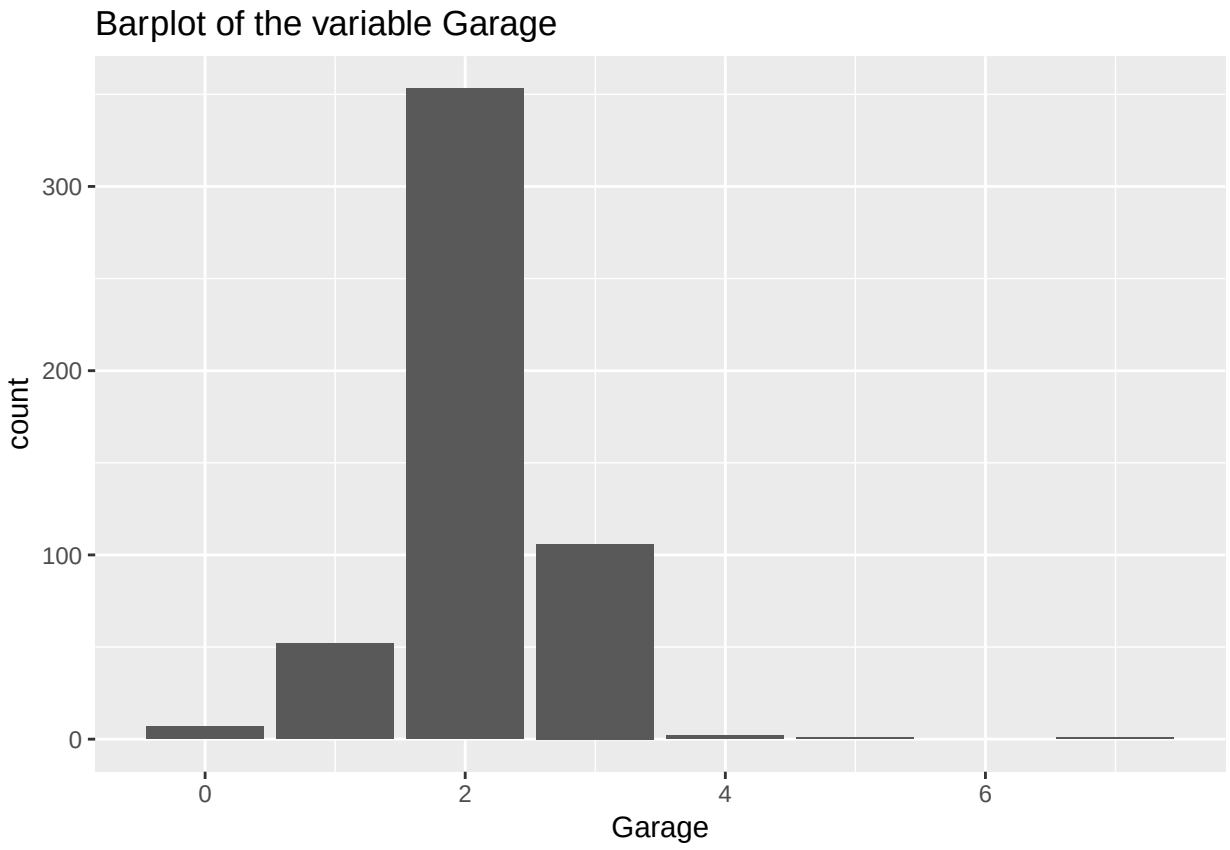
```
#mean for the two values:
house %>%
  group_by(Airconditioning) %>%
  summarise(m=mean(Price))
```

```
## # A tibble: 2 x 2
##   Airconditioning       m
##             <int>   <dbl>
## 1               0 189583.
## 2               1 295801.
```

83.12% (i.e. 434) of the houses have airconditioning. There are no missing values. The mean price for houses with airconditioning is higher (i.e. 295801) than for houses without airconditioning (i.e. 189583).

"Garage":

```
#barplot
ggplot(data=house, aes(Garage)) +
  geom_bar()+
  labs(title="Barplot of the variable Garage")
```




```
table(Garage)
```

```
## Garage
##   0   1   2   3   4   5   7
##   7  52 353 106   2   1   1
```

```
#null values?
sum(is.na(Garage))
```

```
## [1] 0
```

```
#mean for the different values of the categorical variable:
mean_cat <- house %>%
              group_by(Garage) %>%
              summarise(m=mean(Price))
#plot the mean values for the diffrent bedrooms:
ggplot(data=mean_cat, aes(x=Garage,y=m)) +
  geom_point()+
  labs(title="Plot of the mean values of the price against the number of garages")
```

The range of parking spots goes from 0 to 7, with a majority of the houses having 2 parking spots, followed by 3 and 2. There are no missing values and the data lies within $\{0,\ldots,7\}$. There is a slight positive relatinship between the number of garages and the price of a house. However, it is very weak.

"Pool":

```
#generate a frequency table and calculate percentage:
pool_table <-table(Pool)
pool_table
```

```
## Pool
##   0   1
## 486  36
```

```
100*pool_table[2]/(pool_table[1]+pool_table[2])
```

```
##        1
```

```
## 6.896552
```

```
#null values?
sum(is.na(Pool))
```

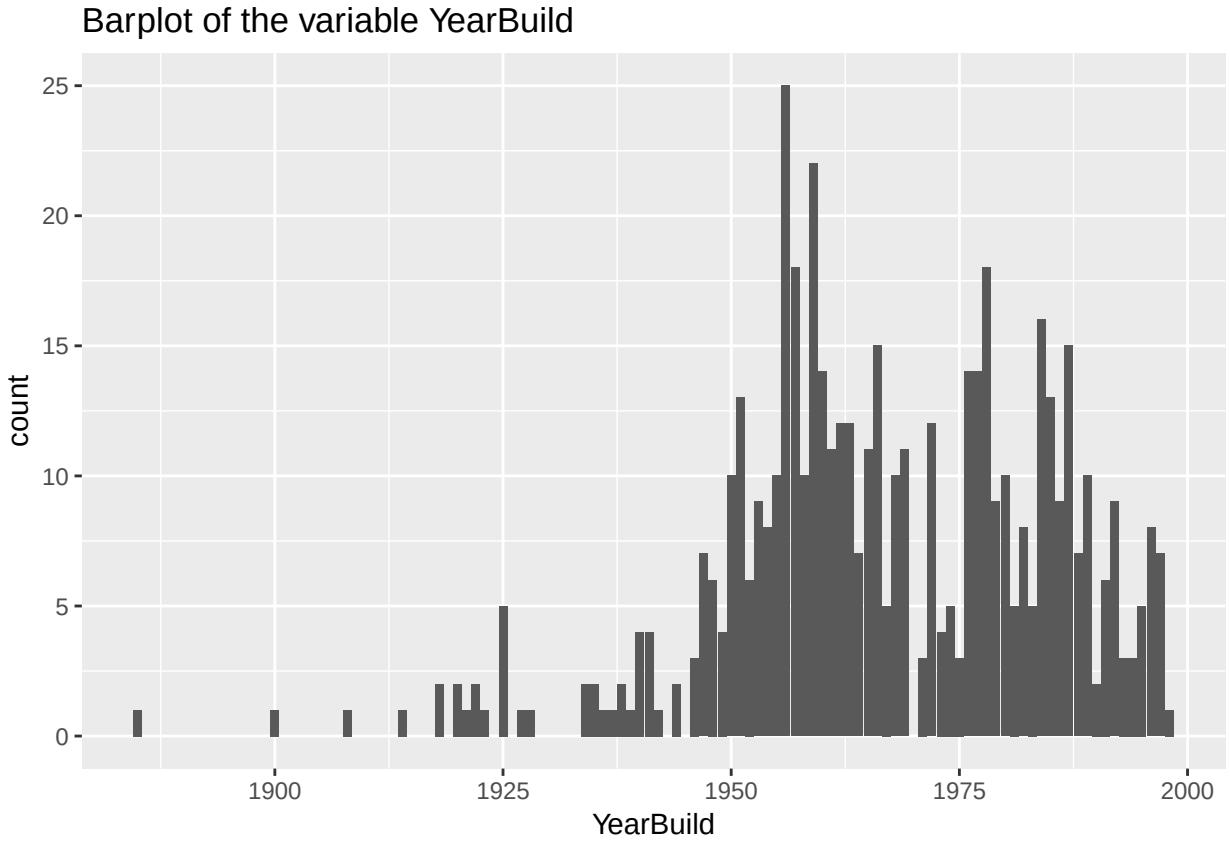
```
## [1] 0
```

```
#Is a house with a pool worth more?
house %>%
    group_by(Pool) %>%
    summarise(m=mean(Price))
```

```
## # A tibble: 2 x 2
##    Pool       m
##   <int>   <dbl>
## 1     0 272396.
## 2     1 352120.
```

Almost 7% of the properties have a pool. A house with a pool is worth on average 352120, *wileahousewithoutapoolisworthonave*

```
#barplot
ggplot(data=house, aes(YearBuild)) +
  geom_bar()+
  labs(title="Barplot of the variable YearBuild")
```



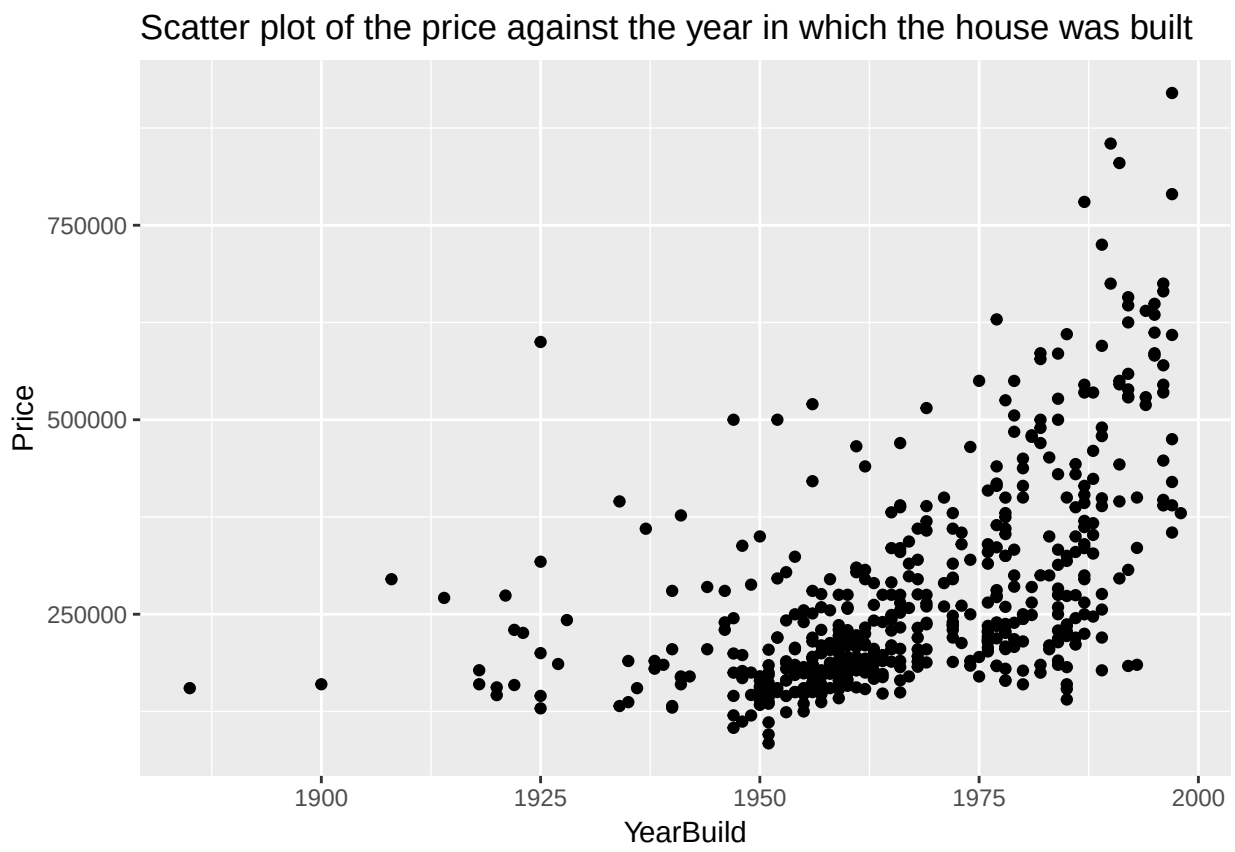

```
summary(YearBuild)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1885    1956    1966    1967    1981    1998
```

```
#null values?
sum(is.na(YearBuild))
```

```
## [1] 0
```

```
#scatterplot between the year and the price:
ggplot(data=house, aes(x=YearBuild,y=Price))+
  geom_point()+
  labs(title="Scatter plot of the price against the year in which the house was built")
```



The build-year for the different houses is spread from 1885 to 1998. The mean year is 1967. The number of houses built increases and reaches a spike around 1960.It then decreases again reaching a low in 1975, before rising sharply again and then slowly decreasing again. There seems to be a relationship between the year in which the house was built and its price. A new house has a higher price tag than an old one. However, this relationship is not very strong.

"Quality":

```
#table for the quality:
table(Quality)
```

```
## Quality
##   1   2   3
##  68 290 164
```

```
#mean values for different quality houses:
house %>%
  group_by(Quality) %>%
  summarise(mean(Price))
```

```
## # A tibble: 3 x 2
##   Quality `mean(Price)`
##     <int>         <dbl>
## 1       1       543611.
## 2       2       273766.
## 3       3       175018.
```

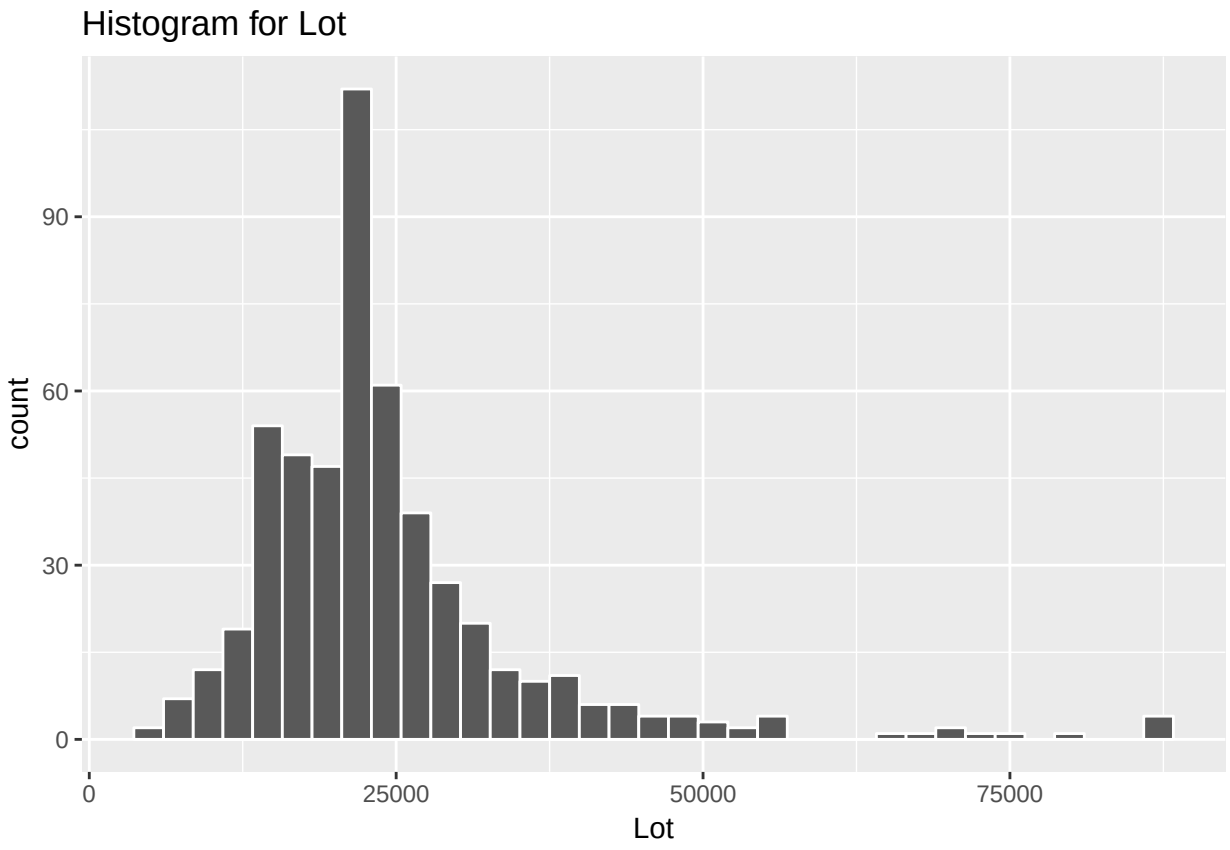
```
#null values?
sum(is.na(Quality))
```

```
## [1] 0
```

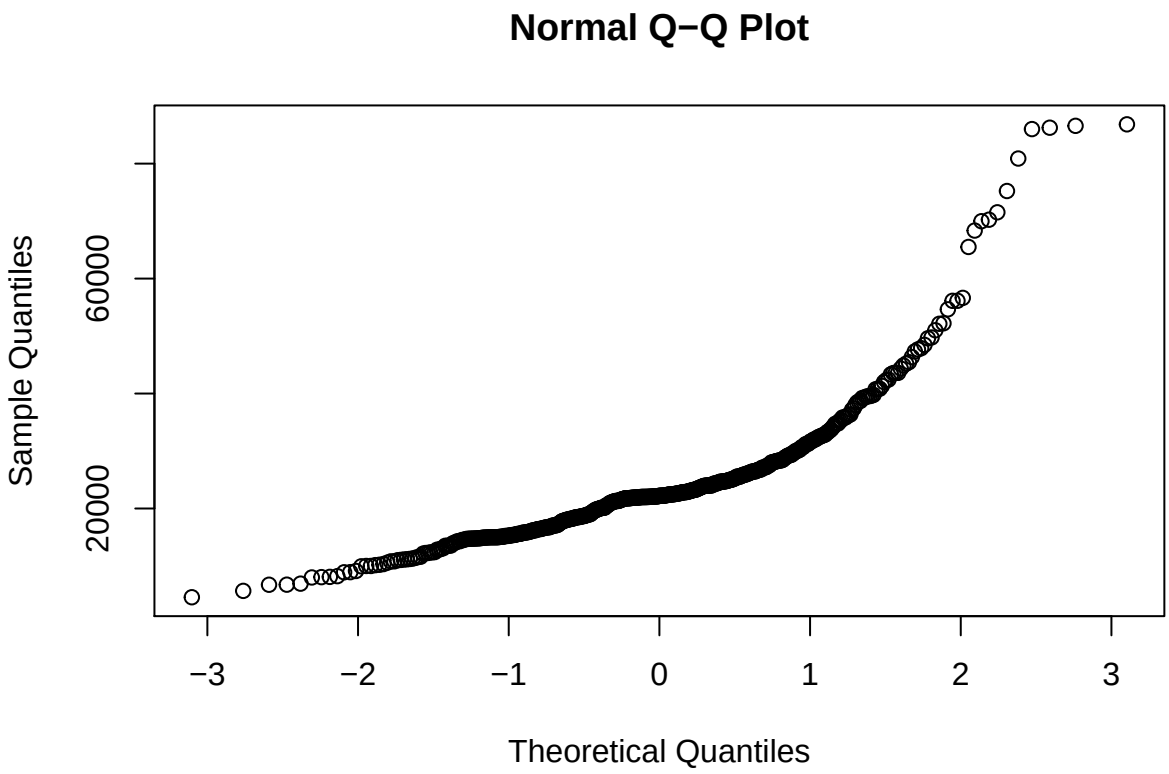
In our data set, 290 houses are built n quality 2, 164 in quality 3 and 68 in quality 1. Following those numbers, it seems that the high quality houses are denoted with 1, medium quality with 2 and low quality/cheap with 3. This seems to be confirmed by the mean of the prices for the different build-qualities. The mean of the prices for build-qulity 1 is 543611$, for build-quality 2 it is 273766, and for built-quality 3, it is 175018. We conclude that the quality has an influence on the housing prices.

"Lot":

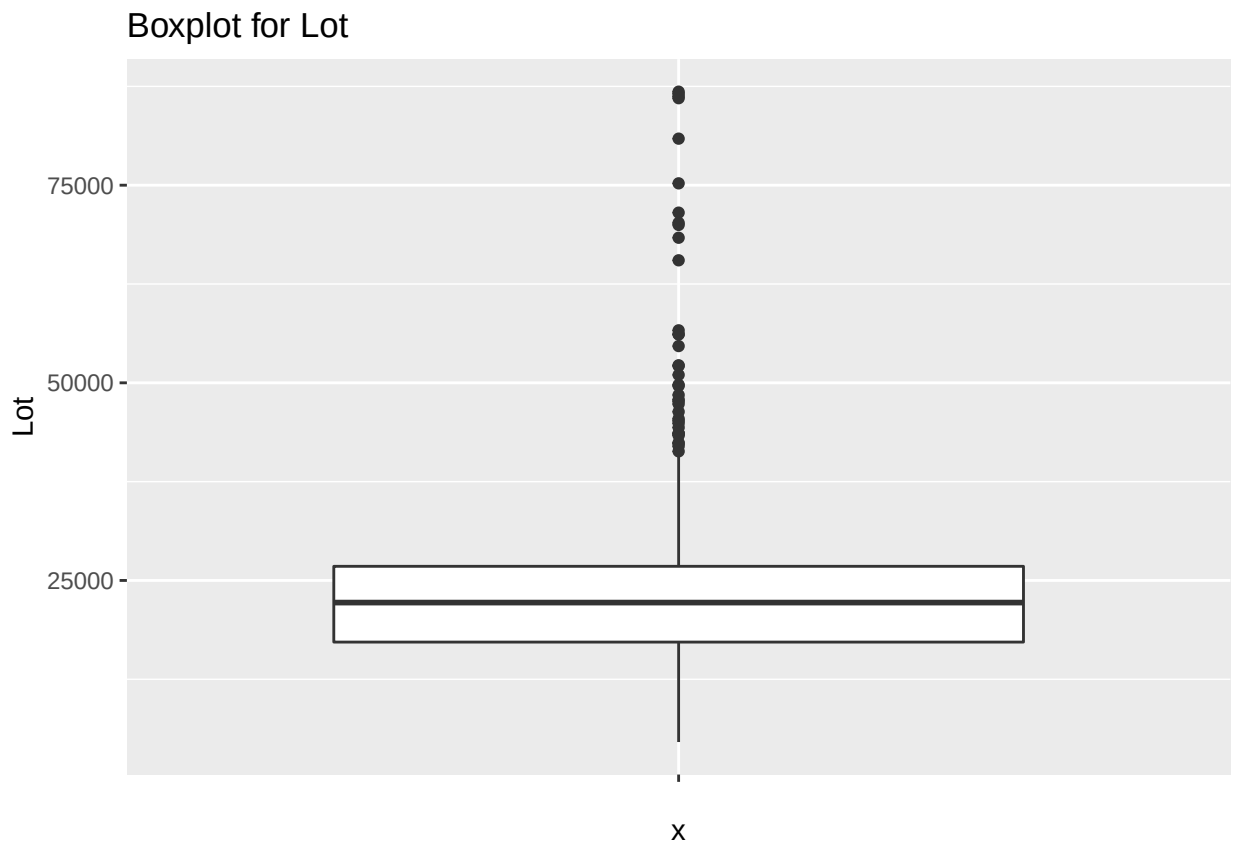
```
#histogram:
ggplot(data=house, aes(Lot)) +
  geom_histogram(col="white",
                 bins=35) +
  labs(title="Histogram for Lot")
```




```
#is it normal?
qqnorm(Lot)
```




```
#boxplot:
ggplot(data=house, aes(x="", y=Lot)) +
  geom_boxplot()+
  labs(title="Boxplot for Lot")
```



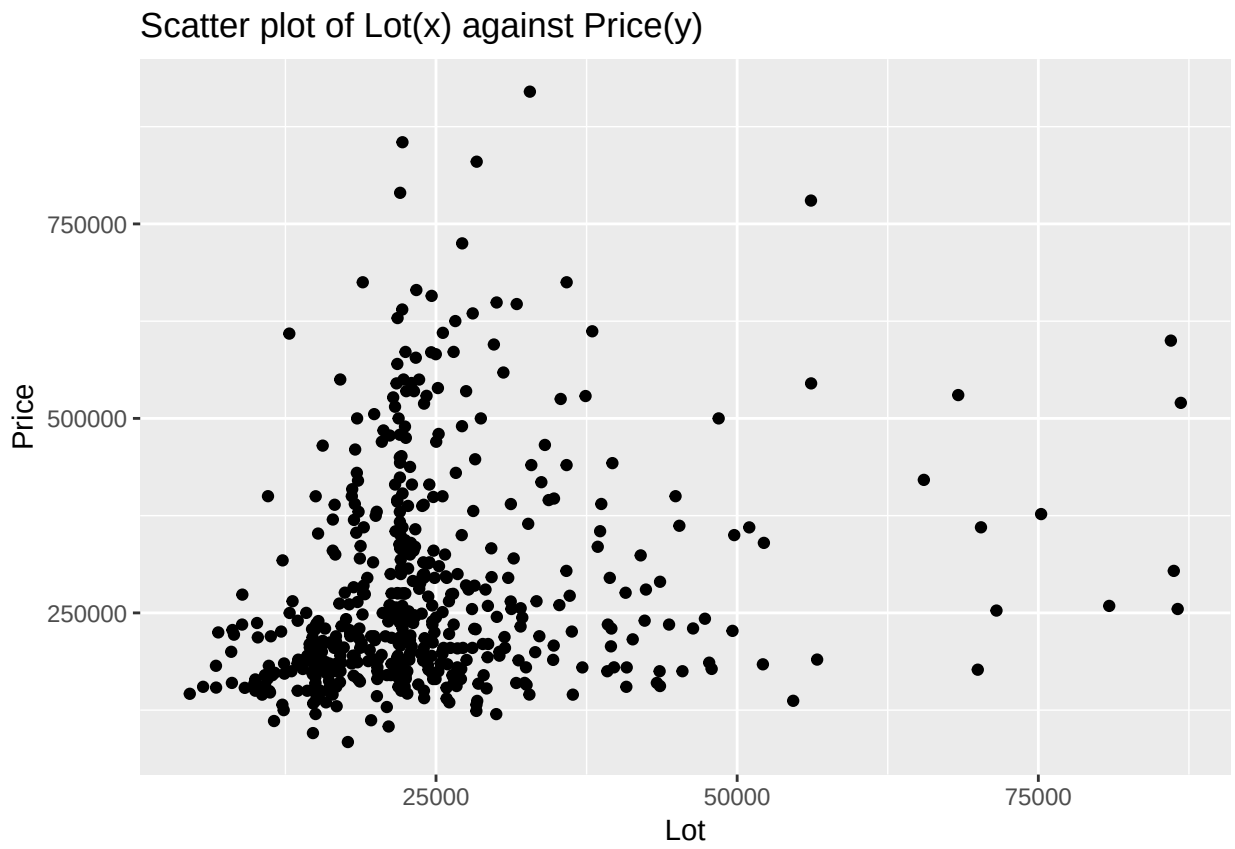

```
#summary statistics:
summary(Lot)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4560   17205   22200   24370   26787   86830
```

```
#null values?
sum(is.na(Lot))
```

```
## [1] 0
```

```
#scatter plot with Price
ggplot(data=house,aes(x=Lot,y=Price))+
  geom_point() +
  labs(title="Scatter plot of Lot(x) against Price(y)")
```




```
#are they correlated?
cor(Price,Lot)
```

```
## [1] 0.2241685
```

The median size of the land is 22200, while the mean size is 24370. This is due to the fact that the values in "Lot" are right-skewd. The values last from 4560 to 86830, and there are no null values. The scatter plot of "Lot" and "Price" shows a slight positive correlation (which is confirmed by the Pearson's correlation which is 0.224). There is a slight tendency that suggests that houses on larger land surfaces are sold for more money than properties on small lands. However, this correlation is not very strong.

AdjHighway:

```
#table for the quality:
table(AdjHighway)
```

```
## AdjHighway
##   0   1
## 511  11
```

```
#mean for the different values of the categorical variable:
house %>%
  group_by(AdjHighway) %>%
  summarise(mean(Price))
```

```
## # A tibble: 2 x 2
##   AdjHighway `mean(Price)`
##        <int>         <dbl>
## 1          0       278925.
```

```
## 2           1       230027.
```

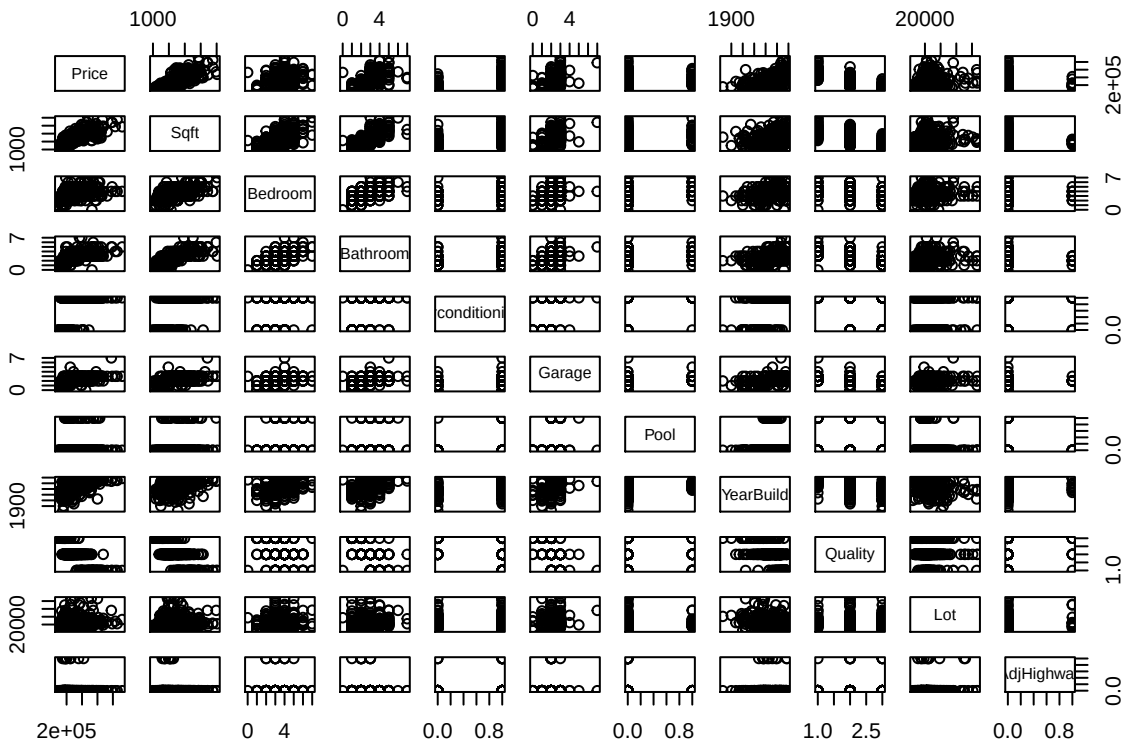
```
#null values?
sum(is.na(AdjHighway))
```

```
## [1] 0
```

A large maority of the properties in our data set is not next to a highway. Only 11 out of 522 houses are next to one. Looking at the mean prices of houses next to a highway (2300027)$compared to prices of houses that are not next to a highway$(2 we can assume that there is a relationship between those to types of properties. However, the number of houses next to the highway is small, so it is difficult to say if the price difference is caused by the fact of being close to the highway.

First model:

```
plot(house)
```




```
cor(house)
```

```
##                      Price        Sqft     Bedroom    Bathroom
## Price           1.00000000  0.81947006  0.41332387  0.68368539
## Sqft            0.81947006  1.00000000  0.55783780  0.75527294
## Bedroom         0.41332387  0.55783780  1.00000000  0.58344693
## Bathroom        0.68368539  0.75527294  0.58344693  1.00000000
## Airconditioning 0.28859624  0.26794957  0.23465143  0.32476037
## Garage          0.57778626  0.53376649  0.31681372  0.48989809
```

```
## Pool             0.14661162  0.16239624  0.13454242  0.18415278
## YearBuild        0.55551645  0.44119674  0.26869237  0.51284096
## Quality         -0.75807834 -0.69555291 -0.37832178 -0.68221493
## Lot              0.22416852  0.15752472  0.12653841  0.14700662
## AdjHighway      -0.05096821 -0.06062519 -0.02874435 -0.05092832
##                 Airconditioning       Garage        Pool   YearBuild
## Price                0.28859624  0.577786264  0.14661162  0.55551645
## Sqft                 0.26794957  0.533766490  0.16239624  0.44119674
## Bedroom              0.23465143  0.316813720  0.13454242  0.26869237
## Bathroom             0.32476037  0.489898090  0.18415278  0.51284096
## Airconditioning      1.00000000  0.319281063  0.10236094  0.42558816
## Garage               0.31928106  1.000000000  0.10893151  0.46176035
## Pool                 0.10236094  0.108931505  1.00000000  0.05982913
## YearBuild            0.42558816  0.461760355  0.05982913  1.00000000
## Quality             -0.41376829 -0.547096684 -0.12530318 -0.61752597
## Lot                 -0.10530489  0.152219307 -0.03685063 -0.10045191
## AdjHighway          -0.04081391 -0.001955469 -0.03993181  0.02577660
##                     Quality         Lot   AdjHighway
## Price            -0.75807834  0.22416852 -0.050968213
## Sqft             -0.69555291  0.15752472 -0.060625186
## Bedroom          -0.37832178  0.12653841 -0.028744349
## Bathroom         -0.68221493  0.14700662 -0.050928324
## Airconditioning -0.41376829 -0.10530489 -0.040813909
## Garage           -0.54709668  0.15221931 -0.001955469
## Pool             -0.12530318 -0.03685063 -0.039931811
## YearBuild        -0.61752597 -0.10045191  0.025776602
## Quality           1.00000000 -0.11605691  0.020336285
## Lot              -0.11605691  1.00000000  0.078446995
## AdjHighway        0.02033629  0.07844699  1.000000000
```

```
#a lot of linear reationships
```