MS5318 Homework 3

HO Yin Shan (57487297)

```
# library packages & read data
library(tidyverse)
home_prices <- read.csv("home_prices.csv")
```

Question 1

What are the mean and median home price in the data set?

```
summary(home_prices$Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    77.0   240.0   318.0   322.8   392.0   826.0
```

Mean home price : 322.8 thousands
Median home price: 318 thousands

Question 2

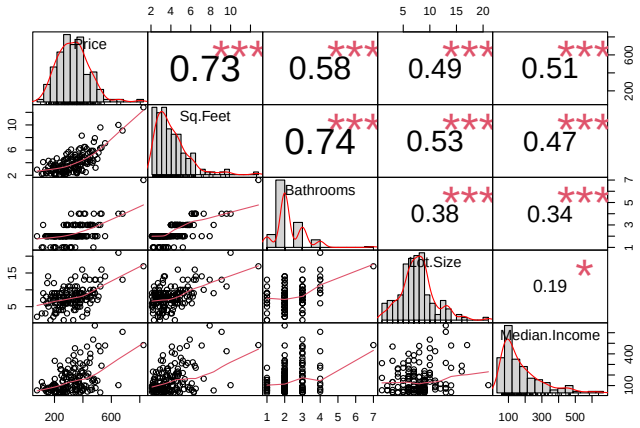
Make a histogram of the response variable Price. When you use the R function hist(), include the following arguments: breaks=15, xlab="Home Price", ylab="Counts", col="grey". Use comma to separate different arguments in the R function. Try to understand the meanings of those arguments.

```
hist(home_prices$Price,
     breaks = 15, xlab = "Home Price", col = "grey",
     ylab = "Count", main = "Histogram of Home Prices")
```

Question 3

Examine the scatterplots of the pairs of variables in the data set. Attach the scatterplots to this assignment.

```
PerformanceAnalytics::chart.Correlation(home_prices)
```



Square Feet has **strong positive correlation** with home prices. Number of bathrooms, Lot size and Median Income shows a **moderate positive correlation*** with price. Number of bathrooms shows a **strong positive correlation** to Square Feet. Lot size and median income shows **moderate positive correlation** to Square Feet. Bathrooms, Lot Size and Median shows **weak positive correlation** to each others.

Question 4

Fit a multiple regression model, using all four explanatory variables. Include the model summary in your submission (e.g. estimated coefficients, p-value, F-test results, etc.).

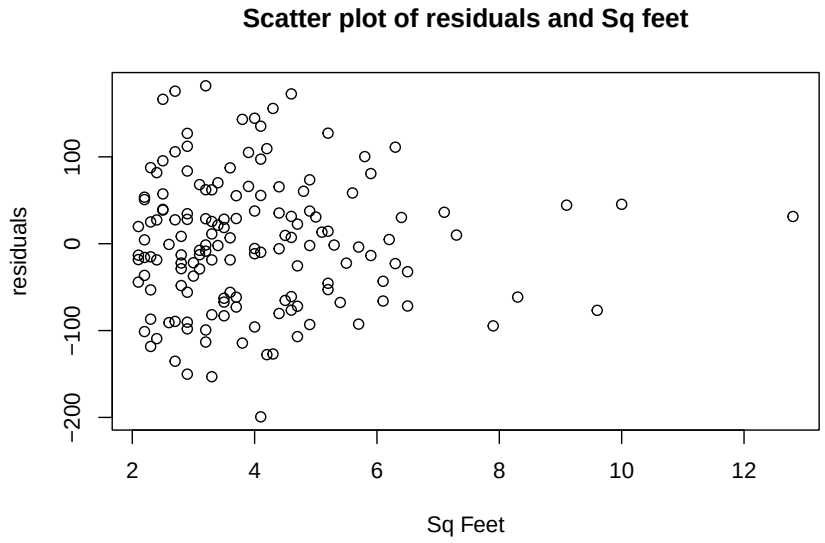
```
model <- lm(Price ~ . ,data = home_prices)
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ ., data = home_prices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -199.324  -59.674   -1.819   45.118  181.802
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   81.81616   20.49680   3.992 0.000104 ***
## Sq.Feet       31.27315    6.35519   4.921 2.31e-06 ***
## Bathrooms     15.55418   11.09373   1.402 0.163032
## Lot.Size       5.99914    2.29707   2.612 0.009959 **
## Median.Income  0.21052    0.05525   3.810 0.000204 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.5 on 145 degrees of freedom
## Multiple R-squared:  0.5899, Adjusted R-squared:  0.5785
## F-statistic: 52.13 on 4 and 145 DF,  p-value: < 2.2e-16
```

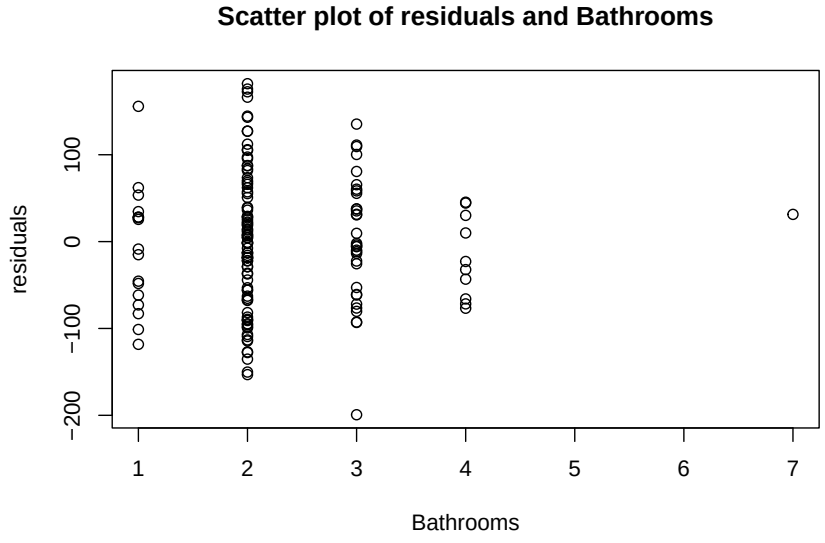
Question 5

Does the estimated model appear to meet the conditions of multiple regression model? (Check model conditions: residual plots, normal quantile plot.)

```
plot(home_prices$Sq.Feet, model$residuals,
     ylab = "residuals", xlab = "Sq Feet",
     main = "Scatter plot of residuals and Sq feet")
```

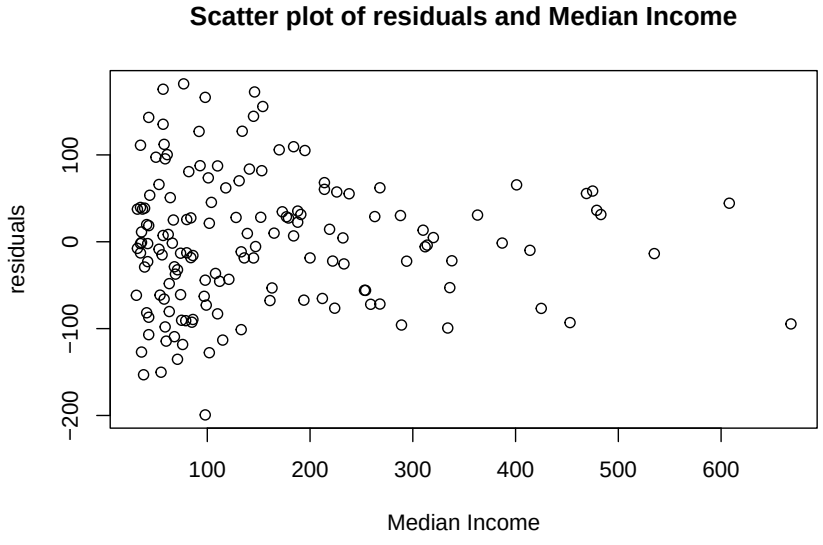



```
plot(home_prices$Bathrooms, model$residuals,
     ylab = "residuals", xlab = "Bathrooms",
     main = "Scatter plot of residuals and Bathrooms")
```

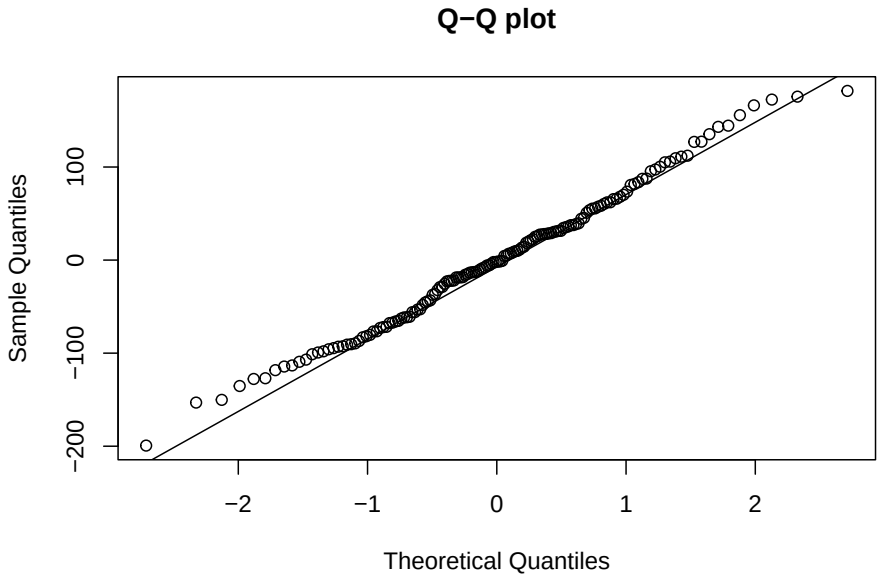


```
plot(home_prices$Lot.Size , model$residuals,
     ylab = "residuals", xlab = "Lot Size",
     main = "Scatter plot of residuals and Lot Size")
```


```
plot(home_prices$Median.Income, model$residuals,
     ylab = "residuals", xlab = "Median Income",
     main = "Scatter plot of residuals and Median Income")
```



```
qqnorm(model$residuals, main = "Q-Q plot")
qqline(model$residuals)
```



Based on the plots, it appears that to basically meet the conditions of multiple regression model. From the **residual plots**, it is found that it is basically even distributed.
In addition, from the **qq-plot** it is fount that the data **most of the data fitted well except some from the tails**.

Question 6

Does this model explain statistically significant variation in the prices of homes? Give your reasons.

Based on the model summary, the **p-value** of the whole model is $<$ 2.2e-16. Which shows that the model shows statistically significant variation in the prices of homes.
In addition, from the **Multiple R-squared**, it shows that the **dependent variables can explain 58.99% of the price**.

Question 7

Interpret the estimated coefficient for Sq.Feet. What does this coefficient mean? What does its p-value mean?

The coefficient of Sq.Feet is 31.27315. It means that **when there is increase in 1000 square feet, the home price will increase for 31.2732 thousands when all other variables are unchanged**. It's **p-value is 2.31e-06** that the **variable is statistically significant**.

Question 8

Compare the marginal coefficient for the number of bathrooms to the partial coefficient. Explain why these are so different.

```
lm(Price~ Bathrooms, data = home_prices) %>%
  summary()
```

```
##
## Call:
## lm(formula = Price ~ Bathrooms, data = home_prices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -262.623  -68.057   -7.307   64.614  256.010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  135.673     22.826   5.944 1.92e-08 ***
## Bathrooms     82.317      9.429   8.730 4.96e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.06 on 148 degrees of freedom
## Multiple R-squared:  0.3399, Adjusted R-squared:  0.3355
## F-statistic: 76.21 on 1 and 148 DF,  p-value: 4.958e-15
```

According to the summary above, the **marginal coefficient of Bathrooms to Price is 82.317**. However, the **partial coefficient is 15.55418**. The reason of such differences is that the marginal regression only consider bathroom as the only coefficient which may over estimate the effect of the bathrooms. However, the full model consider other variables' effects on the home prices.

Question 9

A homeowner wants to sell her home with: Sq.Feet = 3, Bathrooms=3, Lot.Size=9, Median.Income= 10. Give a 95% prediction interval for the price of her home.

```
new_data <-
  data_frame(Sq.Feet = 3, Bathrooms = 3,
             Lot.Size = 9, Median.Income = 10)
predict.lm(model, new_data,
           interval = "prediction",
           level = 0.95)
```

```
##        fit      lwr      upr
## 1 278.3957 123.8469 432.9444
```

The predicted price of her home is [123.8469, 432.9444]thousand.

Question 10

A homeowner asked the realtor if she should spend $40,000 to convert a walk-in closet into a small bathroom in order to increase the sale price of her home. What does your analysis indicate?

Based on the analysis, it is found that when the number of bathrooms increase by 1, keeping all other variables unchanged, the home price would increase for 15.55418 thousands. However, converting a the closet into bathroom required 40 thousands which doesn't worth as it is more expensive than the price increase.