STAT5002 Assignment

Name: Xing Xing
Student_ID: 500390560
Date: 1 Jun 2021

Question 1: Summary table

Preparation

For the preparation step, the *real_estate.csv* which contains the house data need to be imported to the R file by '*read.csv*' function.

```
#=============
# Import data
#=============
file_name = "Real_Estate.csv"
Path = "/Users/xingxing/Desktop/STAT5002/5002Assignment/"
data = read.csv(paste(Path,file_name, sep=''))
```

a) In this data frame is consists of totally 8 numerical columns which are *ID, Price, Bedrooms, Size, Pool, Distance, Suburbs* and *Garage* respectively. We may use R in-built function *mean()* and *sd()* to calculate the sample mean and standard deviation. A R in-built *t.test()* function is utilised to obtain T-test 95% confidential interval for all these 4 columns since the dataset we analysed could be considered as a sample of the population (These suburbs will have more than 170 houses, so we might not know the actual standard deviation). Related with these data, further details to describe *Price, Bedrooms, Size* and *Distance* is shown below:

Column name	Mean(μ)	Standard deviation(σ)	95% Confidence Interval
Price	427.18	91.01	[413.3977, 440.9552]
Bedrooms	2.98	1.08	[2.8181, 3.1466]
Size	164.18	32.80	[159.2157, 169.1490]
Distance	9.43	4.82	[8.7032, 10.1615]

b) According to the summary table above, the typical property in that certain area of Melbourne could have 3 bedrooms with a size of 164 square meters, 9.4 km from city centre and the price would be around $427,180 Australian dollar. However, any 3 bedroom houses within 413 to 440 Australian dollar, have 159 to 169 square meters, distance to the city is in 8.7 to 10 km, could be considered as a typical house.

Question 2: Hypothesis Testing

According to the question, the agency believe that the t test is used to verify if the mean of suburb 2 house price is higher than $420000, We might assume that the distribution for the house price in this area is in Normal distribution. Based on that, we have:

Define terms:

We use the upper bound of the claim which is:

H0: $\mu = 420000$ VS H1: $\mu > 420000$

H0: The mean of Suburb 2 house price is equal to $420000

H1: The mean of Suburb 2 house price is greater than $420000

Assumption:

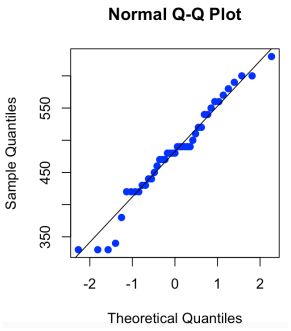
- This sample is randomly selected from the population.
- There are more than 30 instances in dataset (n > 30).
- The population standard deviation is unknown.
- The population house Price is Normally distributed (Price~N(μ, σ)).

Normality test:

Using R in-built *qqnorm()* and *shapiro.test()* function to test if the Price is normally distributed. Firstly, by using *qqnorm()* we could obtain a qqplot and if the dots is. likely linear, means the data is fairly normal distributed.After that, use *shapiro.test()* to test the normality of data further, *shapiro.test()* (shapiro-wilk test) is a W-test which mainly use to test if a dataset is normally distributed. In the test content, if the p-value is greater than 0.05 we considered that the dataset is normally distributed. Related R code and diagram is below:

qqnorm normality test:

```
> test_price = data$Price[data$Suburbs == 2]
> qqnorm(test_price, col = "blue", pch = 19)
> qqline(test_price)
```



By observation, we noticed that there is a straight line in the diagram which fairly fit all the data dots

Shapiro-wilk test:

```
> shapiro.test(test_price)
```

By observation, we noticed that the p-value is equal to 0.24 which is highly greater than 0.05.

```
        Shapiro-Wilk normality test

data:  test_price
W = 0.96697, p-value = 0.2476
```

According to the observation in two tests above, we could conclude that there are sufficient evidences to support that the dataset(test_price) is normally distributed.

T-test statistic under H0:

As mentioned in lecture the

$$test\_statistic = T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{169} \qquad T = \frac{480.9302 - 420.00}{\frac{74.4444}{\sqrt{43}}} = 5.3670$$

```
> t.test(test_price, mu = 420.000, alternative = "greater")

        One Sample t-test

data:  test_price
t = 5.367, df = 42, p-value = 1.605e-06
alternative hypothesis: true mean is greater than 420
95 percent confidence interval:
 461.8356      Inf
sample estimates:
mean of x
 480.9302
```

Large value of T will argue against H0 for H1, observation of T value ≈ 5.3670, then calculate p-value using *1 - pt(61.2, 43-1)* in R code we have p-value = 1.605207e-06 ≈ 0.

Conclusion:
As the p-value is nearly zero which is very small, the null hypothesis (H0) is rejected which means there are sufficient evidence to support that the house price in suburb 2 could be higher than $420000.

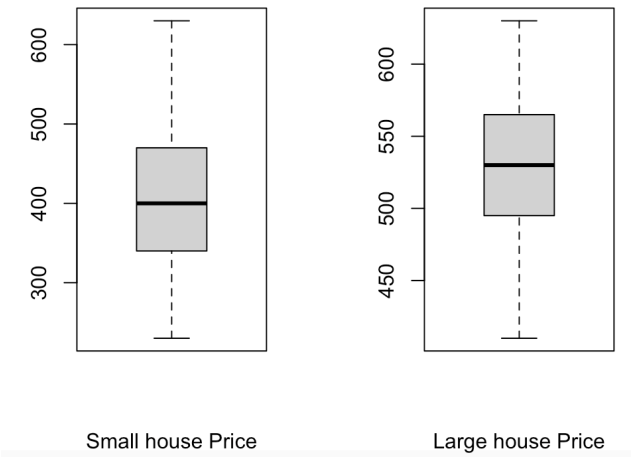
Question 3: Size and Price

a) To divide size into 2 category, a new column is created using conditional statement by R code below. In the new column, there are 27 large houses which size is greater than or equal to 200 m^2. On the other hand, 143 small houses which size is less than 200 m^2.

```
> data2=within(data,{Size_Group=ifelse(data$Size < 200,"Small","Large")})
> table(data2$Size_Group)

Large Small
   27   143
```

b) I choose to use box plot to graphically demonstrate of the Price related to different size session since it is an easier way to protrude the data mean and range. Related R code and output are shown below:

```
> small_price = data$Price[data2$new_Size=='Small']
> large_price = data$Price[data2$new_Size=='Large']
>
>
> par(mfrow = c(1,2))
> boxplot(small_price, xlab = 'Small house Price')
> boxplot(large_price, xlab = 'Large house Price')
```



From the diagram, it is obvious to see that the average Price of large house could be higher than the price of a small house. Also the price range for small houses is wider than big house price range.

c) The 95% confidential interval of small houses and large houses could be calculated by R in-built T test since the standard deviation is still considered as unknown. Related R code and output is below:

```
> t.test(small_price, conf.level=0.95)

        One Sample t-test

data:  small_price
t = 58.361, df = 142, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 394.6262 422.2969
sample estimates:
mean of x
 408.4615
```

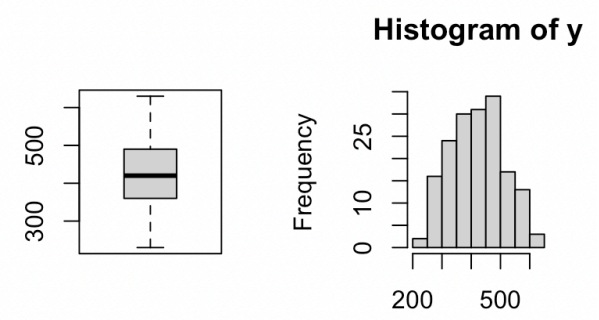
```
> t.test(large_price, conf.level=0.95)

        One Sample t-test

data:  large_price
t = 46.577, df = 26, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 503.0699 549.5226
sample estimates:
mean of x
 526.2963
```

According to the result, the confidence interval of small houses is [396.6262, 422.2969], and the confidence interval of large houses is [503.0699, 549.5226]. By the observation, we could find that there is no overlap between the confidence intervals of large house and small house. No overlap between the confidence intervals of large house and small house could means that the difference between these to data is significant.

Question 4: Regression and Analysis

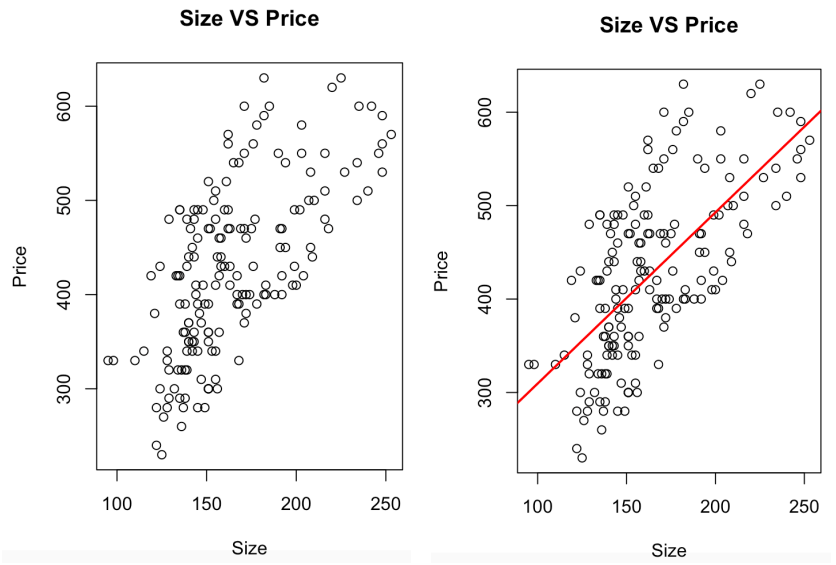
By using boxplot and histogram, we could find that there are no significant outliers in our data:

```
> y = data$Price
> x = data$Size
> par(mfrow = c(1,2))
> boxplot(y)
> hist(y)
```



Below is a scatter plot for Price and Size, and a fitted linear model by using R in-built function:

```
> plot(y~x, main = 'Size VS Price', xlab = 'Size', ylab = 'Price')
> fit = lm(y~x)
> abline(fit, col = 'red', lwd = 2)
```



From these 2 diagrams we could see that there is fairly a linear relationship between the Price and Size. (Price is the dependent variable and Size is the independent variable)

Below, a summary of fit details is concluded:

```
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-125.39  -53.90  -12.19   48.75  170.18

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.3629    26.9055   4.697 5.47e-06 ***
x             1.8322     0.1607  11.400  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68.54 on 168 degrees of freedom
Multiple R-squared:  0.4362,    Adjusted R-squared:  0.4328
F-statistic:   130 on 1 and 168 DF,  p-value: < 2.2e-16
```

The result above will be utilised to answer the questions in Question 4.

a) The regression model should be a simple liner regression model since there are only 1 independent variable and a dependent variable. The theoretical simple linear regression equation is: $Y = \beta_0 + \beta_1 X + \epsilon$ where assume E(X|ϵ) = 0. However for this specific case, the Regression equation (Price VS Size) is: $Y = 126.3629 + 1.8322X$.

b) Coefficient of determination (R^2): measure what is the proportion of the explained data over all the data.

$$R^2 = \frac{RegSS}{TSS} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = 0.4362$$

Standard error of estimate (SER): measure how well a model fit the relationship.

$$\sigma = \sqrt{\frac{\sum(y - \hat{y})^2}{n}} = 68.54$$

c) The gradient/slope in this case which is β_1 equal to 1.8322. For slope in this case, it means that when independent variable (x) increase 1 unit, corresponding dependent variable (y) will increase 1.8322 unit. On the other hand, when independent variable (x) decrease 1 unit, corresponding dependent variable (y) will decrease 1.8322 unit.

d) The intercept in our case β_0 which is 126.3629. For the significance of the intercept, a T test should be used where null hypothesis (H0) is: $\beta_0 = 0$ VS alternative hypothesis (H1) is: $\beta_0 \neq 0$. As mentioned in lecture, $t_{\beta_0} = \frac{\beta_0}{\sigma_{\beta_0}} = \frac{126.3629}{26.9055} = 4.697$ hence that related p-value = 5.47e-06. Since p-value is very small, H0 is rejected, which means $\beta_0 \neq 0$ and β_0 is significant.

e) Firstly, an overall F-test for the model usefulness where null hypothesis (H0) is: $\beta_1 = 0$ VS alternative hypothesis (H1) is: $\beta_1 \neq 0$. As mentioned in the lecture, the t-value for this test is:

$$t = \frac{\beta_1}{Se} = \frac{\beta_1}{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}} \quad \text{Where } \hat{\sigma} = SER = \sqrt{\frac{RSS}{n-2}} \text{ and } S_{xx} = \sum (x - \bar{x})^2$$

By using R code manually calculate the t-value, we find that the t value for this usefulness F-test is equal to 0, related code and output is shown below:

```
> sig.hat = sqrt(sum(fit$res^2) /(n-2))
> Sxx = var(x)*(n-1)
> t = b1/(sig.hat / sqrt(Sxx))
> p = 2*(1-pt(abs(t), n-2));p
[1] 0
```

Also, another summary of model 'fit' also demonstrated that p-value < 2.2e-16 (at the beginning of question 4). According to this test we believe that the $\beta_1 \neq 0$ which means that the model is useful. To be more specific, the overall usefulness of model is significant. (2.2e-16 is the smallest possible number that R can show, so the result between these 2 method could be considered as the same)

After that, another estimation of the model's goodness of fit, which is about R^2,

$$R^2 = \frac{RegSS}{TSS} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = 0.4362$$

According to the fit summary at the beginning of Question 4, $R^2 = 0.4362$ which means that the model is kind of fairly fit the data.

Overall, base on the F test the model seems is useful, however, the R^2 is a little bit low which means that there are a small portion data are explained by the model. It could be considered as: this model is just fairly a good model.