UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SEMESTER 2 2020/2021

SECI2143: PROBABILITY AND STATISTICAL DATA ANALYSIS

REPORT PROJECT 2

SECTION 9
GROUP 3

NAME	MATRIC NUMBER
THERESA LAU XIN YI	A20EC0167
SIEW YU XUAN	A20EC0146
BEH MING YAU	A20EC0020

LECTURER: DR AZURAH BINTI A SAMAH

1.0 INTRODUCTION & BACKGROUND

In statistical mathematics, testing concepts such as hypothesis test, correlation and Chi Squared test are important and useful to prove whether a claim is rejected or not. Nowadays, most people do not fully understand how to make a choice when buying houses as they do not know which aspect is important to be considered. This situation had aroused our curiosity to carry out a project about buying houses. In addition, we also wanted to know whether the existence of infrastructure will affect the preference of a house by people or not. In this project, we would like to use those concepts to conduct an investigation to houses from a variety of aspects. First and foremost, we wanted to know the average house prices using hypothesis test concept. Besides, we would like to investigate the relationship between house prices and size of houses with correlation and regression. Furthermore, we also carried out our investigation into the relationship between the existence of infrastructure and preference by people by applying Chi Square test of independence.

Before starting our project, we had some expectations. First, we expected that the average house price will be greater than $45000. Second, we thought that house prices would affect the size of houses. In addition, we believed that the existence of infrastructure in a house will affect preferences. Through achieving the objectives, we can analyze how we should buy a house with the right choice such as consideration to price, size, infrastructure and also recommendations by others. Thus, we believe that the project will be a great advice to those who want to buy a house and solve our doubts to factor which leads to the preference of houses by people.

2.0 DATASET

In order to carry out the investigation, we had searched for a dataset named "HousePrices.xls" which shows house prices and properties in Canada. In the dataset, it consists of 12 variables with 546 observations of houses that describe properties of the houses. The 12 variables are house prices, lot size, number of bedrooms, number of bathrooms, number of stories, existence of driveway, existence of recreation, existence of fullbase, existence of gasheat, existence of air conditioner, number of garage, and preference. Before writing our program, we had imported the dataset into R program to enable us to assign values from the dataset easily. Next, we had chosen 4 variables from the dataset to carry out investigation which were house prices, lot size, existence of air conditioner, and preference. House prices are used to test the average house price and also do correlation and regression to test the relationship between house price and size of houses. Lot size represents size of houses when testing the relationship using correlation and regression. Existence of air conditioner and preference are used to investigate the independence between the existence of infrastructure and preference by people. The other variables are not being chosen as they are not related to our objectives of this project. In addition, the variables are only the basic things in a house.

3.0 DATA ANALYSIS

3.1 Hypothesis 1 sample test

We had chosen and used the dataset named "HousePrices.xls". According to the data, we would like to conduct a hypothesis test to determine whehter the mean of house prices is greater than $45000. It is a right-tailed t-test. The population variances are unknown. Since the dataset is too large, we have set the sample as the first 30 data from the dataset. We had also assumed the significance level is 95%.

For test on mean, variance population unknown:
test statistic, $t=(\bar{x}-\mu)/(s/\sqrt{n})$

The null and alternative hypothesis is:
H_0: $\mu=45000$; H_1: $\mu>45000$
μ represents population mean for house prices.

Significance level, $\alpha=0.05$
Sample size, $n=30$
The sample mean, $\bar{x}=50198.33$, while the sample standard deviation in four decimal places, s= 16582.0867.

```
> cat("The test statistic is: ", t)
The test statistic is:  1.71706
> cat("The P-value is: ", pval)
The P-value is:  0.04831593
> cat("The critical value is: ", t.alpha)
The critical value is:  1.699127
> res <- t.test(HousePrices$price, mu=45000, alternative =
 "greater")
> #Printing the results
> res

        One Sample t-test

data:  HousePrices$price
t = 1.7171, df = 29, p-value = 0.04832
alternative hypothesis: true mean is greater than 45000
95 percent confidence interval:
 45054.29      Inf
sample estimates:
mean of x
 50198.33
```

Figure 3.1: The results of Hypothesis testing 1 sample in RStudio

In conclusion, we will reject H_0 since the test statistic is more extreme than the critical value. There is sufficient evidence that the mean house prices exceed $45000.

3.2 Correlation

For the correlation test, we measure the relationship between the lot size of the house unit and the price of the house with the sample size of 60. We use Pearson's correlation to calculate the correlation coefficient since both of these data are in ratio type.

Sample correlation coefficient :

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n]}}$$

where,
r = sample correlation coefficient
n = sample size (60)
x = value of independent variable (lot size of the house unit)
y = value of dependent variable (price of the house)

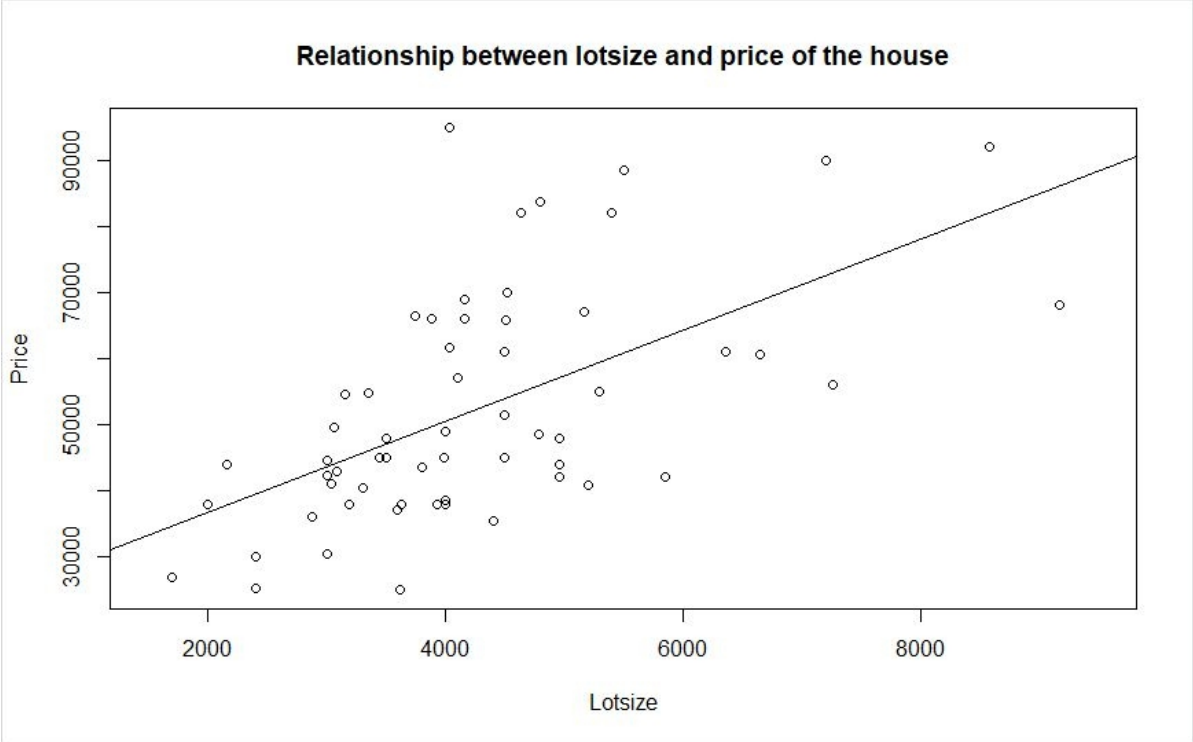

Figure 3.2.1 : Scatter plot of lot size of the house unit against price of the house

```
> library(readxl)
> HousePrices <- read_excel("D:/Downloads/HousePrices.xls", sheet = "C&R", range = "A1:C61")
> x<-c(HousePrices$lotsize)
> y<-c(HousePrices$price)
> plot(x,y,main = "Relationship between lotsize and price of the house",xlim=c(1500,9500),ylim=c(25000,95000),xlab="Lotsize",
ylab="Price")
> abline(lm(y ~ x))
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 5.5317, df = 58, p-value = 7.95e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3922947 0.7323219
sample estimates:
      cor
0.5876788
```

Figure 3.2.2 : Result of correlation coefficient, r and test statistic, t

Based on Figure 3.2.2, we found that the correlation coefficient, r = 0.5876788. The scatter plot and the correlation coefficient we obtained shows that there is a positive relationship between the lot size of the house unit and the price of the house. Since r is within 0.3922947 < r < 0.7323219, it is a moderate positive relationship which means when the lot size increases, the price of the house also increases.

Significance Test for correlation

Besides, we also want to test whether there is any evidence to prove that a linear relationship exists between the lot size of the house unit and the price of the house at the significance level of 95%.

Null hypothesis, H_0 : $\rho = 0$ ---not linear correlation

Alternative hypothesis, H_1 : $\rho \neq 0$ ---linear correlation exists

$\alpha = 0.05$, df = 58

Test statistic :

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

From Figure 3.2.2, we found that the test statistic, t = 5.5317, p-value which is the significance level of the t-test is 7.95e-07 and the confidence interval of the correlation coefficient at 0.05 significance level is between 0.3922947 and 0.7323219. In conclusion, Since the p-value(7.95e-07) is less than the significance level of 0.05, the null hypothesis is rejected. There is sufficient evidence to prove that a linear relationship exists between the lot size of the house unit and price of the house at 0.05 significance level.

3.3 Regression

For the regression test, we measure the relationship between the lot size of the house unit and the price of the house with the sample size of 60. We set the lot size of the house unit as an independent variable and the price of the house as a dependent variable. The sample regression line provides an estimate of the population regression line.

Estimated Regression Model :

$$Y = b_o + b_1 x$$

where,
Y = estimated Y value
b_0 = estimate of the regression intercept
b_1 = estimate of the regression slope
X = independent variable

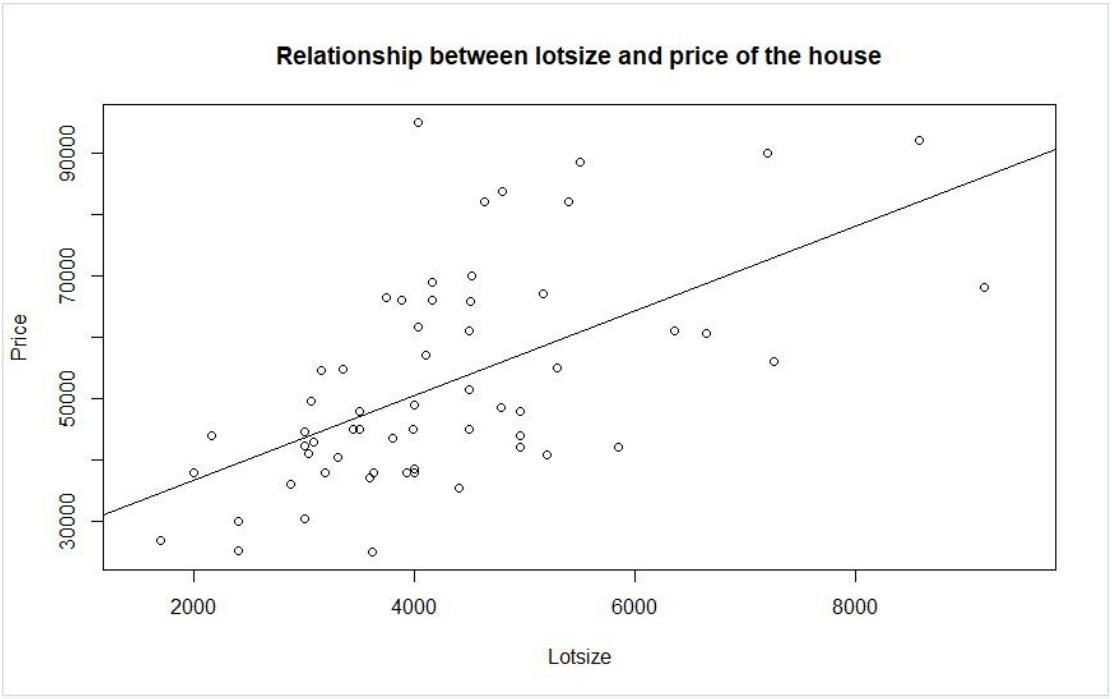

Figure 3.3.1 : Scatter plot of lot size of the house unit against price of the house

```
> library(readxl)
> HousePrices <- read_excel("D:/Downloads/HousePrices.xls", sheet = "C&R", range = "A1:C61")
> x<-c(HousePrices$lotsize)
> y<-c(HousePrices$price)
> plot(x,y,main = "Relationship between lotsize and price of the house",xlim=c(1500,9500),ylim=c(25000,95000),xlab="Lotsize",
ylab="Price")
> abline(lm(y ~ x))
> (lm(y ~ x))

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
  22876.495        6.895

> summary(lm(y ~ x))

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-22836  -9545  -2621   9095  44268

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22876.495   5657.755   4.043 0.000157 ***
x               6.895      1.246   5.532 7.95e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14090 on 58 degrees of freedom
Multiple R-squared:  0.3454,    Adjusted R-squared:  0.3341
F-statistic:  30.6 on 1 and 58 DF,  p-value: 7.95e-07
```

Figure 3.3.2 : Summary of the graph

From the summary of the graph, we can get the complete formula for estimated regression model :

$$\mathbf{Y = 22876.495 + 6.895X}$$

b_1 measures the estimated change of the average value of Y because of a one-unit change in X. Since $b_1 = 6.895$, it means that the average value of price of the house increases by 6.895 for every 1 unit of lot size of the house.

The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{sum of square explaind by regression}}{\text{total sum of squares}}$$

Based on Figure 3.3.2, we obtained the coefficient of determination, $R^2 = 0.3454$.
The coefficient of determination shows a weak linear relationship between x and y since $0 < R^2 < 1$ and it means that some but not all of the variation in y is explained by variation in x.

Regression Statistical Test

Next, we also want to test whether there is any evidence to prove that the lot size of the house unit can affect the price of the house.

Null hypothesis, H_0 : $B_1 = 0$ ---no linear relationship
Alternative hypothesis, H_1 : $B_1 \neq 0$ ---linear relationship exists

Test statistic :

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

where,
b_1 = sample regression slope coefficient
B_1 = hypothesized slope
S_{b1} = estimator of the standard error of the slope

From Figure 3.3.2, we found that the test statistic, t = 5.532. p-value which is the significance value of the t-test is equal to 7.95e-07. Since p-value(0.000000795) is less than the significance level of 0.05, the null hypothesis is rejected. Thus, there is sufficient evidence to prove that the lot size of the house unit can affect the price of the house.

3.4 Chi Square Test of Independence

Furthermore, we would also like to conduct Chi Square test of independence to investigate the relationship between the existence of air conditioner and preference by people at 0.05 significance level. In order to conduct the test, we need to state the test hypothesis first.

Null hypothesis and alternative hypothesis:
H_0: The existence of air conditioner and preference by people are independent.
H_1: The existence of air conditioner and preference by people are independent.

Next, we tabulate our dataset as observed value, O in R program as shown in the table below:

Existence of air conditioner / Preference by people	YES	NO	TOTAL
YES	53	75	128
NO	120	298	418
TOTAL	173	373	**546**

The expected values, E are calculated and recorded in the program as below:

Existence of air conditioner / Preference by people	YES	NO
YES	40.55678	87.44322
NO	132.4432	285.5568

Then, we proceed with the test with R programming. In the program, the test statistic,X^2 will be also calculated using the formula below:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

```
> #Chi-Square Test of Independence
> #H0:Existence of aircon and preference by people are indepe
ndent.
> #H1:Existence of aircon and preference by people are not in
dependent.
>
> #Make table
> tbl=table(HousePrices$aircon,HousePrices$prefer)
>
> #Apply Chi-Square Test
> #It will show Test Statistic with Chi Square value and P-va
lue
> #It will also show degree of freedom
> chisq.test(tbl,correct=FALSE)

        Pearson's Chi-squared test

data:  tbl
X-squared = 7.2997, df = 1, p-value = 0.006897
```

Figure 3.4: The result of Chi-Square Test(X-squared represents test statistic,df represents degree of freedom, p-value represent probability)

From the program, we had found that the test statistic is 7.2997 which is same as P-value 0.006897. Since the P-value is lower than the significance level (0.006897<0.05), H_0 is rejected. In conclusion, the existence of air conditioner and preference by people are not independent.

4.0 CONCLUSION

By doing this project, we all have learned how to code in R programming by using RStudio. Before that, we only learned the C++ programming language in the course Programming Technique. But now, we know more about R programming and improve our skills. Furthermore, we are able to analyse the huge dataset and produce a conclusion statement. We can get the conclusion of tables and scatter plots by debugging the program.

According to the test we have done, we can conclude that the mean house prices are greater than $45000 from the hypothesis 1 sample test. Next, the correlation test shows that a linear relationship exists between the lot size of the housing unit and the price of the house. From the regression test, we also know that the lot size of the housing unit can affect the price of the house. Moreover, we can also summarize that the existence of air conditioner and preference by people are not independent.

5.0 APPENDIX

The raw dataset:
https://drive.google.com/file/d/1ZcLg5wq7eiCvXDgNt-x0iEnfh1jNbucg/view?usp=sharing

6.0 DEMO VIDEO

Video Link:

https://drive.google.com/file/d/1ib5-3_NM9w9j9VIUydCEWjkdUDH4o5ea/view?usp=sharing