SCHOOL OF COMPUTING

SEMESTER II 2020/2021

SECI2143 – PROBABILITY AND STATISTIC DATA ANALYSIS

PROJECT 2: SALES OF SUPERSTORE IN CHICAGO

NAME	MATRIC NUMBER
Muhammad Aniq Aqil Bin Azrai Fahmi	**A20EC0083**
Muhammad Naim Bin Abdul Jalil	**A20EC0096**
Khairul Izzat Bin Hashim	**A20EC0058**
Nur Afikah Binti Mohd Hayazi	**A20EC0220**

DR. NOR AZIZAH ALI

SUBMISSION DATE: 3RD JULY 2021

TABLE OF CONTENT

1. Introduction

Superstores are a common place for human being to buy their essential stuff. There are always be a massive number of sales every year for the Superstore. Usually, Superstore has a big facility which can store a massive number of products. Superstore meet many consumers as their already cover many categories. Their unique retail merchandisers. Our Dataset is about Sales of Superstore in Chicago in the year of 2019 that contains 3 category of product sales which is Furniture, Office Supplies and Technology can be used to analyse their sales, profits and other details of store in Chicago. This superstore is operating through online and every item that has been ordered will be delivered through shipping. It stores every information about the shipment such as ship mode, ship status and days to ship.

The aim of this study is to do an inference statistical analysis based on our chosen dataset and data variables. By making this project, we will be able to know the relationship between all variables of superstore record and how to conduct the analysis using Hypothesis testing, Correlation, Regression and ANOVA. The outcomes obtained from this project is it can be used to increase the profit and sales every superstore.

2. Dataset (Description)

For this project, we choose the Superstore Record in 2019 dataset in USA. Due to the big amount of data, we only choose the Superstore record in Chicago that only contained 113 sample sizes. In this dataset, it shows every Superstore's item that already sold through online in the Chicago area. In this data, we can see the category of each item that has been sold with the date of order. Besides that, we can see the quantity, amount of discount, total sales and total profit from each item.

From all of the variables, we will do various hypothesis tests on the inferential statistics. The table below shows the data type, description and method of analysis for each variable.

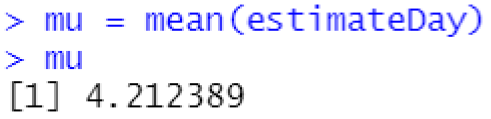
Variables	Data Type	Description	Method of analysis
Category	Qualitative	Category of each item	ANOVA
Date	Quantitative	Date of order from customer	ANOVA
Quantity	Quantitative	Quantity of item sold	Regression
Day of shipping	Quantitative	Total day of shipping	Hypothesis Test
Sales	Quantitative	Sales from sold item	Corelation
Profit	Quantitative	Profit gained from sold item	Corelation, Regression

Table 1 List of variables

3. Data Analysis

3.1 Hypothesis Test

In statistics, a hypothesis test is a standard procedure for testing a claim about a property of a population. Based on the sample of 113 product, the average of estimate days to ship the products is 4.212389.



```
> mu = mean(estimateDay)
> mu
[1] 4.212389
```

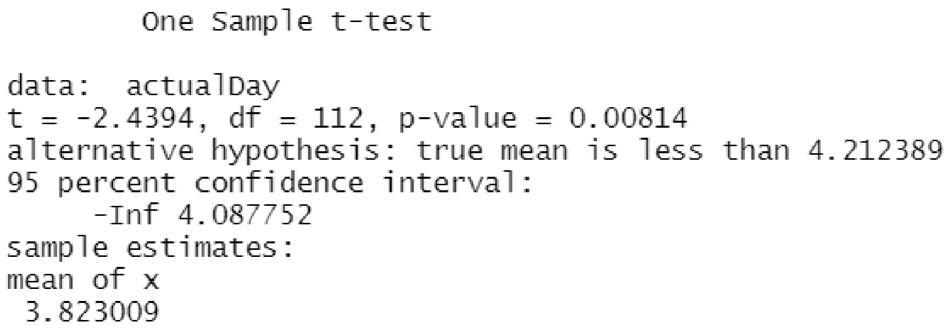
Figure 1 Mean of estimate days to ship

Also, from the sample of 113 products, we claim that the mean of actual days to ship is less than the mean of estimate days to ship. Hence, the null hypothesis, H_0 and alternative hypothesis, H_1 is:

$$H_0: \mu = 4.212389$$

$$H_1: \mu < 4.212389$$

A 95% level of confidence is used to test the claim in R-studio and below is the output by using R-studio.



```
	One Sample t-test

data:  actualDay
t = -2.4394, df = 112, p-value = 0.00814
alternative hypothesis: true mean is less than 4.212389
95 percent confidence interval:
     -Inf 4.087752
sample estimates:
mean of x 
 3.823009 
```

Figure 2 Output One Sample T-test from R studio

Since the p-Value < 0.05, we decide to reject the null hypothesis. There is sufficient evidence that $\mu < 4.212389$. Therefore, we can conclude that average of actual days to ships is less than average of estimate days to ship.

3.2 Correlation

Generally, correlation is measuring statistical that extend to which two variable is linearly related or not. By analysing the data, we get to explain whether two variables are a linear relationship between each other. The two variables that will be analyse are:

- Profit of item sold.
- Sales of item sold.

The dataset in Sales of Superstore, we have examined the relationship between profit of item sold and the sales of item sold in the category of Technology. The observation of number of in the category of technology sample is 21. Hence, this data is measured in a ratio data so it is suitable to use Pearson's Product-Moment Correlation. From the data sample, we make a scatter plot to determine the existent of correlation by using R-studio.

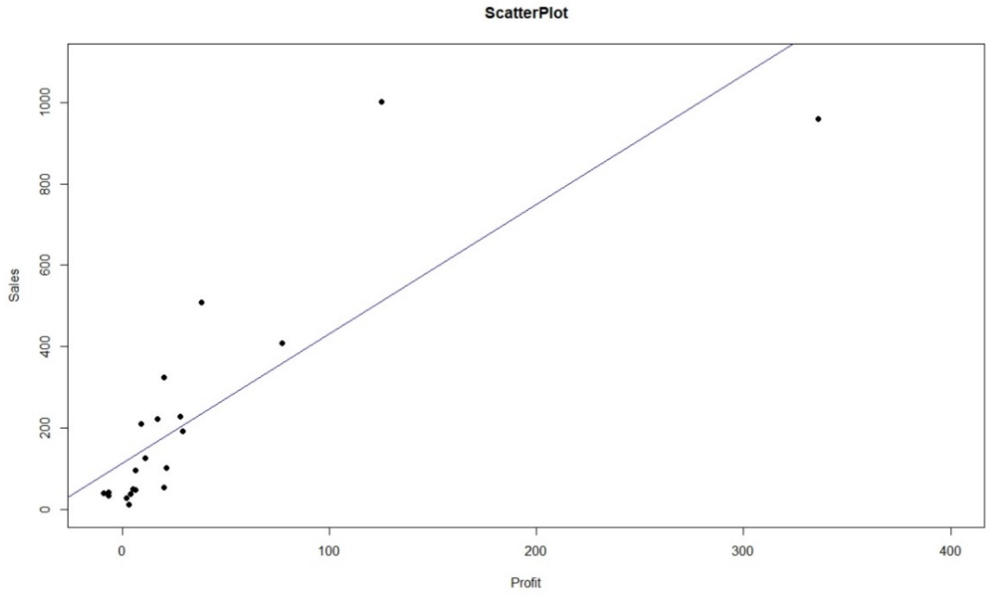
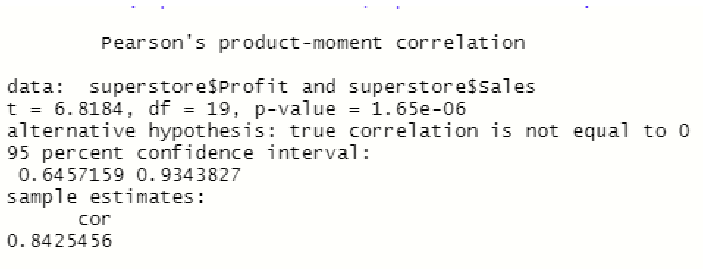


Figure 3 Scatter Plot of Profit and Sales



```
        Pearson's product-moment correlation

data:  superstore$Profit and superstore$Sales
t = 6.8184, df = 19, p-value = 1.65e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6457159 0.9343827
sample estimates:
      cor
0.8425456
```

Figure 4 Output of Pearson's Product-moment Correlation

As the result shows, the sample correlation coefficient which are equal to 0.8425456. Using the 95% confidence interval, then $\alpha = 0.05$. The value of test statistic, t= 6.8184, compare to the $t_{0.05,19} = \pm 1.729$, is greater, and also p-Value 1.65e-06 < significant value, so we reject null hypothesis.

This define to claim the test that the relationship between profit of sold items and sales of sold items in the category of Technology is positive linear relationship. The value of r=0.8425456 represent the moderate strength of a linear relationship between profit and sales in the category of Technology. The closer the value of r to 1, the stronger the relationship. When the profit increases, sales also increase.

3.3 Regression

For regression, we will explain the impact of changes in an independent variable on the dependent variable.

- Dependent variable: Profit from sold item.
- Independent variable: Quantity of item sold.

From the dataset of Sales of Superstore, we wish to examine the relationship between the quantity of item sold and profit from sold items. So, we select 63 number of items of office supplies as sample. We picked quantity and profit from category office supplies only. From the data sample, we make **a scatter plot** with **a linear regression model** by using R-Studio to know the type of regression model.

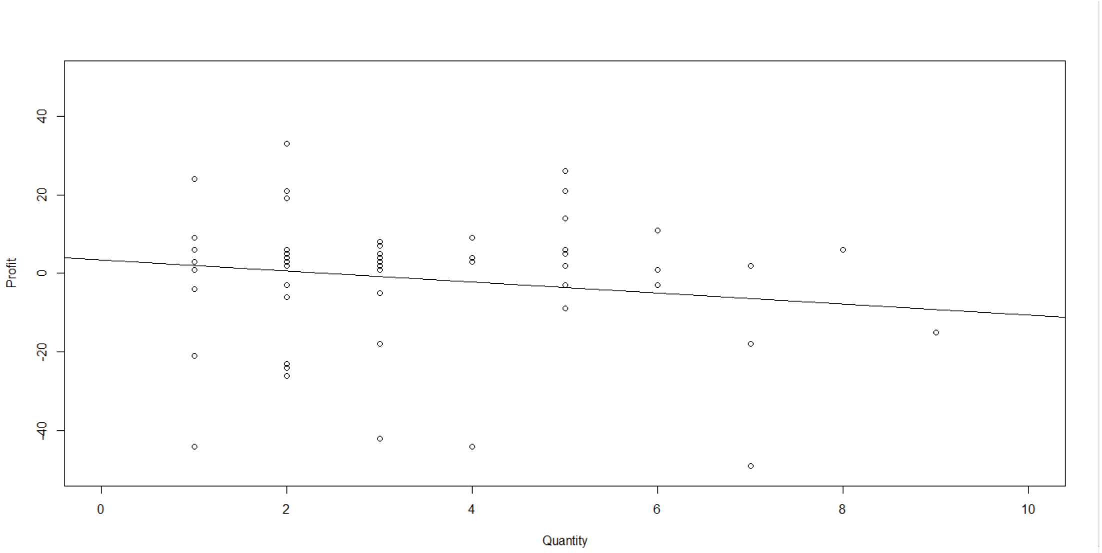
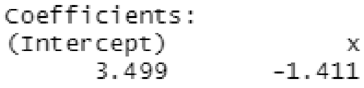


Figure 5 Scatter Plot of Profit and Quantity



```
Coefficients:
(Intercept)            x
      3.499       -1.411
```

Figure 6 Output of R script

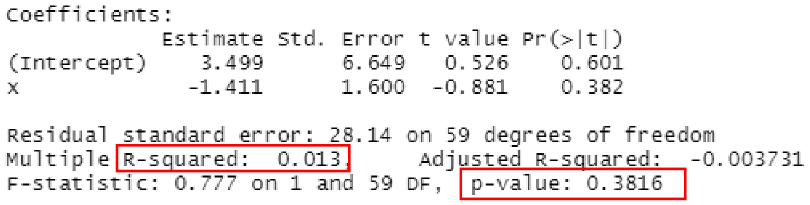
We can see that the quantity and the profit are negative linear relationship and the linear regression equation is:

$$\hat{y} = \mathrm{b}_0 + \mathrm{b}_1 x$$

$$\hat{y} = 3.499 - 1.411x$$

b_0 measures the estimated change in average value of y a result of a one-unit changes in x. Hereby, we can see that $\hat{y}$ is \$3.499 if quantity, x is 0. This means, the profit not explained by the quantity of sold item.

Finally, we calculate the R-squared the coefficient of determination by using R programming.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.499      6.649   0.526    0.601
x             -1.411      1.600  -0.881    0.382

Residual standard error: 28.14 on 59 degrees of freedom
Multiple R-squared:  0.013,     Adjusted R-squared:  -0.003731
F-statistic: 0.777 on 1 and 59 DF,  p-value: 0.3816
```

Figure 7 Summary of Regression

The value of R-squared is 0.013. Thus, 1.3% of the variation of profit is explained by the variation of the quantity. So, the quantity of the item sold is less likely to affect the profit.

To investigate the linear relationship between quantity, x and profit, y, The null and alternative hypothesis:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 < 0$$

By using the 95% confidence interval then $\alpha = 0.05$. The P-value from the output of R Studio is 0.3816. Thus, P-value 0.3816 > 0.05 significance level. Hereby, fail to reject the Hypothesis null. Therefore, there is no sufficient evidence that the quantity of item sold affects the profit of each item sold.

3.4 ANOVA

A one-way ANOVA was conducted to compare the number of sales of the superstore at Chicago in average in month on 2019. Significance level of 0.05 is used to test the data.

The hypothesis statement for this test is:

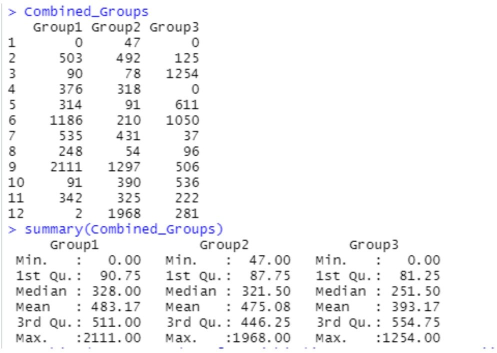
$$H_0 : \mu_{Furniture} = \mu_{Office\ Supplies} = \mu_{Technology}$$

$$H_1 : \mu_{Furniture} \neq \mu_{Office\ Supplies} \neq \mu_{Technology}$$

	Sales in 2019		
Month	Furniture	Office Supplies	Technology
January	0	47	0
February	503	492	125
March	90	78	1254
April	376	318	0
May	314	91	611
June	1186	210	1050
July	535	431	37
August	248	54	96
September	2111	1297	506
October	91	390	536
November	342	325	222
December	2	1968	281

Table 2: Data Set for Sales in 2019 of the Superstore

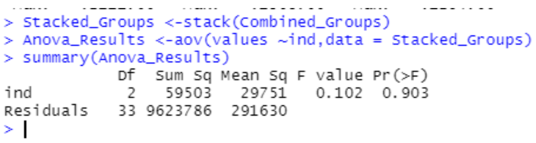
Next the data has already been divided to three category of groups which is Furniture, Office Supplies and Technology and the sales already been divide to 12 months in a year sale of 2019. We decided to use our significance level, 0.05 to test the null hypothesis that different category of sales have the same mean. For one way ANOVA, testing the sample size, n should be the same which is from the table we have 12. The means are already calculated in the R Script as shown below:



```
> Combined_Groups
   Group1 Group2 Group3
1       0     47      0
2     503    492    125
3      90     78   1254
4     376    318      0
5     314     91    611
6    1186    210   1050
7     535    431     37
8     248     54     96
9    2111   1297    506
10     91    390    536
11    342    325    222
12      2   1968    281
> summary(Combined_Groups)
     Group1              Group2              Group3       
 Min.   :   0.00   Min.   :  47.00   Min.   :   0.00  
 1st Qu.:  90.75   1st Qu.:  87.75   1st Qu.:  81.25  
 Median : 328.00   Median : 321.50   Median : 251.50  
 Mean   : 483.17   Mean   : 475.08   Mean   : 393.17  
 3rd Qu.: 511.00   3rd Qu.: 446.25   3rd Qu.: 554.75  
 Max.   :2111.00   Max.   :1968.00   Max.   :1254.00  
```

Figure 78: Data Summary for Data Set in R studio

After that we proceed on findings of One Way Anova Analysis which is to find the F- value.



```
> Stacked_Groups <-stack(Combined_Groups)
> Anova_Results <-aov(values ~ind,data = Stacked_Groups)
> summary(Anova_Results)
            Df  Sum Sq Mean Sq F value Pr(>F)
ind          2   59503   29751   0.102  0.903
Residuals   33 9623786  291630
> 
```

Figure 89: Code for ANOVA analysis in R Studio

$$F(2{,}33) = 0.102\ , p > 0.05$$

The analysis based on the result we gain for the degree of Freedom for Numerator is 2 and degree of Freedom for Denominator is 33, with 0.05 significance level.

Then we also got the F -Value which is 0.102 and also the P-Value is 0.903.

Since P-value 0.903 > significance level 0.05 and the F-value statistics 0.102 < 3.285, we fail to reject H_0. Hence, there are sufficient evidence to claim that the mean value for the sales of the superstore at Chicago in 2019 is same.

4. Conclusion

Based on our best findings, we decide to reject null hypothesis. There is sufficient evidence to support the claim average of actual days to ships is less than average of estimate days to ship. In addition, the analysis had been found that there is a moderate positive linear relationship between profit of sold items and sales of sold items within a correlation coefficient of 0.8425456.

Moreover, the estimated regression model is then obtained as $\hat{y} = 3.499 - 1.411x$ and the regression model is helpful in estimating the estimated change in average value of y a result of a one-unit change in x. Hence, the best finding analysis can get through using R-Studio where we can get the precise analysis and also learning implementing R-Studio by doing project also gave us great experience to manage data set and having high critical thinking to sort data variables to use for each analysis such as Regression, Anova , Correlation, Hypothesis one sample and Hypothesis two sample.