UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI2143

PROBABILITY & STATISTICAL DATA ANALYSIS

SEMESTER 2019/2020-2

PROJECT 2

Report

Name: Siti Najwa binti Apandi

Matrix no: A19EC0163

Section: 06

Lecturer: Dr. Chan Weng Howe

Table of Contents

1.0 INTRODUCTION

Starting on 1st April 2015, the government has introduced Goods and Services Tax (GST) to replace the Sales and Services Tax (SST) which has been used in the country for several decades. Prime Minister decided to change this system tax to increase the economy in this country since many countries have implement GST. This change has lead various reaction from citizen where there are some people who are not agree with this GST implementation and state that GST only burden them. The introduction of GST had contributed to the rising cost of living in Malaysia especially for B40 category which income range is below RM4000. A survey was made by Nor Fatimah Che Sulaiman, Nur Azura Sanusi and Suriyani Muhamad from Universiti Malaysia Terengganu to investigate the perception of Malaysian households about the increasing cost of living and how much their income per month. Many variables have been collected by the researcher including state, locality, gender, household income, income strata category, total expenditure and net income. A total of 735 respondents were selected in this survey. This paper consists of hypothesis testing single sample, correlation analysis, regression analysis and chi-square test.

2.0 CONTENT

2.1 Hypothesis Testing Single Sample

A total of 735 Malaysians were surveyed by researcher to find out their household income. From the dataset, mean and standard deviation of household income is RM4549.246 and RM3715.124 respectively. Let say that we claim the mean of household income is less than RM5000. Is there sufficient evidence to support the claim? To test this claim, hypothesis testing single sample on mean with 0.05 level of significance is used and the testing is left-tailed test.

$H_0: \mu = 5000$

$H_1: \mu < 5000$

$\alpha = 0.05$

```
> n = 735              #sample size
> xbar = 4549.246      #mean
> sd = 3715.124        #standard deviation
> mu = 5000            #null hypothesis value
> alpha = 0.05         #significance value
> z = (xbar-mu)/(sd/sqrt(n))     #calculate z test statistics
> z.alpha = qnorm(1-alpha)       #critical value
> z
[1] -3.289349
> -z.alpha
[1] -1.644854
```

The calculation is done by using Rstudio. Based on the result above, test statistic is -3.289349 and critical value is -1.644854. Since test value < critical value and fall within critical region, we reject null hypothesis. Therefore, there is sufficient evidence to support the claim that the mean of household income is less than RM5000.

```
> #P-value
> pval = pnorm(z)
> pval
[1] 0.0005020972
```

By using p-value method, the lower tail p-value of the z test statistic is 0.0005020972. If p-value < α, null hypothesis is rejected while if p-value > α, fail to reject null hypothesis. Since p-value < α = 0.05, we reject null hypothesis. There is sufficient evidence to conclude that the mean of household income is less than RM5000.

From the hypothesis testing above, mean of the household income is less than RM5000. It is shows that most of the respondent has income less than RM5000 and they are under B40 category. This category may be affected by GST.

2.2 Correlation

Correlation analysis is used to measure the strength of the linear relationship between two variables which are independent and dependent variable. From the dataset, the variable used is household income(total) for independent variable. Dependent variables for analysis purpose are transport expenditure, housing expenditure, food expenditure and net income, hence there are four correlation analysis will be done. Is there evidence of a linear relationship between household income and each type of dependent variable above at the 0.05 level of significance?

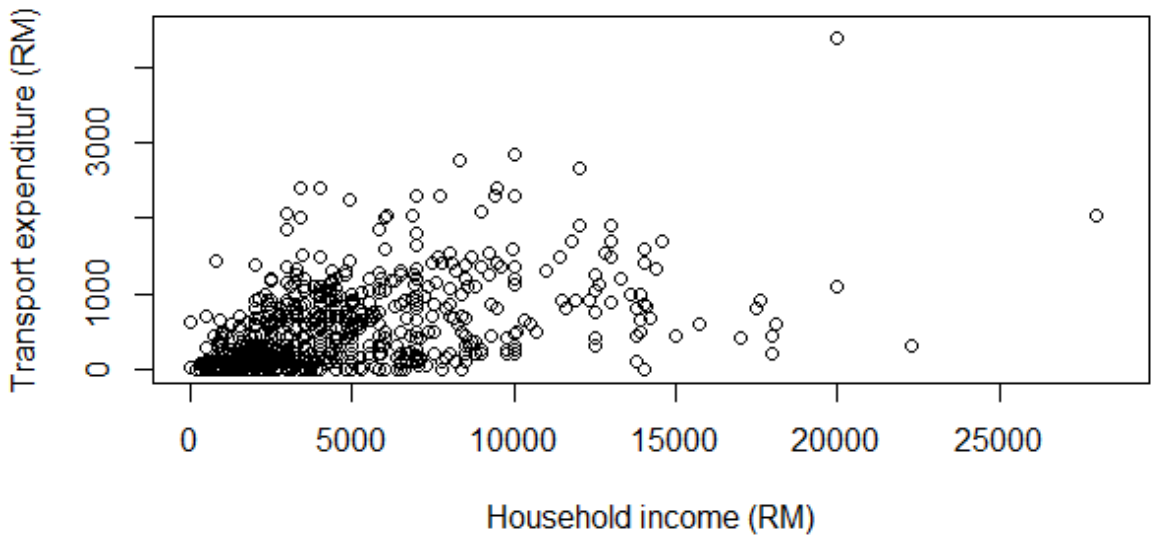
1. Relationship between household income(total) and transport expenditure

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

```
> cor(`Household income (total)`, `Transport expenditure`)
[1] 0.4921441
> n = 735
> r = 0.4921441
> df = n-2
> alpha = 0.05
> div = (1-(r^2))/(n-2)
> t = r/sqrt(div)
> t
[1] 15.30624
> t.alpha = qt(alpha/2, df)
> c(-t.alpha, t.alpha)
[1]  1.963206 -1.963206
```

The result obtained is r = 0.4921441. Test statistic value is t = 15.30624 and critical value is $\pm t_{0.025,733} = \pm 1.963206$. Since test statistic value > critical value and fall within the critical region, we reject null hypothesis. Therefore, there is a linear relationship between household income and transport expenditure. The correlation is between 0 and 0.5, therefore the strength of the linear relationship is weak.



The scatter plot above shows the relationship between household income and transport expenditure. Based on the scatter plot above, it can be seen that the household income increases as the transport expenditure increases. A scatter plot and correlation analysis of the data indicates that there is positive relationship between household income and transport expenditure.

Based on the result, correlation for relationship between household income and transport expenditure is relatively weak. Transport expenditure is including money spent to buy vehicle, the accessories and the service for their vehicle. Family with higher household income are able to buy their own vehicles such as car or motorcycle while family with lower household income use bus or taxi to go somewhere. However, there are some low income households able to spend their money to buy vehicle.

2. Relationship between household income(total) and housing expenditure

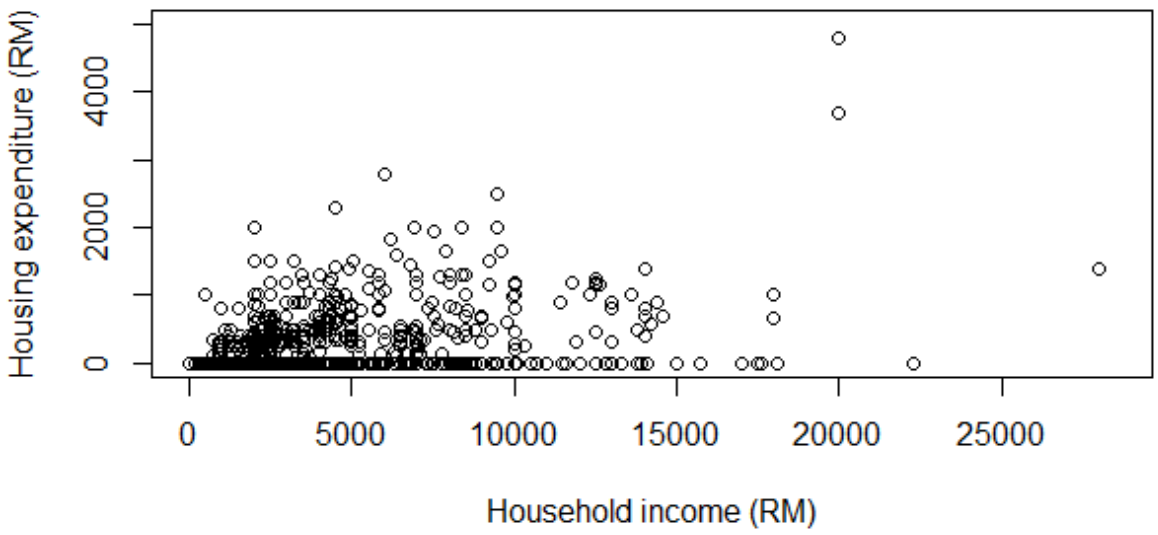
$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

```
> cor(`Household income (total)`, `housing expenditure`)
[1] 0.3423051
```

```
> n = 735
> r = 0.3423051
> df = n-2
> alpha = 0.05
> div = (1-(r^2))/(n-2)
> t = r/sqrt(div)
> t
[1] 9.863419
> t.alpha = qt(alpha/2, df)
> c(-t.alpha, t.alpha)
[1]  1.963206 -1.963206
```

Correlation coefficient for relationship between household income and housing expenditure is r = 0.3423051. Test statistic value is t = 9.863419 and critical value is $t_{0.025,733}$= 1.963206. Since test statistic value > critical value and fall within the critical region, null hypothesis is rejected. Therefore, there is sufficient evidence that there is linear relationship between household income and housing expenditure. The correlation is positive and lies between 0 and 0.5. Thus the correlation is relatively weak positive linear association between household income and housing expenditure.



The scatter plot above shows that household income increases as housing expenditure increases inconsistently. A scatter plot and correlation analysis of the data indicates that there is positive relationship between the household income and housing expenditure.

From the analysis above, the result obtained for relationship between household income and housing expenditure is weak correlation. Not all high income households expend more money for housing even their income is high while there are some low income households expend more for housing. Therefore, the relationship of the two variables is relatively weak.

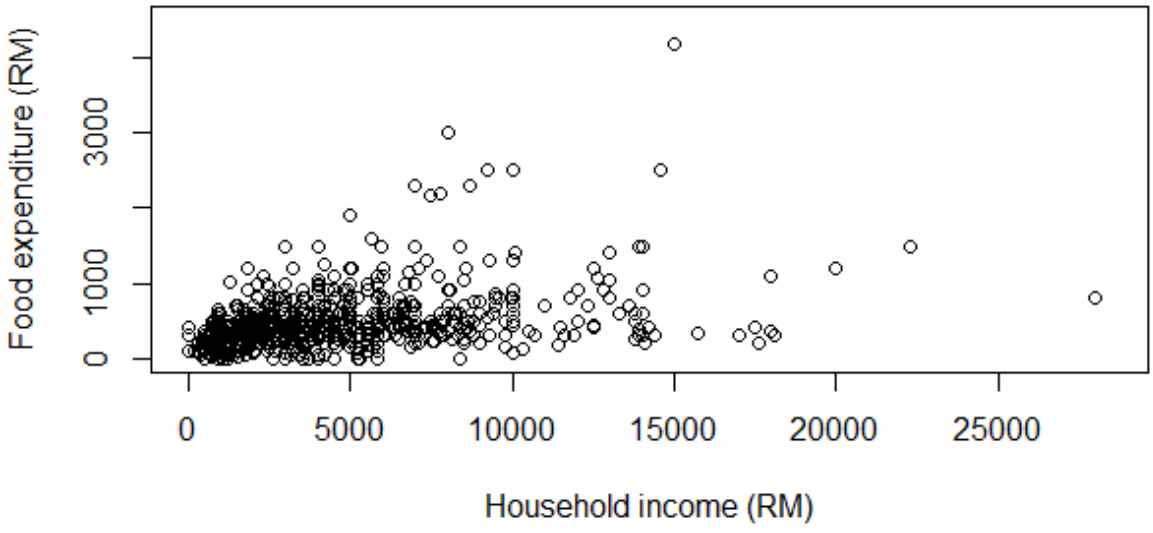
3. Relationship between household income(total) and food expenditure

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

```
> cor(`Household income (total)`, `Food expenditure`)
[1] 0.3744173
> n = 735
> r = 0.3744173
> df = n-2
> alpha = 0.05
> div = (1-(r^2))/(n-2)
> t = r/sqrt(div)
> t
[1] 10.93217
> t.alpha = qt(alpha/2, df)
> c(-t.alpha, t.alpha)
[1]  1.963206 -1.963206
```

Correlation coefficient for above relationship is $r = 0.3744173$. Test statistic value $t = 10.93217$ and critical value is $t_{0.025,733} = 1.963206$. Since test statistic value > critical value, we reject null hypothesis. Therefore, there is linear relationship between household income and food expenditure. Similar with previous correlation, the correlation coefficient is lies between 0 and 0.5 which means that it is relatively weak positive linear association.



From the scatter plot above, it can be seen that as household income increases, food expenditure will also increase and lead to the positive linear correlation. A scatter plot and correlation analysis of the data indicates that there is positive relationship between the household income and food expenditure.

The result shows that the relationship above is relatively weak. Not all high income households spend more money for food and vice versa. This may be due to different number of children. If the high income households have less number of children, they will spend less money for food compared to low income households that have many children. However, the correlation is positive that means when households income is increase, the food expenditure will increase.

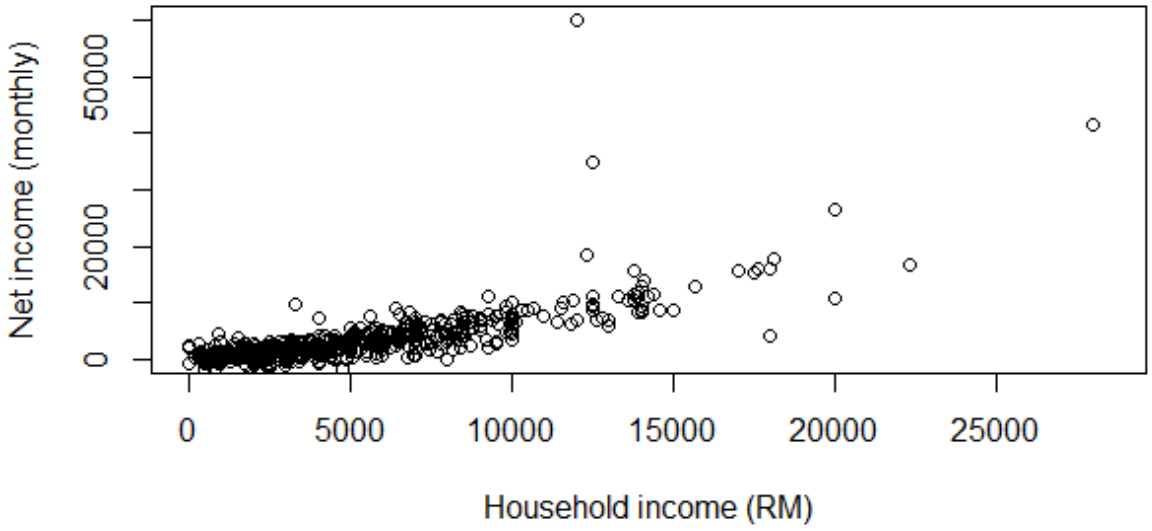
4. Relationship between household income and net income

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

```
> cor(`Household income (total)`, `Total income - total expenditure (monthly)`)
[1] 0.7586219
> n = 735
> r = 0.7586219
> df = n-2
> alpha = 0.05
> div = (1-(r^2))/(n-2)
> t = r/sqrt(div)
> t
[1] 31.52406
> t.alpha = qt(alpha/2, df)
> c(-t.alpha, t.alpha)
[1]  1.963206 -1.963206
```

Net income is a difference of total income and total expenditure per month. The result obtained is 0.7586219. Test statistic value is t =31.52406 and critical value is $t_{0.025,733}$= 1.963206. Since test statistic value > critical value, null hypothesis is rejected. Therefore, there is linear relationship between household income and net income. Unlike the previous relationship, the correlation coefficient for relationship between household income and net income is lies between 0.5 and 0.8 which is relatively moderate positive linear association.



From the scatter plot above, it can be seen clearly that as household income increase, the net income increase and the data points lie almost nearly on a straight line that slopes upward. A scatter plot and correlation analysis of the data indicates that there is positive relationship between the household income and net income

The correlation for all analysis are linear relationship. When variable in x-axis increase, there is increasing of variable in y-axis although the plotting of the data point is not consistent. The first three relationship above are weak correlation while the fourth relationship shows a moderate correlation.

2.3 Regression

Regression analysis is used to predict the value of a dependent variable based on the value of at least one independent variable and explain the impact of changes in an independent variable on the dependent variable. From the dataset, variable household income(total) and housing expenditure is used for regression analysis where household income(total) is independent variable while housing expenditure is dependent variable. Person who has higher income usually afford to spent more money for housing. Is there a linear relationship between household income and housing expenditure? Does household income affect the housing expenditure?

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

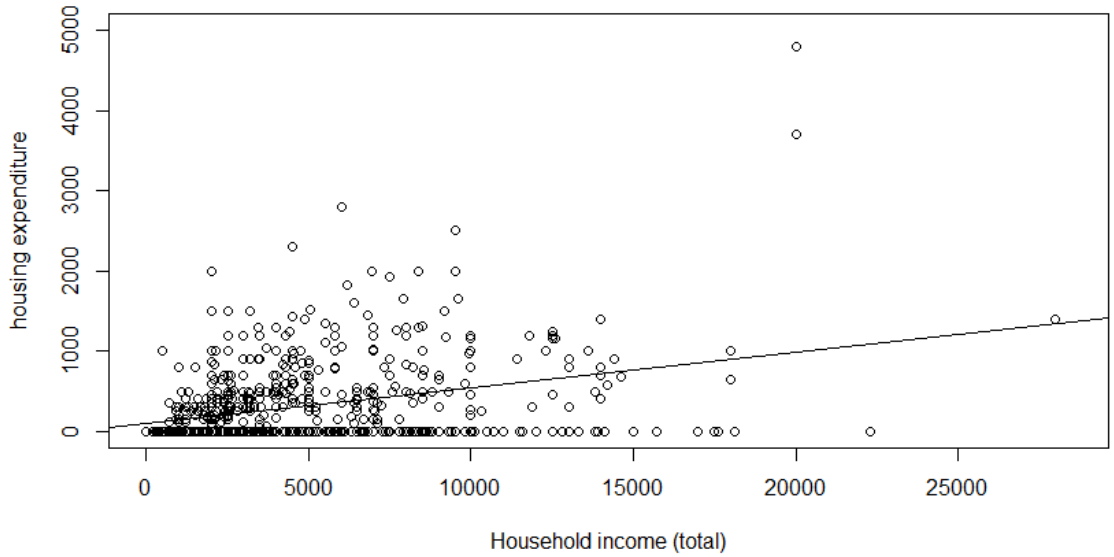
$\alpha = 0.05$

```
> x <- c(`Household income (total)`)
> y <- c(`housing expenditure`)
> model <- lm(y~x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   92.23895      0.04469
```

Assume that x is household income(total) and y is housing expenditure. From the result obtained, estimate of the regression intercept, b_0 is 92.23895 and estimate of the regression slope, b_1 is 0.04469. Therefore, the estimated regression equation is $\hat{y} = 92.239 + 0.045x$.



```
> summary(model)

Call:
lm(formula = `housing expenditure` ~ `Household income (total)`)

Residuals:
    Min      1Q  Median      3Q     Max
-1088.7  -244.2  -132.7   157.5  3814.0

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                92.238950  26.604745   3.467 0.000557 ***
`Household income (total)`  0.044690   0.004531   9.863  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 456 on 733 degrees of freedom
Multiple R-squared:  0.1172,    Adjusted R-squared:  0.116
F-statistic: 97.29 on 1 and 733 DF,  p-value: < 2.2e-16
```

Estimate of the standard error of the least squares slope (S_{b1}) is 0.004531. The result above shows that the test statistic value is t = 9.863.

```
> #calculate T critical value
> n = 735
> alpha = 0.05
> df = n - 2
> t.alpha = qt(alpha/2, df)
> c(-t.alpha, t.alpha)
[1]  1.963206 -1.963206
```

By using significance level of 5%, critical value is $t_{\alpha/2}$ = 1.963 and $-t_{\alpha/2} = -1.963$. Test statistic value (t = 9.863) > critical value ($t_{\alpha/2}$ = 1.963). Since test statistic fall within critical region, null hypothesis is rejected. There is sufficient evidence that household income affects the housing expenditure. To expend some money for housing, household income is the one of the factor to consider it.

2.4 Chi Square Test of Independence

Locality of Malaysians can be classified into two categories which are urban and rural. Urban area is the region surrounding a city and can refer to towns, cities and suburbs. Rural areas are the opposite of urban areas where usually have low population density and large amounts of undeveloped land. There are three categories for income strata category which are B40, M40 and T20. Income range for B40 is below RM4,360 while M40 is between RM4,360 and RM9,619. Income range for T20 is the highest which is more than RM9,619. From this info, is there any evidence on the relationship between locality and income strata category? To test this claim, Chi-Square test of independence is used with 0.05 level of significance.

	Income Strata Category			
Locality	B40	M40	T20	Total
Rural	200	70	35	305
Urban	188	167	75	430
Total	388	237	110	735

H_0 : Income strata category is independent of locality

H_1 : Income strata category is not independent of locality

α = 0.05

```
> Rural <- c(200, 70, 35)
> Urban <- c(188, 167, 75)
> d <- data.frame(Rural, Urban)
> # perform chi-square test on the data table
> chisq.test(d, correct=FALSE)

        Pearson's Chi-squared test

data:  d
X-squared = 34.352, df = 2, p-value = 3.472e-08

> alpha <- 0.05      #critical value
> x2.alpha <- qchisq(alpha, 2, lower.tail=FALSE)
> x2.alpha
[1] 5.991465
```

The solution can be obtained by using Rstudio. Based on result above, test statistic is x^2 = 34.352 and critical value $x^2_{2,0.05}$ = 5.991465. Since test statistic value > critical value, we reject null hypothesis at α = 0.05. The result shows p-value is 3.472×10^{-8} and p-value < α which

means that null hypothesis is rejected. Therefore, there is evidence of a relationship between locality and income strata category.

From the dataset, total of rural community sample is 305 people while urban community is 430 people. Most of the respondents are from B40 category and living in rural area. Job opportunities in rural area are limited and their income is not as high as in urban area. Compared to urban people, they live in city where job opportunities are higher and the income offered is also high. Thus, the number of urban people who under M40 and T20 category is higher than rural people.

Next, we want to find out Malaysian households' perception about GST whether there is rising cost of living or not. Is there any evidence on relationship between income strata category and their perception toward rising cost of living? The claim is test by using Chi-Square test of independence with 5% level of significance.

	Rising cost of living		
Income strata category	No	Yes	Total
B40	67	321	388
M40	42	195	237
T20	23	87	110
Total	132	603	735

H_0 : No relationship between income strata category and rising cost of living

H_1 : Income strata category and rising cost of living has relationship.

$\alpha = 0.05$

```
        Pearson's Chi-squared test

data:  d
X-squared = 0.78455, df = 2, p-value = 0.6755

> alpha <- 0.05       #crtitical value
> x2.alpha <- qchisq(alpha, 2, lower.tail = FALSE)
> x2.alpha
[1] 5.991465
```

Test statistic value is x^2 = 0.78455 and critical value is x^2 = 5.991465. Since test statistic < critical value and fall outside critical region, we fail to reject H_0. Besides, p-value for the test is 0.6755 which is greater than α = 0.05. If p-value > α, null hypothesis cannot be rejected. Therefore, there are no relationship between income strata category and rising cost of living.

From the test analysis, we can conclude that most of Malaysian households are affected with GST no matter which category they are either B40, M40 or T20 because the rising cost of living is not depend on income strata category.

3.0 CONCLUSION

Based on the analysis above, we can simply conclude that majority of Malaysian households agreed that the introduction of GST cause the increasing in cost of living especially for low income households and B40 category. The analysis shows that mean of Malaysian household's income is less than RM5000 and there are from B40 category. Family with lower household income have to spend less for their expenditure in order to survive their living compared to high income households. It can be proven by correlation analysis for the relationship between household income and each type of expenditure. Although most of Malaysian agreed that GST make their life burden, in fact that GST is implemented by 160 countries shows that GST is more effective than SST and higher collection of GST will be resulted in higher revenue because it is comprehensive and more transparent.

4.0 REFERENCE

Nor Fatimah Che Sulaiman, Nur Azura Sanusi, & Suriyani Muhamad (2020). Survey dataset of Malaysian perception on rising cost of living. *Journal Data in Brief, 28* https://doi.org/10.1016/j.dib.2019.104910

Urban Area. Retrieved June 18, 2020, from https://www.nationalgeographic.org/encyclopedia/urban-area/#:~:text=%22Urban%20area%22%20can%20refer%20to,towns%2C%20cities%2C%20and%20suburbs.&text=Rural%20areas%20are%20the%20opposite,an%20urban%20area%20is%20clear.