UTM UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

Semester II 2021/2022

SECI 2143 - 01

(Probability & Statistical Data Analysis)

TITLE :

CASES OF HIV AND AIDS INFECTED (2001 - 2018)

Group 8

NAME	MATRIC NUMBER
FARAH NABILAH BINTI NAJMUDIN	A21EC0023
MAATHUREE A/P VEERABALAN	A21EC0051
MUHAMMAD SAIFUDDIN BIN ISMAIL	A21EC0093
NUR SYAFIKA BINTI MOHD SALMIZI	A21EC0115

Submitted to:

Dr Azurah A Samah

1.0 Introduction

The dataset chosen in this study is “Kes Jangkitan HIV dan Aids yang dilaporkan dari 2001 hingga 2018” was collected from a long term study of cases through 18 years. The main purpose in this project is to identify the increment and how widespread the reported cases of the virus are. Based on the dataset chosen , it included several records from each gender, male, female, total cases per year and percentage of infected females.

As we all know, HIV and Aids infection is a virus that seems to have become obsolete in our society. Moreover, it is a very dangerous virus as it is often infected by women . The first HIV/AIDS case in Malaysia was first detected in 1986 and since then, the virus has become one of the most serious health and development challenges in the country. As of the end of 2015, there are an estimated 93,000 people living with HIV in the country.

Malaysia ranks seventh highest in adult HIV and Aids prevalence in Asia after Thailand, Papua New Guinea, Myanmar, Cambodia, Vietnam and Indonesia with an average value rate of 0.45%. In 2015, Malaysia recorded a rate of 10.9 new cases per 100,000 population, which is below the target set by the World Health Organization (WHO).This shown, Malaysia is one of the top 10 highest cases of HIV and Aids .

(infection ways) - In 2013, heterosexual transmission recorded the highest rate (51%), followed by Injectable Drug Users (22%) and Homo/Bisexual transmission (22%). According to the Ministry of Health, in 2016, 70% of HIV infections were spread through heterosexual contact.

HIV/AIDS di Malaysia - Wikipedia Bahasa Melayu, ensiklopedia bebas

2.0 Objective

1.0 To data collection	1.1.finding the appropriate secondary data for the project. 1.2.learn and understand the nature of the data. 1.3.determine the data analysis suitable for the data .
2.0 To apply suitable method	2.1 .perform data preprocessing.
3.0 Analyze the data	3.1.learn the trend of the data. 3.2.find the relation between the data. 3.3.generate the hypothesis based on the result.
4.0 To apply suitable statical method	4.1 perform statistical analysis of the data using R software 4.2Create Data summarization
5.0 To interpret the results	5.1determining the conclusions, significance, and implications of the result.

3.0 Methodology

Steps of Data Analysis

The purpose of research is to identify the seriousness of HIV and AIDS infected cases in Malaysia between genders using sample data from 2001 until 2018.

We collected the secondary data from online sites of the Ministry of Women , Family and Community Development which were provided in kaggle.

The importance of the data has been collected to analyze the increment or decrement of infected cases of HIV and AIDS through indicated years.

4.0 Data chosen

- Dataset is obtained from the Ministry of Women , Family and Community Development based on cases in Malaysia from 2001 to 2018.
- The dataset has 18 years of observations towards different genders.
- The dataset also contains the total of infected people and percentages of women infected.
- Attached links of secondary data : Kes Jangkitan HIV dan AIDS yang dilaporkan Mengikut Jantina Dan Peratus Wanita - Set Data - MAMPU

	A	B	C	D	E
1		Kes Jangkitan HIV dan Aids yang Dilaporkan 2001			
2					
3	Tahun	Lelaki	Perempuan	Jumlah	Peratusan Perempuan
4	2001	5472	466	5938	7.85
5	2002	6349	629	6978	9.01
6	2003	6083	673	6756	9.96
7	2004	5731	696	6427	10.83
8	2005	5383	737	6120	12.04
9	2006	4955	875	5830	15.01
10	2007	3804	745	4549	16.38
11	2008	2988	704	3692	19.07
12	2009	2527	553	3080	17.95
13	2010	2984	668	3652	18.29
14	2011	2744	735	3479	21.13
15	2012	2704	734	3438	21.35
16	2013	2665	729	3394	21.48
17	2014	2820	697	3517	19.82
18	2015	2818	512	3330	15.38
19	2016	2984	413	3397	12.16
20	2017	2890	457	3347	13.65
21	2018	2894	399	3293	12.12
22					

5.0 Data Analysis

5.1 Hypothesis Testing - 1 Sample Test

a) Hypothesis testing is an act of assumption regarding a population parameter. Two hypotheses will be made to be justified and define the null hypothesis and alternative hypothesis. Then proceed to find out the level of significance. Lastly find the z test value using the formula provided.

Given by:

- H0 (null hypothesis)
- Alternate Hypothesis (Ha)

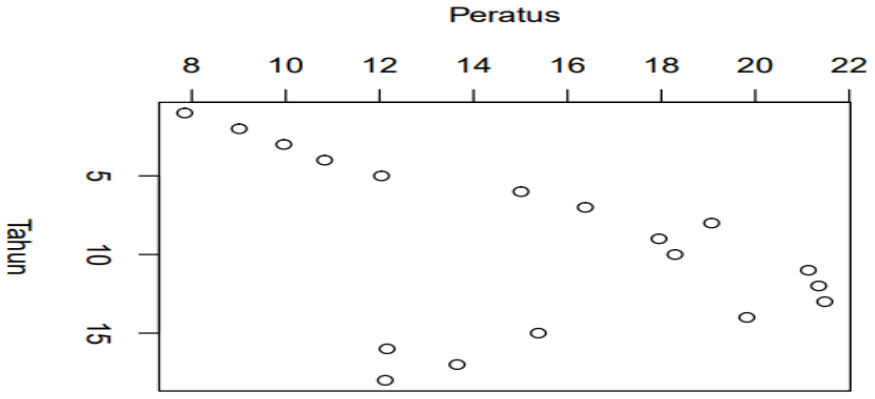
$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

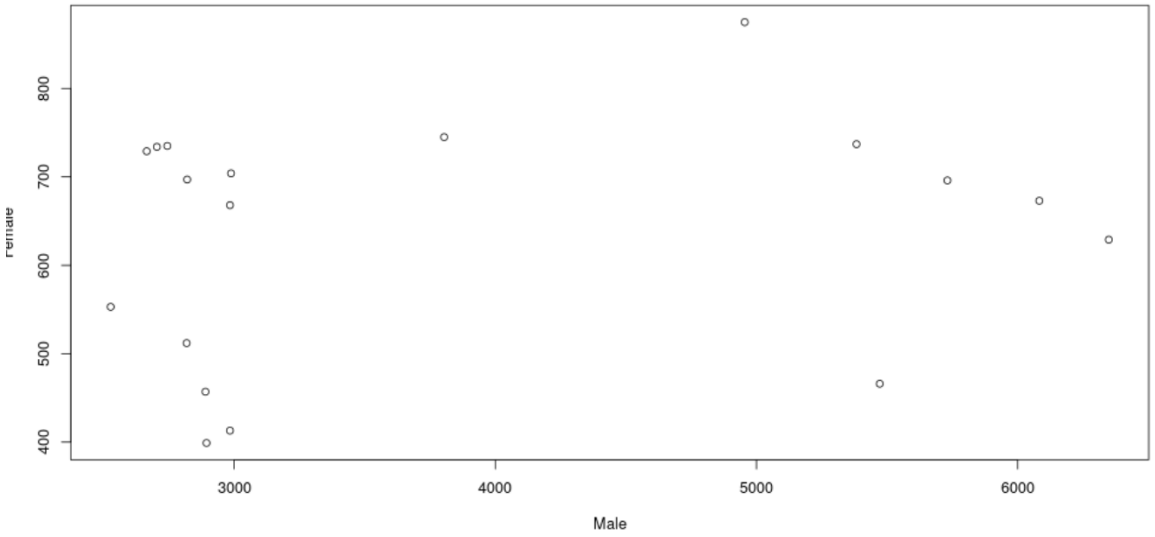
$\hat{p}$ = point estimate of population proportion (sample proportion)
p = population proportion (claimed value)
q = 1 – p
n = sample size

There are many factors that can be the spread of the HIV and AIDS viruses from sharing the same equipment, blood transferring or unprotected intercourse. A survey of a representative sample of 80217 cases among females and males reported that males surveyed had higher cases of infection in each year from 2001 to 2018. However, the result from opposite gender, females, has lower cases than expected. Figure 1 shows the percentage of infected females reported throughout the years.

H0 : p = 0.89

H1 : p ≠ 0.89






```
> #One sample hypothesis testing for population proportion
> Male <- Male
> n = 80217     #sample size
> alpha = 0.05  # significance level
> p = 0.89      #claimed population proportion
> q = 1-p
>
> Male=c(5472,6349,6083,5731,5383,4955,3804,2988,2527,2984,2744,2704,2665,2820,2818,2984,2890,2894)
> Female=c(466,629,673,696,737,875,745,704,553,668,735,734,729,697,512,413,457,399)
> data=data.frame(Male,Female)
> View(data)
> names(data)
[1] "Male"   "Female"
> plot(Male,Female)
>
> #Display male infected data that have total more than female infected
> subset(Male,Male>=Female)
 [1] 5472 6349 6083 5731 5383 4955 3804 2988 2527 2984 2744 2704 2665 2820 2818 2984 2890 2894
>
> #calculate and display the frequency of infected Male than greater than amount of infected female
> k = sum(table(subset(Male,Male>=Female)))
> k
[1] 18
>
> #Calculate and display the point estimate of infected population (sample propotion)
> phat = k/n
> phat
[1] 0.0002243913
>
> #calculate and display z statistics
> z = (phat-p)/sqrt((p*q)/n)
> z
[1] -805.4199
>
> #Calculate critical value
> z.alpha = qnorm(1-alpha/2)
> z.alpha
[1] 1.959964
>
> #Display critical value for two-tailed test
> c(-z,alpha,z.alpha)
[1] 805.419905   0.050000   1.959964
>
```

Figure 5.1

The included code, Figure 5.1 is for the graph . In conclusion, the proportion of infected male and female each year of observation shows that from 2001 to 2018, the population of Male infected cases of HIV and AIDs are more than the Female population. As the point estimate of infected population in sample proportion is 0.0002243913. Within Z statistics achieved are -805.4199, critical value = 1.959964, and critical value of two tailed tests are 805.419905.

5.2 Correlation

Correlation analysis used to measure strength of association between two variables. As we are concerned that the amount of male infected may have the same effect towards the female infected reported cases, since the risk of chances can be related in between both genders.

Pearson's Product-Moment Correlation Coefficient

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

where:

r = Sample correlation coefficient

n = Sample size

x = Value of the independent variable

y = Value of the dependent variable

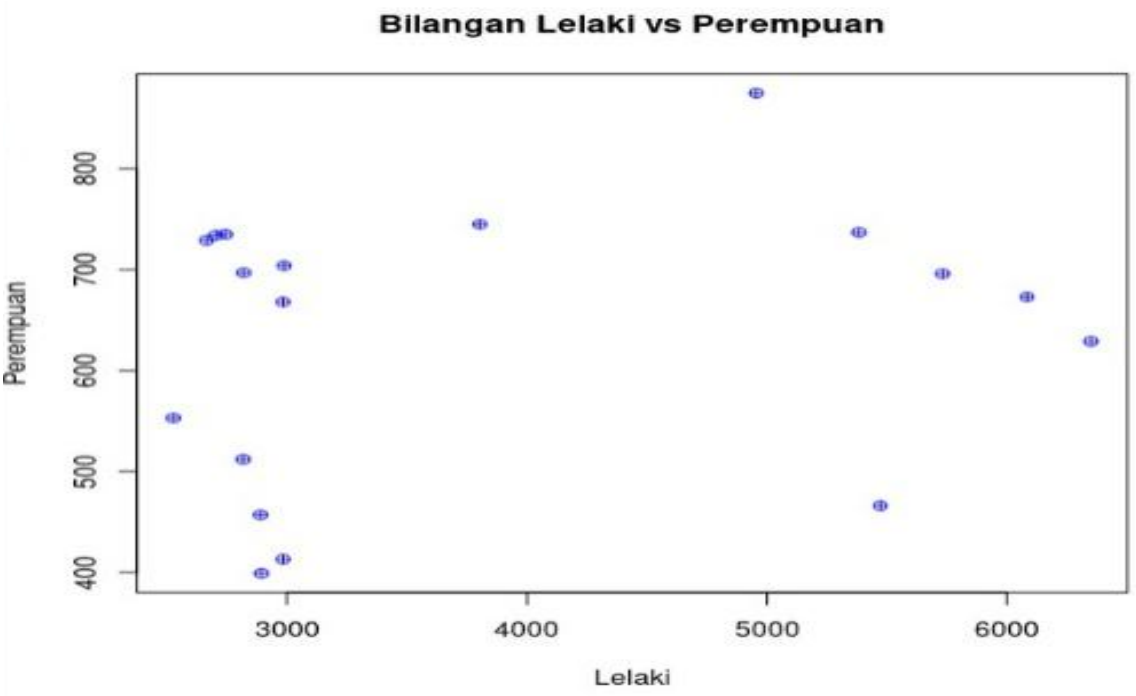
Based on code included below by using Pearson's product-moment correlation, we calculated the correlation correlation, r to know the strength of linear relationship between them based on formula for the test statistic.

```
> library(psych)
> Data <-read.csv("HIV Cases.csv")
> names(Data)
[1] "Tahun"               "Lelaki"              "Perempuan"
[4] "Jumlah"              "Peratusan.Perempuan" "X"
[7] "X.1"
> describe(Data$Perempuan)
   vars  n   mean     sd median trimmed   mad min max range  skew
X1    1 18 634.56 135.19  684.5  634.25 80.06 399 875   476 -0.37
   kurtosis    se
X1    -1.09 31.87
> describe(Data$Lelaki)
   vars  n    mean     sd median trimmed    mad  min  max range skew
X1    1 18 3821.94 1390.2   2984 3744.94 444.04 2527 6349  3822 0.68
   kurtosis     se
X1     -1.4 327.67
> cor.test(Data$Lelaki , Data$Perempuan)

        Pearson's product-moment correlation

data:  Data$Lelaki and Data$Perempuan
t = 0.84927, df = 16, p-value = 0.4083
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2870116  0.6149323
sample estimates:
      cor
0.2076881
```

The plot for correlation below shows that scatterplot has weak positive correlation . As we can see, the relationship between infected females does not depend on the infected male cases reported.

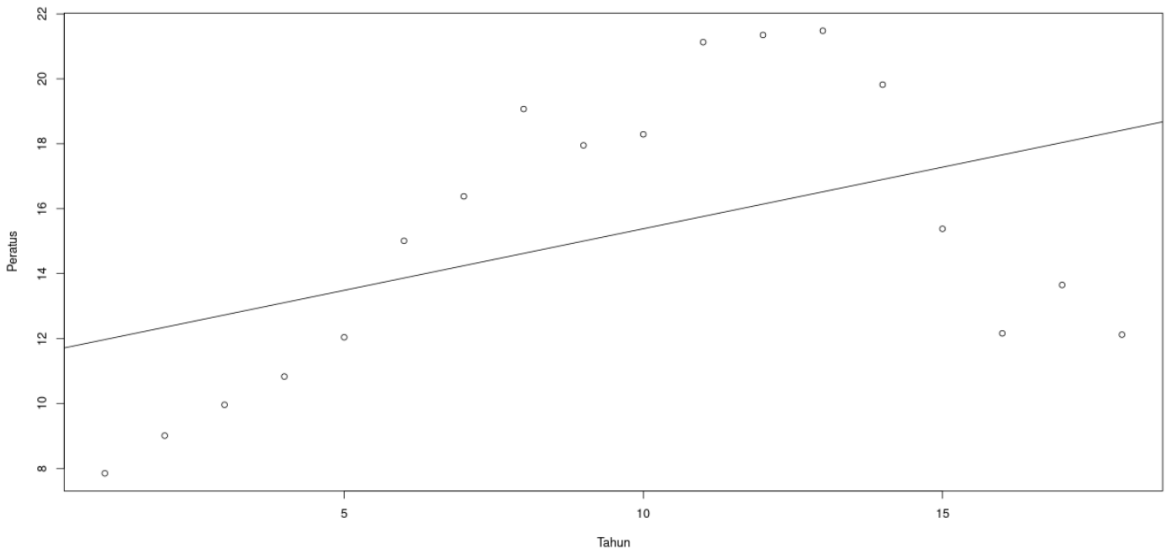


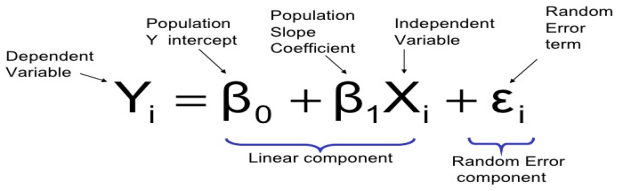
5.3 Regression

Regression analysis is used to predict the value of a dependent variable based on the values of at least one independent variable and it is also to explain impact of changes in an independent variable on the dependent variables.

Regression analysis is primarily used to find equations that fit the data. Linear analysis is a type of regression analysis. The linear equation is y = a + bX. Y is the dependent variable of the expression that is trying to predict what the future value will be if the independent variable X changes by a certain value. The "a" in the formula is the intercept, which is a value that remains constant regardless of changes in the independent variable, and the "b" term in the formula is the gradient that shows how different the dependent variable is from the independent variable is.

```
Console  Terminal  Jobs
R 4.2.0 · /cloud/project/
> Tahun=1:18
> Peratus=c(7.85,9.01,9.96,10.83,12.04,15.01,16.38,19.07,17.95,18.29,21.13,21.3
5,21.48,19.82,15.38,12.16,13.65,12.12)
> data=data.frame(Tahun,Peratus)
> View(data)
> names(data)
[1] "Tahun"   "Peratus"
> plot(Tahun,Peratus)
> model1=lm(Peratus~Tahun)
> abline(model1)
>
```





The figure of the scatter plot shows that the value of coefficient of determination, indicates a weak positive relationship.

5.4 Goodness Fit Test

A Goodness-Fit test is used to describe how well it matches a set of data. Goodness of fit is a measure of how well the data match the predicted values in a model.These measures can be used for statistical hypothesis testing, such as testing the normality of residuals, testing whether two samples come from the same distribution, or whether the frequency of results follows a specified distribution. In the analysis of variance, one of the factors into which the variance is partitioned may be the sum of squares indicating a lack of fit.

The Pearson chi-square test measures the goodness of fit of a data set by calculating the sum of differences between observed and expected outcome frequencies, each squared and divided by the expectation.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i = an observed count for bin i

E_i = an expected count for bin i, asserted by the null hypothesis.

The expected frequency :

$$E_i = \left(F(Y_u) - F(Y_l) \right) N$$

F = the cumulative distribution function for the probability distribution being tested.

Y_u = the upper limit for class i,

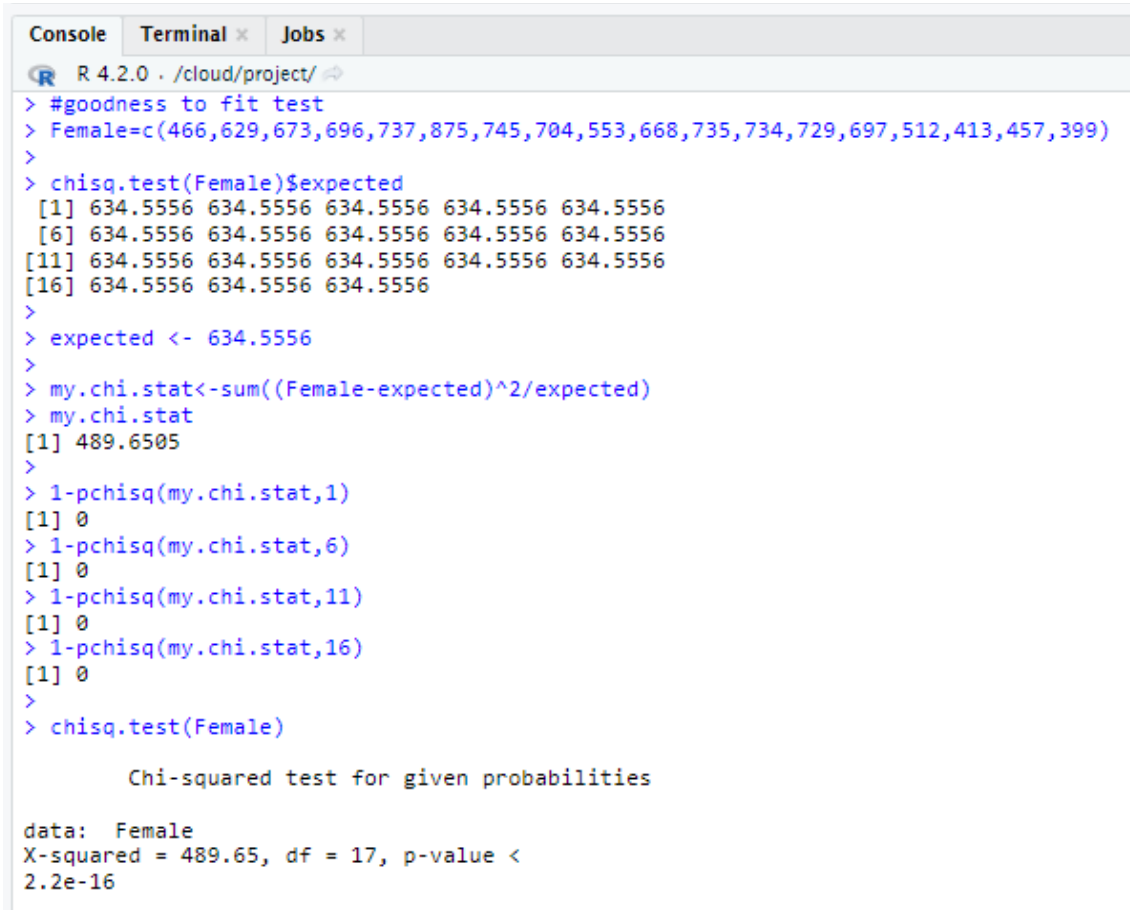
Y_l = the lower limit for class i, and

N = the sample size

The chi-square distribution can be used to determine the goodness of fit of the resulting value. The chi-square distribution has a number of degrees of freedom, depending on the number of non-empty cells in the data set and the number of estimated parameters

H0 : The female population are not fair

H1 : The female population are fair



In statistics for goodness-fit test, we set the significance level,a = 0.05. we calculated a test statistic that is equal to 489.6505. we find the theoretical value from the chi square distribution based on our significance level. The theoretical value is the value we would expect if the female has the same number of infected for each year. For short : Every year the number of females infected has the same amount of cases}. For degree of freedom : we have 18 years, so 18 - 1 = 17 degrees of freedom. The chi square value with a=0.05 and 17 degrees of freedom is 2.2E^-16. We compare the value of our test statistic(489.6505) to the chi square value. since 489.6506 > 2.2E^-16. We reject the null hypothesis that the proportions of females are equal.We set the significance level for the goodness-of-fit test at a = 0.05 in statistics. A test statistic equal to 489.6505 was calculated. Based on our significance threshold, we extract the theoretical value from the chi square distribution. The theoretical value is what we might anticipate if a female contracts the same number of infections every year. In brief: The number of infected females has an equal number of instances every year. We have 18 years, thus 18 - 1 Equals 17 degrees of freedom in terms of the degree of freedom. With a=0.05 and 17 degrees of freedom, the chi square value is 2.2E-16. We contrast the chi square value with the value of our test statistic (489.6505). since 2.2E-16 > 489.6506.We reject the null hypothesis that the proportions of females are equal.

5.5 Chi Square test of Independence

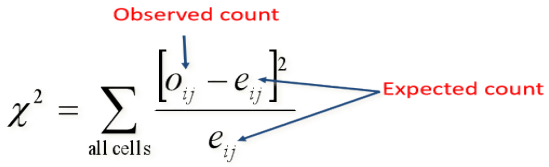
The chi-square (χ 2) test of independence is used to test for a relationship between two categorical variables that influence gender on disease at the significant level of 0.05 with sample size of 18 years . Recall that if two categorical variables are independent, then P (A) = P (A | B).

- Test hypothesis:

H0 : variables are independent that gender not the influence of infection

H1 : variables are related (dependent) that gender can be the influence of infection

- Test Statistic:



- Degrees of Freedom: (r - 1)(c - 1) | r = row , c = column
- Find critical value

$\alpha = 0.05$, df = (18 - 1) (2 - 1) = 17

R 4.2.0 . /cloud/project/

```
> #chisquare test  of independence (two way contingency table)
> library(MASS)
> #Get the two way contingency table
> tbl <- table(data$Male,data$Female)
> tbl

      399 413 457 466 512 553 629 668 673 696 697 704 729 734 735 737 745 875
  2527   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
  2665   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0
  2704   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0
  2744   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0
  2818   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
  2820   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0
  2890   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  2894   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  2984   0   1   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
  2988   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0
  3804   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0
  4955   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1
  5383   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0
  5472   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  5731   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0
  6083   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
  6349   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0
>
> #Perform chi-square test on the data table
> result <- chisq.test(tbl, correct = FALSE)
Warning message:
In chisq.test(tbl, correct = FALSE) :
  Chi-squared approximation may be incorrect
> result

        Pearson's Chi-squared test

data:  tbl
X-squared = 288, df = 272, p-value = 0.2415

>
> #expected frequency
> result$expected
```

```
> #expected frequency
> result$expected

           399       413       457       466       512       553       629       668       673       696       697       704       729       734       735
2527 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2665 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2704 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2744 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2818 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2820 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2890 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2894 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2984 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111
2988 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
3804 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
4955 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
5383 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
5472 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
5731 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
6083 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
6349 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556

           737       745       875
2527 0.05555556 0.05555556 0.05555556
2665 0.05555556 0.05555556 0.05555556
2704 0.05555556 0.05555556 0.05555556
2744 0.05555556 0.05555556 0.05555556
2818 0.05555556 0.05555556 0.05555556
2820 0.05555556 0.05555556 0.05555556
2890 0.05555556 0.05555556 0.05555556
2894 0.05555556 0.05555556 0.05555556
2984 0.11111111 0.11111111 0.11111111
2988 0.05555556 0.05555556 0.05555556
3804 0.05555556 0.05555556 0.05555556
4955 0.05555556 0.05555556 0.05555556
5383 0.05555556 0.05555556 0.05555556
5472 0.05555556 0.05555556 0.05555556
5731 0.05555556 0.05555556 0.05555556
6083 0.05555556 0.05555556 0.05555556
6349 0.05555556 0.05555556 0.05555556
>
> #critical value
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=1,lower.tail = FALSE)
> x2.alpha
[1] 3.841459
>
```

Overall, the chi-square test of independence is to check whether the female infected is related by dependent or independent towards the male infected data. We use the formula of Pearson Chi-Square test as a solution. Based on r-programming, we received the x-squared = 288 with degrees of freedom = 272, while p-value = 0.2415. For the expected frequency, the male and female results are equal to 0.55556 and 0.11111. The critical value with alpha = 0.05,df =1, we get x2-alpha = 841459.

6.0 Work Coordination

NAME	TASKS
FARAH NABILAH BINTI NAJMUDIN (A21EC0023)	**R-programming** **Explanations of test statistics** **The output graph and plot for chi-square, goodness-fit test.** **Reflection** **Details of data chosen** **Appendices sample group photos** **Methodology** **Host video presenter**
MAATHUREE A/P VEERABALAN (A21EC0051)	**Conclusion** **Slides for Presentation** **Explanations for each test statistics** **Reflection** **Appendices** **Hypothesis Testing 1 Sample**
MUHAMMAD SAIFUDDIN BIN ISMAIL (A21EC0093)	**Explanations for each test statistics** **Include formulas needed** **Conclusion** **Slide explanation** **Slide editing** **Objective**
NUR SYAFIKA BINTI MOHD SALMIZI (A21EC0115)	**MANAGED TO CREATE THE CODING AND GOT THE OUTPUT FOR CORRELATION,REGRESSION,HYPOTHESIS TESTING 1 SAMPLE** **Introduction**

7.0 Reflections

From the group's view, we gain a lot of knowledge from this project. Overall, we got to work in a team to study this case data which was collected for 18 years. This project is a new experience for us to understand more about r- programming. Extraction of data was a very important thing in this project where we used various methods such as regression, correlation, hypothesis testing, goodness fitness test and chi square test. As a team we managed to finish off this project with the knowledge that we had.

We would like to thank Dr Azurah, our course lecturer for providing us with enough knowledge in doing this project. Dr taught us from scratch what we have to do and how to do. Each and every knowledge that was provided by Dr Azurah made us come out with a full report. Dr Azurah had never failed to explain to us again even though Dr just explained to us in class. Examples that are used by Dr as a guidance to us make us understand more deeply about the project. At first we felt it was hard to do the report because our understanding towards this project was slightly different from the expectation and atlast we managed to sort all of it with help of Dr Azurah. However, we all learn a lot from the difficulties together as a team and also as we figure out the Rstudio application. We as a team would like to thank Dr Azurah again for giving us the opportunity to do this project and guiding us throughout this project.

We do gain a lot of knowledge from this project. We had to learn more deeply how to use R-Programming. It was hard at first and later on we managed to work it out. We did prefer R-programming than usual programming because it's easier to understand human language but first of all we should acknowledge the appropriate formulas to figure the correct r-programming. Apart from that we also learned the most understanding in this project is from learning by ourselves how to do the solution rather than understanding it by theoretical and slides given only.

8.0 Conclusion

In the nutshell,we learnt to extract data from various method such as regression, correlation, hypothesis testing, goodness fitness test and chi square test. Based on the result of one sample hypothesis testing by the population,we can testify the claim of that the percentage of females infected cases As the point estimate of infected population in sample proportion is 0.0002243913. the proportion of infected male and female each year of observation shows that from 2001 to 2018, the population of Male infected cases of HIV and AIDs are more than the Female population. Within Z statistics achieved are -805.4199, critical value = 1.959964, and critical value of two tailed tests are 805.419905.From the plot from correlation graf shows that scatterplot has weak positive correlation . As we can see, the relationship between infected females does not depend on the infected male cases reported.The figure of the scatter plot from the code in regression shows that the value of coefficient of determination, indicates a weak positive relationship.We set the significance level for the goodness-of-fit test at a = 0.05 in statistics. A test statistic equal to 489.6505 was calculated. Based on our significance threshold, we extract the theoretical value from the chi square distribution. The theoretical value is what we might anticipate if a female contracts the same number of infections every year. In brief: The number of infected females has an equal number of instances every year. We have 18 years, thus 18 - 1 Equals 17 degrees of freedom in terms of the degree of freedom. With a=0.05 and 17 degrees of freedom, the chi square value is 2.2E-16. We contrast the chi square value with the value of our test statistic (489.6505). since 2.2E-16 > 489.6506 We reject the null hypothesis that the proportions of females are equal .The relationship between two categorical variables that influence gender on disease at the significant level of 0.05 with sample size of 18 years in which the value are expected

We can say that the objectives of the project have been fufill.This is based on the result and the interpretation that we gain and the solution that been written above.The objectives also has been perfectly be lay out in the report.

We faced alot of challenges in doing this project. We were still new to use R-programming. We had difficulties in figuring out how to use this language. There were some situations where we couldnt extract the data. Apart from that, we also had some confussions in extracting the data in terms of how to extract and what are the methods to be used. In the end with the help of sources and guidance of Dr Azurah we managed to overcome this challenges.

In the future,we hope that our work will be much more efficient and faster which is a perfect and understandable project that can be use to bring a better future to human kind.We will try much harder in the future by managing our time so that we can have a lot more time to discuss the project to improve our project.Ultimately,achieve our dream of our project become a useful material towards the human society.

9.1 References

9.1.1 Sources of Data

https://www.data.gov.my/data/en_US/dataset/kes-jangkitan-hiv-dan-aids-yang-dilaporkan-mengikut-jantina-dan-peratus-wanita

9.1.2 List of articles and References Used

https://www.investopedia.com/terms/c/correlation.asp

https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/

https://www.youtube.com/watch?v=26hP6woQob0 - How To Perform a Chi-Square Test for Independence in R #92

https://www.youtube.com/watch?v=YbhqM0E0Pno - Correlation analysis using Rstudio

Chi-Square Test: Analysis & Interpretation I StudySmarter - Explanation for Chi-square test

Goodness-of-Fit Test – Introductory Statistics (opentextbc.ca) - Explanation for Goodness-fit test

https://www.youtube.com/watch?v=kvmSAXhX9Hs&t=91s - One-Sample t Test & Confidence Interval in R with Example | R Tutorial 4.1| MarinStatsLectures

9.3 List of e-portfolios

https://eportfolio.utm.my/user/maathuree-veerabalan - (MAATHUREE A/P VEERABALAN)

https://eportfolio.utm.my/blocktype/wall/wall.php?id=1132773 - (NUR SYAFIKA BINTI MOHD SALMIZI)

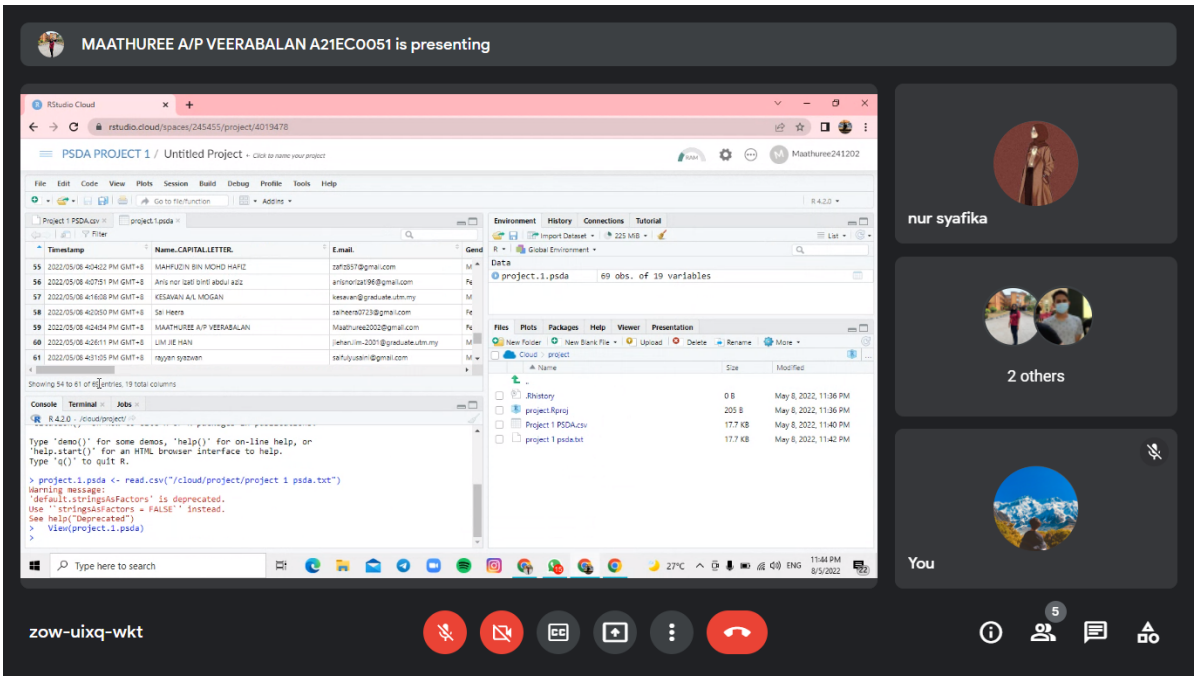
https://eportfolio.utm.my/user/muhammad-saifuddin-bin-ismail/project-2-psda-(MUHAMMAD SAIFUDDIN ISMAIL)

FARAH NABILAH BINTI NAJMUDIN - MyePortfolio@UTM - (FARAH NABILAH BINTI NAJMUDIN)

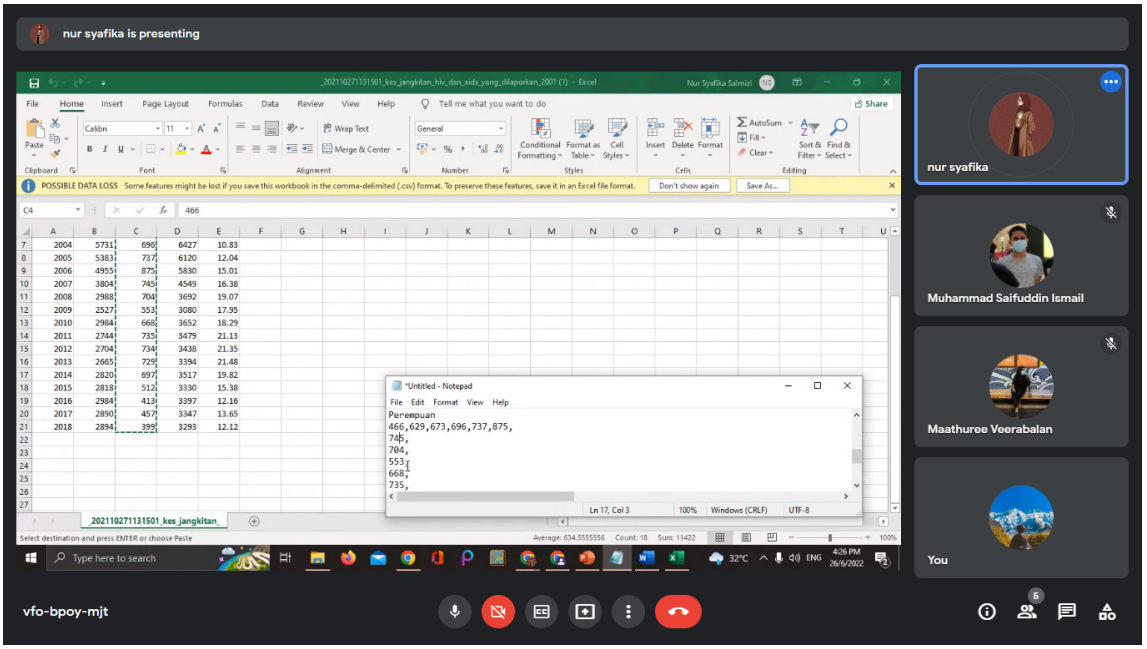
9.4 Link Video Presentation and Slides

https://drive.google.com/drive/folders/1EbdT7scJKykDnlrqLlTuSOTBa_gQrox4?usp=sharing

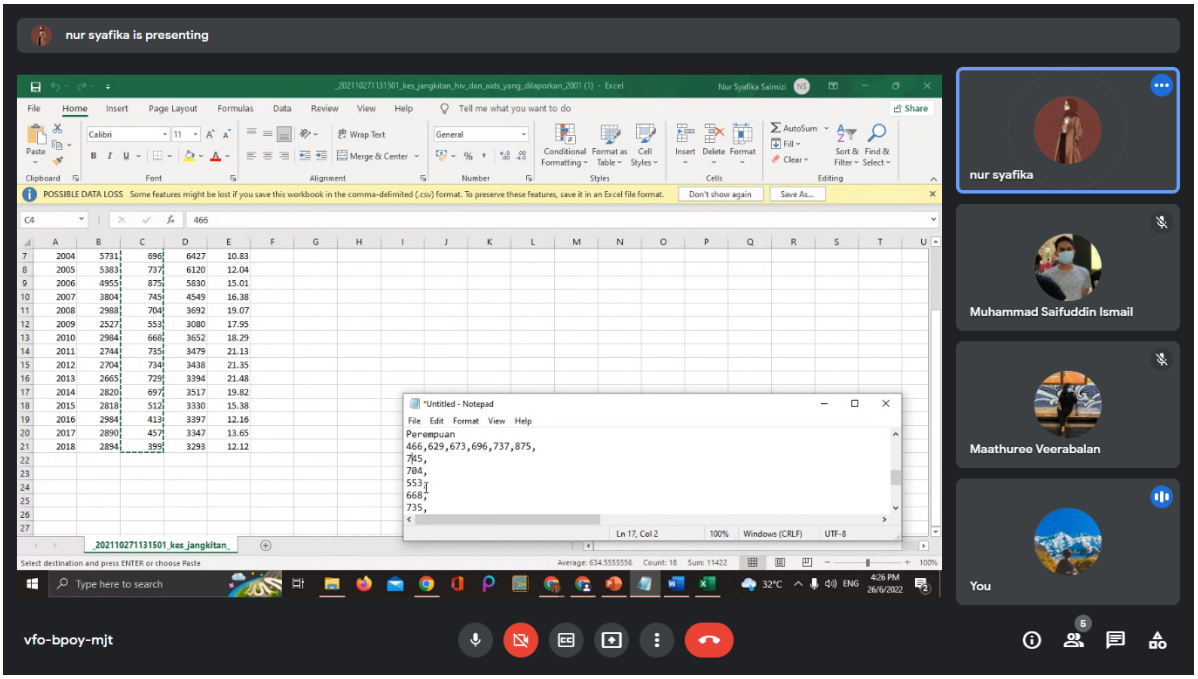
10.0 Appendices

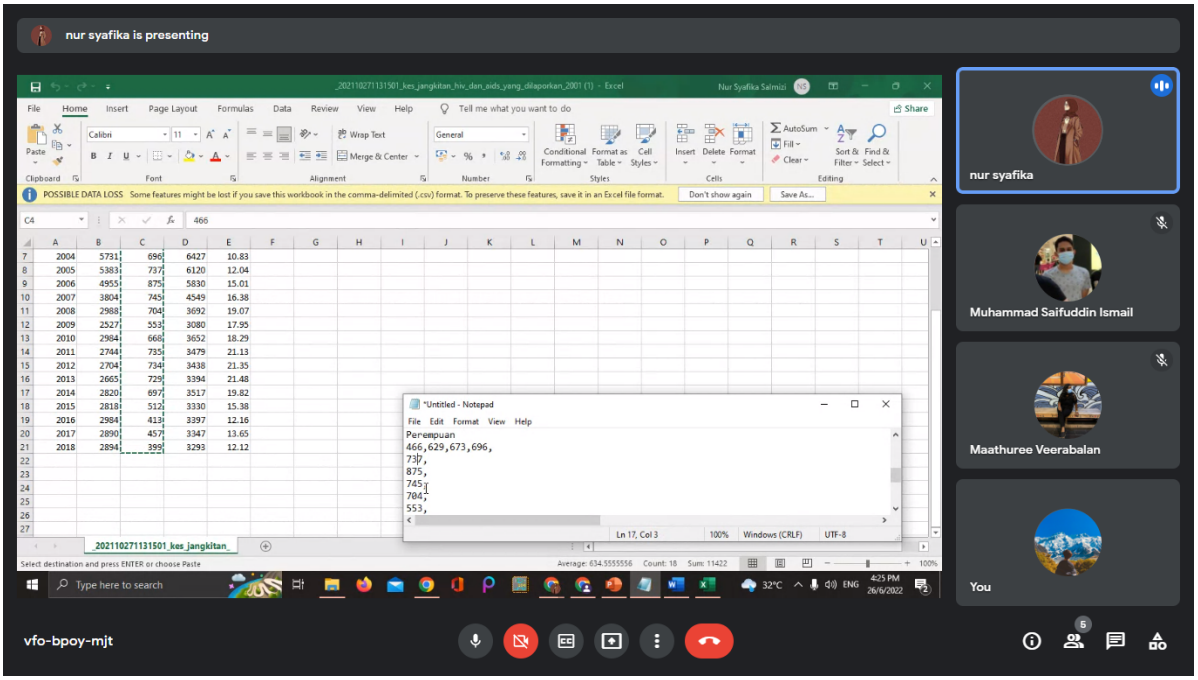


8/5/2022 - Figuring how to use Rstudio cloud together. Lots of arguments just how to open the file in Rstudio then try to figure out the codes.

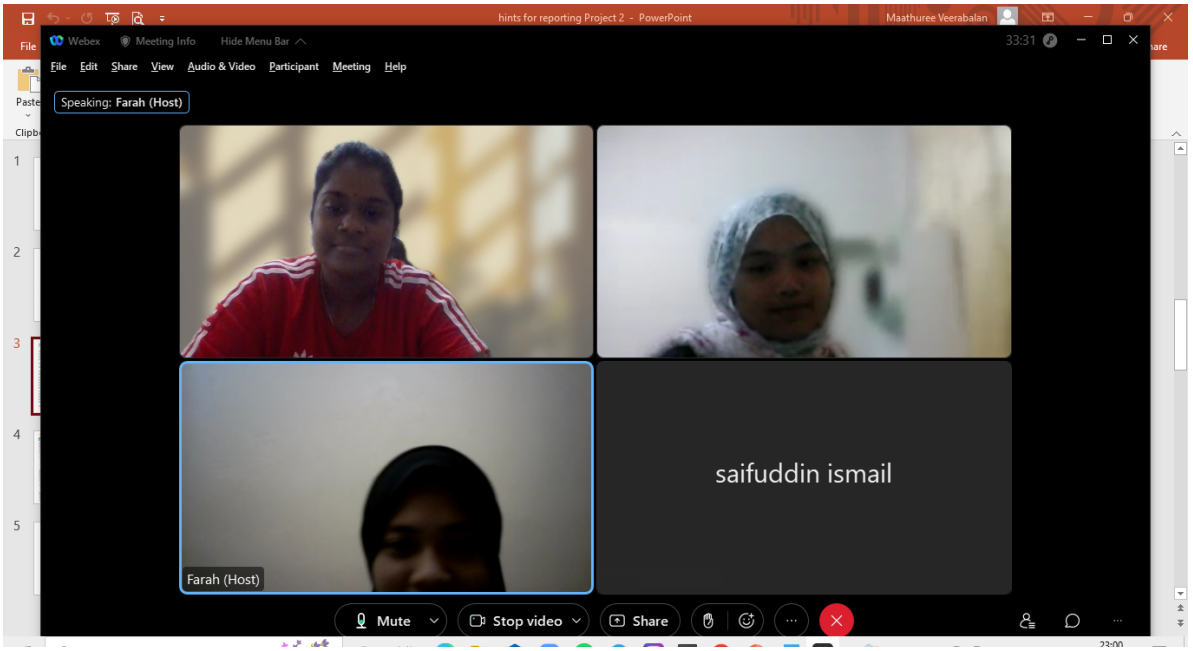


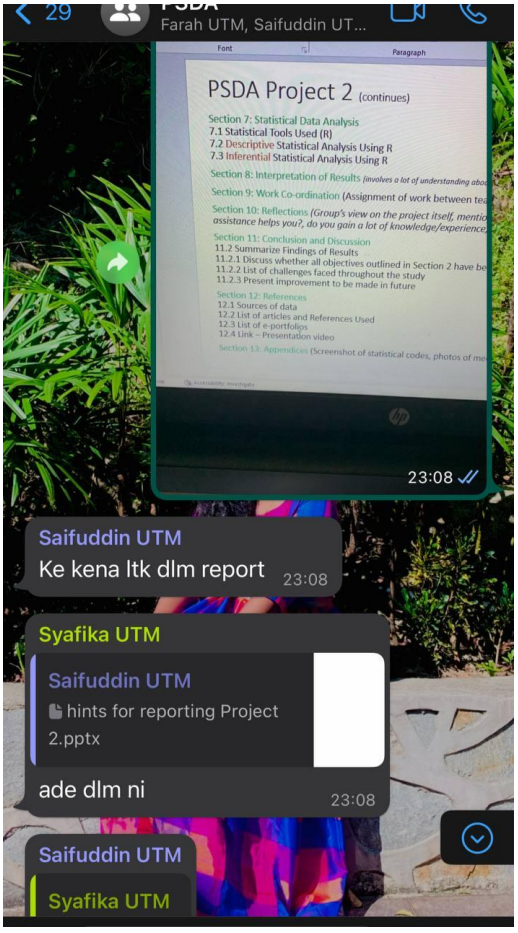
26/6/2022 - All analyzing the data together after finalizing the data chosen.

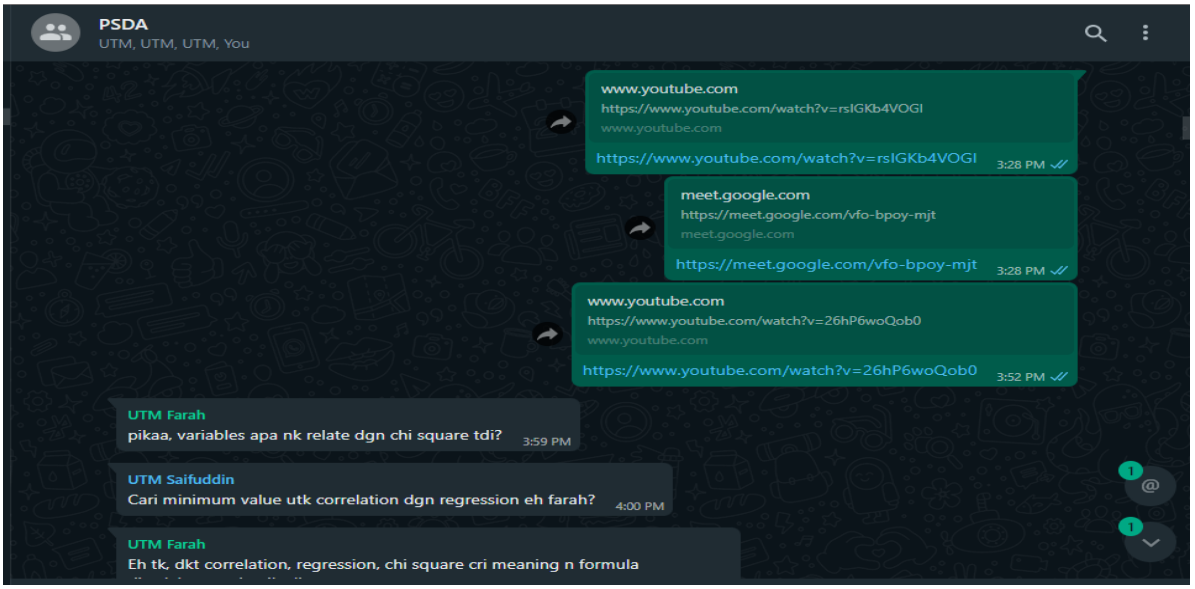

nur syafika is presenting
nur syafika
Muhammad Saifuddin Ismail
Maathuree Veerabalan
You
vfo-bpoy-mjt

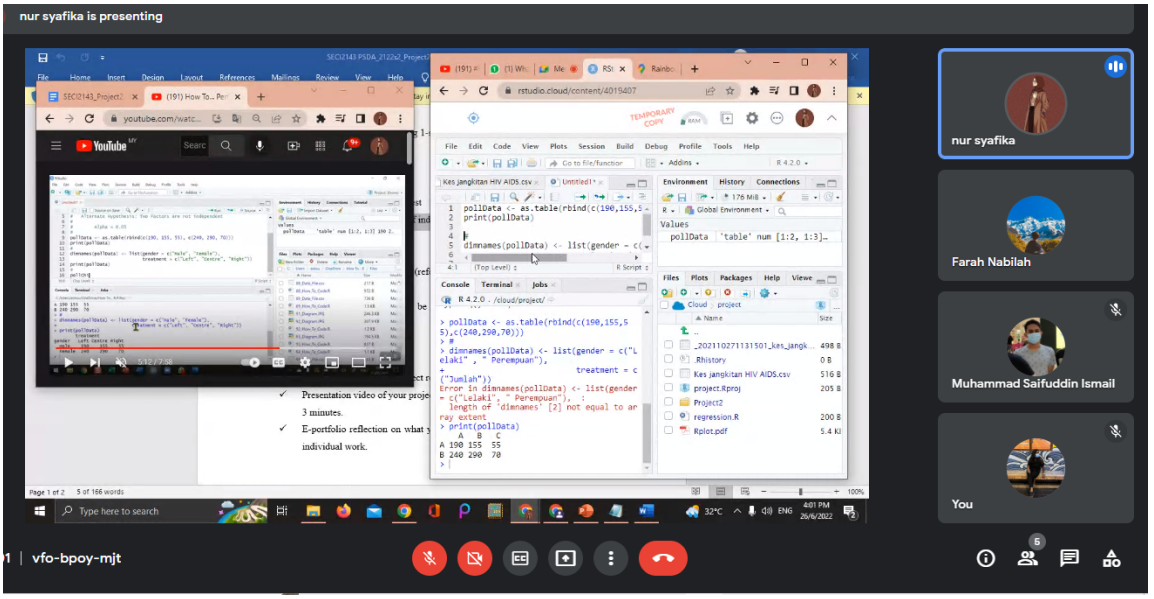

nur syafika is presenting
nur syafika
Muhammad Saifuddin Ismail
Maathuree Veerabalan
You
vfo-bpoy-mjt

Day of video presentations:









The full day of watching youtube learning and understanding rstudio.