



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI2143: PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2 PART 1

Section 01

Group 2

NO.	NAME	MATRIC NUMBER
1.	AFIQAH IZZATI BINTI AZZEROL EFFENDI	A21EC0004
2.	LEE RONG XIAN	A21EC0043
3.	LU QI YAN	A21EC0049
4.	NOOR HANNANI SYAMIMI BINTI MOHD SUFFIAN	A21EC0104

Table of Content

Section	Description of Contents	Page
1	Brief introduction about the project	1
2	Stating the objective and goals of project of the selected data	2
3	Listing the steps of the data analysis conducted	3 – 4
4	Describing the chosen dataset by listing the sample size	5
5	Summarizing and describing the data using descriptive statistics	6 – 7
6	Drawing conclusions about the population based on data observed in sample	8
7	Data Analysis using R	9
8	Describing and interpreting results based on findings	10 – 16
9	Assigned task of each member	17
10	Reflection and knowledge gained from the project	18
11	Conclude and summarize results	19 – 20
12	References of data	21
13	Proof and evidences of the process of the project	22 - 25

Section 1 Introduction

The dataset that we have chosen for our project is “Go To College Dataset”. This dataset contains data from 1000 sample data sets. However, for this project we only use 154 sample data sets. We only use the data from Academic school and parents age within 40 to 50 years old. Data was gathered to predict whether students will continue to go to college or not. It was collected to help school counsellors to identify the factors and causes in order to help the students.

For the dataset, several demographic factors are included such as parents age, parents' salary, house area, average grade of students and type of school. Not only that, the dataset also includes behavioural risk factors associated with the level of interest of going to college in which whether the students are not interested, less interested, uncertain, quite interested, and very interested.

Moreover, it can also help to identify the future population and ensure they continue their studies by going to college. Hence, it will actually be the stepping stone for the poor family to be successful, by continuing their education in college. For example, one can change his or her family's life by going to college with a loan, scholarship and many more. Therefore, it is vital to be able to predict the level of interest and associated factors, in order to help them change their life by helping them.

Section 2 Scope and Objective of project

Data analysis

The dataset was collected by Saddam Sinatrya Jalu Mukti on Kaggle, it might be regarded as secondary data. Based on input characteristics like, level of interested, parent's age, parent's salary and house area, this dataset is used to help school counsellors to identify the factors and causes in order to help the students to go to college or not.

The total number of students in the dataset is 1000. There are 10 variables in the original dataset, but we only utilise 5, including level of interest, parent's age, parent's salary, house area, average grade of students. We prefer to use ratio and ordinal types of variables because they make analysis easier.

The following five statistical tests will be run:

1. Hypothesis testing using variable of average of grades in scale
2. Correlation analysis using 2 variables of parent's age and parent's salary
3. Regression analysis using 2 variables of parent's salary and house area
4. Goodness of fit test using variable of level of interest

Section 3 Methodology

Steps of Data Analysis:

1. The first step that we will do is to understand the nature of the problem. This is in order to decide the goal of the project which is to help school counsellors to identify the factors and causes in order to help the students.
2. The second step is deciding what to measure and how to measure it. For this project, we measure:
 - The average(mean) of grades of the academic school student is equal to grades of 86.0 at 5% of significance level.
 - The linear relationship exists between the parent's age and the parent's salary using Pearson's Product-Moment Correlation Coefficient at 5% of significance level.
 - Whether the house area depends on the parent's salary, using house area as dependent variable x and parents' salary as independent variable y.
 - Whether the level of interest is the same as the given percentage.
3. The third step is data collection. For this project, we are not collecting data but instead we are using a secondary data that we retrieved from Kaggle.com. The dataset that we found is about "Go To College".
4. The fourth step is data summarization and preliminary analysis. All the data that is collected is then summarized and present in graphical formats as shown in Section 5.
5. The fifth step is formal data analysis. In this step, we apply the appropriate inferential statistics methods by implementing (Hypothesis Test 1 Sample, Correlation Analysis, Regression, and Goodness of Fit Test) as shown in Section 6. Analysis of the data is conducted using R programming in RStudio.

6. The sixth step which is also the last step is the interpretation of results. The conclusion of the study is conducted from all of the data and information collected through the analysis process.

Section 4

Dataset used

4.1 resources of the dataset

[Go To College Dataset | Kaggle](#)

4.2 Sample size

	A	B	C	D	E	F
1	Type of School	Level of Interested	Parent's Age	Parent's Salary	House Area	Average Grade of Students
2	Academic	Not Interested	41	3020000	50.6	77.18
3	Academic	Not Interested	44	7790000	85.3	85.73
4	Academic	Not Interested	46	7200000	76	92.99
5	Academic	Not Interested	46	4810000	76.5	94.28
6	Academic	Not Interested	48	6090000	74.4	95.81
7	Academic	Not Interested	48	3320000	72.8	83.63
8	Academic	Not Interested	49	6220000	84	82.48
9	Academic	Not Interested	49	7570000	62.3	89.95
10	Academic	Not Interested	49	2600000	100.8	88.86
11	Academic	Not Interested	49	6340000	77.3	92.68
12	Academic	Not Interested	49	5340000	80.4	94.06
13	Academic	Not Interested	50	5630000	69.9	80.5
14	Academic	Not Interested	50	8290000	67.4	89.16
15	Academic	Not Interested	50	4910000	76.2	93.01
16	Academic	Not Interested	50	6050000	106.1	84.28
17	Academic	Not Interested	50	6050000	107.5	85.46
18	Academic	Not Interested	50	4920000	113.1	82.93
19	Academic	Less Interested	40	3810000	44.7	81
20	Academic	Less Interested	44	5410000	36.6	83.48
21	Academic	Less Interested	44	3280000	62.8	82.58
22	Academic	Less Interested	46	2730000	62.4	83.08
23	Academic	Less Interested	46	4980000	47.8	84.72

https://drive.google.com/file/d/1vEFqb_PY7boYJ_fv1lMqhgIL-8yfBItg/view?usp=sharing

Section 5

Descriptive statistics

The dataset that we have chosen for our project is “Go to College Dataset”. This dataset contains data from 1000 sample data sets. However, for this project we only use 154 sample data sets. We use the data from Academic school and parents age within 40 to 50 years old.

Processed Data	
Variables	Data type
Level of interest	Ordinal
Parent’s age	Ratio
Parent’s salary	Ratio
House area	Ratio
Average grade of students	Ratio

Selected Variable(s)	Test	Description
Average of grades in scale of 0-100	Hypothesis testing (one sample test)	Explanation: The variable is used to test whether the average(mean) of grades of the academic school student is equal to grades of 86.0 at 5% of significance level.
Parent’s age and parent’s salary	Correlation analysis	Explanation: The variables are selected to test whether the linear relationship exists between the parent’s age and the parent’s salary using Pearson’s Product-Moment Correlation Coefficient at 5% of significance level.

Parent's salary and house area	Regression analysis	Explanation: The variables are selected to test whether the house area depends on the parent's salary, using house area as dependent variable x and parents' salary as independent variable y.
Level of interest	Goodness of fit test	Explanation: To test whether the level of interest is the same as the given percentage.

Section 6

Inferential statistics

Selected Variable(s)	Test	Description about Test
Average of grades in scale of 0-100	Hypothesis testing (one sample test)	H₀: $\mu = 86.0$ H₁: $\mu > 86.0$
Parent's age and parent's salary	Correlation analysis	H₀: $\rho = 0$ (no linear correlation) H₁: $\rho \neq 0$ (linear correlation exists)
Parent's salary(x) and house area(y)	Regression analysis	Depend on the coefficient of determination, $R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y_i - \bar{y})^2}$ If $R^2 = 1$, it indicates there is a perfect linear relationship between x and y. If $0 < R^2 < 1$, it indicates there is a weak relationship between x and y. <u>If $R^2 = 0$, it indicates there is no linear relationship between x and y.</u>
Level of interest	Goodness of fit test	H₀: $p_1=0.11, p_2=0.21, p_3=0.29, p_4=0.05, p_5=0.34$ H₁: At least one of the proportions is different from the claimed value.

Section 7

Statistical Data Analysis

7.1 Hypothesis Testing One Sample

<https://drive.google.com/file/d/1f37TIGdVqkiyxIMU1lyMY05ljfPF59bX/view?usp=sharing>

7.2 Correlation Analysis

https://drive.google.com/file/d/19XnECXeVxDXKY7_sKsW4Jfxwn1bc2cW3/view?usp=sharing

7.3 Regression Analysis

https://drive.google.com/file/d/12jnVN8nnrrzcM_0rMotTtsn_GKF0nnV_/view?usp=sharing

7.4 Goodness of Fit Test

<https://drive.google.com/file/d/1BRuJaXEY7s3tNCnkJmairS21lLdAq8mf/view?usp=sharing>

Section 8

Interpretation of Result

8.1 Hypothesis Testing – 1 sample

Based on the test, we wish to determine whether the mean of average grade in academic school is greater than 86.0.

Let μ = mean of average grade in academic school

$$H_0: \mu = 86.0$$

$$H_1: \mu > 86.0$$

Significance level, $\alpha = 0.05$

Sample size, $n = 154$

$$Z_{0.05} = 1.645$$

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Reject H_0 if $Z > Z_{0.05} = 1.645$

Computations: Since $\bar{x} = 86.1741$, $\sigma = 4.0020$, $\mu_0 = 86.0$ and $n = 154$, we have

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{86.1741 - 86.0}{\frac{4.0020}{\sqrt{154}}} = 0.5398$$

Conclusion: Since the size of sample is relatively large and variance is unknown, we assume that the sample is normally distributed. The test statistic we applied is Z-test. We obtained Z-value for test statistic using R. It is one-tailed test, the rejection of test statistic value is located on the right side of $Z = 0.5398$. Since test statistic value ($Z = 0.5398$) $< Z_{0.05} = 1.645$, we fail to reject the null hypothesis, H_0 . Thus, there is insufficient evidence to reject the claim that the mean of average grade in academic school is 86.0.

8.2 Correlation Analysis

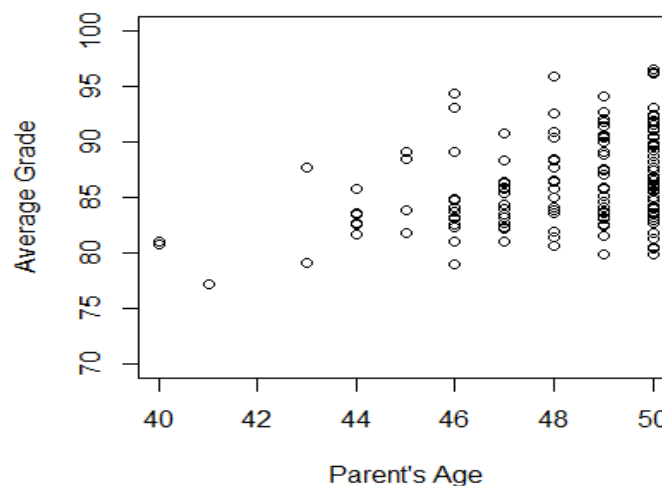
In this correlation analysis, the variables that we used are parent's age of the students and average grade of academic school. We will test whether there is a linear relationship between parent's age of the students and average grade of academic school using the significance level, $\alpha=0.05$.

H₀: $\rho = 0$ (no linear correlation)

H₁: $\rho \neq 0$ (linear correlation exists)

By using Rstudio, we obtain correlation coefficient, $r = 0.3471$, $t = 56.2610$ and $t_{0.025, 152} = 1.9757$. Thus, H_0 will be rejected if $t < -1.9757$ or $t > 1.9757$. Otherwise, fail to reject H_0 .

Conclusion: Since $t = 56.2610 > 1.9757$, we reject the null hypothesis, H_0 . There is sufficient evidence to conclude that there is a linear relationship between parent's age of the students and average grade of academic school at 5% level of significance.



Graph 1: Average grade against parent's age

From the scatter plot graph, we can see that the points slope upward, which denotes that there is a positive relationship between parent's age and average grade, such that the higher the parent's age, the higher the average grade.

Since we got $r = 0.3471$, which indicates that there is a relatively weak positive linear relationship between parent's age of the students and average grade of academic school. Through the result of the scatter plot graph and the correlation test, it is clearly shown that the parent's age is correlated with the average grade pf academic school. They have a positive linear relationship, so when the parent's age increases, the average grade increases.

8.3 Regression Analysis

For the regression test, we used the parent's salary and house area. As we used only a single independent variable, this linear regression model. Through this test, we wanted to find out whether there is a linear relationship between the parent's salary and the house area. The dependent variable, which is denoted as y, is the house area, while the independent variable, x is the parent's salary. The following is the scatter plot of house area against the parent's salary.



Graph 2 House area against parent's salary

Based on the plot, we can see the line is nearly horizontal which indicates that there is an extremely weak and even no linear relationship between the independent variable and the dependent variable. Through analysis, we obtain the value of intersection coefficient (b_0) is $4.899e+01$ and the value of estimated change in the average house area (b_1) is $-4.111e-06$.

The estimated regression model is as below:

From R studio, the least-square regression line is

$$\hat{y} = 4.899e+01 - 4.111e-06 x$$

From R studio, the coefficient of determination,

$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 0.1441$$

The coefficient of determination, R^2 value which is nearly equal to zero shows that there are less than 0.14% of parent's salary related to their house area. Therefore, we can conclude that

house area is not dependent on the parent's salary as the dependent of the house area to parent's salary is nearly absent.

8.4 Goodness of Fit Test

It was claimed that students from an academic school rated the level of interest in going to school. Of 154 students, 11% rated "Not Interested", 21% rated "Less Interested", 29% rated "Uncertain", 5% rated "Quiet Interested" and 34% rated "Very Interested".

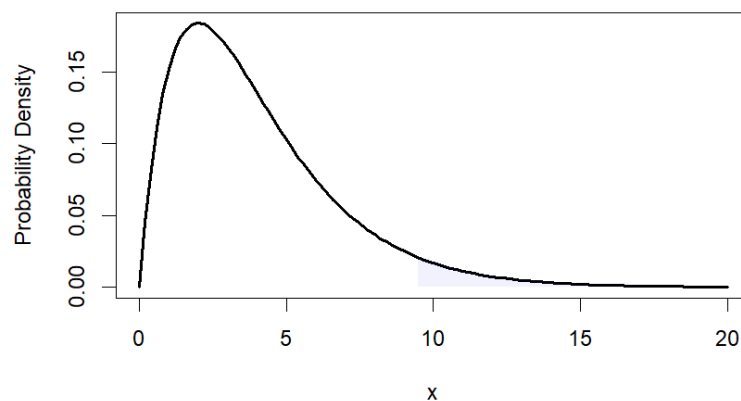
Let p_1 = proportion of "Not Interested", p_2 = proportion of "Less Interested", p_3 = proportion of "Uncertain", p_4 = proportion of "Quiet Interested" and p_5 = proportion of "Very Interested".

H₀: $p_1=0.11$, $p_2=0.21$, $p_3=0.29$, $p_4=0.05$, $p_5=0.34$

H₁: At least one of the proportions is different from the claimed value.

Level of Interested	Not Interested	Less Interested	Uncertain	Quiet Interested	Very Interested
Observed Frequency	17	33	44	8	52
Expected Frequency	154(0.11) =16.94	154(0.21) =32.34	154(0.29) =44.66	154(0.05) =7.70	154(0.34) =52.36

Chi-Square Distribution (df = 4, α = 0.05)



Graph 3 Chi-square Distribution

Test statistic:

By using R Studio, $\chi^2 = \sum \frac{(O-E)^2}{E} = 0.037599$

(Degree of freedom = $k-1 = 4$ and $\alpha = 0.05$)

Critical Value, $\chi^2_{4, 0.05} = 9.487729$

$\chi^2 = 0.037599$ (< 9.487729) that is it does not fall within the critical region. Thus, we do not reject H_0 at $\alpha = 0.05$. There is sufficient evidence to support the claim that $p_1=0.11$, $p_2=0.21$, $p_3=0.29$, $p_4=0.05$, $p_5=0.34$. We fail to reject the H_0 because the claimed proportion is a good fit to the level of interest.

Section 9

Work Coordination

No	Group member	Task
1.	AFIQAH IZZATI BINTI AZZEROL EFFENDI	<ul style="list-style-type: none">• SECTION 01• SECTION 03• SECTION 10
2.	LEE RONG XIAN	<ul style="list-style-type: none">• SECTION 04• SECTION 05• SECTION 07<ul style="list-style-type: none">➤ 7.3➤ 7.4• SECTION 08<ul style="list-style-type: none">➤ 8.3➤ 8.4
3.	LU QI YAN	<ul style="list-style-type: none">• SECTION 06• SECTION 07<ul style="list-style-type: none">➤ 7.1➤ 7.2• SECTION 08<ul style="list-style-type: none">➤ 8.1➤ 8.2
4.	NOOR HANNANI SYAMIMI BINTI MOHD SUFFIAN	<ul style="list-style-type: none">• SECTION 02• SECTION 11

Section 10

Reflections

In the process of completing this project, we have learned a lot especially on how to do statistical analysis by using R programming. Even though R programming is new to us, for this project, we tried our best to explore and learn to use R. From the efforts of doing this project, we also gain valuable knowledge by implementing what we have learned in class, such as hypothesis testing, correlation, regression analysis and goodness of fit test. By implementing all of these, it makes us understand more about the topics that we have learned. We also retrieved more information about the statistical tools that R provides.

Moreover, from the data set that we use for this project, we have learned that there are still people who are considering not going to college to continue their education due to several factors. The curiosity on why there are still students who do not continue their education is our main motivation in completing this project.

Our deepest gratitude goes to our SECI2143 lecturer, Dr Azurah binti Abu Samah for her guidance, enthusiastic encouragement and support throughout this project. We would also like to thank Dr Azurah for her assistance and advice on our project. Dr Azurah briefed us and guided us on steps to do the project, she also outlined for us so that we can complete our project excellently. Without her encouragement, we couldn't finish our project. Furthermore, we appreciated Dr Azurah's effort to spend her time checking on the progress of our project and gave us an extension of the due date so that we have more time to do our project.

Section 11

Conclusion and Discussion

In conclusion, for hypothesis testing in section 8, we fail to reject the null hypothesis. There is insufficient evidence to reject the claim that the mean of the average grade in academic school is 86.0.

Next, for correlation analysis test, it is found that there is a relatively weak positive linear relationship between parent's age of the students and the average grade of academic school with a correlation coefficient (r) of 0.3471. We will reject the null hypothesis at 5% level of significance since the test statistic (t) = 56.2610 is greater than the critical value (t) = 1.9757. Due to the fact that there is sufficient evidence to claim that there is a linear relationship between parent's age of the students and average grade of academic school.

For regression analysis test, parent's salary and house area is used as independent variable and dependent variable. We can see in the scatter plot that the line is nearly horizontal which indicates that there is an extremely weak and even no linear relationship between the independent variable and the dependent variable. Moreover, we also obtain the least-square regression line is $\hat{y} = 4.899e+01 - 4.111e-06x$. The coefficient determination value that we get is nearly equal to zero shows that there are less than 0.14% of parent's salary related to their house area. We can conclude that the house area is not dependent on the parent's salary as the dependent of the house area to parent's salary is nearly absent.

Moreover, the Goodness of Fit Test shows that the students from an academic school rated the level of interest in going to school. Since the test value (0.037599) does not exceed the critical value (9.487729) that is it does not fall within the critical region. Thus, here we failed to reject the null hypothesis at 0.05 significance level and there is sufficient evidence to support that $p_1 = 0.11$, $p_2 = 0.21$, $p_3 = 0.29$, $p_4 = 0.05$ and $p_5 = 0.34$. Here, we fail to reject the null hypothesis because the claimed proportion is a good fit to the level of interest.

In a nutshell, the goal of this project is to investigate the variables' relationships and how they may influence whether a student decides to enrol in college or not. Additionally, we hope that the project's findings will serve as a guide to aid school counsellors in determining the variables and causes so they may better assist the pupils. We have also learned a number of other things, like clearing out extra data before beginning the analysis process. Our practise with R programming in the R studio is also aided by this project.

Section 12

References

12.1 Source of Data

1. Mukti, S. S. J. (2022, May 20). *Go to college dataset*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/saddamazyazy/go-to-college-dataset>

12.2 List of Articles/References

1. Scholarships.com. (n.d.). *Why students don't go to college*. 2022 Scholarship Search Find Scholarships for College. Retrieved from <https://www.scholarships.com/resources/college-prep/preparing-for-college/why-students-dont-go-to-college/>
2. *Factors that Influence Students' Desires to Attend Higher Education*. (n.d.). Retrieved from <https://scholarship.shu.edu/cgi/viewcontent.cgi?article=1420&context=dissertations>

12.3 List of E-portfolio

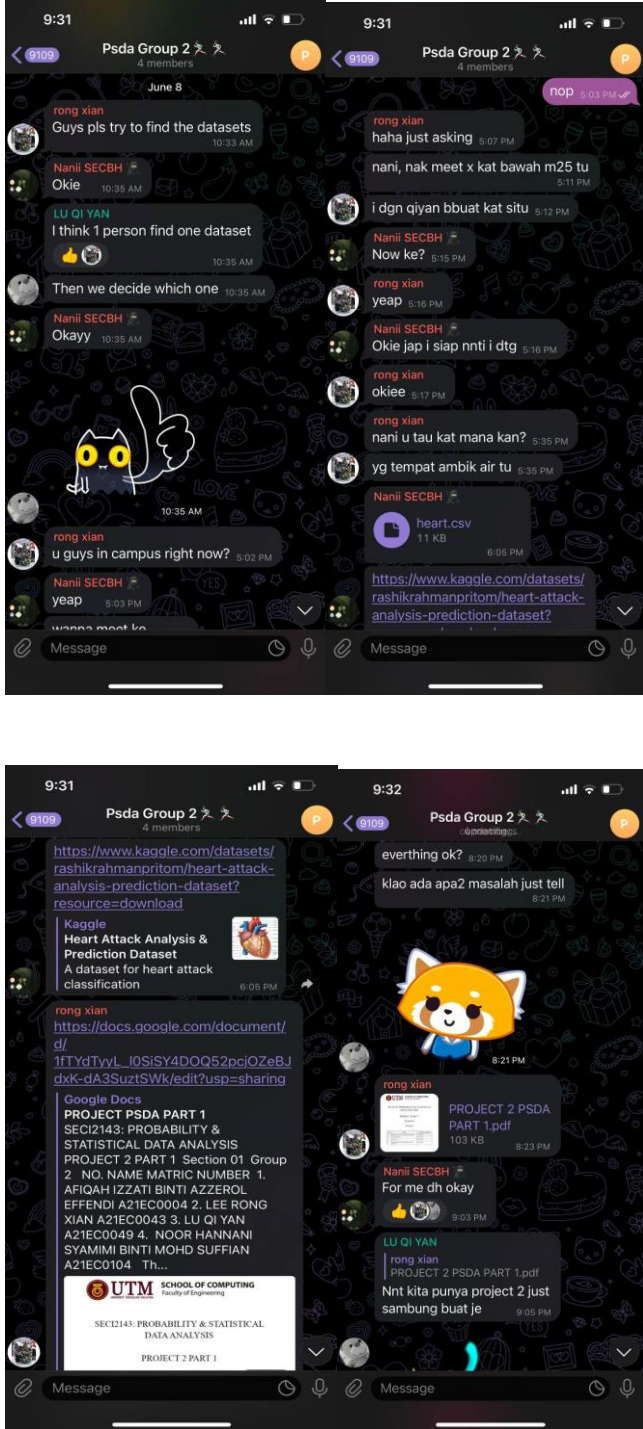
No	Group member	Link of eportfolio
1.	AFIQAH IZZATI BINTI AZZEROL EFFENDI	https://eportfolio.utm.my/view/view.php?t=pVJPoIutkUnTILY24RHg
2.	LEE RONG XIAN	https://eportfolio.utm.my/user/lee-rong-xian
3.	LU QI YAN	https://eportfolio.utm.my/user/lu-qi-yan
4.	NOOR HANNANI SYAMIMI BINTI MOHD SUFFIAN	https://eportfolio.utm.my/user/noor-hannani-syamimi

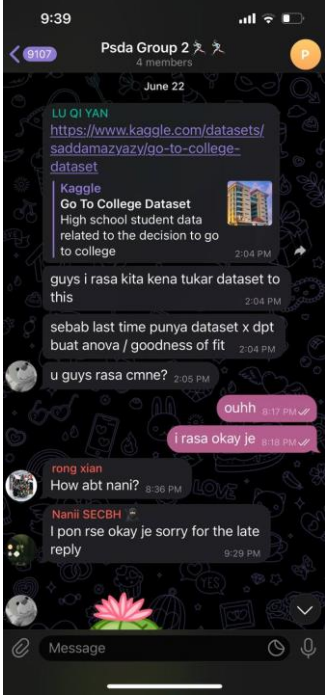
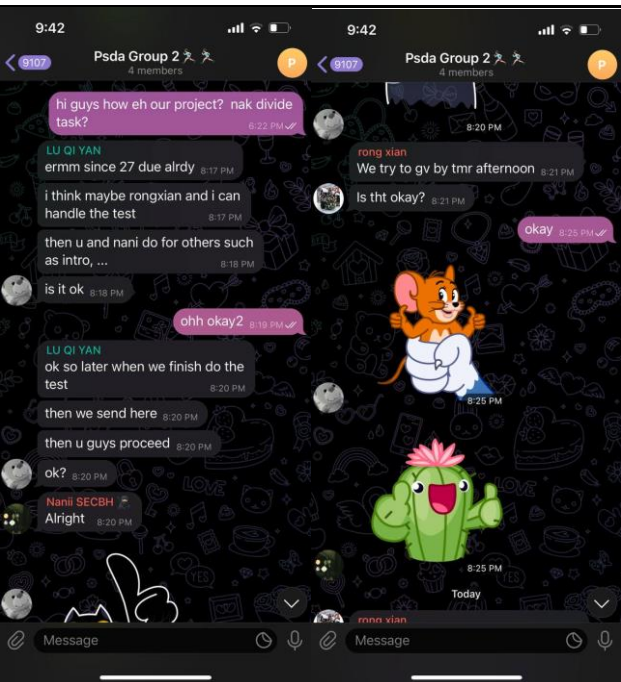
12.4 Link of Presentation Video

<https://drive.google.com/file/d/1BAfkzVJRw5LsWdID0R1TBTbNeL7mvh0v/view?usp=sharing>

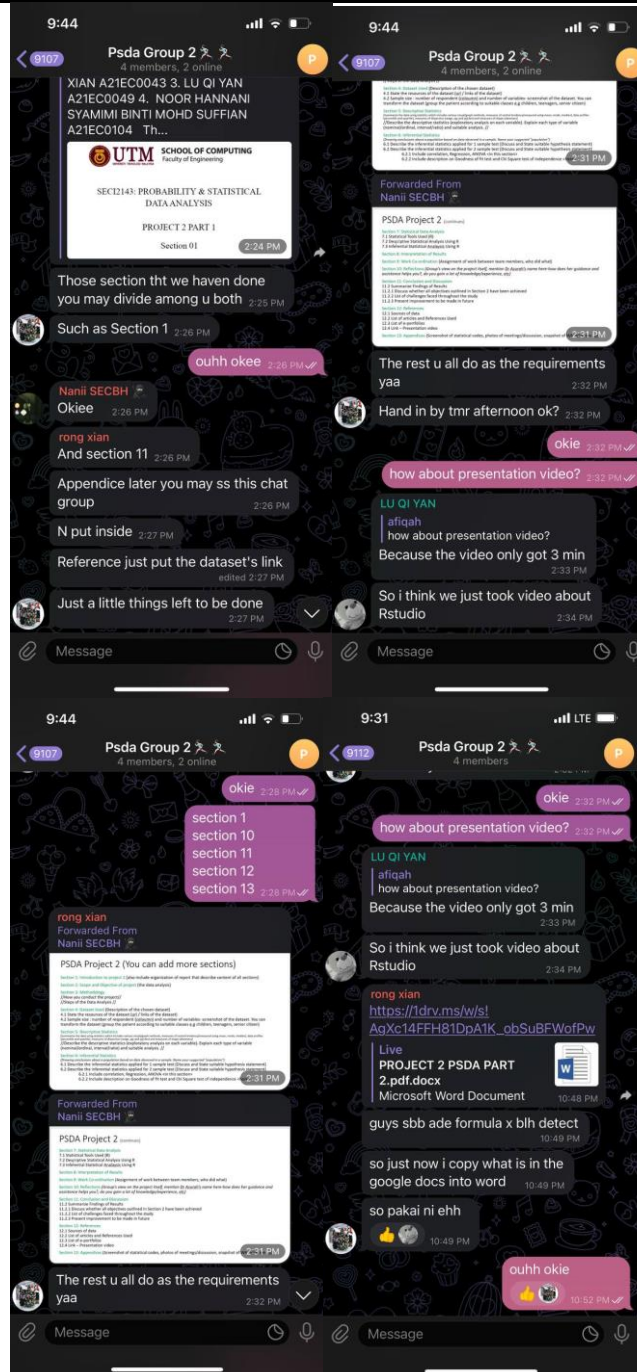
Section 13

Appendices

Dates	Photos	Description
8/6/2022	 <p>The first screenshot shows a WhatsApp chat group 'Psda Group 2' with 4 members. The chat history includes messages from 'rong xian' asking members to find datasets, 'Nanii SECBH' replying 'Okie', and 'LU QI YAN' stating 'I think 1 person find one dataset'. They decide to choose one dataset. 'Nanii SECBH' says 'Okay'. A sticker of a cat with a thumbs up is shared. 'rong xian' asks 'u guys in campus right now?', and 'Nanii SECBH' replies 'yeap'. The second screenshot continues the chat. 'rong xian' shares a link to a Kaggle dataset: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?resource=download. 'Nanii SECBH' shares a file named 'heart.csv' (11 KB). 'rong xian' shares a Google Docs link: https://docs.google.com/document/d/1fTYdTyvL_I0SiSY4DOQ52pciOZeBJdxK-dA3SuztSWk/edit?usp=sharing. The document is titled 'PROJECT PSDA PART 1 SEC12143: PROBABILITY & STATISTICAL DATA ANALYSIS'. 'rong xian' shares a PDF file 'PROJECT 2 PSDA PART 1.pdf' (103 KB). 'Nanii SECBH' says 'For me dh okay'. 'LU QI YAN' says 'Nnt kita punya project 2 just sambung buat je'.</p>	<ol style="list-style-type: none"> 1. Each member assign to find one dataset for project 2. Decided to do meet face to face to discuss the suitable dataset 3. Choose the dataset and start to do an initial report (by listing the selected variables and chosen test) based on the dataset.

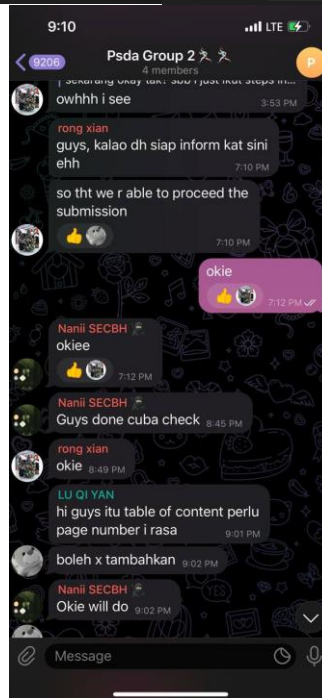
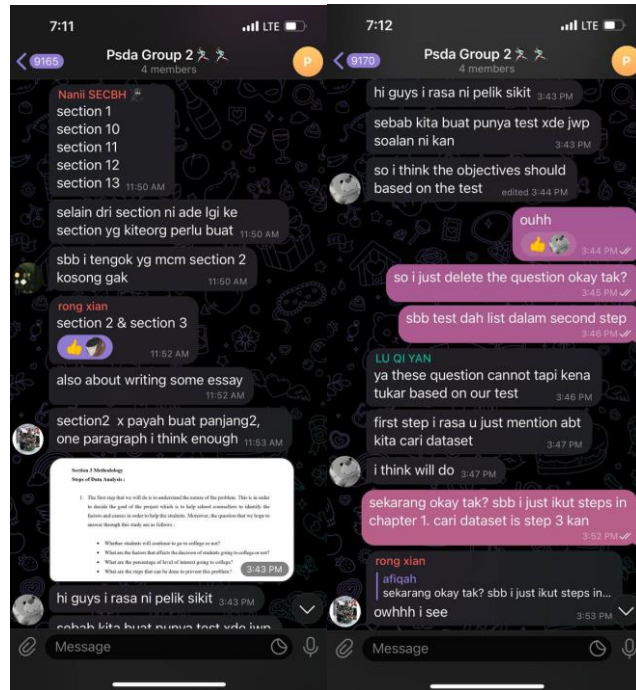
22/6/2022		<p>4. Decided to change dataset because the previous dataset is not suitable to do certain test</p> <p>5. Started to do data analysis based on the dataset</p>
25/6/2022		<p>6. Task confirmation</p>

26/6/2022



7. Completed several part of the statistical analysis
8. Continue with few more section

27/6/2022



9. Last check before submitting
10. Completed the report
11. Completed the video
12. Completed eportfolio
13. Submitted all