PROBABILITY AND STATISTICAL DATA ANALYSIS

PROJECT 2

SECI 2143

LECTURER: DR SHARIN HAZLIN BINTI HUSPI

NAME	MATRIC NO
RISEEBAA SARAVANAN	B20EC3019
FATIN SYAZLIANA BINTI NAZRI	A20EC0036
NUR DINIE SAJEEDA BINTI AZMAN	A20EC0112
ANATASYA HUMAIRA	A20EC0261

Introduction

Student performance is where the students are valued through examination or assessments and evaluate how much the student has achieved the goals. Student performance is basically the result of a student's hard work to achieve excellent results or grades. It was important as it will encourage students to become better in achieving their goals. Today, most of the society assumes that students who have good grades will become successful and the students who did not excel in their studies will fail. So one of the initiatives of all education institutions is to evaluate their student performance to encourage students to have a better result. Other than that, student performance can evaluate students strengths and weaknesses so that they can fix the weaknesses and have better results. Therefore, here we will study what are the factors that affect a student's grade whether the student's grade is high or low.

Dataset

The data set that our group has chosen is student performance. The data set is the secondary data where the data has been collected beforehand. There are a few variables that are represented by columns which are gender, race/ethnic, parental level of education, test preparation, math score, reading score and writing score. This data set contains both continuous and discrete data. In order to fulfill the requirement of inference statistical analysis for this project, few variables have been chosen. In this case we use variables gender and math score for hypothesis testing 2-sample. This statistical data analysis is to test whether the sample mean between female and male students on their mathematics score is equal or not. Other than that, reading score and writing score are chosen to conduct correlation tests. This test is to identify if there were any relationship between the two variables. Next for regression analysis, we use math scores to predict the reading score of the students. The last test we conduct is the chi-square test where we use it to determine whether the variables that we use in this report have any significant correlation to each other or not, which include the percentage of overall scores and between the students who have accomplished the test preparation and those who have not.

Data Analysis

1. **Hypothesis 2-Sample Test (Test on Difference Between TwoMean - assume UNEQUAL VARIANCE)**

We wish to determine if there is any difference in mean among students on their mathematics score between male and female. With the Significance level, $\alpha = 0.05$. Since the size of the sample is relatively large, we can assume that both samples are normally distributed. Z-value for test statistic is obtained by using R. The number of mathematics scores for male students is 9, while the number of mathematics scores for female students is 11. We assume both variance of both variables is not equal, since there is no way to say that variance of both variables is equal. Variables used in this test are mathematics score and gender (male and female).

Two sample tests on mathematics score among students .

Let μ 1 = sample mean for male students on their mathematics score among students.

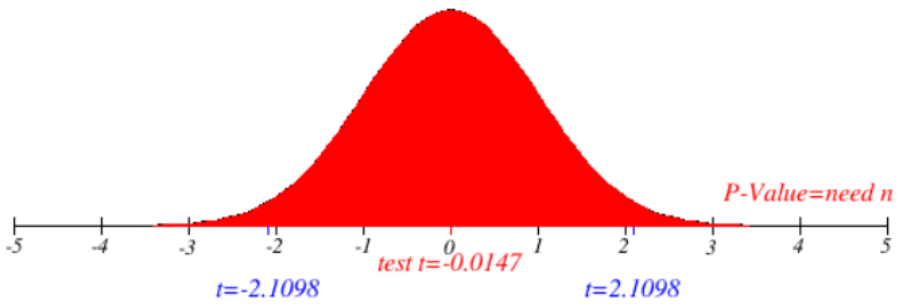
Let μ 2 = sample mean for female students on their mathematics score among students.

$H_0 : \mu_1 - \mu_2 = 0$
$H_1 : \mu_1 \neq \mu_2$

We reject H0 if t-stat < critical-value = -2.1098 or if t-stat* >critical-value =2.1098

If the t stat is larger than the critical two tail value we reject H0.

Since t-stat = -0.0416> critical-value = -2.1098 and t-stat* -0.0416 <critical-value =2.1098 , therefore fail to reject H0 and say data supports H0 so the sample mean of the math score among male and female students can be considered equal.



2. Correlation Test

Correlation test between variable reading score and writing score. This test is to determine if there is any linear relationship between the variable reading score and writing score whether it is positive correlation, negative correlation or no correlation. For this test we use Pearson's Product-Moment Correlation Coefficient method to determine the sample correlation coefficient, r.

Pearson's Product-Moment Correlation Coefficient	
n	1000
Sum Y	69169
Sum X	68054
Sum XY	4918802
Sum Y^2	4862024
Sum X^2	4997303
r	0.9545981

Table 3

Test for Correlation	
t	101.23
d.f	998
t(α/2,d.f)	1.962344

Table 4

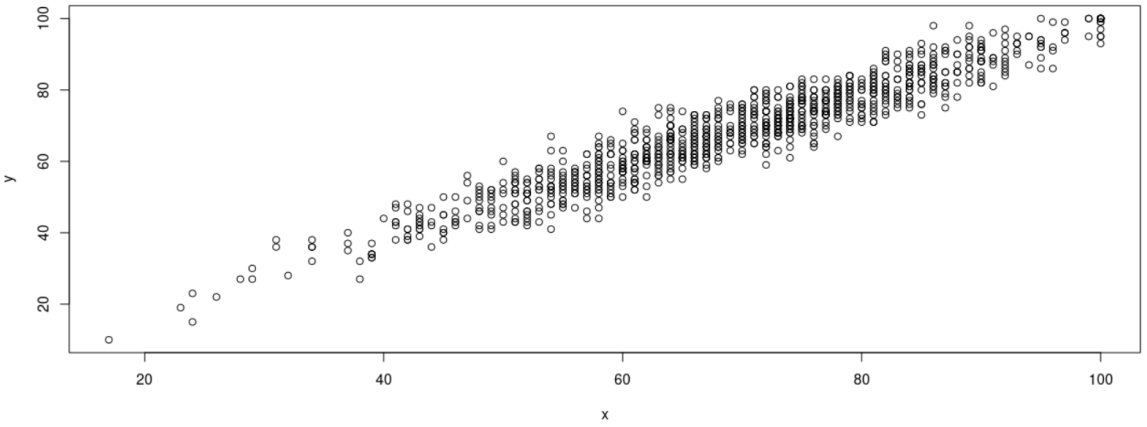


Figure 1

Based on the table 3 and scatter plot we can conclude that the linear relationship between reading score and writing score is strongly positive correlation as the reading score increases, the writing score also increases and the sample correlation coefficient, r = 0.9545981where it is close to 1(+ve).

H0: ρ = 0 (does not have linear correlation relationship between reading and writing score)
HA: ρ ≠ 0 (have linear correlation relationship between reading and writing score)

Based on the result we obtain, t0 = 101.23, so we conclude that we reject the null hypothesis since t0 > t(0.025/998) = 1.962344. There is sufficient evidence to claim that there is a linear relationship between the reading and writing score of the students at 5% of level of significance.

3. Regression Test

For this test, a random sample of 1000 students and their scores have been selected. We want to test if math score are related to reading score.

The null hypothesis, Ho: $\beta 1 = 0$
The alternative hypothesis, H1: $\beta 1 \neq 0$

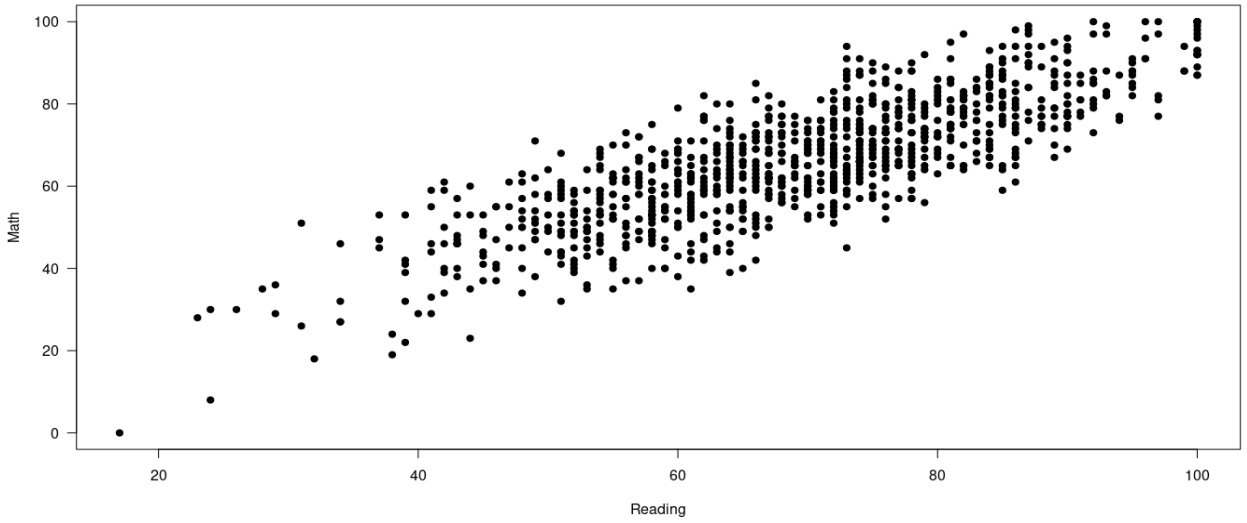


Figure 2 : Plot of Math Score against Reading Score

Based on the graphs generated, the regression model involves a single independent variable and it is called simple regression. The regression models are positive linear model which have straight-line relationships between math score and reading score.

Least Squares Regression Equation : **y = 7.3576 + 0.8691x**

The value of intersection coefficient, β0 is the estimated average value of Y when the value of X is zero (if x = 0 is in the range of observed x values). Here, β0 = 7.3576 which indicates that for math score is 7.3576. While the value of slope coefficient, β1 measures the estimated change in the average value of Y as a result of a one-unit change in X. Here b1 = 0.8691 which indicates that the average value of math score increases by 0.8691 on average. Therefore, based on the result we have we can conclude that there is a positive relationship between the dependent variable (y) and the independent variable (x).

4. Chi Square Test

By using the chi-squared test, we manage to know whether the data between two variables, which are the overall score percentage and test preparation completion, have any certain or significant correlation to each other or not. The test itself is essentially a data analysis based on random observations of a set of variables. However, in this project we tried to sum up all of the scores for each student first before determining the hypothesis to find the percentage since we would like to know if both the amount of students who have completed and have not completed the test preparation could actually impact the percentage of total scores, or else. Then, we divided the overall score percentage into 2 categories, which are above 60% or less than 60%. The table shown below is the result that we obtain after calculating several variables according to the percentage:

	TEST PREPARATION	
SCORE	**COMPLETED**	**NONE**
ABOVE 60%	293	414
BELOW 60%	65	228

Table 1: Observed value

Thus, we could write the hypothesis in this report by :

Hypothesis

H_0: $x^2 = \alpha$ (There is no association between test preparation completion and the overall score percentage | independent).

H_1 : $x^2 > \alpha$ (There is an association between test preparation completion and the overall score percentage | dependent).

Additionally, we could achieve this following expected data based on the observed test statistic above by using several formulas, such as :

SCORE	COMPLETED	NONE	Total
ABOVE 60%	358*707/1000	642*707/1000	707
BELOW 60%	358*293/1000	642*293/1000	293
Total	358	642	1000

Table 2. Expected value

And the final outcome of the calculation is :

SCORE	COMPLETED	NONE	Total
ABOVE 60%	253.11	453.89	707
BELOW 60%	104.89	188.11	293
Total	358	642	1000

Table 3 : Expected value

According to the data set that we have collected so far, we manage to find the X-square by inputting the number stated above into the formula which is $\chi 2 = \sum(Oi - Ei)2/Ei$, where Oi is the observed value and Ei is the expected value. Therefore, the implementation of the formula could be written in :

$$x^2 = \frac{(298 - 253.11)^2}{253.11} + \frac{(414 - 453.89)^2}{453.89} + \frac{(65 - 104.89)^2}{104.89} + \frac{(228 - 188.11)^2}{188.11} = 33.428$$

We could obtain the critical value by firstly finding the number of degree of freedom with the formula of *df=(Rows-1)*(Columns-1).* Thus, as a result, we could come up with the result of 1 by the multiplication of *(2-1)*(2-1) = 1.* However, we also assign the significance level of alpha (α) as 0.05 since it is of numbers that are commonly used to find the critical value. Then next, the number of critical values is finally constructed according to the Chi-Square distribution table, resulting 3.841.

$x^2_{df=1,\alpha=0.05} = 3.841$

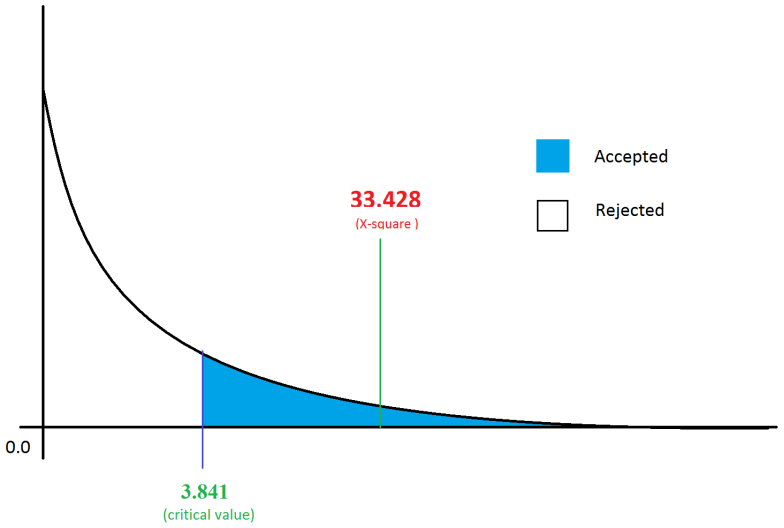


Figure 1 : Graph of the X^2 distribution.

Based on the graph and data that we have been determined earlier, p value ≈ 0 meaning that there is almost 0 chance of an error that can contradict the correct hypothesis.

Thus, we could also conclude that the final decision for this data analysis is to reject the null hypothesis since 33.428>3.841. On the other hand, there is an association between the test preparation completion and the overall score percentage with the significance of 0.05.

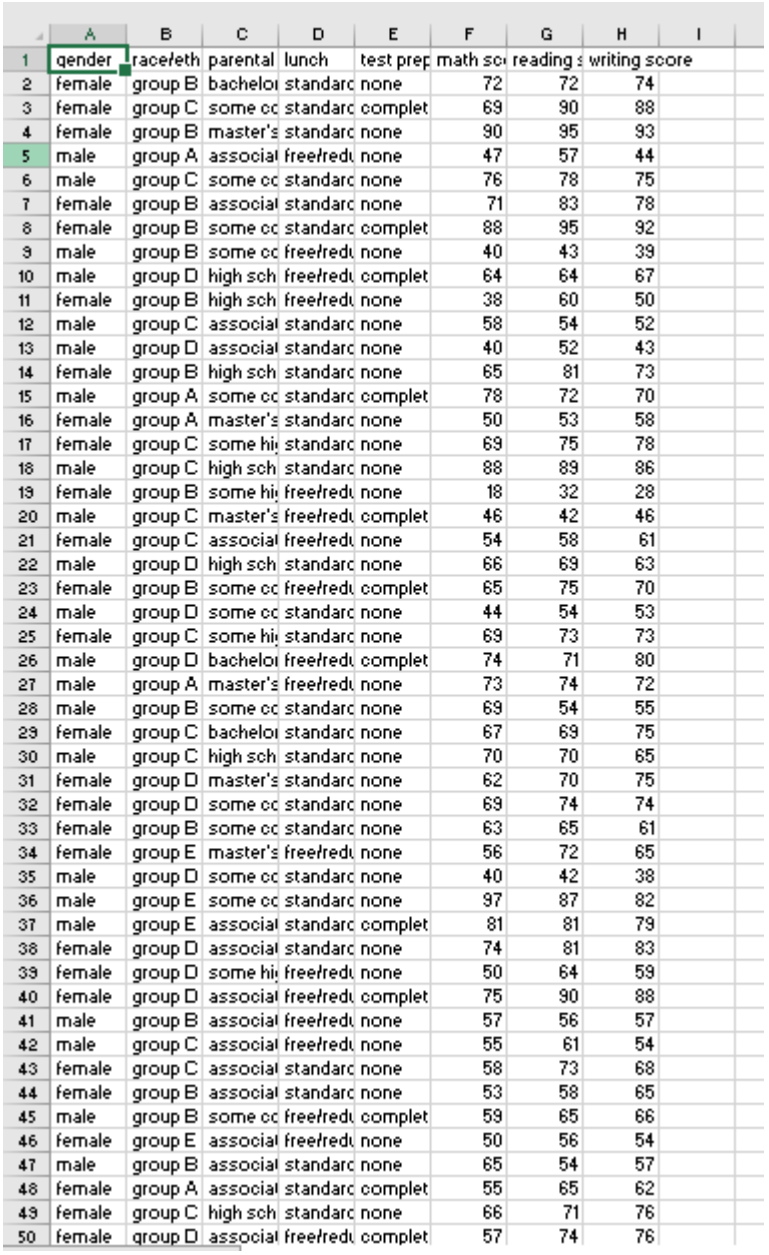
Conclusion

Based on our findings, we have found many factors that affect student performances. So, to conclude the study,we have found that the math score among the students is affected by gender. Other than that, we identify that reading score and math score are related to each other as the reading score is high, the writing score also high. We also identify that reading scores have an impact of change on math scores. And lastly, we could determine that the test preparation completion also has a significant impact on the total score percentage of each student.

Overall, this project has achieved its objectives. The project provided a sample of dataset for students performance and was successfully built using RStudio Cloud. By finishing this project, we understand more about what we learnt especially in choosing a dataset, doing a pre-processing and making analysis from the data.

Appendix

Sample of original/raw dataset



	A	B	C	D	E	F	G	H	I
1	gender	race/eth	parental	lunch	test prep	math sc	reading s	writing score	
2	female	group B	bachelo	standarc	none	72	72	74	
3	female	group C	some co	standarc	complet	69	90	88	
4	female	group B	master's	standarc	none	90	95	93	
5	male	group A	associat	free/redu	none	47	57	44	
6	male	group C	some co	standarc	none	76	78	75	
7	female	group B	associat	standarc	none	71	83	78	
8	female	group B	some co	standarc	complet	88	95	92	
9	male	group B	some co	free/redu	none	40	43	39	
10	male	group D	high sch	free/redu	complet	64	64	67	
11	female	group B	high sch	free/redu	none	38	60	50	
12	male	group C	associat	standarc	none	58	54	52	
13	male	group D	associat	standarc	none	40	52	43	
14	female	group B	high sch	standarc	none	65	81	73	
15	male	group A	some co	standarc	complet	78	72	70	
16	female	group A	master's	standarc	none	50	53	58	
17	female	group C	some hi	standarc	none	69	75	78	
18	male	group C	high sch	standarc	none	88	89	86	
19	female	group B	some hi	free/redu	none	18	32	28	
20	male	group C	master's	free/redu	complet	46	42	46	
21	female	group C	associat	free/redu	none	54	58	61	
22	male	group D	high sch	standarc	none	66	69	63	
23	female	group B	some co	free/redu	complet	65	75	70	
24	male	group D	some co	standarc	none	44	54	53	
25	female	group C	some hi	standarc	none	69	73	73	
26	male	group D	bachelo	free/redu	complet	74	71	80	
27	male	group A	master's	free/redu	none	73	74	72	
28	male	group B	some co	standarc	none	69	54	55	
29	female	group C	bachelo	standarc	none	67	69	75	
30	male	group C	high sch	standarc	none	70	70	65	
31	female	group D	master's	standarc	none	62	70	75	
32	female	group D	some co	standarc	none	69	74	74	
33	female	group B	some co	standarc	none	63	65	61	
34	female	group E	master's	free/redu	none	56	72	65	
35	male	group D	some co	standarc	none	40	42	38	
36	male	group E	some co	standarc	none	97	87	82	
37	male	group E	associat	standarc	complet	81	81	79	
38	female	group D	associat	standarc	none	74	81	83	
39	female	group D	some hi	free/redu	none	50	64	59	
40	female	group D	associat	free/redu	complet	75	90	88	
41	male	group B	associat	free/redu	none	57	56	57	
42	male	group C	associat	free/redu	none	55	61	54	
43	female	group C	associat	standarc	none	58	73	68	
44	female	group B	associat	standarc	none	53	58	65	
45	male	group B	some co	free/redu	complet	59	65	66	
46	female	group E	associat	free/redu	none	50	56	54	
47	male	group B	associat	standarc	none	65	54	57	
48	female	group A	associat	standarc	complet	55	65	62	
49	female	group C	high sch	standarc	none	66	71	76	
50	female	group D	associat	free/redu	complet	57	74	76	

Processed Data Sample

	A	B	C	D	E	F	G	H
1	gender	race/ethn	parental l	lunch	test prepa	math score	reading sc	writing scor
2	female	group B	bachelor's	standard	none	72	72	74
3	female	group C	some coll	standard	completec	69	90	88
4	female	group B	master's c	standard	none	90	95	93
5	female	group B	associate'	standard	none	71	83	78
6	female	group B	some coll	standard	completec	88	95	92
7	female	group B	high scho	free/redu	none	38	60	50
8	female	group B	high scho	standard	none	65	81	73
9	female	group A	master's c	standard	none	50	53	58
10	female	group C	some high	standard	none	69	75	78
11	female	group B	some high	free/redu	none	18	32	28
12	female	group C	associate'	free/redu	none	54	58	61
13	male	group A	associate'	free/redu	none	47	57	44
14	male	group C	some coll	standard	none	76	78	75
15	male	group B	some coll	free/redu	none	40	43	39
16	male	group D	high scho	free/redu	completec	64	64	67
17	male	group C	associate'	standard	none	58	54	52
18	male	group D	associate'	standard	none	40	52	43
19	male	group A	some coll	standard	completec	78	72	70
20	male	group C	high scho	standard	none	88	89	86
21	male	group C	master's c	free/redu	completec	46	42	46