PROBABILITY & STATISTICAL DATA ANALYSIS

(SECI2143-SECTION 01)

PROJECT 2-Graphique

Lecturer:

DR. SHARIN HAZLIN BINTI HUSPI

Group Members:

No	Name	Matric No
1	Omar Hamed Abdellatif Ibrahim	A18CS4061
2	Tasnia Hoque Nidhi	A18CS9010
3	Hanis Rafiqah binti Hisham Razuli	A20EC0041
4	Nik Syahdina Zulaikha binti Badrul Hisham	A20EC0108

Table of contents:

Introduction

This study considers how well the students from different races and ethnicities perform in mathematics, reading and writing according to their gender. The analysis reveals the considerable variation in the relative standing of countries in terms of their students' capacity to put knowledge and skills to functional use. However, the analysis also suggests that differences between the qualification of the paternal level of education variation in student test performance. The purpose of the study is to compare males and females in average scoring. Moreover, how the ethnic background contributes to their children obtaining better scores. Parent's educational qualification in student's profile. Besides analysing whether students who received free or reduced lunch perform better than students who get standard lunch, whereas students get higher marks when they had taken a preparation course before or not.

Variation in student performance within males and females can have a variety of causes including the socio-economic backgrounds of students and schools. From the data set by examining more closely the performance gap is shown. The gap has been reflected on students' total and average scores and last of all positioned an effect on the overall grade. Furthermore, the main objective of the study is to examine student's profiles who are struggling or need more attention with the help of selected variable test preparation courses, reading score, writing score, lunch availability, and parental level of education. This study applies two sampled hypothesis testing in order to get a proper view of who has an overall good performance and who has not. We have applied 4 data analysis testing here. Such as hypothesis testing, correlation testing, regression testing, chi-square test of independence. Through testing, the result will give us a proper idea about the interconnections between the fields and what are the reasons behind good and bad performances.

Data Description

Name	Data Type	Level of Measurement
Gender	Categorical	Nominal
Race/ethnicity	Categorical	Nominal
Parental level of education	Categorical	Nominal
Lunch	Categorical	Nominal
Test Preparation Course	Categorical	Nominal
Math Score	Quantitative	Ratio
Reading Score	Quantitative	Ratio
Writing Score	Quantitative	Ratio
Total Score	Quantitative	Ratio
Average Score	Quantitative	Ratio
Grade	Categorical	Ordinal

Table 1

Table 1 above shows a portion of the Student's Performance dataset. Data shown above in a tabular manner is a secondary data collection. Secondary data collection is data that already exists and facts that were already recorded prior to the project. We had to add .csv to the Excel file as a typical extension so that it would be easier to access

through the R programming language. This dataset shows the details of Student's Performance. Six of the variables are categorical while another five are quantitative. We took a variable test preparation course, math score, reading score and writing score and used two-sample hypothesis testing to see if students who had completed the test preparation course had performed better than those who did not. Variable's gender, math score, reading score and writing score were selected to carry out a two-sample hypothesis test to see if female students have a higher score than male students. Variable lunch and average score were selected to carry out a two-sample hypothesis test to see if the type of lunch affects the student's average score. Variables reading scores and writing scores were selected to conduct a correlation test to see the relationship between writing score and reading score. Variable reading scores and writing scores were again selected to conduct a regression test to check if reading scores affect writing scores. Lastly, variables race and average score were taken to carry out a chi-square test of independence and see if race will affect the students' performance.

Data Analysis

Hypothesis Testing

First Test - Two Sample Test

Statement: "Students who took test preparation course performed better than students who did not take the preparation course"

Parameter name	Parameter Value
Mean for test preparation Completed Total Score	218.00
Standard Deviation For Test Preparation Completed Total Score	39.11
Mean for test preparation none Total Score	195.11
Standard Deviation For Test Preparation none Total Score	42.56
CL	0.95
T.Test	-8.5945
P value	2.2e-16

Table 2

- Null Hypothesis is that there is no difference between the means for completed test preparation and none completed test preparation.
- Alternative Hypothesis is that the students who do not undergo test preparation courses got less mean than students that undergo test preparation courses.

Result:
From the test we can conclude that the p.value is 2.2e-16 and p.value <0.05 so we reject the null hypothesis because there is no difference in means and accept that students who undergo test preparation course has higher mean than students who does not undergo test preparation course which resulted in lower mean and we can make sure that our test is correct from the figure below.

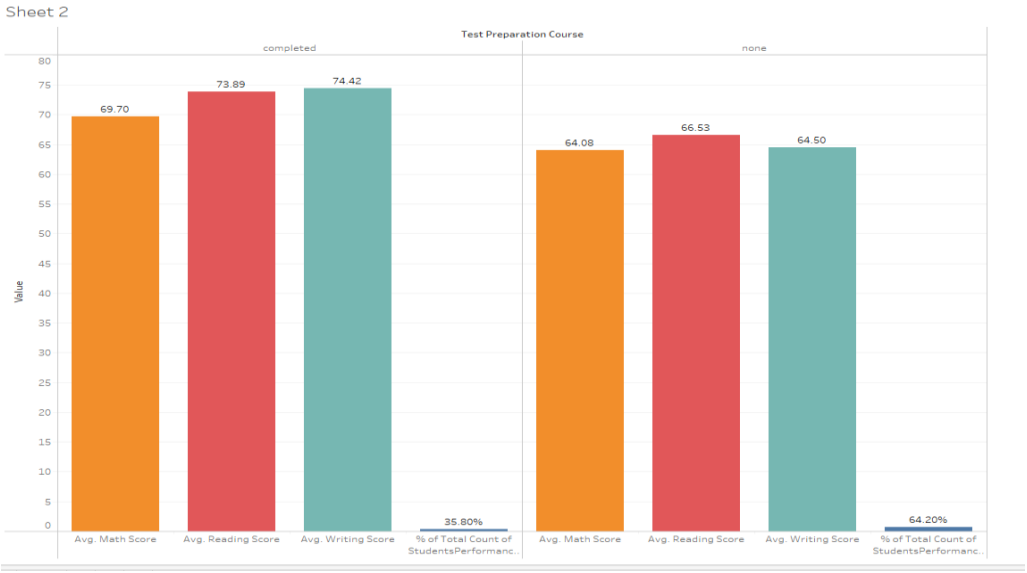

Figure 1: The comparison between students who took the test preparation and who did not and their average scores.

Second Test - Two Sample Test

Statement: "Female students have higher average scores than male students"

Parameter name	Parameter Value
Total Score mean for female	208.70
Total score standard deviation for female	43.62
Total Score mean for male	197.51
Total score standard deviation for male	41.09
CL	0.95
T.Test	4.1789
P value	1.593e*-05

Table 3

- Null Hypothesis is that there is no difference in the means for the two sample female and male average scores.
- Alternative hypothesis that female sample mean has greater average score than male sample mean.

Result:

After performing two sample test the p-value is 1.5593e *-05 form the result we can conclude that the p-value < 0.05 adding on we reject the null hypothesis because there is difference between the means and that a significance different exists and accept the alternative hypothesis that female mean has greater average score than male mean average score and we can make sure that our test is correct from the figures below.

Figure 2: The total number of students according to their gender.

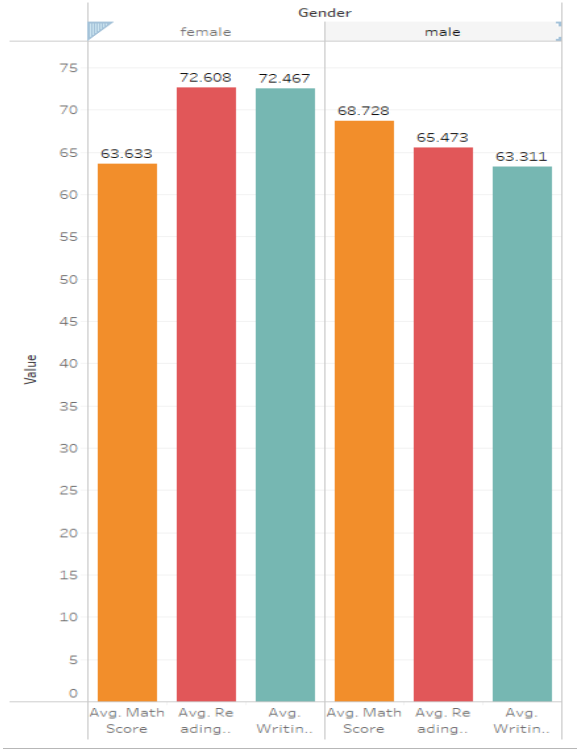

Figure 3: Female students have higher scores than male students.

Third Test - Two Sample Test

Statement: "Type of lunch affects students total average score"

Parameter name	Parameter Value
Mean for Reduced/free lunch	186.597
standard deviation for Reduced/free lunch	43.37
Mean for Standard lunch	212.511
standard deviation for Standard lunch	39.55

CL	0.95
T.Test	-9.3232
P value	2.2e *-16

Table 4

- Null Hypothesis is that there's no difference between the means of the two samples Reduced/free and Standard and that the two samples will have the same performance (average score) for students who take reduced/free meals and students who take standard meals.
- Alternative Hypothesis is that lunch type will affect student performance (total average score)

Result:
After Performing a two-sample two-sided test the result was that p.value is 2.2e*-16, and from the test, we can conclude that p.value < 0.05, Adding on we reject the null hypothesis because there's no difference between the mean for Reduced/free lunch and Standard lunch and accept the Alternative hypothesis which is that the lunch type will affect students' performance (total average score) and we can make sure from the figures below.

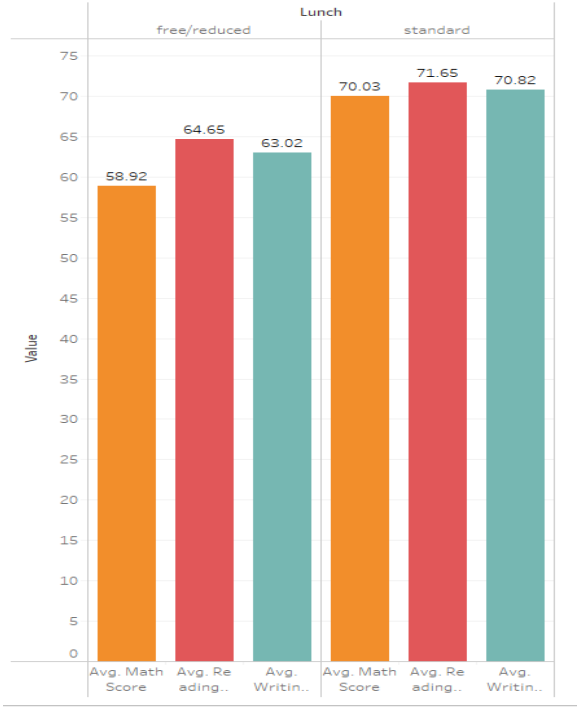

Figure 4: The average scores for students with standard lunch are much higher than students with Reduced/free lunch.

Correlation Testing

Parameters used:

r	For all the Reading Scores
w	For all the Writing Scores

Table 5

After applying the Correlation our result for the Correlation coefficient is r= 0.9545981

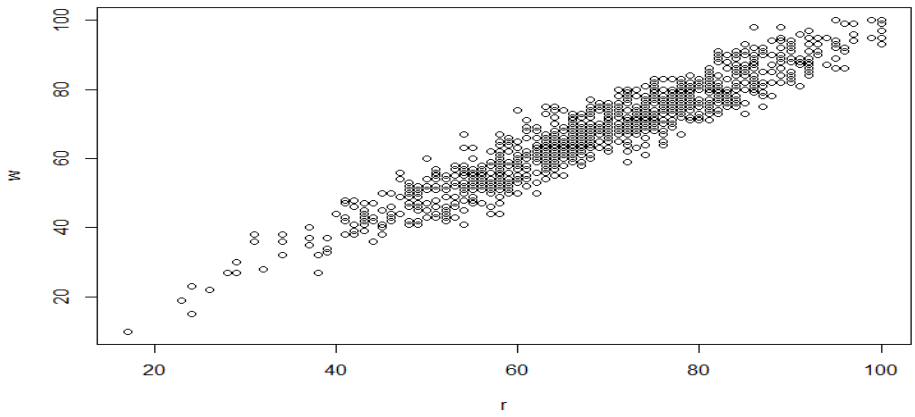

Figure 5: The Scatter plot for the correlation between reading score and writing score

As we can conclude from the plot and that our coefficient r = 0.9545981 which is close to 1, adding on the form the graph we can see a straight-line pattern and the values are near each other, so we can conclude that there's a strong relationship between writing score and reading score, in relation we can observe that students that are good in reading will also be good in writing.

Regression Testing

Parameters Used:

r	For all the Reading Scores
w	For all the Writing Scores

Table 6

After applying the Correlation our result for the Correlation coefficient is r= 0.9545981

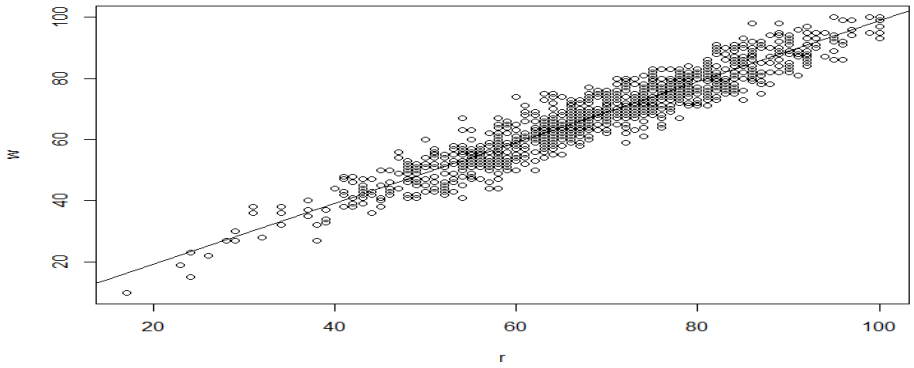

Figure 6: The Scatter plot for the Regression between reading score and writing score

As we can conclude from the scatter plot that we have a positive Linear Relationship between Reading Scores and Writing scores changes that occurs in reading scores are being reflected on writing scores positively for the Coefficient which is -0.6676 and r=0.9935 we can conclude that Reading scores do affect the writing scores due to the strong correlation between them.

Chi-Square test of independence

Forth Test - Statement:"Race does not have an effect on students' performance"

Parameter name	Parameter Value
X-squared	49.926
P-value	0.0002269

Table 6

Since p-value = 0.0002269 < 0.05 so we reject the null hypothesis that mentions that there's no difference in means and accepts the alternative hypothesis that says that Race will have an effect on student performance Average score

Conclusion

To be put in a nutshell, students' performance can be affected by many factors. Based on the study, we could conclude that students with test preparation have more scores than students that did not do preparation. Other than that, based on gender, the female has a greater average score than the male. Then, the next factor is students who take a standard meal can get a better average score than those who are reduced or have free mealtime. Students that excel in reading will also excel in writing tests because reading scores can affect writing scores. Lastly, race does affect the average score. This project makes our understanding of statistical analysis better and we managed to use the R language.

Appendix

Processed Dataset:

	A	B	C	D	E	F	G	H	I	J	K
1	gender	race.ethnicity	parental.level.of.education	lunch	test.preparation.course	math.score	reading.score	writing.score	total.score	avg.score	grade
2	female	group B	bachelor's degree	standard	none	72	72	74	218	73	C
3	female	group C	some college	standard	completed	69	90	88	247	82	B
4	female	group B	master's degree	standard	none	90	95	93	278	93	A
5	male	group A	associate's degree	free/reduced	none	47	57	44	148	49	F
6	male	group C	some college	standard	none	76	78	75	229	76	C
7	female	group B	associate's degree	standard	none	71	83	78	232	77	C
8	female	group B	some college	standard	completed	88	95	92	275	92	A
9	male	group B	some college	free/reduced	none	40	43	39	122	41	F
10	male	group D	high school	free/reduced	completed	64	64	67	195	65	D
11	female	group B	high school	free/reduced	none	38	60	50	148	49	F
12	male	group C	associate's degree	standard	none	58	54	52	164	55	E
13	male	group D	associate's degree	standard	none	40	52	43	135	45	F
14	female	group B	high school	standard	none	65	81	73	219	73	C
15	male	group A	some college	standard	completed	78	72	70	220	73	C
16	female	group A	master's degree	standard	none	50	53	58	161	54	E
17	female	group C	some high school	standard	none	69	75	78	222	74	C
18	male	group C	high school	standard	none	88	89	86	263	88	B
19	female	group B	some high school	free/reduced	none	18	32	28	78	26	F
20	male	group C	master's degree	free/reduced	completed	46	42	46	134	45	F
21	female	group C	associate's degree	free/reduced	none	54	58	61	173	58	E
22	male	group D	high school	standard	none	66	69	63	198	66	D
23	female	group B	some college	free/reduced	completed	65	75	70	210	70	C
24	male	group D	some college	standard	none	44	54	53	151	50	E
25	female	group C	some high school	standard	none	69	73	73	215	72	C
26	male	group D	bachelor's degree	free/reduced	completed	74	71	80	225	75	C