UTM

UNIVERSITI TEKNOLOGI MALAYSIA

PROBABILITY & STATISTICAL DATA ANALYSIS

SECI2143

Project 2

Lecturer:

Rozilawati Dollah @ Md Zain

Group Members:

No	Name	Matric No
1	AIMAN HAIKAL BIN ZAINUDDIN	A21EC0154
2	ARSYAD WIRATAMA	A20EC0290
3	MUHAMAD AMSYAR BIN IBRAHIM	A21EC0058
4	MUHAMMAD DANIAL WAJDI BIN SAFIAY	A21EC0071

Table of Contents:

INTRODUCTION

Education is very important in our life. Children have to come to school to learn many types of subjects such as Science, Mathematics, Physics, Biology, Chemistry and many more in order to build a country with civilized and educated citizens. To scale their understanding about the subjects, they have to undergo a test where they try to answer the question about the topics learned in the subjects and the results will give them a chance to enroll themselves into universities with more advanced education systems based on their desired course.

This report is to understand the influence of the parent's background of education and other factors on a student's performance. The reason why we choose this topic is because education is very important in our life. As a student, we want to know the reason whether parents' background in educational level and other factors will influence their child's performance in exams. We are hoping that with this data, we are able to identify the factors that affect students' performance in exams and also solve the significant factors on students' performance.

DATASET

This report uses the secondary data retrieved from a website Kaggle about Students Performance in Exams. This data was collected by Jakki Seshapanpu who is the owner of this data set. This data has many different types of variables such as Gender, students performance in exams such as Math score, Reading score and Writing score, Parental level of education, Test preparation course and many more.

At first, for the variable 'Parental level of education', there are a lot of parental levels of education, ranging from Masters' Degree, Bachelor's Degree, Associate's Degree, College, Some College, High School and Some High School. But in data pre-processing, we decided to combine the data that is similar such as combining College with Some College and High School with Some High School into one variable respectively which is College and High School. Therefore, there are only 5 levels of Parental level of education which are Masters' Degree, Bachelor's Degree, Associate's Degree, College and High School. The reason why we chose these variables is so that we can identify whether it is an independent variable or not with other variables such as Gender using the Chi Square Test of Independence.

ANALYSIS AND RESULT

2.1 HYPOTHESIS TESTING 1 SAMPLE

Test Hypothesis :

To test whether the hypothesis is true that the average score of students in math subjects is above 65.

- Null Hypothesis(H_0) : $\mu_0 = 65$
- Alternative hypothesis(H_1) : $\mu_1 > 65$

Test statistic:

α	**0.05**
Sample size, n	**400**
Sample mean, $\bar{x}$	**65.278**
Sample standard deviation,σ	**15.408**

By using this formula :

$$\boxed{z = \frac{\bar{x} - \mu}{s/\sqrt{n}}}$$

Test statistic, z	0.361
Critical value, $z_{0.05}$	1.645

Since the test statistic value is less than the critical value(0.361 < 1.645), it falls outside the rejection region. Therefore, we fail to reject the null hypothesis. There is sufficient evidence that the average score of math subject is 65,

2.2 CORRELATION

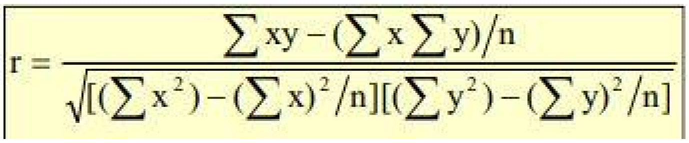
Test Hypothesis :

- $(H_0) : p_1 = 0$ (no linear relationship)

  $(H_1) : p_1 \neq 0$ (linear relationship does exist)

Test Statistics:

Total score (y)	Hours of studying (x)	XY	y^2	x^2
218	10.9	2376.2	47524	118.81
247	12.35	3050.45	61009	152.52
278	13.9	3864.2	77284	193.21
148	7.4	1095.2	21904	54.76
229	11.45	2622.05	52441	131.10
232	11.6	2691.2	53824	134.56
275	13.75	3781.25	75625	189.06
122	6.1	744.2	14884	37.21
195	9.75	1901.25	38025	95.06
164	8.2	1344.8	26896	67,.24
= 2018	=105.4	=23470.8	=469416	= 1173.53



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

r = 1.115 -> relatively strong positive linear association between x and y

2.3 REGRESSION

Test Hypothesis :

- (H_0) : $\beta_1 = 0$ (no linear relationship)
- (H_1) : $\beta_1 \neq 0$ (linear relationship does exist)

Test Statistic :

Total score (y)	**Hours of studying** (x)
218	**10.9**
247	**12.35**
278	**13.9**
148	**7.4**
229	**11.45**
232	**11.6**
275	**13.75**
122	**6.1**
195	**9.75**
164	**8.2**

y	**x**	**xy**	x^2
218	**10.9**	**2376.2**	**118.81**
247	**12.35**	**3050.45**	**152.52**
278	**13.9**	**3864.2**	**193.21**
148	**7.4**	**1095.2**	**54.76**
229	**11.45**	**2622.05**	**131.10**
232	**11.6**	**2691.2**	**134.56**

275	**13.75**	**3781.25**	**189.06**
122	**6.1**	**744.2**	**37.21**
195	**9.75**	**1901.25**	**95.06**
164	**8.2**	**1344.8**	**67.24**
$\Sigma y \ = \ 2108$	$\Sigma x \ = \ 105.4$	$\Sigma xy \ =23470.8$	$\Sigma x^2 \ = \ 1173.53$

$$b_1 \ = \ \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

$$= \ \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$= \ \frac{23470.8 - \frac{(105.4)(2108)}{10}}{1173.53 - \frac{(105.4)^2}{10}}$$

$$= \ \frac{1252.48}{62.61}$$

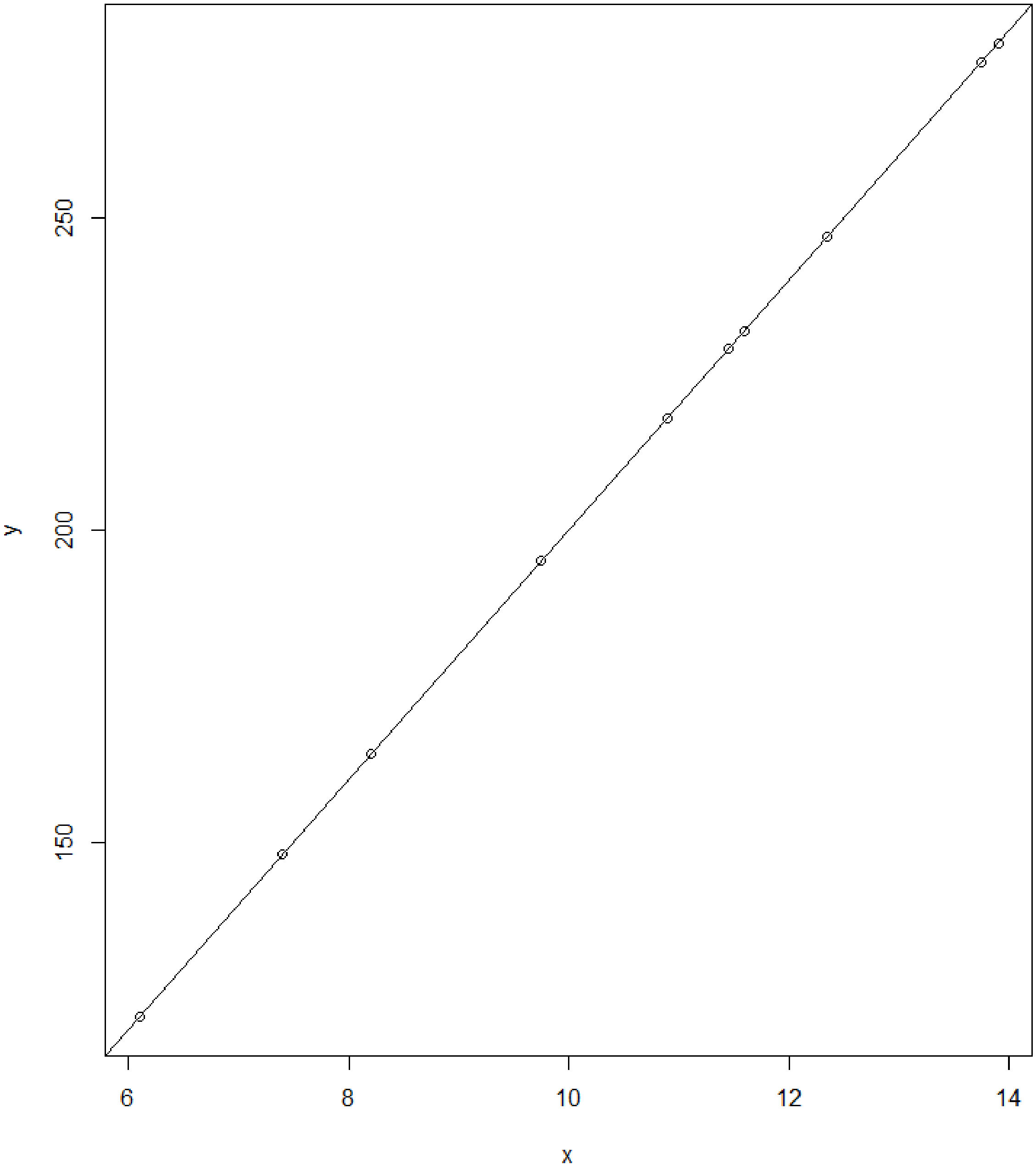
$$= \ 20.00$$

$$b_0 \ = \ \bar{y} \ - \ b_1\bar{x}$$

$$= \ 210.8 \ - \ 20.00(10.54)$$

$$= \ 0$$

Graphical Presentation



Regression equation : $\hat{y} = 20.00x$

$R^2 = +1$

$\therefore$ perfect linear relationship between x and y means 100% of the variation in y is explained by variation in x. Hence, reject H_0 as there is sufficient evidence that hours of study affects total marks

2.4 CHI SQUARE TEST OF INDEPENDENCE

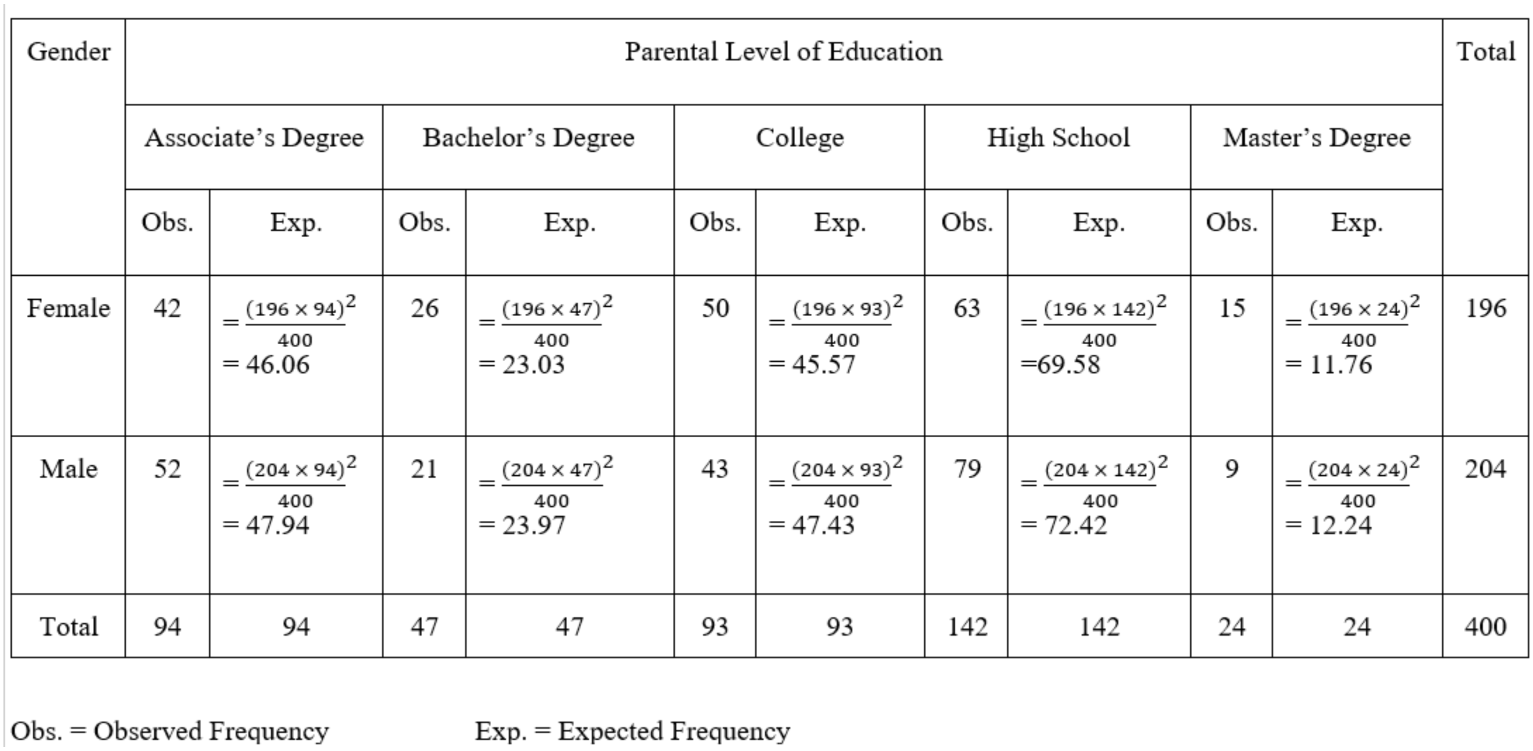
Test Hypothesis :

Null Hypothesis(H_0) = The variables are independent

Alternate Hypothesis(H_1) = Variables are related (Dependent)

Test Statistic :

Gender	Parental Level of Education					Total
	Associate's Degree	Bachelor's Degree	College	High School	Master's Degree	
Female	42	26	50	63	15	196
Male	52	21	43	79	9	204
Total	94	47	93	142	24	400



Gender	Parental Level of Education										Total
	Associate's Degree		Bachelor's Degree		College		High School		Master's Degree		
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	
Female	42	$= \frac{(196 \times 94)^2}{400}$ = 46.06	26	$= \frac{(196 \times 47)^2}{400}$ = 23.03	50	$= \frac{(196 \times 93)^2}{400}$ = 45.57	63	$= \frac{(196 \times 142)^2}{400}$ =69.58	15	$= \frac{(196 \times 24)^2}{400}$ = 11.76	196
Male	52	$= \frac{(204 \times 94)^2}{400}$ = 47.94	21	$= \frac{(204 \times 47)^2}{400}$ = 23.97	43	$= \frac{(204 \times 93)^2}{400}$ = 47.43	79	$= \frac{(204 \times 142)^2}{400}$ = 72.42	9	$= \frac{(204 \times 24)^2}{400}$ = 12.24	204
Total	94	94	47	47	93	93	142	142	24	24	400

Obs. = Observed Frequency Exp. = Expected Frequency

Cell (i,j)	**Observed Count,** o_{ij}	**Expected Count,** e_{ij}	$\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
1,1	42	$= \frac{196 \times 94}{400}$ = 46.06	0.36
1,2	26	$= \frac{196 \times 47}{400}$ = 23.03	0.38
1,3	50	$= \frac{196 \times 142}{400}$ = 45.57	0.43
1,4	63	$= \frac{196 \times 24}{400}$ = 69.58	0.62
1,5	15	$= \frac{196 \times 24}{400}$ = 11.76	0.89
2,1	52	$= \frac{204 \times 94}{400}$ = 47.94	0.34
2,2	21	$= \frac{204 \times 47}{400}$ = 23.97	0.37
2,3	43	$= \frac{204 \times 93}{400}$ = 47.93	0.41
2,4	79	$= \frac{204 \times 142}{400}$ = 72.42	0.60
2,5	9	$= \frac{204 \times 24}{400}$ = 12.24	0.86
		$x^2 =$	5.26

Degree of freedom : (2-1)(4-1) = 4
Confidence interval, α = 0.05
From the chi square table: $x^{2}_{4,\, 0.05}$ = 9.488

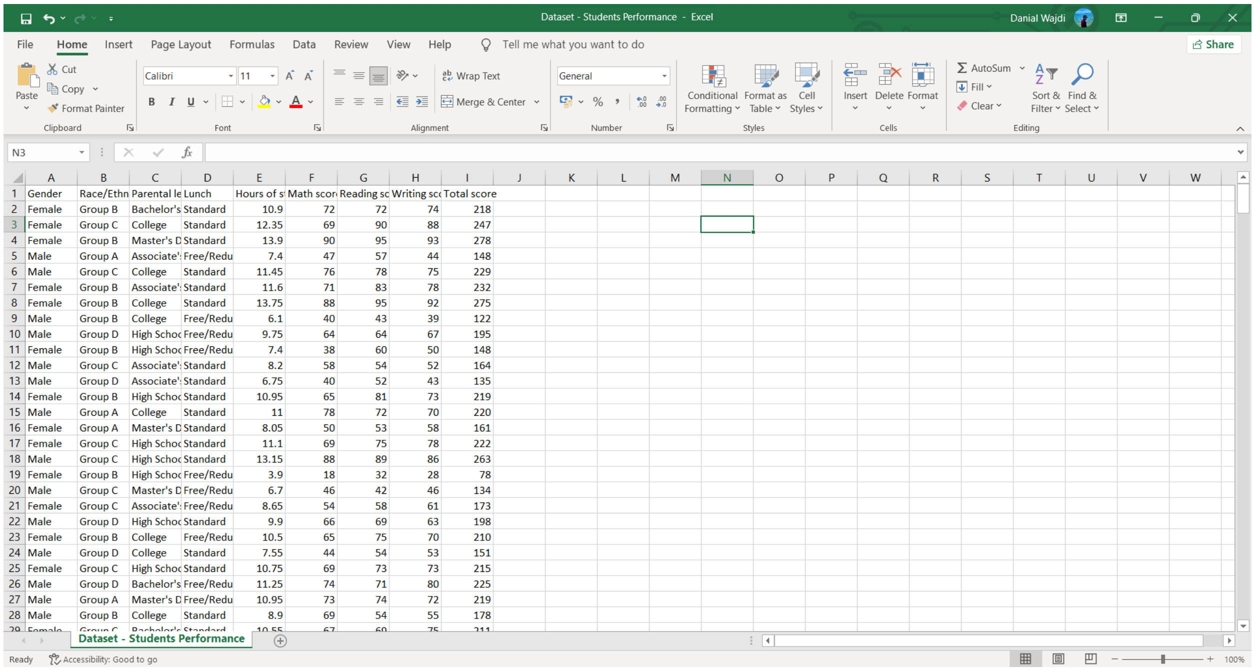
Since the statistic value is less than the critical value(5.26 < 9.488), we fail to reject the null hypothesis. Therefore, there is sufficient evidence that the variable is related(dependent)

Discussion and Result

Appendix A

Link to the dataset:
https://www.kaggle.com/datasets/spscientist/students-performance-in-exams/metadata

Data before data cleaning(Raw data)



	A	B	C	D	E	F	G	H	I
1	Gender	Race/Ethn	Parental le	Lunch	Hours of s	Math scor	Reading sc	Writing sc	Total score
2	Female	Group B	Bachelor's	Standard	10.9	72	72	74	218
3	Female	Group C	College	Standard	12.35	69	90	88	247
4	Female	Group B	Master's D	Standard	13.9	90	95	93	278
5	Male	Group A	Associate's	Free/Redu	7.4	47	57	44	148
6	Male	Group C	College	Standard	11.45	76	78	75	229
7	Female	Group B	Associate's	Standard	11.6	71	83	78	232
8	Female	Group B	College	Standard	13.75	88	95	92	275
9	Male	Group B	College	Free/Redu	6.1	40	43	39	122
10	Male	Group D	High Schoc	Free/Redu	9.75	64	64	67	195
11	Female	Group B	High Schoc	Free/Redu	7.4	38	60	50	148
12	Male	Group C	Associate's	Standard	8.2	58	54	52	164
13	Male	Group D	Associate's	Standard	6.75	40	52	43	135
14	Female	Group B	High Schoc	Standard	10.95	65	81	73	219
15	Male	Group A	College	Standard	11	78	72	70	220
16	Female	Group A	Master's D	Standard	8.05	50	53	58	161
17	Female	Group C	High Schoc	Standard	11.1	69	75	78	222
18	Male	Group C	High Schoc	Standard	13.15	88	89	86	263
19	Female	Group B	High Schoc	Free/Redu	3.9	18	32	28	78
20	Male	Group C	Master's D	Free/Redu	6.7	46	42	46	134
21	Female	Group C	Associate's	Free/Redu	8.65	54	58	61	173
22	Male	Group D	High Schoc	Standard	9.9	66	69	63	198
23	Female	Group B	College	Free/Redu	10.5	65	75	70	210
24	Male	Group D	College	Standard	7.55	44	54	53	151
25	Female	Group C	High Schoc	Standard	10.75	69	73	73	215
26	Male	Group D	Bachelor's	Free/Redu	11.25	74	71	80	225
27	Male	Group A	Master's D	Free/Redu	10.95	73	74	72	219
28	Male	Group B	College	Standard	8.9	69	54	55	178
29	Female	Group C	Bachelor's	Standard	10.55	67	69	75	211