



**UTM**  
UNIVERSITI TEKNOLOGI MALAYSIA

**Razak Faculty of Technology  
and Informatics**

COURSE CODE: MCSD1113

COURSE: STATISTIC FOR DATA SCIENCE

PROGRAMME: MASTER OF DATA SCIENCE

REPORT: GROUP PROJECT

SESSION: SEMESTER 2 SESSION 2022/2023

DATE: 15<sup>th</sup> JULY 2023

<b>Member's Name</b>	1. FAN CHIN WEI 2. ABIR DUTTA 3. ASURUMUNI PASINDU MANISHKA DE SILVA 4. MUHAMAD AFIQ ASYRAF BIN MD ISHAK
<b>Matric No.</b>	1. MCS221024 2. MCS221032 3. MCS221029 4. MCS221026
<b>Lecturer's Name</b>	DR. MOHAMAD SHUKOR BIN TALIB

## **TABLE OF CONTENTS**

<b>TITLE</b>	<b>PAGE</b>
<b>TABLE OF CONTENTS</b>	<b>1</b>
<b>1.0 Introduction</b>	<b>2</b>
<b>2.0 Data Collection</b>	<b>2</b>
<b>3.0 Data Analysis</b>	<b>4</b>
<b>3.1 Exploratory Data Analysis</b>	<b>4</b>
<b>3.2 Hypothesis Testing</b>	<b>7</b>
<b>3.2.1 Hypothesis Testing 1</b>	<b>7</b>
<b>3.2.2 Hypothesis Testing 2</b>	<b>8</b>
<b>3.2.3 Hypothesis Testing 3</b>	<b>8</b>
<b>3.3 Goodness of Fit Test</b>	<b>9</b>
<b>3.4 Chi-Square Test of Independence</b>	<b>10</b>
<b>3.4.1 Gender vs Math Score</b>	<b>11</b>
<b>3.4.2 Gender vs Reading Score</b>	<b>12</b>
<b>3.4.3 Gender vs Writing Score</b>	<b>12</b>
<b>3.4.4 Parental Level of Education vs Average Score</b>	<b>13</b>
<b>3.4.5 Race vs Average Score</b>	<b>13</b>
<b>3.5 Correlation</b>	<b>14</b>
<b>3.6 Regression</b>	<b>16</b>
<b>3.7 ANOVA</b>	<b>18</b>
<b>4.0 Conclusion</b>	<b>24</b>
<b>APPENDIX</b>	<b>26</b>

## 1.0 Introduction

Students are getting set to return to campus now that the Covid-19 pandemic has subsided. However, we know that there is a significant difference in student performance before and during the pandemic in terms of teaching and learning. Some things can have an impact on a student's performance. It is not only how much a student studies or understands that determines how well he or she scores on an exam. External influences can also have an impact on exam performance. For example, if the student's parents are well educated, the student may receive extra tutoring outside of school hours from his parents, resulting in the student performing well in exams. In this study, the factors that may influence student performance are examined in order to improve the students' exam performance. In addition, the effectiveness of exam preparation will be evaluated. As a student, we should be aware of the aspects that influence performance in order to improve our performance on the exam and pass with flying colors. As a result, the student performance in exams dataset is chosen to investigate the component that influenced the performance outcome.

## 2.0 Data Collection

The dataset was collected and obtained from Kaggle in CSV files which is 'StudentsPerformance.csv'. Generally, the dataset provides students information and their marks in various subjects. It consists of students' gender which are male or female, race and ethnicity of the students that are grouped to 5 groups which are group A, group B, group C, group D and group E, and level of education of students' parents which are some college, associate's degree, high school, some high school, bachelor's degree and master degree. It also consists of information on the type of lunch the students have, which are standard or free/reduced and whether the students complete the test preparation course or not. In addition, the dataset includes students' scores for math, writing and reading.

Data source: <https://www.kaggle.com/code/spscientist/student-performance-in-exams/notebook>

No	Variables	Type	Description
1	gender	Qualitative,Nominal	This attribute indicates the gender of each student, whether they are male or female
2	race	Qualitative,Nominal	This attribute categorizes the students into different racial or ethnic groups, such as group A, B, C, D, etc
3	parental_level_of_education	Qualitative,Nominal	This attribute represents the educational attainment of the students' parents or guardians. It includes categories like bachelor's degree, master's degree, associate's degree, high school, etc.
4	lunch	Qualitative,Nominal	This attribute describes the type of lunch the students have, indicating whether it is a standard lunch or a lunch provided at a reduced cost (free/reduced)
5	test_preparation_course	Qualitative,Nominal	This attribute indicates whether the student completed a test preparation course or not. It specifies if they received additional preparation before taking the exams
6	math_score	Quantitative, Ratio	This attribute represents the score the student achieved in the math subject
7	reading_score	Quantitative, Ratio	This attribute represents the score the student achieved in the reading subject
8	writing_score	Quantitative, Ratio	This attribute represents the score the student achieved in the writing subject

Table 1: Data Description

### 3.0 Data Analysis

#### 3.1 Exploratory Data Analysis

Exploratory data analysis (EDA) was performed on the 'StudentsPerformance.csv' dataset to gain initial understandings and insights into the dataset. The dataset was loaded as 'student'. Then, an overview of the data was obtained. The first few rows were examined to get a glimpse of the variables, and their values as depicted in figure 1. Then, the structure of the data was observed to get information on number of rows, columns, and data type. As depicted in figure 1, the data has 100 observations (rows) and 8 variables (columns) whereas for the data type, 5 variables are characters, and 3 variables are integers. For missing values in dataset, since the output displays 0, it means that there are no missing values in the dataset.

```
> head(student)
  gender race.ethnicity parental.level.of.education lunch test.preparation.course math.score reading.score writing.score
1 female      group B      bachelor's degree      standard                none          72          72          74
2 female      group C          some college      standard                completed        69          90          88
3 female      group B          master's degree      standard                none          90          95          93
4 male        group A      associate's degree free/reduced                none          47          57          44
5 male        group C          some college      standard                none          76          78          75
6 female      group B      associate's degree      standard                none          71          83          78

> str(student)
'data.frame': 1000 obs. of 8 variables:
 $ gender      : chr  "female" "female" "female" "male" ...
 $ race.ethnicity: chr  "group B" "group C" "group B" "group A" ...
 $ parental.level.of.education: chr  "bachelor's degree" "some college" "master's degree" "associate's degree" ...
 $ lunch        : chr  "standard" "standard" "standard" "free/reduced" ...
 $ test.preparation.course : chr  "none" "completed" "none" "none" ...
 $ math.score   : int   72 69 90 47 76 71 88 40 64 38 ...
 $ reading.score: int   72 90 95 57 78 83 95 43 64 60 ...
 $ writing.score : int   74 88 93 44 75 78 92 39 67 50 ...

> mv_df<-sum(is.na(student))
> mv_df
[1] 0
```

Figure 1: Data Overview.

Next, each variable in the dataset was explored. Firstly, as shown in figure 2 and figure 3, for gender, most of the students are female with 518 students (51.8%) in comparison with male with 482 (48.2%). For race and ethnicity, the majority of the students are in group C with 319 students (31.9%) followed by group D with 262 students (26.2%), group B with 190 students (19%), group E with 140 students (14%) and group A with 89 students (8.9%). For level of education of parents, many parents of the students have some college with 226 people (22.6%), followed by associate's degree with 222 people (22.2%), high school with 196 people (19.6%), some high school with 179 people (17.9%), bachelor's degree with 118 people (11.8%) and master degree with 59 people (5.9%). Moreover, most of the students have standard lunch with 645

students (645%) in comparison with free or reduced lunch with 355 (35.5%). For test preparation course, most of the students did not complete the course with 642 students (64.2%) in comparison with students that completed the course with 358 students (35.8%).

```
> gender_counts
female  male
   518   482

> race_counts
group A group B group C group D group E
    89    190    319    262    140

> education_counts
associate's degree  bachelor's degree  high school  master's degree  some college
               272                118             196                59             226
some high school
               179

> lunch_counts
free/reduced  standard
           355          645

> prep_counts
completed  none
          358          642
```

Figure 2: Number of students for each variable.

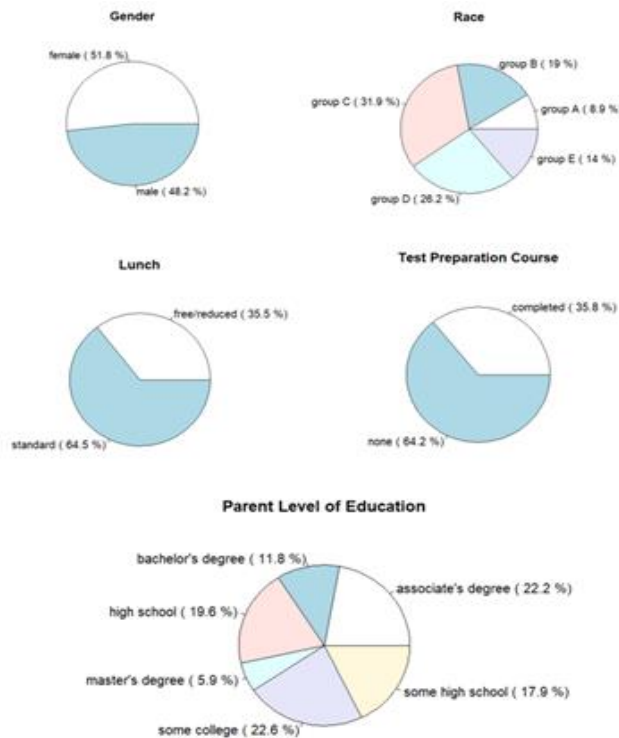


Figure 3: Pie charts for each variable.

For math score, reading score and writing score, descriptive statistics were obtained as shown in figure 4 and 5. For math score, the lowest score is 0, the highest score is 100, the median is 66, and the mean is 66.09. In addition, for math scores, most students score between 60 to 70. For reading score, the lowest score is 17, the highest score is 100, the median is 70, and the mean is 69.17. In addition, for reading scores, most students score between 70 to 80. For writing score, the lowest score is 10, the highest score is 100, the median is 69, and the mean is 68.05. In addition, for writing scores, most students score between 70 to 80.

```
> summary(student$math.score)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.00   57.00   66.00   66.09   77.00   100.00

> summary(student$reading.score)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 17.00   59.00   70.00   69.17   79.00   100.00

> summary(student$writing.score)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 10.00   57.75   69.00   68.05   79.00   100.00
```

Figure 5: Descriptive statistics.

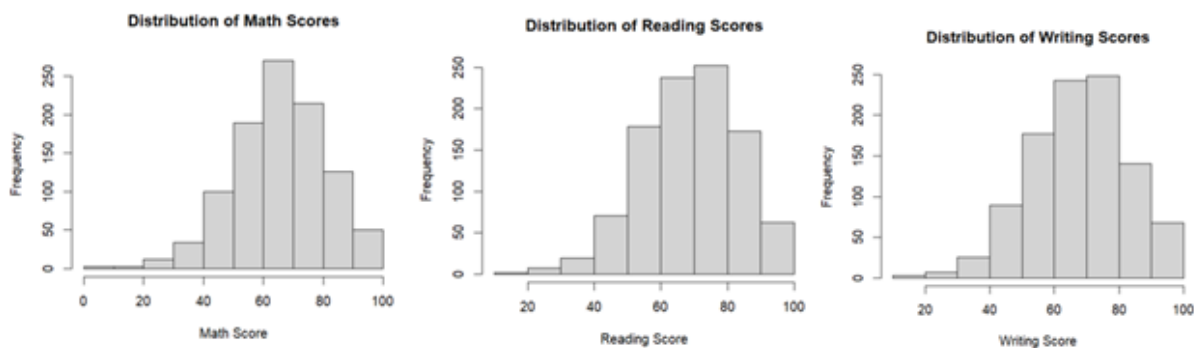


Figure 6: Histogram for math scores, reading scores and writing scores.

## 3.2 Hypothesis Testing

### 3.2.1 Hypothesis Testing 1

The hypothesis testing 1 is the average mathematical score of male students is greater than 65 marks. Is there enough evidence to conclude that the average mathematical score of male students is greater than 62 marks? The hypothesis of this scenario is as shown in figure 7.

$$H_0: \mu = 65 \text{ and } H_1: \mu > 65 \text{ (claim)}$$

```
# Hypothesis Testing 1
Male_student <- student$math_score[student$gender=="male"]
sd(Male_student)
library(BSDA)
z.test(Male_student,mu=65, sigma.x=14.36, alt="greater", conf.level=0.95)
```

Figure 7: Hypothesis testing 1.

As shown in figure 8, since the P-value is less than 0.05, the decision is to reject null hypothesis. There is enough evidence to support the claim that the average mathematical score of male students is greater than 62 marks.

```
one-sample z-Test

data:  Male_student
z = 5.6999, p-value = 5.993e-09
alternative hypothesis: true mean is greater than 65
95 percent confidence interval:
 67.65235      NA
sample estimates:
mean of x
 68.72822
```

Figure 8: Output hypothesis testing 1.



### 3.2.2 Hypothesis Testing 2

The hypothesis testing 2 is that the average writing score of both male and female students are less than 62 marks. Is there enough evidence to conclude that the average writing score of male and female students are less than 62 marks? The hypothesis is as shown in figure 9.

$$H_0: \mu = 62 \text{ and } H_1: \mu < 62 \text{ (claim)}$$

```
# Hypothesis Testing 2  
  
sd(student$writing_score)  
library(BSDA)  
z.test(student$writing_score,mu=62, sigma.x=15.20, alt="less", conf.level=0.95)
```

Figure 9: Hypothesis testing 2.

As shown in figure 10, since the P-value is more than 0.05, the null hypothesis is failed to reject. There is not enough evidence to support the claim that the average writing score of male and female students are less than 62 marks.

```
One-sample z-Test  
  
data: student$writing_score  
z = 12.595, p-value = 1  
alternative hypothesis: true mean is less than 62  
95 percent confidence interval:  
NA 68.84463  
sample estimates:  
mean of x  
68.054
```

Figure 10: Output hypothesis testing 2.

### 3.2.3 Hypothesis Testing 3

The hypothesis testing 3 is the average reading score for those who do not have done a test preparation course may not be 75 marks. Is there enough evidence to conclude that the average reading score for those who do not have done a test preparation course may not be 75 marks? The hypothesis is as shown in figure 11.

$H_0: \mu = 75$  and  $H_1: \mu \neq 75$  (claim)

```
# Hypothesis Testing 3

student_no_prepare_course <- student$reading_score[student$test_preparation_course=="none"]
sd(student_no_prepare_course)
library(BSDA)
z.test(student_no_prepare_course,mu=75, sigma.x=14.46, alt="two.sided", conf.level=0.95)
```

Figure 11: Hypothesis testing 3.

As shown in figure 12, since the P-value is less than 0.05, the decision is to reject null. There is enough evidence to support the claim that the average reading score for those who do not have done a test preparation course is not 75 marks.

```
one-sample z-Test

data:  student_no_prepare_course
z = -14.834, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 65.41573 67.65280
sample estimates:
mean of x
 66.53427
```

Figure 12: Output hypothesis testing 3.

### 3.3 Goodness of Fit Test

In this report, goodness of fit test was performed on race and ethnicity data to test whether the observed frequencies significantly deviate from the expected distribution which is equal distribution at  $\alpha = .05$ . The hypothesis is as followed:

$H_0$  : The observed frequencies for race and ethnicity data follow the expected distribution which is equal distribution.

$H_1$  : The observed frequencies for race and ethnicity data do not follow the expected distribution which is equal distribution.

The test was being performed as figure 13.

```
# Define expected probabilities
expected_probs <- rep(0.20, 5)

# Calculate observed frequencies
observed_freqs <- table(student$race.ethnicity)

# Perform chi-square test
chi_sq_test <- chisq.test(observed_freqs, p = expected_probs)

# Print test results
print(chi_sq_test)
```

Figure 13: Goodness of fit test on race and ethnicity data.

The output of the test is as figure 14:

```
Chi-squared test for given probabilities

data:  observed_freqs
X-squared = 170.13, df = 4, p-value < 2.2e-16
```

Figure 14: Output for Goodness of fit test.

The test statistic (X-squared) is calculated as 170.13 with 4 degrees of freedom. The p-value is less than  $2.2e-16$ , indicating a very small value. According to the result of the goodness of fit test, the null hypothesis ( $H_0$ ) is rejected since the p-value being significantly less than the significance level (.05), providing strong evidence to support the alternative hypothesis ( $H_1$ ).

### 3.4 Chi-Square Test of Independence

A chi-square test of independence is a statistical test that is used to determine whether there is a significant relationship between two categorical variables. It is useful to assess whether the occurrence of one variable is related to the occurrence of another variable, or if they are independent of each other. It is based on the principle of comparing the observed frequencies in a contingency table with the expected frequencies that would be expected under the assumption of independence between variables. The steps below depict how to perform a chi-square test of independence.

1. Form the Null Hypothesis ( $H_0$ ) and the Alternative Hypothesis ( $H_A$ ):

- a.  $H_0$  assumes that there is no association between the variables.
  - b.  $H_A$  suggests that there is a significant association between the variables.
2. Form the Contingency Table,
3. Calculate the expected frequencies.
4. Obtain the test statistics.
5. Obtain the P-value.
6. Interpret the result.
  - a. If the p-value is below the predetermined significance level (0.05), the  $H_0$  is rejected suggesting a significant relationship between the variables.
  - b. If the p-value is above the significance level, the  $H_0$  is not rejected, implying that there is insufficient evidence to conclude significant association.

The chi-square test of independence was performed on the following pairs of variables:

1. Gender vs Math Score
2. Gender vs Reading Score
3. Gender vs Writing Score
4. Parental Level of Education vs Average Score
5. Race vs Average Score

#### **3.4.1 Gender vs Math Score**

The test of independence was performed on the 'gender' variable versus the 'math\_score' variable and the following output was obtained as shown in figure 15. From the result, with a significance level of 0.05, it could be seen that the p-value of 0.1475 is greater than the significance

level. As a result, there is not enough evidence to state that there is a significant association between the two variables.

```
> result_gender_math

Pearson's Chi-squared test

data:  contingency_table_gender_math
X-squared = 93.257, df = 80, p-value = 0.1475
```

Figure 15: Output for gender vs math score.

### 3.4.2 Gender vs Reading Score

The test of independence was performed on the '*gender*' variable versus the '*reading\_score*' variable and the following output was obtained as shown in figure 16. From the result, with a significance level of 0.05, it could be seen that the p-value of 1.427e-06 is significantly lesser than the significance level. As a result, there is a significant association between the two variables. Therefore, there is an effect of gender when compared to the reading scores that were earned by students.

```
> result_gender_reading

Pearson's Chi-squared test

data:  contingency_table_gender_reading
X-squared = 141.28, df = 71, p-value = 1.427e-06
```

Figure 16: Output for gender vs reading score.

### 3.4.3 Gender vs Writing Score

The test of independence was performed on the '*gender*' variable versus the '*writing\_score*' variable and the following output was obtained as shown in figure 17. From the result, with a significance level of 0.05, it could be seen that the p-value of 2.206-08 is significantly lesser than the significance level. As a result, there is a significant association between the two variables. Therefore, there is an effect of gender when compared to the writing scores that were earned by students.

```
> result_gender_writing

Pearson's Chi-squared test

data:  contingency_table_gender_writing
X-squared = 163.75, df = 76, p-value = 2.206e-08
```

Figure 17: Output for gender vs writing score.

#### 3.4.4 Parental Level of Education vs Average Score

The test of independence was performed on the '*parental\_level\_of\_education*' variable versus the '*average\_score*' variable. The '*average\_score*' variable was calculated using the mean of '*math\_score*', '*reading\_score*' and '*writing\_score*'. From the result as shown in figure 18, with a significance level of 0.05, it could be seen that the p-value obtained is 0.1659. Therefore, there is not enough evidence to suggest that there is an effect of parents' education on the child's scores.

```
> result_education_avg

Pearson's Chi-squared test

data:  contingency_table_education_avg
X-squared = 1007.6, df = 965, p-value =
```

Figure 18: Output for education vs average score.

#### 3.4.5 Race vs Average Score

The test of independence was performed on the '*race*' variable versus the '*average\_score*' variable. The '*average\_score*' variable was calculated using the mean of '*math\_score*', '*reading\_score*' and '*writing\_score*'. From the result as shown in figure 19, with a significance level of 0.05, it could be seen that the p-value obtained is 0.05381. Therefore, there is not enough evidence to suggest that there is an effect of race on a child's scores.

```
> result_race_avg

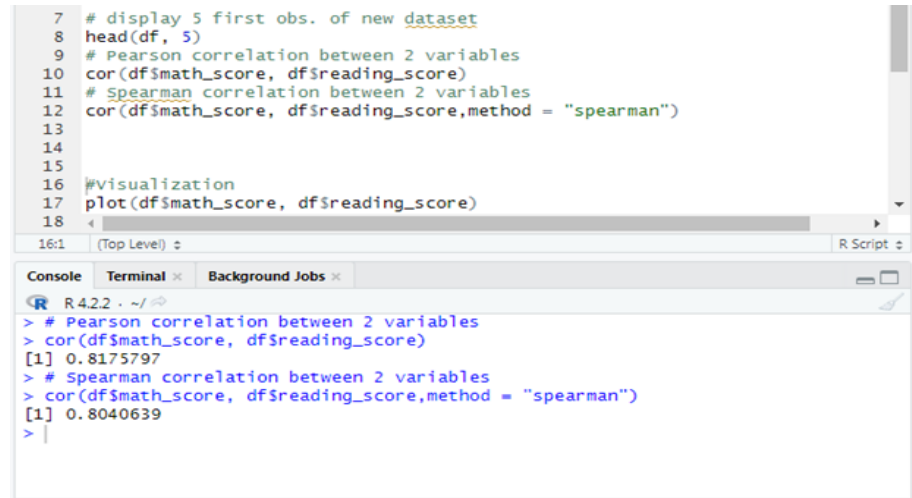
Pearson's Chi-squared test

data:  contingency_table_race_avg
X-squared = 836.26, df = 772, p-value = 0.05381
```

Figure 19: Output for race vs average score.

### 3.5 Correlation

The Pearson correlation and the Spearman test was performed with two variables 'math score' and 'reading score' and the correlation obtained from each method is presented in figure 20.



```
7 # display 5 first obs. of new dataset
8 head(df, 5)
9 # Pearson correlation between 2 variables
10 cor(df$math_score, df$reading_score)
11 # Spearman correlation between 2 variables
12 cor(df$math_score, df$reading_score, method = "spearman")
13
14
15
16 #Visualization
17 plot(df$math_score, df$reading_score)
18
```

```
R 4.2.2 ~ /
> # Pearson correlation between 2 variables
> cor(df$math_score, df$reading_score)
[1] 0.8175797
> # Spearman correlation between 2 variables
> cor(df$math_score, df$reading_score, method = "spearman")
[1] 0.8040639
> |
```

Figure 20: Output for Pearson correlation and Spearman test.

Both the results from the Pearson and Spearman correlation are close to each other. The values are 0.81 and 0.804, which shows a strong correlation between the variables. Then, the visualization of the two variables and multiple variables was performed. The visualization is represented in figure 21.

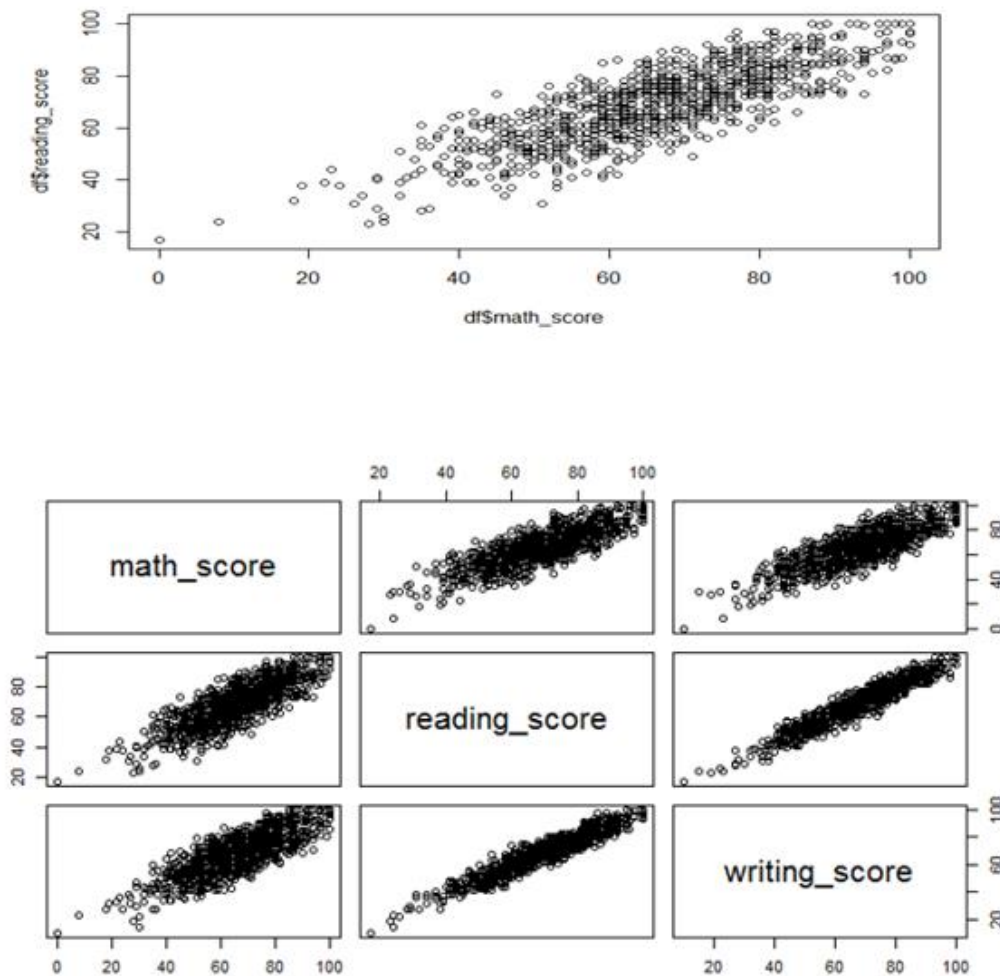


Figure 21: Visualization for Pearson correlation and Spearman test.

The visualizations show the strong correlation between the variables where all the data points follow nearly straight line. Then, the correlation test is performed between the variables and the P value is observed as shown in figure 22.

```
Pearson's product-moment correlation
data: df$math_score and df$reading_score
t = 44.855, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7959276 0.8371428
sample estimates:
cor
0.8175797
```

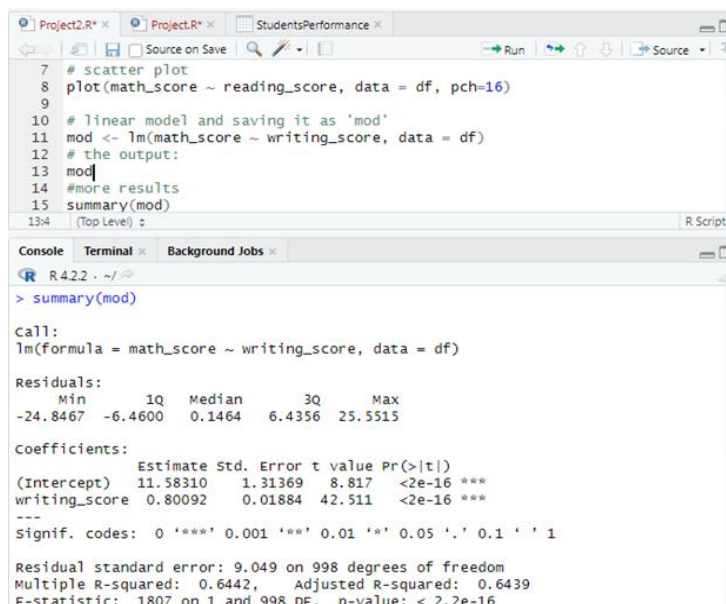
Figure 22: Output for correlation test.



As  $P > 0$ , which is the indication of strong correlation. With the 95% confidence level, the hypothesis test which represented the alternate hypothesis is true. So strong correlation is presented between the variables.

### 3.6 Regression

For the regression analysis, a linear model was built for the variables after visualizing the scatter plot of the variables.



```

7 # scatter plot
8 plot(math_score ~ reading_score, data = df, pch=16)
9
10 # linear model and saving it as 'mod'
11 mod <- lm(math_score ~ writing_score, data = df)
12 # the output:
13 mod
14 #more results
15 summary(mod)

```

```

R 4.2.2 ~. /
> summary(mod)

call:
lm(formula = math_score ~ writing_score, data = df)

Residuals:
    Min       1Q   Median       3Q      Max
-24.8467  -6.4600   0.1464   6.4356  25.5515

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.58310    1.31369   8.817  <2e-16 ***
writing_score  0.80092    0.01884  42.511  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.049 on 998 degrees of freedom
Multiple R-squared:  0.6442,    Adjusted R-squared:  0.6439
F-statistic: 1807 on 1 and 998 DF,  p-value: < 2.2e-16

```

Figure 23: Output for linear model.

As shown in figure 23, the summary of our linear model with the different indicators of the model was obtained. The Residual standard error is 0.6442. The Multiple and Adjusted R squared values are 0.6442 and 0.6439 which is not below the 0% zone. This represents a strong linear regression between the variables. After that combination of regression with scatter plot is executed and the fitted line is plotted. The observation is demonstrated as shown in figure 24.

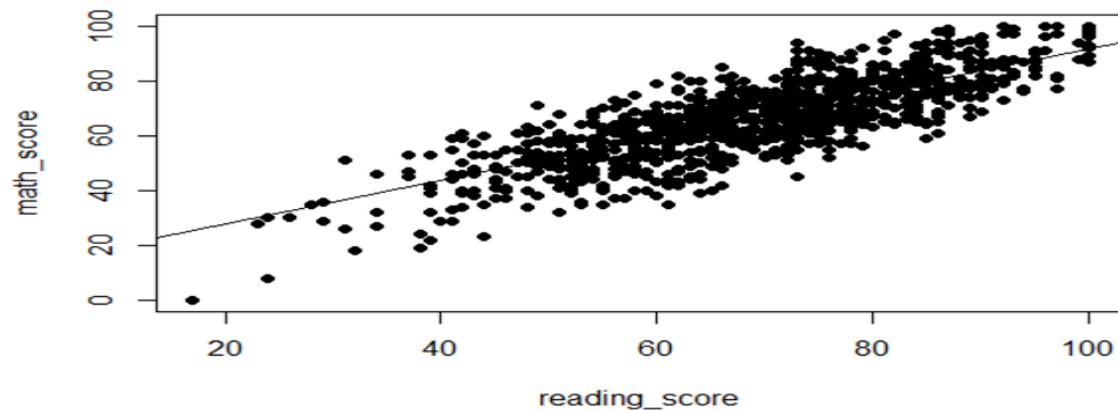


Figure 24: Output for regression with scatter plot.

The fitted line is observed around the variables data point. The observed values are almost fitted around the straight line. This indicates a strong relationship between the variables. Then, the regression table with the corresponding values to observe our linear model regression values was checked.

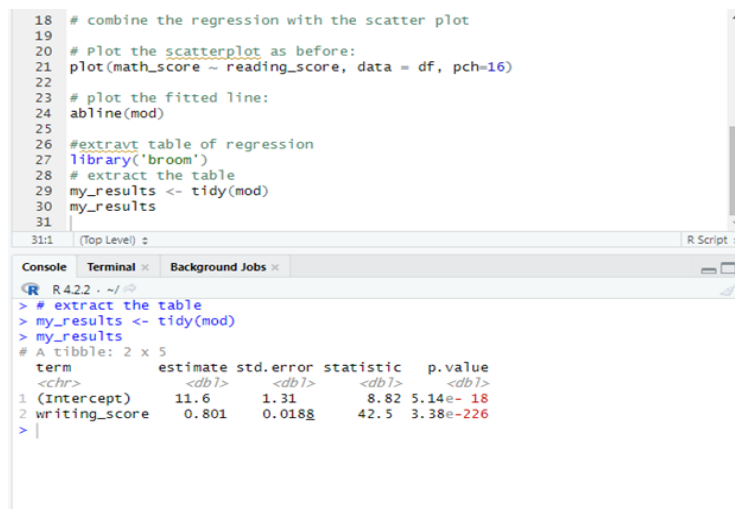


Figure 25: Output for regression table.

As shown in figure 25, the P value clearly indicates that there is a strong relationship obtained from the linear regression between the variables.

### 3.7 ANOVA

A one-way analysis of variance (ANOVA) is performed to determine if there are any statistically significant differences in the means of three or more independent variables. First of all, factors are frequently interpreted as quantitative variables when importing a dataset into R software. To overcome this issue, apply the function of `read.csv()` command to read in the data, indicating whether each variable should be quantitative (numerical) or categorical (factor) within the command as shown in figure 26.

```
> #student <- read.csv("StudentsPerformance.csv", header=TRUE, colClasses = c("factor", "factor", "factor", "factor", "factor", "numeric", "numeric", "numeric"), quote = "\"")
> student <- read.csv("StudentsPerformance.csv", header = TRUE,
+                   colClasses = c("factor", "factor", "factor", "factor", "factor", "character", "character", "character"),
+                   quote = "\"")
> student[, 6:8] <- lapply(student[, 6:8], function(x) as.numeric(as.character(x)))
> summary(student)
```

gender	race	parental_level_of_education	lunch	test_preparation_course
female:518	group A: 89	associate's degree:222	free/reduced:355	completed:358
male :482	group B:190	bachelor's degree :118	standard :645	none :642
	group C:319	high school :196		
	group D:262	master's degree : 59		
	group E:140	some college :226		
		some high school :179		

math_score	reading_score	writing_score
Min. : 0.00	Min. : 17.00	Min. : 10.00
1st Qu.: 57.00	1st Qu.: 59.00	1st Qu.: 57.75
Median : 66.00	Median : 70.00	Median : 69.00
Mean : 66.09	Mean : 69.17	Mean : 68.05
3rd Qu.: 77.00	3rd Qu.: 79.00	3rd Qu.: 79.00
Max. :100.00	Max. :100.00	Max. :100.00

Figure 26: Output 1 for ANOVA.

According to figure 26, `gender`, `race`, `parental_level_of_education`, `lunch`, and `test_preparation_course` variables are classified as categorical variables while `math_score`, `reading_score` and `writing_score` are classified as quantitative variables.

```
> one.way <- aov(math_score ~ gender, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gender	1	6481	6481	28.98	9.12e-08 ***
Residuals	998	223208	224		

---  
signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Figure 27: Output 2 for ANOVA.

Based on figure 27, the math\_score and gender data is analyzed. From the result, The gender variable has a low p-value which is less than 0.001. As a result, it appears that gender has a real impact on the math\_score.

```
> one.way <- aov(reading_score ~ gender, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gender	1	12711	12711	63.35	4.68e-15 ***
Residuals	998	200242	201		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 28: Output 3 for ANOVA.

Based on figure 28, the reading\_score and gender data is analyzed. From the result, the gender variable has a low p-value which is less than 0.001. As a result, it appears that gender has a real impact on the reading\_score.

```
> one.way <- aov(writing_score ~ gender, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gender	1	20931	20931	99.59	<2e-16 ***
Residuals	998	209746	210		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 29: Output 4 for ANOVA.

Based on figure 29, the writing\_score and gender data is analyzed. From the result, The gender variable has a low p-value which is less than 0.001. As a result, it appears that gender has a real impact on the writing\_score.

```
> one.way <- aov(math_score ~ race, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
race	4	12729	3182	14.59	1.37e-11 ***
Residuals	995	216960	218		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 30: Output 5 for ANOVA.

Based on figure 30, the math\_score and race data is analyzed. From the result, The race variable has a low p-value which is less than 0.001. As a result, it appears that race has an effect on the math\_score.

```
> one.way <- aov(reading_score ~ race, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
race	4	4706	1176.6	5.622	0.000178 ***
Residuals	995	208246	209.3		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 31: Output 5 for ANOVA.

Based on figure 31, the reading\_score and race data is analyzed. From the result, The race variable has a low p-value which is less than 0.001. As a result, it appears that race has a significant effect on the reading\_score.

```
> one.way <- aov(writing_score ~ race, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
race	4	6456	1614.0	7.162	1.1e-05 ***
Residuals	995	224221	225.3		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 32: Output 6 for ANOVA.

Based on figure 32, the ANOVA test is applied to estimate writing\_score variables changes according to the race data. From the result, the p-value of the race variable is low, which is less than 0.001, so that it appears that the type of race has a significant impact on writing\_score.

```
> one.way <- aov(math_score ~ parental_level_of_education, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
parental_level_of_education	5	7296	1459.1	6.522	5.59e-06 ***
Residuals	994	222394	223.7		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 33: Output 7 for ANOVA.

Based on figure 33, the ANOVA test is applied to estimate math\_score variables changes according to the parental level of education data. From the result, the p-value of the parental level of education variable is low, which is less than 0.001, so that it appears that the type of parental level of education has a significant impact on math\_score.

```
> one.way <- aov(reading_score ~ parental_level_of_education, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
parental_level_of_education	5	9506	1901.3	9.289	1.17e-08 ***
Residuals	994	203446	204.7		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 34: Output 8 for ANOVA.

Based on figure 34, the ANOVA test is applied to estimate reading\_score variables changes according to the parental level of education data. From the result, the p-value of the parental level of education variable is low, which is less than 0.001, so that it appears that the type of parental level of education has a significant impact on reading\_score.

```
> one.way <- aov(writing_score ~ parental_level_of_education, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
parental_level_of_education	5	15623	3124.6	14.44	1.12e-13 ***
Residuals	994	215054	216.4		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 35: Output 9 for ANOVA.

Based on figure 35, the ANOVA test is applied to estimate writing\_score variables changes according to the parental level of education data. From the result, the p-value of the parental level of education variable is low, which is less than 0.001, so that it appears that the type of parental level of education has an impact on writing\_score.

```
> one.way <- aov(math_score ~ lunch, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lunch	1	28278	28278	140.1	<2e-16 ***
Residuals	998	201411	202		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 36: Output 10 for ANOVA.

Based on figure 36, the ANOVA test is applied to estimate math\_score variables changes according to the lunch data. From the result, the p-value of the lunch variable is low, which is less than 0.001, so that it appears that the type of lunch has a real impact on math\_score.

```
> one.way <- aov(writing_score ~ lunch, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lunch	1	13933	13933	64.16	3.19e-15 ***
Residuals	998	216744	217		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 37: Output 11 for ANOVA.

Based on figure 37, the ANOVA test is applied to estimate writing\_score variables changes according to the lunch data. From the result, the p-value of the lunch variable is low, which is less than 0.001, so that it appears that the type of lunch has a real impact on writing\_score.

```
> one.way <- aov(reading_score ~ lunch, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lunch	1	11222	11222	55.52	2e-13 ***
Residuals	998	201730	202		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 38: Output 12 for ANOVA.

Based on figure 38, the ANOVA test is applied to estimate reading\_score variables changes according to the lunch data. From the result, the p-value of the lunch variable is low, which is less than 0.001, so that it appears that the type of lunch has a real impact on reading\_score.

```
> one.way <- aov(math_score ~ test_preparation_course, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
test_preparation_course	1	7253	7253	32.54	1.54e-08 ***
Residuals	998	222436	223		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 39: Output 13 for ANOVA.

Based on figure 39, the ANOVA test is applied to estimate math\_score variables changes according to the test preparation course data. From the result, the p-value of the test preparation course variable is low, which is less than 0.001, so that it appears that the type of test preparation course has a real impact on math\_score.

```
> one.way <- aov(writing_score ~ test_preparation_course, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
test_preparation_course	1	22591	22591	108.4	<2e-16 ***
Residuals	998	208086	209		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 40: Output 14 for ANOVA.

Based on figure 40, the ANOVA test is applied to estimate writing\_score variables changes according to the test preparation course data. From the result, the p-value of the test preparation



course variable is low, which is less than 0.001, so that it appears that the type of test preparation course has a real impact on writing\_score.

```
> one.way <- aov(reading_score ~ test_preparation_course, data = student)
> summary(one.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
test_preparation_course	1	12449	12449	61.96	9.08e-15 ***
Residuals	998	200504	201		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 41: Output 15 for ANOVA.

Based on figure 41, the ANOVA test is applied to estimate reading\_score variables changes according to the test preparation course data. From the result, the p-value of the test preparation course variable is low, which is less than 0.001, so that it appears that the type of test preparation course has a real impact on reading\_score.

## 4.0 Conclusion

In this report, a variety of methods to analyze the student performance dataset were used. These methods included exploratory data analysis, hypothesis testing, ANOVA, goodness of fit, correlation, and regression.

Exploratory data analysis helped us to get a better understanding of the data. We were able to identify the distributions of the data, the relationships between the variables, and the outliers. Hypothesis testing allowed us to test specific claims about the data. We were able to test whether there was a significant difference between the means of two groups, whether there was a significant correlation between two variables, and whether a particular model fit the data well. ANOVA allowed us to test whether there was a significant difference between the means of three or more groups. We were able to use ANOVA to test whether there was a difference in the performance of students in different classes, different genders, or different levels of academic achievement. Goodness of fit allowed us to assess how well a particular model fits the data. We were able to use goodness of fit to test whether a normal distribution, a binomial distribution, or a Poisson distribution provided a good fit to the data. Correlation allowed us to assess the strength of the

relationship between two variables. We were able to use correlation to determine whether there was a positive or negative correlation between two variables, and whether the correlation was statistically significant. Regression allowed us to predict the value of one variable based on the value of another variable. We were able to use regression to predict the performance of students in a future test based on their performance in a previous class test.

The results of our analysis showed that there were a number of factors that influence student performance. These factors included the gender of the student, the level of academic achievement of the student, and the class that the student was taking. We also found that there was a positive correlation between the performance of a student in a previous class test and their performance in a future class test. The methods that we used in this report were effective in helping us to understand the factors that influence student performance. Moreover, these methods can be used to help educators to improve the performance of their students.

## APPENDIX

### ***R* Script**

```
#import packages

library(dplyr)

library(stats)

library(ggplot2)

library(tidyverse)

# Read csv file

student <- read.csv("StudentsPerformance.csv")

summary(student)

# Observe data

head(student)

View(student)

# Check dataset structure

str(student)

#missing values

mv_df<-sum(is.na(student))

mv_df

#Exploratory data analysis

# Gender

gender_counts <- table(student$gender)

gender_counts

gender_percentages <- round(100 * gender_counts / sum(gender_counts), 2)

pie(gender_percentages, labels = paste(names(gender_percentages), "(", gender_percentages, "%)",

    main = "Gender")

# Race

race_counts <- table(student$race.ethnicity)

race_counts
```

```

race_percentages <- round(100 * race_counts / sum(race_counts), 2)
pie(race_percentages, labels = paste(names(race_percentages), "(", race_percentages, "%)"),
    main = "Race")

# Parent Level of Education
education_counts <- table(student$parental.level.of.education)
education_counts
edu_percentages <- round(100 * education_counts / sum(education_counts), 2)
pie(edu_percentages, labels = paste(names(edu_percentages), "(", edu_percentages, "%)"),
    main = "Parent Level of Education")

# Lunch
lunch_counts <- table(student$lunch)
lunch_counts
lunch_percentages <- round(100 * lunch_counts / sum(lunch_counts), 2)
pie(lunch_percentages, labels = paste(names(lunch_percentages), "(", lunch_percentages, "%)"),
    main = "Lunch")

# Test Preparation Course
prep_counts <- table(student$test.preparation.course)
prep_counts
prep_percentages <- round(100 * prep_counts / sum(prep_counts), 2)
pie(prep_percentages, labels = paste(names(prep_percentages), "(", prep_percentages, "%)"),
    main = "Test Preparation Course")

# Math Score
summary(student$math.score)
hist(student$math.score, xlab = "Math Score", ylab = "Frequency",
    main = "Distribution of Math Scores")

# Reading Score
summary(student$reading.score)
hist(student$reading.score, xlab = "Reading Score", ylab = "Frequency",
    main = "Distribution of Reading Scores")

```

```

# Writing Score
summary(student$writing.score)

hist(student$writing.score, xlab = "Writing Score", ylab = "Frequency",
      main = "Distribution of Writing Scores")

# Hypothesis Testing 1
Male_student <- student$math_score[student$gender=="male"]
sd(Male_student)

library(BSDA)
z.test(Male_student,mu=65, sigma.x=14.36, alt="greater", conf.level=0.95)

# Hypothesis Testing 2
sd(student$writing_score)

library(BSDA)
z.test(student$writing_score,mu=62, sigma.x=15.20, alt="less", conf.level=0.95)

# Hypothesis Testing 3
student_no_prepare_course <- student$reading_score[student$test_preparation_course=="none"]
sd(student_no_prepare_course)

library(BSDA)
z.test(student_no_prepare_course,mu=75, sigma.x=14.46, alt="two.sided", conf.level=0.95)

#Goodness of fit
# Define expected probabilities
expected_probs <- rep(0.20, 5)

# Calculate observed frequencies
observed_freqs <- table(student$race.ethnicity)

# Perform chi-square test
chi_sq_test <- chisq.test(observed_freqs, p = expected_probs)

# Print test results
print(chi_sq_test)

#Chi-Square test for independence
# Create contingency tables for chi-square test for gender vs score

```

```

contingency_table_gender_math <- table(student$gender, student$math_score)
contingency_table_gender_reading <- table(student$gender, student$reading_score)
contingency_table_gender_writing <- table(student$gender, student$writing_score)

# Perform chi-square test
result_gender_math <- chisq.test(contingency_table_gender_math)
result_gender_reading <- chisq.test(contingency_table_gender_reading)
result_gender_writing <- chisq.test(contingency_table_gender_writing)

# Interpret the results
result_gender_math
result_gender_reading
result_gender_writing

# Create a new column for average score
student $average_score <- rowMeans(student[, c("math_score", "reading_score", "writing_score")])

# Create contingency table
contingency_table_education_avg <- table(student$parental_level_of_education, student
$average_score)
contingency_table_race_avg <- table(student$race, student $average_score)

# Perform chi-square test
result_education_avg <- chisq.test(contingency_table_education_avg )
result_race_avg <- chisq.test(contingency_table_race_avg)

# Interpret the results
result_education_avg
result_race_avg

#Correlation

# Pearson correlation between 2 variables
cor(student$math_score, student$reading_score)

# Spearman correlation between 2 variables
cor(student$math_score, student$reading_score,method = "spearman")

#multiple scatter plots

```

```

pairs(student[, c("math_score", "reading_score", "writing_score")])

# Pearson correlation test
test <- cor.test(student$math_score, student$reading_score)

#Regression

# combine the regression with the scatter plot

# Plotting the scatter plot
plot(math_score ~ reading_score, data = student, pch=16)

# plotting the fitted line:
abline(mod)

# linear model and saving it as 'mod'
mod <- lm(math_score ~ writing_score, data = student)

# the output:
mod

#more results
summary(mod)

#extract table of regression
library('broom')

# extract the table
my_results <- tidy(mod)

my_results

#ANOVA

student <- read.csv("StudentsPerformance.csv", header = TRUE,
                    colClasses = c("factor", "factor", "factor", "factor", "factor", "character", "character",
                                   "character"),
                    quote = "\"")

student[, 6:8] <- lapply(student[, 6:8], function(x) as.numeric(as.character(x)))

summary(student)

# 1.
one.way <- aov(math_score ~ gender, data = student)

```

```

summary(one.way)

# 2.
one.way <- aov(reading_score ~ gender, data = student)
summary(one.way)

# 3.
one.way <- aov(writing_score ~ gender, data = student)
summary(one.way)

# 4.
one.way <- aov(math_score ~ race, data = student)
summary(one.way)

# 5.
one.way <- aov(reading_score ~ race, data = student)
summary(one.way)

# 6.
one.way <- aov(writing_score ~ race, data = student)
summary(one.way)

# 7.
one.way <- aov(math_score ~ parental_level_of_education, data = student)
summary(one.way)

# 8.
one.way <- aov(reading_score ~ parental_level_of_education, data = student)
summary(one.way)

# 9.
one.way <- aov(writing_score ~ parental_level_of_education, data = student)
summary(one.way)

# 10.
one.way <- aov(math_score ~ lunch, data = student)
summary(one.way)

# 11.

```



```
one.way <- aov(writing_score ~ lunch, data = student)
```

```
summary(one.way)
```

```
# 12.
```

```
one.way <- aov(reading_score ~ lunch, data = student)
```

```
summary(one.way)
```

```
# 13.
```

```
one.way <- aov(math_score ~ test_preparation_course, data = student)
```

```
summary(one.way)
```

```
# 14.
```

```
one.way <- aov(writing_score ~ test_preparation_course, data = student)
```

```
summary(one.way)
```

```
# 15.
```

```
one.way <- aov(reading_score ~ test_preparation_course, data = student)
```

```
summary(one.way)
```