Task A – (30%)

Data Acquisition and Exploratory Data Analysis

Task A – Data Acquisition and Exploratory Data Analysis

Much of the attention was given for Student Performance related to gender, maths, reading and writing scores to determine whether there is a relationship between those variables. Hence, this study aims to analyse student performance to the variables of gender, math score, reading score and writing score to identify the relationships. The data set of Student Performance for maths, reading and writing scores has been used for the analysis. Accordingly, below charts and output tables for Summary Statistics was generated for Exploratory Data Analysis.

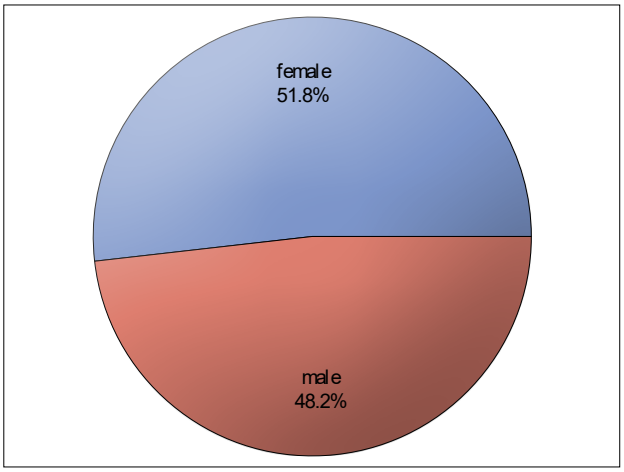


Figure 1 – Gender-wise distribution

The given data set was analysed for gender-wise distribution, in which 48.2% proportion is consist with male and 51.8% female. Further, 35.8% of students in the sample has completed the Test preparation course and out of that, 51.4% is female students. Further, the data for parent education level represent 37.5% is for high school education, 22.6% for college and rest 39.9% for graduate/degree level.

Following describes the summary statistics with relevant to the different variable.

Variable	Mean	Std Dev	Minimum	Maximum	Median	Coeff of Variation	Skewness	Kurtosis
math score	66.0890000	15.1630801	0	100.0000000	66.0000000	22.9434249	-0.2789351	0.2749641
reading score	69.1690000	14.6001919	17.0000000	100.0000000	70.0000000	21.1079992	-0.2591045	-0.0682655
writing score	68.0540000	15.1956570	10.0000000	100.0000000	69.0000000	22.3288227	-0.2894440	-0.0333646

Table 1 – Summary Statistics for math, reading and writing scores

According to the Table 1, reading score has the highest average (69.1690000), the lowest standard deviation (14.6001919) and the lowest coefficient of variation (21.1079992) than the math and writing scores. Further, reading score has the highest minimum value and the lowest range. Accordingly, that the data points of reading score are more compact around the mean, thus less variation than the others. As such, it indicates that the students are consistently performing better for reading in comparison to math and writing scores.

Skewness is the degree of asymmetry of a distribution in which negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. As per the above table (Figure 2), math, reading and writing scores have negative values. That indicate it is a slight negative or left skewed distribution. Further, it could also be mentioned as a petty symmetric distribution, since the values for skewness is close to zero.

Kurtosis represents the degree of pointer/peakdness of the data in which positive values for Kurtosis indicate that data more closely grouped around the mean with huge peak in the middle and negative values indicate more flatter data distribution than to a normal distribution. Kurtosis value of the math score is 0.2749641, which indicate that the data is closely grouped around the mean and distributed with a high peak (Leptokurtic distribution) than to a normal distribution. Thus, there is a low variability in math score. Further, the kurtosis for reading and writing scores are slightly flatter which indicate that the data points are spread out from the mean (negative/platykurtic distribution) than to a normal distribution.

Below table indicates the gender wise comparison of summary statistics.

gender	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	Median	Coeff of Variation	Skewness	Kurtosis
female	518	math score	63.6332046	15.4914532	0	100.0000000	65.0000000	24.3449208	-0.3319983	0.5878742
		reading score	72.6081081	14.3782453	17.0000000	100.0000000	73.0000000	19.8025340	-0.4221371	0.2977608
		writing score	72.4671815	14.8448418	10.0000000	100.0000000	74.0000000	20.4849167	-0.5592048	0.5964611
male	482	math score	68.7282158	14.3562772	27.0000000	100.0000000	69.0000000	20.8884765	-0.1452703	-0.3659926
		reading score	65.4730290	13.9318321	23.0000000	100.0000000	66.0000000	21.2787346	-0.1714356	-0.1973495
		writing score	63.3112033	14.1138318	15.0000000	100.0000000	64.0000000	22.2927872	-0.1549173	-0.1620305

Table 2 – Summary Statistics for gender-wise math, reading and writing scores

The above table (Table 2) has unequal observations count for male or female. In gender-wise, male has the highest average (68.7282158), lowest standard deviation (14.3562772) and lowest coefficient of variation (20.8884765) for the math score than the female. Accordingly, the male students are consistently performing better than female students on math, since the data points are more closely grouped around the mean value of male than female for math score.

However, female has the highest average and lower coefficient of variation for reading and writing scores than the male. As such, female students consistently performing better than male students for reading and writing. Further, in both categories of male and female, the male students have the highest minimum value and lower range for all of the score types which indicates data points are more compact and less variation in male than female.

Additionally, as per the Figure 4, both categories indicate a slightly negative skewed distribution for all the three scores. However, the data distribution is relatively petty symmetric or slight negative skewed/left skewed for male than female. Further, female has positive values for Kurtosis which indicates that the data points of female is pointer than a perfect normal distribution and more closely grouped around the mean by leading to a positive/Leptokurtic distribution. However, male has the negative values for Kurtosis, which indicates the data points are spread away from the mean. Hence, it is a flat/negative distribution compared to a normal distribution.

Task B – (30%)

Statistical Analysis

Task B – Statistical Analysis

1. HYPOTHESES:

The above data set on student performance was tested for different hypothesis as mentioned below.

Hypothesis 1:

- H_o: there is no difference between the math score when comparing to gender;
- H_1: there is a difference between the math score when comparing to gender.

Hypothesis 2:

- Ho: there is no relationship between the reading score and writing score;
- H1: there is a relationship between the reading score and writing score.

Hypothesis 3:

- Ho: the distribution of score for the categories of parent education level are the same;
- H1: the distribution of score for the categories of parent education level are not all the same.

2. NORMALITY TEST FOR INDIVIDUAL VARIABLES:

Normality test is carried out to determine whether the variables are Normally distributed or not. Therefore, there are a series of goodness-of-fit tests based on the empirical distribution function (EDF) to test the normality of the distribution. These are the Kolmogorov-Smirnov statistic, the Anderson-Darling statistic, and the Cramér-von Mises statistic.

2.1 The hypotheses in this case are:

- H_o: there is no difference between the distribution of the variable and that of the normal distribution;
- H_1: there is a difference between the distribution of the variable and that of the normal distribution.

2.2 Distribution of math score.

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.03085524	**Pr > D**	0.021
Cramer-von Mises	**W-Sq**	0.09237034	**Pr > W-Sq**	0.145
Anderson-Darling	**A-Sq**	0.65721220	**Pr > A-Sq**	0.089

Table 3: Fitted Normal Distribution for math score

The output for all three tests in the above table (Table 3) indicate that the p-value is not significant at the 5% level is reasonably appropriate. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest that the data is normally distributed. Therefore, the distribution of the math score is not significantly different to that of a Normal distribution.

2.3 Distribution of reading score.

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.04387352	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	0.15888281	**Pr > W-Sq**	0.020
Anderson-Darling	**A-Sq**	1.02113052	**Pr > A-Sq**	0.011

Figure 4: Fitted Normal Distribution for reading score

The outcome for all three tests in the above table (Table 4) can conclude, that the p-value is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H_1) and conclude that the data is not normally distributed. Therefore, the distribution of the reading score is significantly different to that of a Normal distribution.

2.4 Distribution of writing score.

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.04158230	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	0.25990092	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	1.44741080	**Pr > A-Sq**	<0.005

Table 5: Fitted Normal Distribution for writing score

The outcome for all three tests in the above table (Table 5) can conclude, that the p-value is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H_1) and conclude that the data is not normally distributed. Therefore, the distribution of the writing score is significantly different to that of a Normal distribution.

3. TESTS/METHODS

Hypothesis 1:

- H_o: there is no difference between the math score when comparing to gender;
- H_1: there is a difference between the math score when comparing to gender.

Two sample t-test could be used to compare two independent variables**.** The assumptions of 2 sample t-test is; the two samples are drawn independently; normally distributed sample means and equal group variances. Hence, it is required to determine whether the assumptions are met in advance/prior to carry out the test. However, if the variances are unequal an alternative t-test exists where a pooled estimate is used, assuming the other two assumptions are satisfied.

Therefore, it is required to evaluate the normality across each gender to determine whether the assumptions are satisfied. According to the normality test for two categories of gender, that the p-value is not significant at the 5% level. Hence, the null hypothesis should not be rejected, since there is enough evidence to suggest that the data is normally distributed is reasonable appropriate. Therefore, the distribution of the math score is not significantly different to that of a Normal distribution.

The following table in the output report provides the outcome of the equality of variances test which tested of homogeneity of variances. This tests the null hypothesis that the group variances are equal ($\sigma_1^2 = \sigma_2^2$), against the alternative that they are different.

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	517	481	1.16	0.0902

Table 6: Outcome of the equality of variances test

The p-value of above test is 0.0902 which is not significant at the 5% level. Hence, the null hypothesis should not be rejected, since that there is enough evidence to suggest that there is no difference in group variances and conclude that the assumption of equal group variances is reasonably appropriate.

As such, all the assumptions (the two samples are drawn independently; normally distributed sample means and equal group variances) of t-test has been satisfied.

The Pooled test should be interpreted when the assumption of equal group variances has been met.

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	998	-5.38	<.0001
Satterthwaite	Unequal	997.98	-5.40	<.0001

Table 7: Outcome of the Pooled test or degrees of freedom

The p-value associated with the t-test for equal variances (Pooled) is less than the α of 0.05. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H_1). Hence, it can conclude that there is a difference between the math score when comparing to gender.

gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
female		63.6332	62.2960	64.9704	15.4915	14.6021	16.4971
male		68.7282	67.4433	70.0131	14.3563	13.5036	15.3248
Diff (1-2)	**Pooled**	-5.0950	-6.9523	-3.2377	14.9551	14.3269	15.6414
Diff (1-2)	**Satterthwaite**	-5.0950	-6.9472	-3.2428			

Table 8: Outcome of the Pooled test or degrees of freedom

Alternatively, the above output table (Table 8), the confidence interval difference does not include zero ('0') and those are not overlapping for each category of gender. As such, it could conclude that there is a difference between the math score when comparing to gender.

Hypothesis 2:

- Ho: there is no relationship between the reading score and writing score;
- H1: there is a relationship between the reading score and writing score.

The above hypothesis could be tested by the correlation coefficient (*r)* which measures the strength of association between two variables. Pearson's correlation coefficient is the most commonly used for continuous variables in a Normal distribution. However, if the data is skewed or if one of the variables is on an ordinal scale and the other is not on an ordinal scale, then a more appropriate measurement is Spearman's rho.

As such, the distribution of two variables namely reading score and writing score is not normally distributed or skewed and significantly different to that of a Normal distribution (please refer section 2.3 and 2.4). Accordingly, it is required to carry out non-parametric test. Thus, Spearman's correlation coefficient has been used to test the given hypothesis.

The following output table was generated for correlation matrix on reading score and writing score by using the Spearman Correlation Coefficient.

Spearman Correlation Coefficients, N = 1000 Prob > |r| under H0: Rho=0		
	reading score	**writing score**
reading score	1.00000	0.94895 <.0001
writing score	0.94895 <.0001	1.00000

Table 9: Spearman's rho for reading and writing score

The output in the above table (Table 9) indicate that the p-value of the test is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H_1) and conclude that there is a significant relationship between the reading score and writing score.

The strength of the relationship/association is 0.94895 which indicates a strong positive relationship between the two variables. Hence, the two variables are positively correlated with a strong association.

Hypothesis 3:

- H_o: there is no difference between the reading score when comparing to writing score;
- H_1: there is a difference between the reading score when comparing to writing score.

Above hypothesis is comparing two variables with the same response for the same subjects as each student in the study tested for both of the scores on writing and reading (each measurement in one sample being matched or paired with a particular measurement in another sample).

Hence, a paired t-test is appropriate to use which requires the assumption of normality of the difference variable. If not, the Wilcoxon signed rank test could be used for non-normal, ranked or scored data. The output of the distribution for normality analysis is as follows:

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.994549	**Pr < W**	0.0011
Kolmogorov-Smirnov	**D**	0.056485	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.41932	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	2.256771	**Pr > A-Sq**	<0.0050

Table 10: Output report of the tests for normality

The output in the above table (Table 10) indicate that the p-value of the test is significant at the 5% level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H_1) and conclude that the data is not normally distributed. Hence, Wilcoxon signed rank test is appropriate to use for testing the hypothesis.

A more direct result is given by the results of the Wilcoxon signed rank test. This analysis shows a p-value is significant at 0.05 alpha level. Hence, the null hypothesis should be rejected, since there is enough evidence in favor of alternative hypothesis (H_1) and conclude that there is a difference between the reading score when comparing to writing score.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	7.787446	**Pr > |t|**	<.0001
Sign	**M**	107.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	60368	**Pr >= |S|**	<.0001

Table 11: Wilcoxon signed rank test

Additionally, the descriptive statistics for the difference between two variables and 95% confidence interval bounds has been determined. Accordingly, the confidence interval difference does not include zero (‘0’). As such, it could conclude that there is a difference between the reading score and writing score.

Annexure:

parental level of education	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	Skewness	Kurtosis
master's degree	59	math score	69.7457627	15.1539152	40.0000000	95.0000000	-0.2342272	-1.0697549
		reading score	75.3728814	13.7751629	42.0000000	100.0000000	-0.0217429	-0.6515802
		writing score	75.6779661	13.7307115	46.0000000	100.0000000	-0.0481119	-0.6355849
associate's degree	222	math score	67.8828829	15.1120932	26.0000000	100.0000000	0.0079796	-0.6780895
		reading score	70.9279279	13.8689482	31.0000000	100.0000000	-0.1717077	-0.4253311
		writing score	69.8963964	14.3111223	35.0000000	100.0000000	-0.1368959	-0.6008644
bachelor's degree	118	math score	69.3898305	14.9437886	29.0000000	100.0000000	-0.1149286	-0.2047535
		reading score	73.0000000	14.2852503	41.0000000	100.0000000	0.0342327	-0.4134249
		writing score	73.3813559	14.7282620	38.0000000	100.0000000	-0.1692236	-0.4065040
high school	375	math score	62.7866667	15.2128335	0	99.0000000	-0.5306230	0.7372425
		reading score	65.7706667	14.8127598	17.0000000	100.0000000	-0.3248195	-0.0432654
		writing score	63.6133333	14.9262835	10.0000000	100.0000000	-0.3762097	0.0935298
college	226	math score	67.1283186	14.3128969	19.0000000	100.0000000	-0.2273277	0.5372718
		reading score	69.4601770	14.0570491	23.0000000	100.0000000	-0.3748608	0.0258513
		writing score	68.8407080	15.0123306	19.0000000	99.0000000	-0.4218311	0.1300501

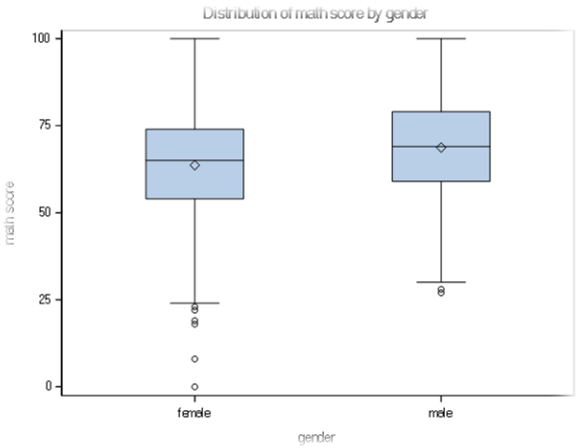
Annexure 1 – Summary Statistics for parent education level-wise math, reading and writing scores

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.991177	**Pr < W**	0.0035
Kolmogorov-Smirnov	**D**	0.043394	**Pr > D**	0.0187
Cramer-von Mises	**W-Sq**	0.123067	**Pr > W-Sq**	0.0564
Anderson-Darling	**A-Sq**	0.705866	**Pr > A-Sq**	0.0686

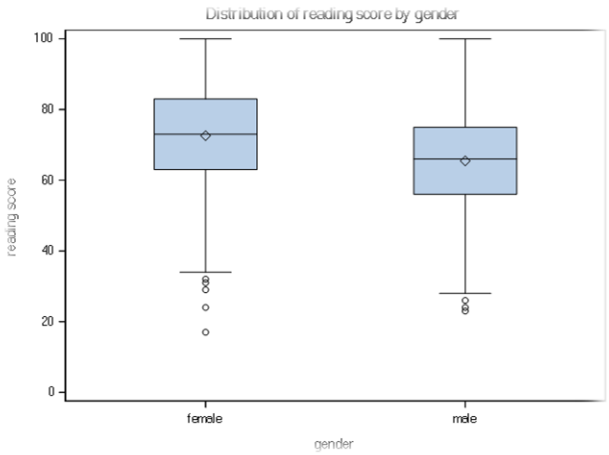
Annexure 2: Normality test for gender: female

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.993565	**Pr < W**	0.0380
Kolmogorov-Smirnov	**D**	0.038781	**Pr > D**	0.0779
Cramer-von Mises	**W-Sq**	0.080635	**Pr > W-Sq**	0.2105
Anderson-Darling	**A-Sq**	0.550841	**Pr > A-Sq**	0.1602

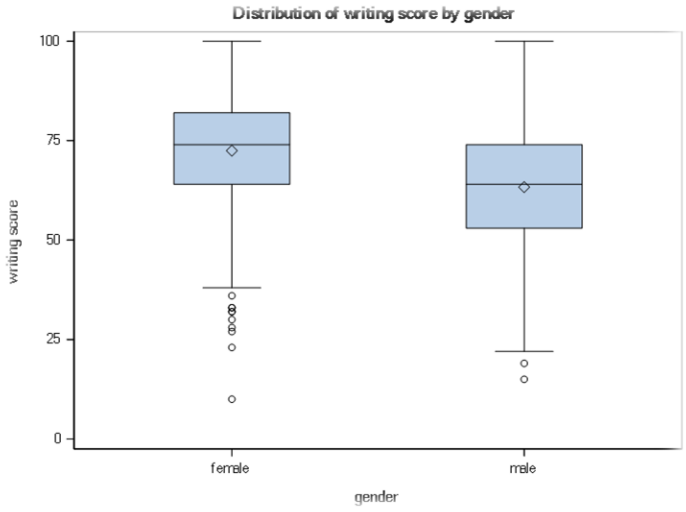
Annexure 3: Normality test for gender: male



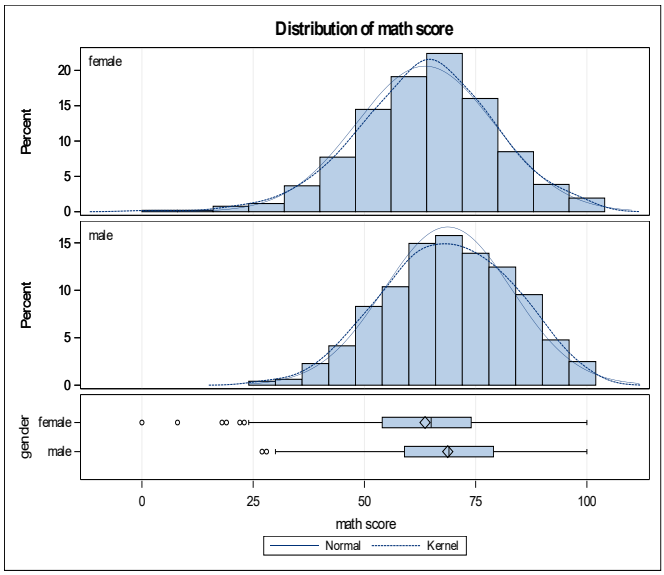
Annexure 4 – Distribution of math score by gender



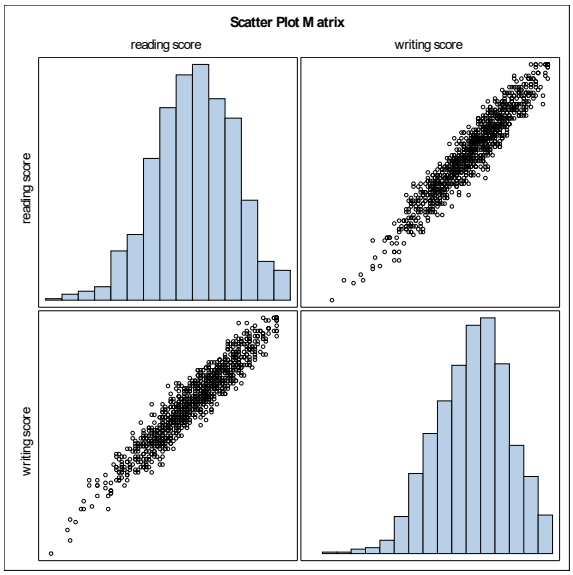
Annexure 5 – Distribution of reading score by gender



Annexure 6 – Distribution of writing score by gender



Annexure 7



Annexure 8