Application Report Master Data Science

A DESCRIPTIVE AND STATISTICAL ANALYSIS OF STUDENT PERFORMANCE IN MATHEMATICS AND LANGUAGE ACROSS DEMOGRAPHIC GROUPS

By

Shafiul Mahmud Partho

11 June, 2026

TABLE OF CONTENTS

CONTENTS	PAGE NO
INTRODUCTION	03
DETAILED DESCRIPTION OF THE PROBLEM	04
METHODOLOGY	05-08
EVALUATION	09-13
SUMMARY	14
BIBLIOGRAPHY	15

LIST OF TABLES

TABLE NO. **PAGE NO.**

Table 1: Description of dataset 4
Table 2: Overall Mean, Median, Standard Deviation, Variance and IQR value of scores by subject 9
Table 3: Descriptive statistics by parental education level 11
Table 4: Result for Mann-Whitney U test for question 1 12

LIST OF FIGURES

FiGURE NO. **PAGE NO.**

Figure 1: Score distribution of both subjects with mean (dashed) and median (solid) lines indicated.....9
Figure 2: Box plot of Math and Language scores by gender.....10
Figure 3: Violin plots of score distribution by gender and subject 10
Figure 4: Q-Q plots for all tested subgroups (Shapiro-Wilk Normality Test) 11
Figure 5: Result for Kruskal-Wallis Test for question 2 in box plot.....13
Figure 6: Heatmap of Bonferroni-corrected p-values from Dunn's post test for both subjects 13

INTRODUCTION

The education system plays a crucial role in every student's life, and, for this reason, it should be effective. However, the system faces some challenges worldwide, particularly in understanding some essential factors in how a student performs academically. There are some demographic variables – such as gender and parental education level—that matter when it comes to the students' performance outcomes. The authorities of educational institute, teachers want to enhance the educational quality. For instance, the students whose parents belong to lower educational backgrounds frequently underperform[1]. To reduce the performance gaps between students, authorities want to take the necessary steps. Likewise, understanding gender-based outcomes from various subjects can rearrange the curriculum and teaching pattern[2]. In this report, the academic scores of 486 students in two subjects are analyzed and investigated. It is being statistically proven whether the gender and parental educational background make any difference in students' academic achievement or not. Every student is classified by two variables: gender and parental level of education. However, there are two tasks given. One is the significant difference between student in their Mathematics and language scores based on their gender, and another is the difference in scores across the parental education level and identify the differ group. To answer this question, a descriptive analysis is conducted, which includes mean, median, standard deviation, etc. A normality check is done by using the Shapiro-Wilk test, and the non-parametric tests are chosen because normality cannot be assumed for all groups. For both tasks, the Mann-Whitney U test and the Kruskal-Wallis Test are run, respectively, with Dunn's Test as a post-hoc test. By running all the hypothesis tests, it is found that there is a significant change in the students' scores based on their gender. The male students' scores in Mathematics are higher, and the p-value is 0.000242, while the female students score better in Language and the p-value is less than 0.001. The null hypothesis is rejected in both cases and follows a reversal pattern. And for the second task, it is statistically found that the parental level of education significantly impacts students' academic performance. Those parents who passed high school, their children scored significantly lower than other groups. So high school level differs but there is no significant variety among higher groups. Later in this report, detailed descriptions of the problem, the statistical calculation method, the evaluation of the result, and finally the summary are added, with used bibliographies at the very end.

DETAILED DESCRIPTION OF THE PROBLEM

There are two tasks. The first task is to determine statistically if there is any noticeable difference between the two genders (male and female) in their scores in two subjects. To analyse this, I grouped the subject and gender and checked the p-value. And, for the second task, I determined if there is any visible difference in scores of different parental education levels (high school, associate's degree, bachelor's degree, master's degree), and found out which group is the most differentiable. To analyse this, I grouped the subjects by parental education level and checked the p-value again. Beyond these initial tasks, I examined whether there is any reversal pattern noticeable across subjects of the gender gap. It will be measured by effect size. Also, alongside the second task, I investigated if any threshold effect exists in the relationship between parental education level and students' scores.

The given dataset consists of 486 unique students' academic performance on different criteria. There are scores for two subjects (Mathematics and Language), the gender of the students, and their parental education level. There are different scores based on gender and parental academic background. There are a total of 972 rows for each score. It means every student has two rows. Each student gave an exam on both subjects. The data was given in semicolon-separated format with European decimal notation. And there are no missing values with no anomalous entries. Table 1 shows the dataset description.

Attributes	**Description**	**Type**	**Scale**
student_id	Unique student identifier (id_001 – id_486)	Identifier	Nominal
gender	Student gender (male, female)	Categorical	Nominal
parental.level.of. education	Highest parental education (high school, associate's, bachelor's, master's)	Categorical	Ordinal
subject	Academic subject (math, language)	Categorical	Nominal
score	Test score (19.5-100)	Continuous	Ratio

Table 1: Description of dataset

Here, the student id is not used in statistical analysis. The outcome variable is score which is measured by grouping the gender and the parental level of education. So, the dataset is complete and suitable for the analysis.

METHODOLOGY

The statistical analysis was executed using Python code. For the descriptive analysis, I applied mean, median, standard deviation, variance, quantiles and IQR[3]. After that, various hypothesis tests such as Shapiro-Wilk Normality test, Mann-Whitney U test, Kruskal-Wallis test, and Dunn's Post-hoc test are applied for the given tasks analysis and figure out the results. I also took the help of some of the Python libraries, such as stats, pyplot, gridspec, etc for a better visualization of the dataset and compared the p-value of the final result. The mathematical explanations of the calculation are given below.

Mean: This is the arithmetic average of any dataset which is used to representing the central tendency of the distribution. It is the ratio of the sum of all values and the total number of values (n). It can be written as,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{\mathrm{n}}$$

Here, $\bar{x}$ = sample mean, x_i = i-th value, n = total number of data points, $\sum x_i$ = sum of all data points. Mean consider all data points, but it is highly sensitive to outliers [3].

Median: This is the middle value of an seriallized dataset. It is done by ordering the data in a descending or ascending order and selecting the data in the middle. There are two rules considering the total number of datasets (n). The mathematical expression is,

When *n* is odd, $$Median = \left(\frac{n+1}{2}\right) th\ term$$

When *n* is even, $$Median\ = \frac{\left(\frac{n}{2}\right) th\ term\ + \left(\left(\frac{n+1}{2}\right)+1\right) th\ term}{2}$$

The median is mostly preferable when the distribution is skewed. However, it does not consider the dataset as a whole [3].

Variance and Standard Deviation: Variance is the average squared deviation of each value from the arithmetic mean. It shows how the data varies. On the other hand, standard deviation is used for measuring the distribution of data points. It is the square root of variance. The mathematical expression is as follows,

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$s = \sqrt{s^2}$$

Here, s^2 = variance, s = standard deviation, n-1 = degrees of freedom, x_i = i-th data, , $\bar{x}$ = mean value. Standard deviation is affected by outliers, but it gives the greater range around the mean.

Quantiles and IQR: Quantiles are used to divide the dataset into equal sized groups. If dataset are divided in four equal parts, then there will be 3 quartile which are 1st quartile Q_1 (p=0.25), 2nd quartile Q_2 (p=0.50, median) and 3rd quartile Q_3 (p=0.75). So, the interquartile range is,

$$IQR = Q_3 - Q_1$$

It is robust to outliers. It is used to measure for non-normal distributions.

Shapiro-Wilk Normality Test: This test is used to check whether the dataset is normally distributed or not [4]. When the total number of data points (n) is less than 5000, this normality test works powerfully. It analyse dataset with null hypothesis and check weather the dataset is normally distributed. The test formula W is,

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Here, $x_{(i)}$ = i-th sample value, a_i = constants which are derived from the expected values of standard normal oder statistics. The value range of W is from 0 to 1. W=1 means perfect normality. The hypotheses are:

Null Hypothesis (H_0): When the p-value is greater than 0.05, then the data follows normal distribuition.

Alternative Hypothesis (H_1): When the p-value is smaller than 0.05, then H_0 is rejected and the dataset does not follow normal distribution.

In order to do the test, the steps are necessary,

1. Oder the value ascendingly.
2. Using coefficients a_i ,measure the sum $\sum a_i x_{(i)}$. Then, measure W value.
3. Get the p-value corresponding to W.

The limitation of the test is that, it is sensitive at large datasets [5].

Mann-Whitney U test: This is a non-parametric test used to compare two independent groups of data, and the data is ordinal or continuous. When the normality test failed, the test is being done [6]. It is an alternative approach to the independent sample t-test where the dataset is normally distributed [7]. The steps for this test are given below:

1. Combining the dataset into both groups and rank them jointly from smallest to largest. In case of tied values, the average rank is taken.
2. Then, calculate the sum of the ranks for each group separately as R_1 and R_2.
3. After that, calculate the U statistics for both groups using the following formulas,

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

   Here, n_1 and n_2 are the total number of sample data of two groups.
4. Taking the smallest value between U1 and U2 as a U statistic value. Then, compute the z-score and get the p-value.

Null Hypothesis (H_0): When the p-value is greater than 0.05, then the data follows equal distribuition.

Alternative Hypothesis (H_1): When the p-value is smaller than 0.05, then H_0 is rejected and the dataset distribution differs.

This test is robust to outliers. However, it can compare only two group and slightly less powerful then t-test.

Kruskal-Wallis Test: This hypothesis test is also a non-parametric test and it is also an extented version of Mann-Whitney U test for comparing more that two independent group [8]. This is an alternative of one-way ANOVA test and used when the dataset is not normally distributed. The formula of this test is,

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

Here, N= total number of observation across all groups, k=total number of groups, n_i= total number of observation of i-th group, R_i= sum of ranks in i-th group. The test hypothesis will be:

Null Hypothesis (H_0): When the p-value is greater than 0.05, then the data follows equal distribuition. In this, the result almost follows a chi-squared distribution with *df = k-1* degree of freedom.

Alternative Hypothesis (H_1): When the p-value is smaller than 0.05, then H_0 is rejected and the dataset distribution differs. A post-hoc test is used to check which group differ the most.

The steps of the test is as follows:

1. Combining all groups and ranks together. Then, caluate the sum of ranks R_i for every group.
2. After that, calculate H value using the above formula. Then, compare H with the chi-squared distribution and get the p-value corresponding to H value.

This test is also robust with outliers and can identify atleast one group differs. However, it can not identify which specific pairs varies and slightly less powerful from ANOVA [8].

Dunn's Post-hoc Test with Bonferroni Correction: The test is done following the result of Kruskal-Wallis test and it is used to identify the signaficant differ of specific pairs of groups [9]**.** Using rank based z-statistics, the test compares all possible combination of all pairs. The test statistics for each pair (i,j) is:

$$z_{ij} = \frac{\left(\bar{R}_\iota - \bar{R}_J\right)}{\sigma_{ij}}$$

and standard error is, $$\sigma_{ij} = \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Here, $\bar{R}_\iota$ and $\bar{R}_J$ are the mean rank of each group, N = total sample size.

The Bonferroni correction is used to control error in order to compare multiple pairs. The following formula is applied, $$\alpha_{adjusted} = \frac{\alpha}{m},\ m = \binom{k}{2} = \frac{k(k-1)}{2}$$

Here, k = total number of groups, m = possible combinations pairwise. When the p-value of this correction is less than $\alpha = 0.05$, then it can be said that a pair is significantly different [9].

EVALUATION

In this section, I am giving an insight about the dataset and its property. To give the proper evaluation of the two tasks, I am adding some statistical graphs and figures alongside some tables. By analysing the hypothesis test, the proper outcome of the two task is shown. But, at first, I am doing the descriptive analysis.

	Total Student(n)	Mean	Median	Std. Devtiation	Variance	IQR
All Subjects	972	67.69	68.00	14.80	219.16	21.00
Math	486	66.38	66.00	15.05	226.64	21.75
Language	486	69.00	70.00	14.45	208.69	20.50

Table 2: Overall Mean, Median, Standard Deviation, Variance and IQR value of scores by subject.

Table 2 represent the overall descriptive value for all 972 students. From it, it can be said that, students tend to score well in language than mathematics inspite of their gender and parental level of education. The dataset is approximately symmetric. This can also be seen by the following histogram of mean and median of the student scores.

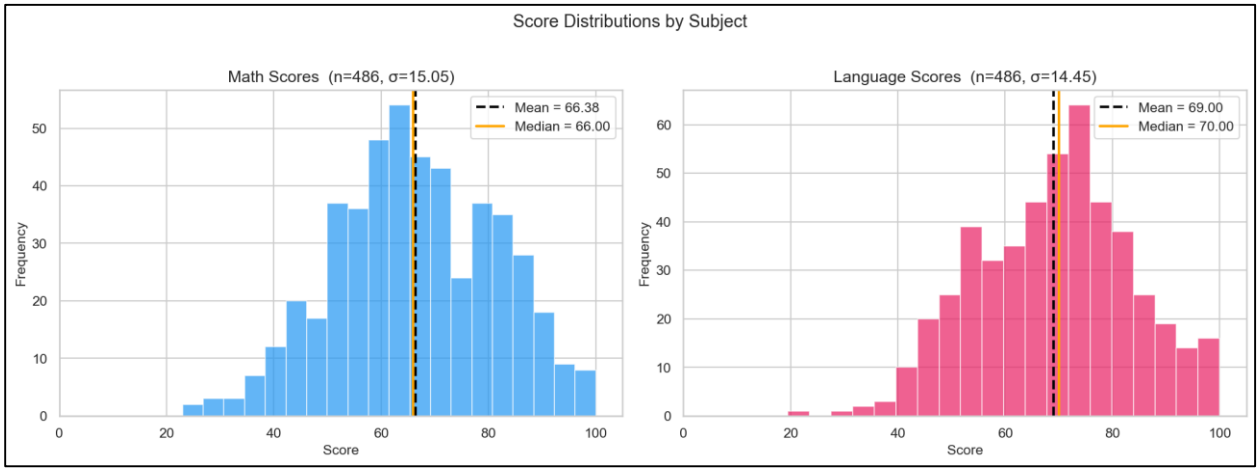


Figure 1: Score distribution of both subjects with mean (dashed) and median (solid) lines indicated.

In figure 1, it can be seen that, the consistancy of both subject is almost similar but math scores is more varied because of the standard deviation value. But, when it comes to gender, than male students are tends to score well and female students are good in language.

By analysing the following box plot, it can be understand very well.

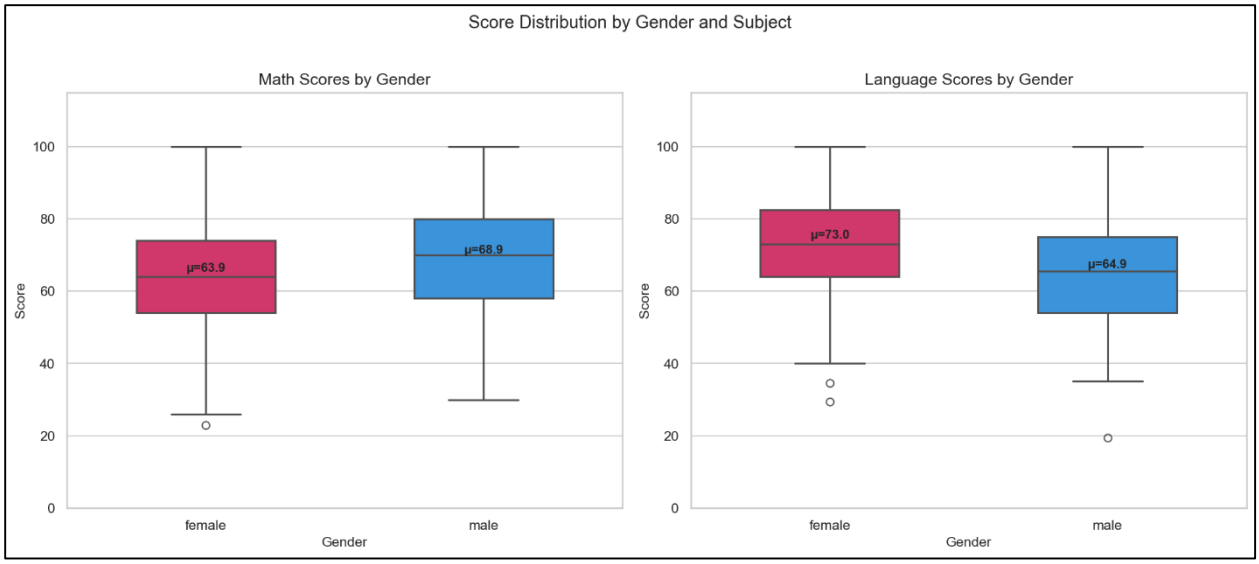


Figure 2: Box plot of Math and Language scores by gender.

In Figure 2, a higher mean score is showed by male student rather than female student and there is 5.09 points of difference. On the other hand, female students have a higher mean than males. This is called an inverse relationship. Where the box plot shows us the mean, the next figure shows the density distribution.

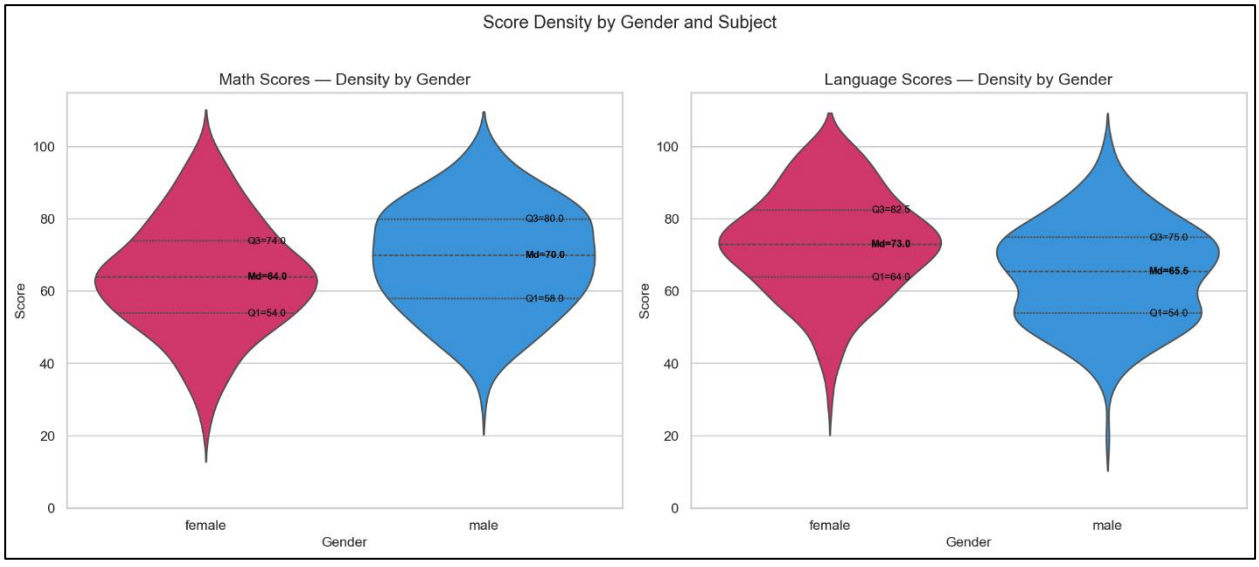


Figure 3: Violin plots of score distribution by gender and subject.

In Figure 3, it shows the full density shapes with quartile lines. For the math scores, a thick diamond shape in the 55-75 range is shown for female students. And for the male, the shape is much blockier and wider near the top. In the right side, the female students achieve high scores in

language so it indicates a very high density. The scores of male students is higher around Q3 = 75.0 as well as it also widening down Q1 = 54.0. Now, the descriptive analysis of students scores by their parental education level is shown via the following table.

	Total Student(n)	Mean	Median	Std. Deviation	IQR
High school	312	62.89	64.00	13.71	18.63
Associate's degree	354	68.92	69.50	14.76	22.50
Bachelor's degree	188	70.25	70.00	14.74	18.13
Master's degree	118	72.64	73.00	14.61	21.75

Table 3: Descriptive statistics by parental education level.

Here, the mean score is increased from 62.89 (high school) to 72.64 (master's). The students of master's parents are most likely to perform better than others. The most improvement is seen in the students who are the children of Associate's degree holders. The performance gap is noticeable between high school and an associate's degree. The mean score jump is almost 6 points. Now, to answer the two questions, I have done the normality test to check the dataset distribution. The Shapiro-Wilk test is done and the following Q-Q plot for each group shows that there are one pair that follow normality assumption or not. The p value should be greater than 0.05 to satisfy the null hypothesis.

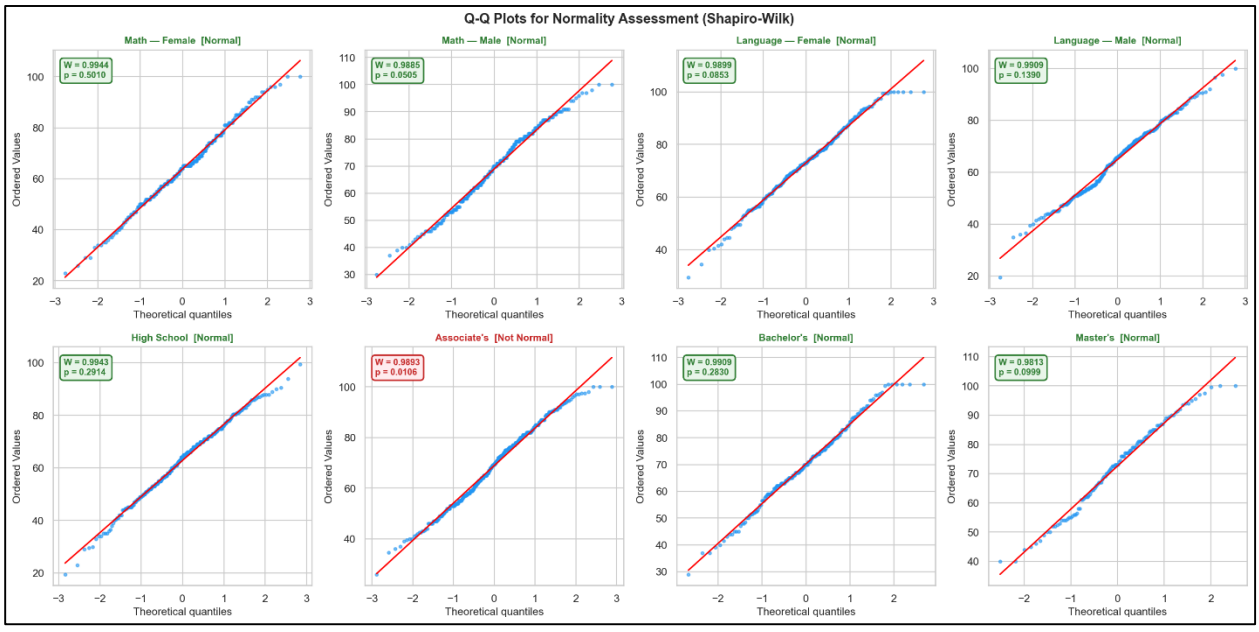


Figure 4: Q-Q plots for all tested subgroups (Shapiro-Wilk Normality Test).

In this figure, it is showed that, the data points which are the blue dots aligned the red line in the center almost every time or not. So, we can see that, the data set is not normally distributed because several blue dots dropping from the theoretical quanties. For this reason, non-parametric hypothesis test are selected to analyse the two given task.

Hypothesis Test 1 (Mann-Whitney U test): For the first question, I have to test if there is a variety between male and female students in their subjects. This test has been used as the data group of male and female are independent, continous and not normally distributed. By pairing male and female students by their subjects, the p-value is calculated to find out the outcome.
Here,
Null Hypothesis: If this satisfies, then there will be no significate differences between male and female students' performance regarding their subject.
Alternative Hypothesis: If the null hypothesis rejects, then there is a difference of performance based on gender.
The results of this test are shown in following table:

	First U statistics U_1	**Second U statistics U_2**	**z-score**	**p-value**
Mathematics	23,842	35,203	-3.6696	0.000242
Language	20,139	29,522.5	6.0618	<0.001

Table 4: Result for Mann-Whitney U test for question 1

Here, the p-value for both subject is less than 0.05 and rejects null hypothesis. So, it can be said that a significant difference is noticable between male and female students in their subject scores. And a reversal pattern is followed because the null hypothesis is rejected. So the gender gap reverses across subjects.

Hypotheis Test 2 (Kruskal-Wallis Test): For the second question, I am testing if there is any variety in scores across the education level of parents. The Kruskal-Wallis test is selected to compare the score by the level groups. This test can handle more than two independent groups and violates normal distribution. Here,
Null Hypothesis: The score distribution are same across all four education levels.
Alternative Hypothesis: At least one of the four education level has difference in scores.

The results of this test are shown in the following box plots:

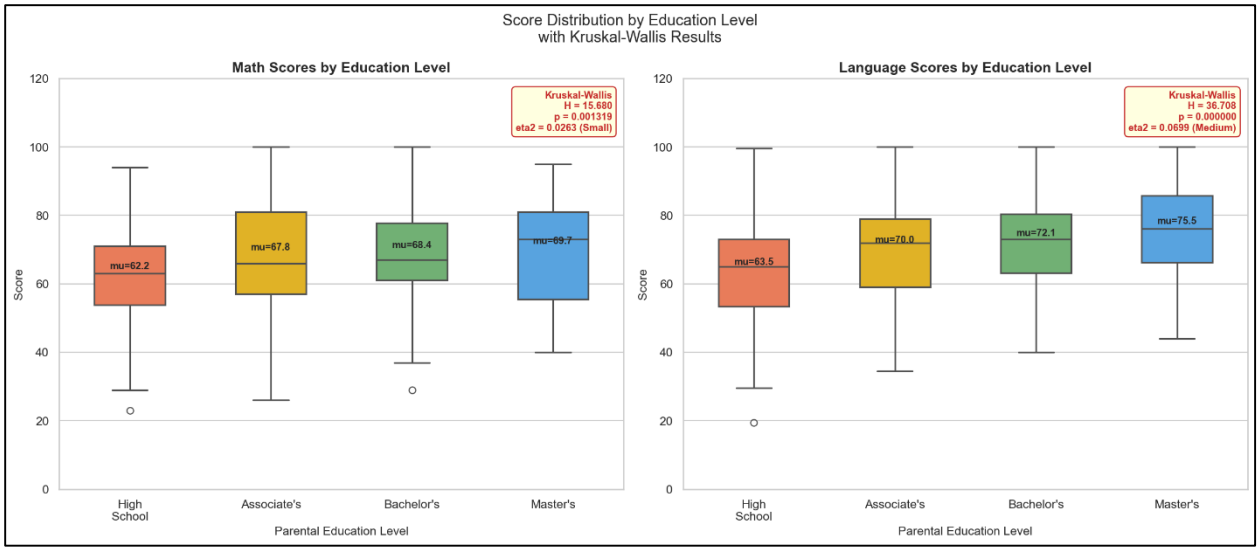


Figure 5: Result for Kruskal-Wallis Test for question 2 in box plot.

Here, p-value for both subjects across different education level is less than 0.05. So, it can be said that, for both subjects, the scores are significantly varies across the parental educational level. Following this test hypothesis and results, Dunn's post-hoc test with Bonferroni correction is selected and implemented to check which parental education level groups significantly vary. The test is used specially for this reason. With four groups and six pairswise comparison, a heat map is shown below:

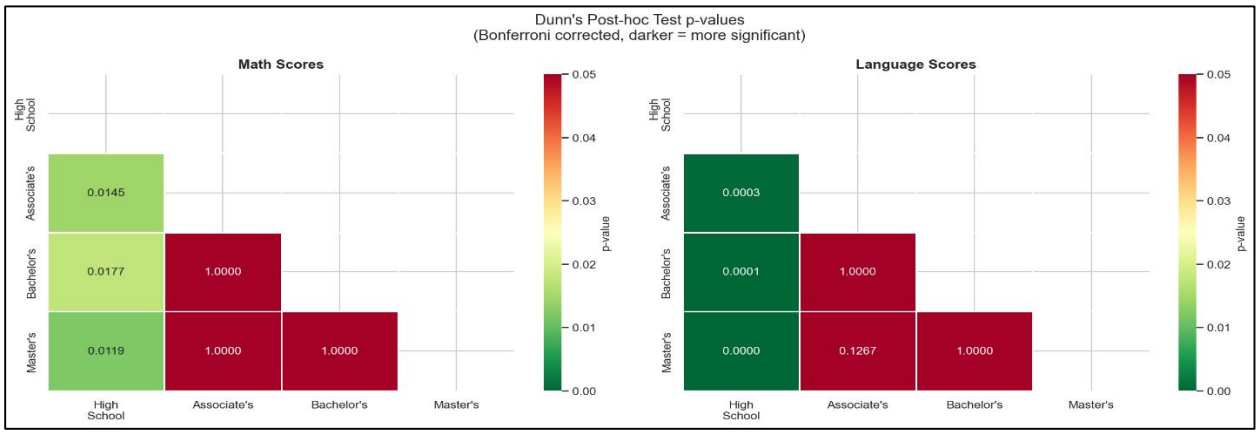


Figure 6: Heatmap of Bonferroni-corrected p-values from Dunn's post test for both subjects.

In this figure, it indicates that, the p-value is less than 0.05 for high school vs every other degrees. So, it can be said that, there is no signaficant difference among higher groups. But, the high school level significantly differs for all other groups. Therefore, it shows a threshold effect.

SUMMARY

From the analysis, it can be concluded that there is a gender bias in various subjects, such as Mathematics and Language. Where female students perform well in Language, male students score well in Mathematics. But the highest score is obtained by female students in Language. So, gender can be a factor in any educational performance. Although there should be many more reasons why gender plays such a crucial role in this. Different genders can perform well in different subjects. For academic performance, parents education background is very essential. According to the analysis, those students can perform well whose parents hold higher academic degrees. It is because of the environment and the influence they are getting. The high school passed parents' children sometimes struggle to get good scores, while the other educational background's students can perform well quite easily. So, there is a significant difference between the high school education level and others. For a student performing well in academia, parents' education level plays a vital role, and according to this analysis, parents must complete an associate's degree for their children's academic future.

BIBLIOGRAPHY

[1] S. R. Sirin, “Socioeconomic status and academic achievement: A meta-analytic review of research,” *Rev. Educ. Res.*, vol. 75, no. 3, pp. 417–453, 2005, doi: 10.3102/00346543075003417.

[2] J. S. Hyde and M. C. Linn, “Gender differences in verbal ability: A meta-analysis.,” *Psychol. Bull.*, vol. 104, no. 1, pp. 53–69, 1988, doi: 10.1037//0033-2909.104.1.53.

[3] R. Wackerly, D., Mendenhall, W., & Scheaffer, “Mathematical Statistics with Applications,” 2008.

[4] S. S. Shapiro and M. B. Wilk, “An Analysis of Variance Test for Normality (Complete Samples),” *Biometrika*, vol. 52, no. 3/4, p. 591, 1965, doi: 10.2307/2333709.

[5] A. Field, “Discovering Statistics Using IBM SPSS Statistics (Electronic Version),” pp. 1–1104, 2017.

[6] H. B. Mann and D. R. Whitney, “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other,” *Ann. Math. Stat.*, vol. 18, no. 1, pp. 50–60, 1947, doi: 10.1214/aoms/1177730491.

[7] N. Nachar, “The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution,” *Tutor. Quant. Methods Psychol.*, vol. 4, no. 1, pp. 13–20, 2008, doi: 10.20982/tqmp.04.1.p013.

[8] W. H. Kruskal and W. A. Wallis, “Use of ranks in one-criterion variance analysis,” *J. Am. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952.

[9] O. Jean, “Multiple Comparisons Using Rank Sums,” *Am. Soc. Qual.*, vol. 6, no. 3, pp. 241–252, 1964, doi: 10.1007/978-1-349-95810-8_109.