UNRAVELLING THE IMPACT OF DEMOGRAPHIC VARIABLES ON PRE-ADOLESCENT ACADEMIC ACHIEVEMENT: A COMPREHENSIVE R-BASED STUDY

Name: Siboham Pattanayak

Roll Number: STUG/132/21

Registration Number: A03-1112-0218-21

Ramakrishna Mission Residential College (Autonomous),

Narendrapur, Kolkata-700103.

Abstract:

This study examines the impact of gender, race, parental education, and completion of an additional preparatory course offered by the institute, on the academic performance of pre-adolescents of a particular school in reading, writing, mathematics, and their overall average scores. The dataset, sourced from Kaggle and anonymized for confidentiality, comprises data from an undisclosed school.

The categorical variables investigated include gender (male or female), race (Group A to Group E), parental education (categorized as school, baseline college, and quality college), and completion of preparatory courses (none or completed). The continuous response variables are scores in reading, writing, mathematics (each scored out of 100), and the overall average score.

To analyse the data comprehensively, four-way ANOVA tests are conducted separately for each response variable. This allows for the assessment of the main effects of each factor (gender, race, parental education, preparatory course) as well as their interactions.

The findings reveal significant associations between certain factors and academic performance across different subjects. Gender, race, and parental education demonstrate notable effects on scores in reading, writing, and mathematics, with varying degrees of significance. Additionally, the completion of preparatory courses influences performance in all subjects. Furthermore, interactions between factors highlight nuanced relationships that contribute to academic achievement.

Identifying optimal combinations of factors that positively impact academic performance provides valuable insights for educators, policymakers, and stakeholders. By understanding the complex interplay of gender, race, parental education, and preparatory courses, interventions can be tailored to support pre-adolescents in achieving their fullest academic potential.

Index:

1. Introduction

Basic education is a fundamental pillar of society, shaping the future of individuals and communities. Understanding the factors that influence academic performance among pre-adolescents is essential for designing effective interventions and educational policies. In this project, we investigate the impact of gender, race, parental education, and completion of the offered preparatory course on the academic marks of pre-adolescents in reading, writing, mathematics, and their overall average scores for a particular anonymous school from USA.

This research endeavour seeks to uncover valuable insights into the dynamics of academic achievement among pre-adolescents and identify the most influential factors and their optimal combinations. The findings derived from this study hold significant implications for educational practitioners, policymakers, and stakeholders of the school, enabling them to implement targeted strategies aimed at enhancing academic outcomes and fostering equitable access to quality education.

The categorical variables under examination include gender (categorized as male or female), race (classified into Group A to Group E), parental education (grouped into school, baseline college, and quality college), and completion of preparatory courses (distinguished as none or completed). By employing a four-way ANOVA analysis with one observation per cell, we aim to dissect the individual and interactive effects of these factors on academic performance across the specified subjects.

The dataset utilized for this study was sourced from Kaggle, ensuring anonymity of the participating school to uphold privacy and confidentiality. The dataset encompasses a diverse set of variables, including sensitive demographic information such as gender and race, as well as socio-economic indicators like parental education. Given the sensitivity of race-related information, ethnic categories have been coded into anonymized groups labelled from A to E. Academic performance is measured through marks scored in reading, writing, and mathematics exams each out of 100, along with an aggregated average score.

2. Definition of Terms

1. ANOVA

ANOVA, or Analysis of Variance, is a statistical method used to compare means among multiple groups to determine if there are statistically significant differences between them. It assesses whether differences observed in the data are likely due to real effects or random variation. ANOVA partitions the total variance in the data into different sources to test hypotheses about group differences. It's commonly employed in experimental and observational studies to analyze categorical variables' impact on a continuous outcome variable.

2. Covariates

Covariates in ANOVA are continuous variables that are included in the analysis to control for their influence on the outcome variable. They are used to adjust for potential confounding factors and reduce variability in the data, enhancing the accuracy of the estimated effects of the main factors. By accounting for covariates, researchers can better isolate and interpret the effects of the categorical independent variables (factors) of interest in the analysis.

3. Response

In statistics and research, a "response" refers to the outcome or dependent variable that is being measured or studied in an experiment or observational study. It is the variable of interest, and its values are typically influenced by one or more independent variables or factors. The response variable is what researchers seek to understand, predict, or analyse in their investigation.

4. Test Statistic

In statistical terms, a test statistic is a function of sample data used to make inferences about the population from which the sample is drawn. It measures the degree to which the observed data deviate from what would be expected under a null hypothesis, providing a basis for evaluating the statistical significance of an effect or relationship. Test statistics are typically compared to critical values or calculated probabilities to determine the likelihood of observing the data if the null hypothesis were true.

5. Null Hypothesis

In our context, the null hypothesis (H_0) is a statistical hypothesis that assumes no significant difference or effect between groups or variables being compared. It is typically the hypothesis that researchers aim to either reject or fail to reject based on the evidence from a statistical test.

6. Alternative Hypothesis

In our context, the alternative hypothesis (Ha) in statistics is a statement that contradicts the null hypothesis, suggesting that there is a significant difference, effect, or relationship between groups or variables being compared. It represents the possibility that observed results are not due to chance alone, but rather reflect a true effect or relationship in the population.

7. Level of Significance

The level of significance, often denoted as α, is a threshold used to determine the statistical significance of research findings. It represents the probability of incorrectly rejecting the null hypothesis when it is true. It is usually denoted by 'α'

8. p-value

In statistics, p-value refers to the probability of obtaining a test statistic as extreme or more extreme than the observed value, assuming that the null hypothesis is true.

9. Fisher's Multiple Comparison Test

Fisher's Multiple Comparison Test, a post hoc analysis, is commonly applied subsequent to an Analysis of Variance (ANOVA) to discern significant differences between specific group means while controlling the overall Type I error rate. This method is utilized under the assumption of normality and involves pairwise comparisons between group means. By calculating individual confidence intervals and comparing them to a critical value derived from the studentized range distribution, Fisher's test identifies significant differences between groups, aiding in the interpretation of ANOVA results and providing insights into the comparative effects of different treatments or conditions using a t-test.

3. Collection of Data

The dataset utilized in this study was obtained from Kaggle, a reputable platform for sharing and accessing datasets for research and analysis. The dataset was selected due to its comprehensive coverage of demographic variables and academic performance metrics, which align with the objectives of our research project.

The dataset contains anonymized data from a specific school, ensuring the privacy and confidentiality of the students and institution involved. The anonymity of the school was preserved to uphold ethical standards and protect the identities of the participants.

The variables included in the dataset are as follows:

1. **Demographic Variables:**

- *Gender*: Categorized as male or female.
- *Race*: Classified into Group A to Group E, representing different racial or ethnic backgrounds.
- *Parental Education*: Grouped into three categories – school, baseline college, and quality college – reflecting the educational attainment of the students' parents or guardians.
- *Preparatory Course*: Distinguished as none or completed, indicating whether the pre-adolescents have undertaken preparatory courses prior to the assessment provided by the school.

2. **Academic Performance Metrics:**

- *Reading Score*: Marks scored in reading, scaled out of 100.
- *Writing Score*: Marks scored in writing, scaled out of 100.
- *Mathematics Score*: Marks scored in mathematics, scaled out of 100.
- *Average Score*: An aggregated average of the scores in reading, writing, and mathematics, providing an overall measure of academic performance.

Data collection procedures adhered to ethical guidelines and regulations governing research involving human subjects. The anonymized nature of the dataset ensures the confidentiality of the participants and the institution involved, safeguarding their privacy rights.

The dataset was thoroughly reviewed and pre-processed to ensure data integrity and reliability. Any missing or erroneous values were addressed through appropriate data cleaning techniques to mitigate potential biases and inaccuracies in the analysis.

Overall, the dataset serves as a robust foundation for investigating the multifaceted demographic influences on academic performance among pre-adolescents, facilitating rigorous analysis and interpretation to derive meaningful insights for educational research and practice.

4. Exploratory Data Analysis

Without meeting the assumption of normality of errors, the use of ANOVA becomes inappropriate for analysis. We have our categorical variables as non-stochastic, hence randomness only lies in the response variable(s). Hence, it is essential for an ANOVA model to have its response variable follow a normal distribution. To assess the normality of our observations, we plot histograms of the response variables for Math, Reading, and Writing scores, accompanied by their respective average scores. Additionally, we overlay a normal curve onto each histogram, utilizing sample mean and sample variance (UMVUEs of population parameters μ and σ^2) as parameters for the curve. This allows us to visually inspect the distribution and evaluate whether it approximates a normal distribution.

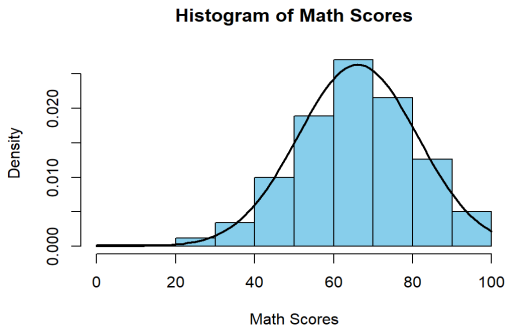


Figure 1.1

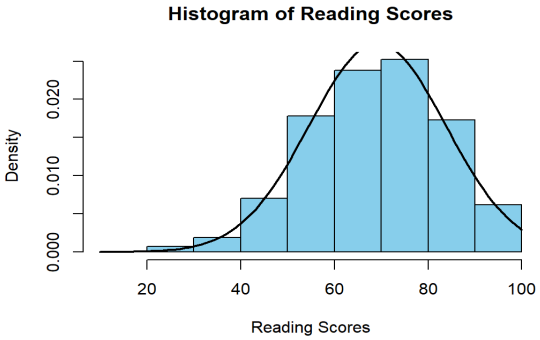


Figure 1.2

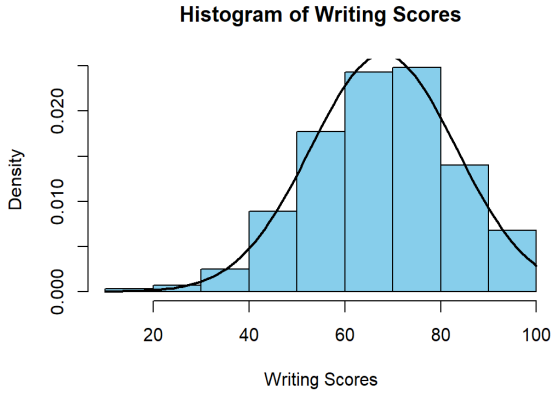


Figure 1.3

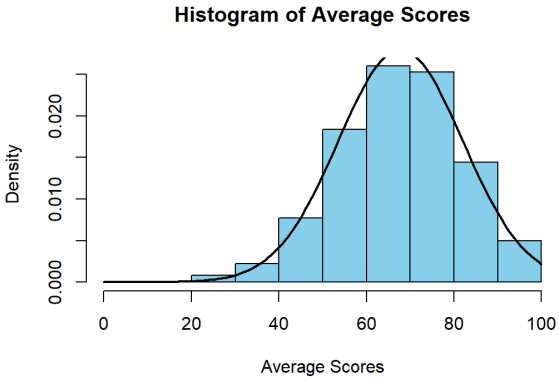


Figure 1.4

Our response variables in all four cases seem to be more or less normally distributed. Hence, the basic assumption of normality in our 4-way ANOVA Layout Linear Model is valid.

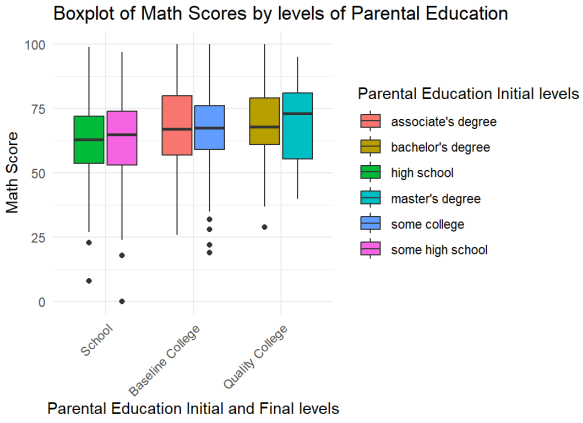


Figure 2.1

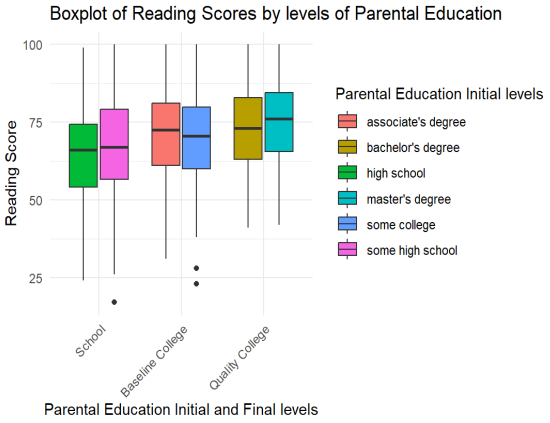


Figure 2.2

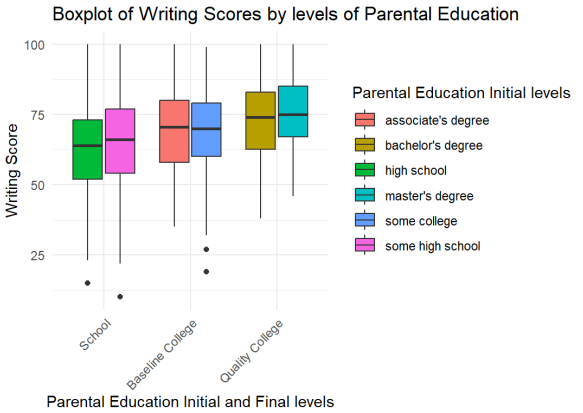


Figure 2.3

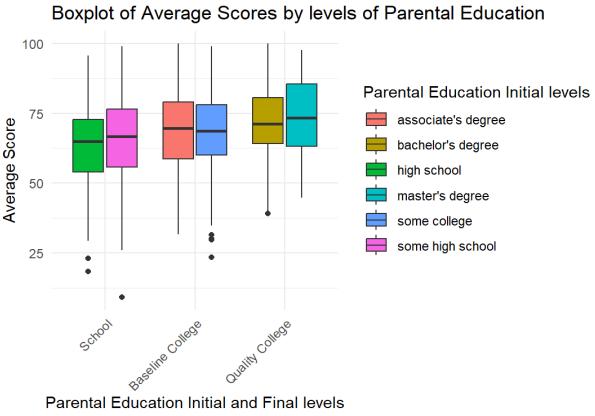


Figure 2.4

The initial parental education levels were given as ‘high school’, ‘some high school’, ‘some college’, ‘associate’s degree’, ‘bachelor’s degree’ , ‘master’s degree’ . We group them into 3 categories, ‘School’ , ‘ Baseline College’ and ‘Quality College’. The quartiles of scores as depicted in the above four boxplots justify this grouping. In our further analysis, we use the final parental education levels as the one mentioned, keeping in mind how close they are.

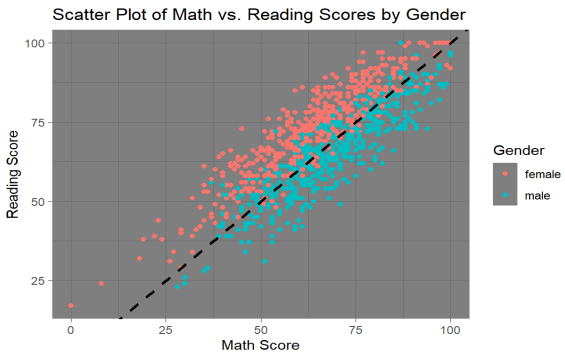


Figure 3.1

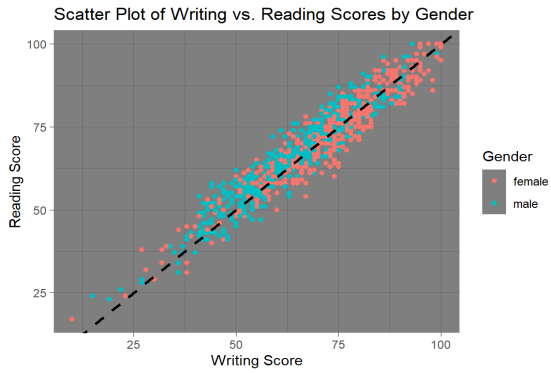
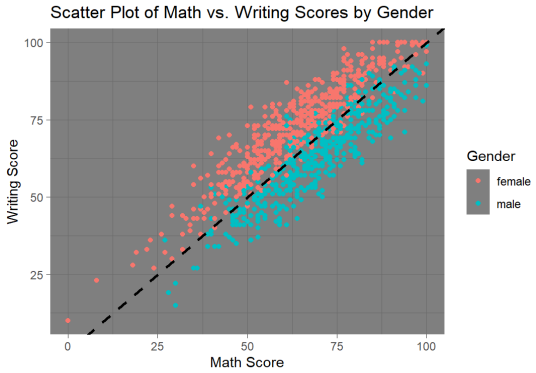


Figure 3.2

Figure 3.3



As seen from the scatter plots, male individuals for our dataset seem to be relatively better scorers in mathematics than reading or writing whereas female individuals seem to be vice versa.

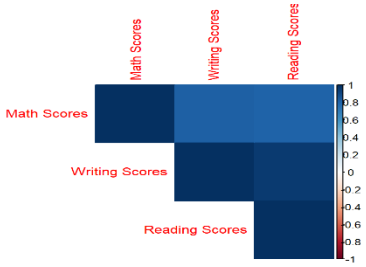


Figure 3.4

Also, Reading scores and Writing scores tend to be highly positively correlated but seem to differ from Math scores, showing distinguished skillsets required to ace these two types of examinations even among equivalent performers.

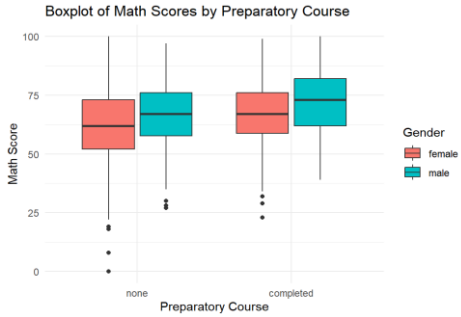


Figure 4.1

The boxplots are hinting towards the fact that taking the preparatory course was beneficial for both males and females for all three subjects and overall average scores.

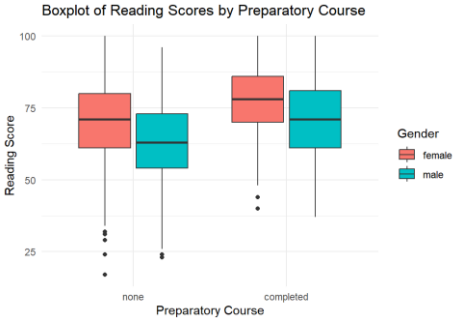


Figure 4.2

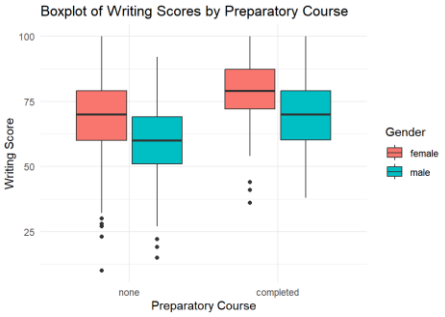


Figure 4.3

It also hints to the fact that males were better performers in mathematics but females were better performers in reading, writing and overall score, for both the categories that did not take the preparatory course and the one that completed it.

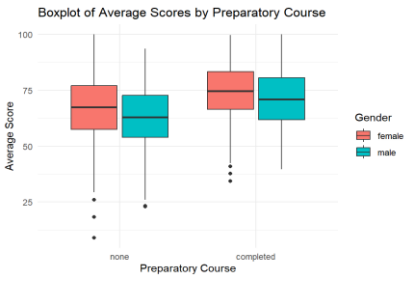


Figure 4.4

From the above exploratory data analysis, we can only hypothesize the fact that individuals belonging to different levels of the demographic variables considered, perform differently in the considered academic metrics. We'll proceed to test the statistical significance of all our hypotheses in the subsequent phase of the project using ANOVA.

5. Introduction to ANOVA Model (4-way layout)

Model:

The ANOVA model for a 4-way layout with one observation per cell, including three-factor and two-factor interactions but excluding four-factor interaction, can be expressed as:

$$Y_{ijkl}=\mu+\alpha_i+\beta_j+\gamma_k+\delta_l+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\alpha\delta)_{il}+(\beta\gamma)_{jk}+(\beta\delta)_{jl}+(\gamma\delta)_{kl}+(\alpha\beta\gamma)_{ijk}+(\alpha\beta\delta)_{ijl}+(\alpha\gamma\delta)_{ikl}+(\beta\gamma\delta)_{jkl}+\epsilon_{ijkl}$$

where

- Y_{ijkl} : Observation corresponding to the i-th level of factor A, j-th level of factor B, k-th level of factor C, and l-th level of factor D.
- μ : General mean.
- α_i: Fixed effect due to i-th level of Factor A.
- β_j : Fixed effect due to j-th level of Factor B.
- γ_k : Fixed effect due to k-th level of Factor C.
- δ_l : Fixed effect due to l-th level of Factor D
- $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$,$(\alpha\delta)_{il}$, $(\beta\gamma)_{jk}$, $(\beta\delta)_{jl}$, and $(\gamma\delta)_{kl}$ are the corresponding two-factor interactions between (A,B) , (A,C), (A,D), (B,C) , (B,D) , (C,D) respectively.
- $(\alpha\beta\gamma)_{ijk}$, $(\alpha\beta\delta)_{ijl}$, $(\alpha\gamma\delta)_{ikl}$, and $(\beta\gamma\delta)_{jkl}$ are the corresponding three-factor interactions among (A,B,C), (A,B,D), (A,C,D), (B,C,D) respectively.
- ϵ_{ijkl} : Random error

[Factors A,B,C & D have n_1 , n_2 , n_3 , & n_4 levels respectively. $i=1$ (|) n_1, $j=1$(|) n_2, $k=1$(|) n_3, $l=1$(|) n_4.]

Assumption:

$$\epsilon_{ijkl} \overset{iid}{\sim} N(0,\sigma^2)\ \forall\ (i,j,k,l)$$

Hypothesis:

We test for the following hypotheses based on our model:

Main Effects:

$$H_{0\alpha}: \alpha_i = 0 \text{ vs } H_{1\alpha}: H_{0\alpha} \text{ is false.}$$
$$H_{0\beta}: \beta_j = 0 \text{ vs } H_{1\beta}: H_{0\beta} \text{ is false.}$$
$$H_{0\gamma}: \gamma_k = 0 \text{ vs } H_{1\gamma}: H_{0\gamma} \text{ is false.}$$
$$H_{0\delta}: \delta_l = 0 \text{ vs } H_{1\delta}: H_{0\delta} \text{ is false.}$$

Two-Factor Interactions:

$$H_{0\alpha\beta}: (\alpha\beta)_{ij} = 0 \text{ vs } H_{1\alpha\beta}: H_{0\alpha\beta} \text{ is false.}$$
$$H_{0\alpha\gamma}: (\alpha\gamma)_{ik} = 0 \text{ vs } H_{1\alpha\gamma}: H_{0\alpha\gamma} \text{ is false.}$$
$$H_{0\alpha\delta}: (\alpha\delta)_{il} = 0 \text{ vs } H_{1\alpha\delta}: H_{0\alpha\delta} \text{ is false.}$$
$$H_{0\beta\gamma}: (\beta\gamma)_{jk} = 0 \text{ vs } H_{1\beta\gamma}: H_{0\beta\gamma} \text{ is false.}$$
$$H_{0\beta\delta}: (\beta\delta)_{jl} = 0 \text{ vs } H_{1\beta\delta}: H_{0\beta\delta} \text{ is false.}$$
$$H_{0\gamma\delta}: (\gamma\delta)_{kl} = 0 \text{ vs } H_{1\gamma\delta}: H_{0\gamma\delta} \text{ is false.}$$

Three-Factor Interactions:

$$H_{0\alpha\beta\gamma}: (\alpha\beta\gamma)_{ijk} = 0 \text{ vs } H_{1\alpha\beta\gamma}: H_{0\alpha\beta\gamma} \text{ is false.}$$
$$H_{0\alpha\beta\delta}: (\alpha\beta\delta)_{ijl} = 0 \text{ vs } H_{1\alpha\beta\delta}: H_{0\alpha\beta\delta} \text{ is false.}$$
$$H_{0\alpha\gamma\delta}: (\alpha\gamma\delta)_{ikl} = 0 \text{ vs } H_{1\alpha\gamma\delta}: H_{0\alpha\gamma\delta} \text{ is false.}$$
$$H_{0\beta\gamma\delta}: (\beta\gamma\delta)_{jkl} = 0 \text{ vs } H_{1\beta\gamma\delta}: H_{0\beta\gamma\delta} \text{ is false.}$$

These hypotheses can be tested using appropriate F-tests within the ANOVA framework to determine the significance of the main effects and interactions.

Test Rule:

Working procedure:

In the analysis of variance (ANOVA), Mean Square X (MSX) and Mean Square Error (MSE) are pivotal statistics used to evaluate the significance of factors and interactions. MSX quantifies the variability attributed to the factor or interaction under consideration, computed as the sum of squares (SSX) for that effect divided by its corresponding degrees of freedom (DFX), expressed mathematically as $MSX = \frac{SSX}{DFX}$. On the other hand, MSE represents the variability unexplained by the factors or interactions in the model, calculated by dividing the sum of squares of errors (SSE) by the degrees of freedom for error (DFE), denoted as $MSE = \frac{SSE}{DFE}$. MSE estimates the variance within groups (or cells) that is not accounted for by the factors or interactions. Comparing MSX to MSE through the F-test aids in determining whether the factor or the interaction has a significant impact on the response variable under consideration.

Test Statistic :

For a four-way layout in ANOVA with one observation per cell, denoting n_1 levels for factor A, n_2 levels for factor B, n_3 levels for factor C, and n_4 levels for factor D, the null distribution of the F-statistic for each effect or interaction can be expressed as:

$$F_X \sim F(df_X, df_E)$$

where:

- X represents the specific effect or interaction being tested, such as A, B, C, D, AB, AC, AD, BC, BD, CD, ABC, BCD, or ACD.
- df_X represents the degrees of freedom associated with the numerator, typically the degrees of freedom for the effect of interest (MSX).
- df_E represents the degrees of freedom associated with the denominator, typically the degrees of freedom for error (MSE).

The specific calculation of degrees of freedom for MSX and MSE depends on the design of the experiment, including the number of levels for each factor and the total number of observations. These degrees of freedom are used to define the null distribution of the F-statistic for each effect or interaction, providing a reference for testing their significance in the ANOVA model.

Decision making (An alternative to critical value approach) :

Using the p-value approach to test the null hypothesis for a specific effect or interaction X in ANOVA:

We compute the p-value p_X associated with the corresponding F-statistic F_X using the cumulative distribution function (CDF) of the F-distribution with degrees of freedom

df_X ,df_E

$p_X = P(F_X > (\text{Observed } F_X) \mid H_{0X})$

where:

- F_X is the F-statistic for the effect or interaction X,
- "Observed F_X" refers to the calculated F-statistic from the sample data under consideration, and
- H_{0X} is the null hypothesis corresponding to the effect or interaction X.

This p-value represents the probability of observing an F-statistic greater than the observed F_X under the assumption that the null hypothesis H_{0X} is true. If this probability is very small, we have evidence against H_{0X}, suggesting that the effect or interaction X has a significant impact on the response variable.

Hence we reject the respective null hypothesis if the p-value is considerably small

i.e. we reject H_0 if $p_X < \alpha$, and accept it otherwise.

, where α is the level of significance.

[For our purpose of analysis, we set α=0.05 and perform all the subsequent tests.]

Note:

- If the effect(s) of k-factor interaction is accepted to be significant for the response variable considered, we similarly perform subsequent (k-1) way ANOVA layouts using the same random sample drawn, for each level of a particular factor fixed beforehand from case to case, to find the respective best level combinations. (k=2,3)

- For finding the best level of a factor through multiple comparison, we use Fisher's Multiple Comparison method whenever required, since all our response variables are normally distributed.

6. ANALYSIS

[All tests are performed at 5% level of significance]

Factors	Levels
gender	'male' , 'female'
race	'group A', 'group B', 'group C', 'group D', 'group E'
Parentaledu	'School', 'Baseline College', 'Quality College'
prepcourse	'none' , 'completed'

Four way layout with response as Math ,Reading, Writing and Average scores

- *Response : Math score*

Table 1

	Df	Sum Sq	Mean Sq	F value
s_data1$gender	1	824	824.4	8.464
s_data1$race	4	1521	380.3	3.905
s_data1$Parentaledu	2	367	183.6	1.885
s_data1$prepcourse	1	1359	1359.5	13.957
s_data1$gender:s_data1$race	4	155	38.8	0.399
s_data1$gender:s_data1$Parentaledu	2	155	77.5	0.796
s_data1$race:s_data1$Parentaledu	8	1929	241.1	2.475
s_data1$gender:s_data1$prepcourse	1	76	76.2	0.782
s_data1$race:s_data1$prepcourse	4	469	117.4	1.205
s_data1$Parentaledu:s_data1$prepcourse	2	114	56.8	0.583
s_data1$gender:s_data1$race:s_data1$Parentaledu	8	4117	514.6	5.284
s_data1$gender:s_data1$race:s_data1$prepcourse	4	1208	301.9	3.099
s_data1$gender:s_data1$Parentaledu:s_data1$prepcourse	2	579	289.6	2.973
s_data1$race:s_data1$Parentaledu:s_data1$prepcourse	8	1806	225.8	2.318
Residuals	8	779	97.4	

	Pr(>F)	
s_data1$gender	0.01961	*
s_data1$race	0.04797	*
s_data1$Parentaledu	0.21347	
s_data1$prepcourse	0.00574	**
s_data1$gender:s_data1$race	0.80460	
s_data1$gender:s_data1$Parentaledu	0.48386	
s_data1$race:s_data1$Parentaledu	0.11069	
s_data1$gender:s_data1$prepcourse	0.40234	
s_data1$race:s_data1$prepcourse	0.37970	
s_data1$Parentaledu:s_data1$prepcourse	0.58028	
s_data1$gender:s_data1$race:s_data1$Parentaledu	0.01496	*
s_data1$gender:s_data1$race:s_data1$prepcourse	0.08118	.
s_data1$gender:s_data1$Parentaledu:s_data1$prepcourse	0.10826	
s_data1$race:s_data1$Parentaledu:s_data1$prepcourse	0.12783	
Residuals		

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

Significant factors at 5% level of significance:

i. gender
ii. race
iii. prepcourse
iv. Three way interaction among gender,race and Parentaledu

i. Since prepcourse is a significant factor in Math scores, we find its best level:

Table 1.1

```
> fisher_lsd_test1 <- LSD.test(anova_result1,"s_data1$prepcourse", alpha = 0.05)
> print(fisher_lsd_test1)
$statistics
   MSerror Df     Mean       CV  t.value      LSD
  97.40017  8 68.59333 14.38792 2.306004 5.876169

$parameters
        test p.ajusted             name.t ntr alpha
  Fisher-LSD      none s_data1$prepcourse   2  0.05

$means
          s_data1$mathS      std  r       se      LCL      UCL Min Max  Q25  Q50   Q75
completed      73.35333 16.12731 30 1.801852 69.19825 77.50841  32  96 63.7 76.0 86.75
none           63.83333 15.03807 30 1.801852 59.67825 67.98841  29  94 54.0 63.5 77.00

$comparison
NULL

$groups
          s_data1$mathS groups
completed      73.35333      a
none           63.83333      b

attr(,"class")
[1] "group"
```

Thus, best level of prepcourse for Math scores is ‘completed’.

Hence, we can conclude that individuals having completed the preparatory course were better performers at math than those who had not taken the preparatory course.

ii. Three way interaction present among Gender,Race and Parental education.

Two way layouts with Parentaledu and gender for each level of race :

- Group A:

Table 1.2.1

```
> summary(anova_result11)
                                 Df Sum Sq Mean Sq F value Pr(>F)
s_data11$Parentaledu              2  141.4   70.68   0.257  0.781
s_data11$gender                   1   45.6   45.63   0.166  0.698
s_data11$Parentaledu:s_data11$gender  2  180.6   90.28   0.329  0.732
Residuals                         6 1648.4  274.73
```

Parental education, gender or their interaction is not a significant factor for Math scores of individuals in race Group A.

- Group B:

Table 1.2.2

```
> summary(anova_result12)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data12$Parentaledu               2 1779.2   889.6   5.891 0.0384 *
s_data12$gender                    1  300.0   300.0   1.987 0.2083
s_data12$Parentaledu:s_data12$gender  2  585.5   292.8   1.939 0.2241
Residuals                          6  906.0   151.0
```

Parental Education is a significant factor for math scores of individuals in race Group B.

For Group B, finding best level of Parental education which will be valid for both Males and Females belonging to Group B.

Table 1.2.2.1

```
> fisher_lsd_test12 <- LSD.test(anova_result12, "s_data12$Parentaledu", alpha = 0.05)
> print(fisher_lsd_test12)
$statistics
  MSerror Df     Mean       CV  t.value     LSD
     151  6 66.33333 18.52493 2.446912 21.2614

$parameters
        test p.ajusted              name.t ntr alpha
  Fisher-LSD      none s_data12$Parentaledu   3  0.05

$means
                s_data12$mathS      std r       se      LCL      UCL Min Max
Baseline College          76.75 12.84199 4 6.144103 61.71592 91.78408  60  91
Quality College           73.00 11.28421 4 6.144103 57.96592 88.03408  62  87
School                    49.25 17.46186 4 6.144103 34.21592 64.28408  32  68
                 Q25  Q50   Q75
Baseline College 72.00 78.0 82.75
Quality College  65.00 71.5 79.50
School           35.75 48.5 62.00

$comparison
NULL

$groups
                s_data12$mathS groups
Baseline College          76.75      a
Quality College           73.00      a
School                    49.25      b

attr(,"class")
[1] "group"
```

Joint best levels of parental education for math scores of individuals in race group B is Baseline College and Quality College.

- Group C:

Table 1.2.3

```
> summary(anova_result13)
                                  Df Sum Sq Mean Sq F value  Pr(>F)
s_data13$Parentaledu               2  103.2    51.6   0.696 0.53499
s_data13$gender                    1  432.0   432.0   5.825 0.05233 .
s_data13$Parentaledu:s_data13$gender  2 1918.5   959.2  12.934 0.00667 **
Residuals                          6  445.0    74.2
---
```

Only Interaction between Parental education and gender is a significant factor in group C.

For Group C, we perform one way anova to find best level of parental education for each level of gender.

- For males in Group C :

Table 1.2.3.1

```
> summary(anova_result131)
                        Df Sum Sq Mean Sq F value Pr(>F)
s_data131$Parentaledu    2   1436   718.2   14.08 0.0299 *
Residuals                3    153    51.0
```

For males in group C, Parental education is a significant factor.

Best level of parental education for males belonging to Race group C:

Table 1.2.3.1.1

```
> fisher_lsd_test131 <- LSD.test(anova_result131,"s_data131$Parentaledu", alpha = 0.05)
> print(fisher_lsd_test131)
$statistics
  MSerror Df     Mean       CV  t.value      LSD
        51  3 75.33333 9.479772 3.182446 22.72721

$parameters
        test p.ajusted                name.t ntr alpha
  Fisher-LSD      none s_data131$Parentaledu   3  0.05

$means
                s_data131$mathS        std r       se      LCL       UCL Min Max
Baseline College            85.0  2.8284271 2 5.049752 68.92943 101.07057  83  87
Quality College             87.5 12.0208153 2 5.049752 71.42943 103.57057  79  96
School                      53.5  0.7071068 2 5.049752 37.42943  69.57057  53  54
                  Q25  Q50   Q75
Baseline College 84.00 85.0 86.00
Quality College  83.25 87.5 91.75
School           53.25 53.5 53.75

$comparison
NULL

$groups
                s_data131$mathS groups
Quality College            87.5      a
Baseline College           85.0      a
School                     53.5      b

attr(,"class")
[1] "group"
```

Joint best levels of parental education for Math scores of male individuals in group C are Quality College and Baseline College.

- For females in group C :

Table 1.2.3.2

```
> summary(anova_result132)
                        Df Sum Sq Mean Sq F value Pr(>F)
s_data132$Parentaledu    2  585.3  292.67   3.007  0.192
Residuals                3  292.0   97.33
```

Parental Education is not a significant factor for Math scores of female individuals in race Group C.

- Group D:

Table : 1.2.4

```
> summary(anova_result14)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data14$Parentaledu               2  101.2   50.58   0.162  0.854
s_data14$gender                    1  154.1  154.08   0.493  0.509
s_data14$Parentaledu:s_data14$gender  2  611.2  305.58   0.977  0.429
Residuals                          6 1876.5  312.75
```

Parental education, gender or their interaction is not a significant factor for Math scores of individuals in race Group D.

- Group E:

Table : 1.2.5

```
> summary(anova_result15)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data15$Parentaledu               2  171.2    85.6   0.339  0.725
s_data15$gender                    1   48.0    48.0   0.190  0.678
s_data15$Parentaledu:s_data15$gender  2  976.5   488.3   1.934  0.225
Residuals                          6 1515.0   252.5
>
```

Parental education, gender or their interaction is not a significant factor for Math scores of individuals in race Group E.

➢ *Response : Reading score*

Table : 2

```
> summary(anova_result2)
                                                    Df Sum Sq Mean Sq F value
s_data2$gender                                       1    325   324.8   2.745
s_data2$Parentaledu                                  2    270   134.9   1.140
s_data2$race                                         4    679   169.8   1.435
s_data2$prepcourse                                   1   1991  1990.7  16.826
s_data2$gender:s_data2$Parentaledu                   2    312   155.9   1.317
s_data2$gender:s_data2$race                          4    149    37.3   0.315
s_data2$Parentaledu:s_data2$race                     8   3023   377.9   3.194
s_data2$gender:s_data2$prepcourse                    1    223   222.7   1.883
s_data2$Parentaledu:s_data2$prepcourse               2    179    89.5   0.757
s_data2$race:s_data2$prepcourse                      4    791   197.7   1.671
s_data2$gender:s_data2$Parentaledu:s_data2$race      8   3914   489.3   4.136
s_data2$gender:s_data2$Parentaledu:s_data2$prepcourse 2    883   441.5   3.731
s_data2$gender:s_data2$race:s_data2$prepcourse       4    935   233.8   1.976
s_data2$Parentaledu:s_data2$race:s_data2$prepcourse  8   2038   254.8   2.153
Residuals                                            8    946   118.3
                                                    Pr(>F)
s_data2$gender                                      0.13613
s_data2$Parentaledu                                 0.36662
s_data2$race                                        0.30683
s_data2$prepcourse                                  0.00343 **
s_data2$gender:s_data2$Parentaledu                  0.32024
s_data2$gender:s_data2$race                         0.86009
s_data2$Parentaledu:s_data2$race                    0.06039 .
s_data2$gender:s_data2$prepcourse                   0.20729
s_data2$Parentaledu:s_data2$prepcourse              0.49998
s_data2$race:s_data2$prepcourse                     0.24856
s_data2$gender:s_data2$Parentaledu:s_data2$race     0.03040 *
s_data2$gender:s_data2$Parentaledu:s_data2$prepcourse 0.07165 .
s_data2$gender:s_data2$race:s_data2$prepcourse      0.19133
s_data2$Parentaledu:s_data2$race:s_data2$prepcourse  0.14932
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Significant factors at 5% level of significance:

i. prepcourse
ii. Interaction among gender,Parentaledu and race

i. Since prepcourse is a significant factor in Reading scores, we find its overall best level:

Table : 2.1

```
> fisher_lsd_test2 <- LSD.test(anova_result2,"s_data2$prepcourse", alpha = 0.05)
> print(fisher_lsd_test2)
$statistics
  MSerror Df     Mean       CV  t.value      LSD
  118.311  8 71.22667 15.27109 2.306004 6.476301

$parameters
        test p.ajusted             name.t ntr
  Fisher-LSD      none s_data2$prepcourse   2
  alpha
   0.05

$means
          s_data2$readingS      std  r       se
completed         76.98667 17.97292 30 1.985875
none              65.46667 13.51814 30 1.985875
              LCL     UCL Min Max   Q25  Q50
completed 72.40723 81.5661  37 100 66.25 78.8
none      60.88723 70.0461  41  91 57.25 63.5
           Q75
completed 90.0
none      73.5

$comparison
NULL

$groups
          s_data2$readingS groups
completed         76.98667      a
none              65.46667      b

attr(,"class")
[1] "group"
```

For reading scores, ‘completed’ is the overall best level for prepcourse.

Thus we can conclude that individuals who completed the preparatory course ended up scoring higher in the reading examination than those who did not take any preparatory course.

ii. Three way interaction present among gender,race and parental education.

Two way layouts with parental education and gender for each level of race :

- Group A: Table 2.2.1

```
> summary(anova_result21)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data21$Parentaledu               2  147.1   73.53   0.296  0.754
s_data21$gender                    1   99.8   99.76   0.402  0.550
s_data21$Parentaledu:s_data21$gender  2  348.1  174.06   0.701  0.533
Residuals                          6 1490.2  248.36
```

Parental education, gender or their interaction is not a significant factor for Reading scores of individuals in race group A.

- Group B: Table 2.2.2

```
> summary(anova_result22)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data12$Parentaledu               2 2845.5  1422.7   7.312 0.0246 *
s_data12$gender                    1   70.1    70.1   0.360 0.5704
s_data12$Parentaledu:s_data12$gender  2   83.2    41.6   0.214 0.8135
Residuals                          6 1167.5   194.6
---
```

Parental education is a significant factor for Reading scores of individuals in Race group B. Hence we find the best level of parental education now.

Table 2.2.2.1

```
> fisher_lsd_test22 <- LSD.test(anova_result22,"s_data12$Parentaledu", alpha = 0.05)
> print(fisher_lsd_test22)
$statistics
   MSerror Df  Mean       CV  t.value      LSD
  194.5833  6 70.25 19.85667 2.446912 24.13549

$parameters
        test p.ajusted                name.t ntr alpha
  Fisher-LSD      none s_data12$Parentaledu   3  0.05

$means
                 s_data12$readingS       std r       se      LCL       UCL Min
Baseline College             84.00 16.573071 4 6.974657 66.93363 101.06637  60
Quality College              78.00 12.027746 4 6.974657 60.93363  95.06637  63
School                       48.75  4.573474 4 6.974657 31.68363  65.81637  44
                 Max   Q25  Q50   Q75
Baseline College  96 79.50 90.0 94.50
Quality College   90 71.25 79.5 86.25
School            54 45.50 48.5 51.75

$comparison
NULL

$groups
                 s_data12$readingS groups
Baseline College             84.00      a
Quality College              78.00      a
School                       48.75      b

attr(,"class")
[1] "group"
```

Hence,we can conclude that 'Baseline College' and 'Quality College' are jointly the best levels of parental education for race group C in accordance with reading scores obtained.

Joint best levels of parental education for reading scores of individuals in race group B are Baseline College and Quality College.

- Group C:

Table 2.2.3

```
> summary(anova_result23)
                                  Df Sum Sq Mean Sq F value
s_data13$Parentaledu               2   99.5    49.7   0.187
s_data13$gender                    1    2.1     2.1   0.008
s_data13$Parentaledu:s_data13$gender  2 1883.2   941.6   3.532
Residuals                          6 1599.5   266.6
                                  Pr(>F)
s_data13$Parentaledu              0.8344
s_data13$gender                   0.9324
s_data13$Parentaledu:s_data13$gender 0.0969 .
Residuals
```

Parental education,gender or their interaction is not a significant factor for Reading scores of individuals in race group C.

- Group D:

Table 2.2.4

```
> summary(anova_result24)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data14$Parentaledu               2  168.5    84.3   0.266  0.775
s_data14$gender                    1   12.0    12.0   0.038  0.852
s_data14$Parentaledu:s_data14$gender  2  580.5   290.2   0.917  0.449
Residuals                          6 1899.0   316.5
```

Parental education,gender or their interaction is not a significant factor for Reading scores of individuals in race group D.

- Group E:

Table 2.2.5

```
> summary(anova_result25)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data15$Parentaledu               2   32.2    16.1   0.053  0.949
s_data15$gender                    1  290.1   290.1   0.951  0.367
s_data15$Parentaledu:s_data15$gender  2 1331.2   665.6   2.183  0.194
Residuals                          6 1829.5   304.9
```

Parental education, gender or their interaction is not a significant factor for Reading scores of individuals in race group E.

- *Response: Writing score*

Table 3

```
> summary(anova_result3)
                                                    Df Sum Sq Mean Sq
s_data3$gender                                       1    590   590.3
s_data3$Parentaledu                                  2    618   309.2
s_data3$race                                         4    355    88.9
s_data3$prepcourse                                   1   2260  2259.5
s_data3$gender:s_data3$Parentaledu                   2    126    63.1
s_data3$gender:s_data3$race                          4    181    45.2
s_data3$Parentaledu:s_data3$race                     8   2663   332.9
s_data3$gender:s_data3$prepcourse                    1     10     9.8
s_data3$Parentaledu:s_data3$prepcourse               2      5     2.7
s_data3$race:s_data3$prepcourse                      4    819   204.8
s_data3$gender:s_data3$Parentaledu:s_data3$race      8   3597   449.7
s_data3$gender:s_data3$Parentaledu:s_data3$prepcourse 2   948   474.0
s_data3$gender:s_data3$race:s_data3$prepcourse       4    872   218.0
s_data3$Parentaledu:s_data3$race:s_data3$prepcourse  8   1475   184.4
Residuals                                            8    919   114.9
                                                    F value  Pr(>F)
s_data3$gender                                        5.137 0.05317 .
s_data3$Parentaledu                                   2.691 0.12773
s_data3$race                                          0.773 0.57209
s_data3$prepcourse                                   19.664 0.00218 **
s_data3$gender:s_data3$Parentaledu                    0.549 0.59766
s_data3$gender:s_data3$race                           0.393 0.80817
s_data3$Parentaledu:s_data3$race                      2.897 0.07681 .
s_data3$gender:s_data3$prepcourse                     0.085 0.77812
s_data3$Parentaledu:s_data3$prepcourse                0.023 0.97718
s_data3$race:s_data3$prepcourse                       1.783 0.22547
s_data3$gender:s_data3$Parentaledu:s_data3$race       3.913 0.03540 *
s_data3$gender:s_data3$Parentaledu:s_data3$prepcourse 4.125 0.05874 .
s_data3$gender:s_data3$race:s_data3$prepcourse        1.897 0.20440
s_data3$Parentaledu:s_data3$race:s_data3$prepcourse   1.605 0.25924
Residuals
```

Significant factors at 5% level of significance:

i. prepcourse
ii. Interaction between gender,Parentaledu and race

i. Since prepcourse is a significant factor for writing scores, we find its overall best level

Table 3.1

```
> fisher_lsd_test3 <- LSD.test(anova_result3,"s_data1$prepcourse", alpha =
0.05)
> print(fisher_lsd_test3)
$statistics
   MSerror Df     Mean      CV  t.value      LSD
  114.9057  8 70.13667 15.2836 2.306004 6.382417

$parameters
        test p.ajusted             name.t ntr alpha
  Fisher-LSD      none s_data1$prepcourse   2  0.05

$means
          s_data1$writingS      std  r       se      LCL      UCL Min Max
completed         76.27333 16.54565 30 1.957087 71.76028 80.78638  40 100
none              64.00000 13.44465 30 1.957087 59.48695 68.51305  40  88
            Q25  Q50   Q75
completed 66.75 77.0 89.50
none      54.50 61.5 75.75

$comparison
NULL

$groups
          s_data1$writingS groups
completed         76.27333      a
none              64.00000      b

attr(,"class")
[1] "group"
```

For writing scores, ‘completed’ is the overall best level for prepcourse.

Thus we can conclude that individuals who had completed the preparatory course ended up scoring higher in the writing examination than those who did not take any preparatory course.

ii. Three way interaction present among Gender,Race and Parental education.

Two way layouts with parental education and gender for each level of race :

- Group A:

Table 3.2.1

```
> summary(anova_result31)
                                Df Sum Sq Mean Sq F value Pr(>F)
s_data11$Parentaledu             2  163.9   81.97   0.385  0.696
s_data11$gender                  1   97.5   97.47   0.457  0.524
s_data11$Parentaledu:s_data11$gender  2  193.3   96.67   0.454  0.655
Residuals                        6 1278.6  213.10
```

Parental education, gender or their interaction is not a significant factor for Writing scores of individuals in race group A.

- Group B:

Table 3.2.2

```
> summary(anova_result32)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data12$Parentaledu               2 2677.2  1338.6   7.734 0.0218 *
s_data12$gender                    1   90.7    90.7   0.524 0.4963
s_data12$Parentaledu:s_data12$gender  2  168.5    84.3   0.487 0.6369
Residuals                          6 1038.5   173.1
```

Parental education is a significant factor for Writing scores of individuals in Race group B. Hence we now find the best level of parental education .

Table 3.2.2.1

```
> #best level of parentaledu for B
> fisher_lsd_test32 <- LSD.test(anova_result32,"s_data12$Parentaledu", alpha = 0.05)
> print(fisher_lsd_test32)
$statistics
   MSerror Df     Mean       CV  t.value      LSD
  173.0833  6 68.58333 19.18267 2.446912 22.76308

$parameters
        test p.ajusted              name.t ntr alpha
  Fisher-LSD      none s_data12$Parentaledu   3  0.05

$means
                 s_data12$writingS       std r       se      LCL      UCL
Baseline College             80.25 13.889444 4 6.578057 64.15407 96.34593
Quality College              78.00 14.988885 4 6.578057 61.90407 94.09593
School                       47.50  3.872983 4 6.578057 31.40407 63.59593
                 Min Max   Q25  Q50   Q75
Baseline College  60  91 77.25 85.0 88.00
Quality College   56  88 74.75 84.0 87.25
School            44  53 45.50 46.5 48.50

$comparison
NULL

$groups
                 s_data12$writingS groups
Baseline College             80.25      a
Quality College              78.00      a
School                       47.50      b

attr(,"class")
[1] "group"
```

Hence, we can conclude that 'Baseline College' and 'Quality College' are jointly the best levels of parental education for race group B in accordance with Writing scores obtained.

- Group C:

Table 3.2.3

```
> summary(anova_result33)
                                   Df Sum Sq Mean Sq F value Pr(>F)
s_data13$Parentaledu                2  234.5   117.2   0.713  0.527
s_data13$gender                     1   18.7    18.7   0.114  0.747
s_data13$Parentaledu:s_data13$gender  2 1786.5   893.2   5.433  0.045 *
Residuals                           6  986.5   164.4
```

Only interaction between parental education and gender is a significant factor for individuals in race group C.

For Group C, one way ANOVA is implemented to find best level of parental education for each level of gender.

- For males in Group C : Table 3.2.3.1

```
> summary(anova_result331)
                      Df Sum Sq Mean Sq F value Pr(>F)
s_data131$Parentaledu  2 1612.0   806.0   4.433  0.127
Residuals              3  545.5   181.8
```

For males in group C, Parental education is not a significant factor for writing scores.

- For females in Group C: Table 3.2.3.2

```
> summary(anova_result332)
                      Df Sum Sq Mean Sq F value Pr(>F)
s_data132$Parentaledu  2    409   204.5   1.391  0.374
Residuals              3    441   147.0
```

For females in group C, parental education is not a significant factor for writing scores.

Hence, parental education is not a significant factor for writing scores in males or females but interaction between gender and parental education is a significant factor for reading scores of individuals in race group C as found earlier!

We perform one way ANOVA for gender with writing scores as response, for each level of parental education now:

- For individuals with parental education as 'School' in race group C: Table 3.2.3.3

```
> summary(anova_result333)
                   Df Sum Sq Mean Sq F value Pr(>F)
s_data1310$gender  1   1369  1369.0    14.8 0.0614 .
Residuals          2    185    92.5
```

Since 'gender' has a considerably small p-value we reject the null hypothesis that the levels of gender are not significantly different from each other. Thus, gender is a significant factor for individuals in race group C having level of parental education as 'School'. Thus we find the best level of gender for those individuals**:**

Table 3.2.3.3.1

```
> fisher_lsd_test333 <- LSD.test(anova_result333,"s_data1310$gender", alpha = 0.1)
> print(fisher_lsd_test333)
$statistics
  MSerror Df Mean       CV  t.value      LSD
     92.5  2   68 14.14366 2.919986 28.08352

$parameters
        test p.ajusted            name.t ntr alpha
  Fisher-LSD      none s_data1310$gender   2   0.1

$means
       s_data1310$writingS      std r       se      LCL       UCL Min Max   Q25
female                86.5  2.12132 2 6.800735 66.64195 106.35805  85  88 85.75
male                  49.5 13.43503 2 6.800735 29.64195  69.35805  40  59 44.75
        Q50   Q75
female 86.5 87.25
male   49.5 54.25

$comparison
NULL

$groups
       s_data1310$writingS groups
female                86.5      a
male                  49.5      b

attr(,"class")
[1] "group"
```

Female individuals having parental education category as 'School' belonging to race group C have a larger mean score in the writing exams, than male individuals belonging to the same category. The difference between means for males and females is large, which may not be significant at 5% level, but is found to be significant at 10% level.

- For individuals with parental education as 'Baseline College':

Table 3.2.3.4

```
> summary(anova_result334)
                    Df Sum Sq Mean Sq F value Pr(>F)
s_data1310$gender  1  240.2   240.2   2.219  0.275
Residuals          2  216.5   108.2
```

Gender is not a significant factor for individuals with parental education as 'Baseline College' in race group C.

- For individuals with parental education as 'Quality College':

Table 3.2.3.5

```
> summary(anova_result335)
                  Df Sum Sq Mean Sq F value Pr(>F)
s_data1310$gender  1    196   196.0    0.67  0.499
Residuals          2    585   292.5
```

Gender is not a significant factor for individuals with parental education as 'Quality College' in race group C.

- Group D:

Table 3.2.4

```
> summary(anova_result34)
                                    Df Sum Sq Mean Sq F value Pr(>F)
s_data14$Parentaledu                 2   24.5    12.3   0.035  0.966
s_data14$gender                      1   70.1    70.1   0.200  0.671
s_data14$Parentaledu:s_data14$gender 2  492.2   246.1   0.701  0.532
Residuals                            6 2105.5   350.9
```

Parental education, gender or their interaction is not a significant factor for Writing scores of individuals in race group D.

- Group E:

Table 3.2.5

```
> summary(anova_result35)
                                   Df Sum Sq Mean Sq F value Pr(>F)
s_data15$Parentaledu                2  181.5    90.7   0.287  0.761
s_data15$gender                     1  494.1   494.1   1.561  0.258
s_data15$Parentaledu:s_data15$gender 2 1083.2   541.6   1.711  0.258
Residuals                           6 1899.5   316.6
```

Parental education, gender or their interaction is not a significant factor for Writing scores of individuals in race group E.

➢ *Response: Average score*

Table 4

```
> summary(anova_result4)
                                                          Df Sum Sq Mean Sq F value
s_data1$gender                                             1     21    20.6   0.222
s_data1$race                                               4    676   169.1   1.825
s_data1$Parentaledu                                        2    398   199.2   2.150
s_data1$prepcourse                                         1   1850  1849.6  19.967
s_data1$gender:s_data1$race                                4    140    34.9   0.377
s_data1$gender:s_data1$Parentaledu                         2    180    89.9   0.971
s_data1$race:s_data1$Parentaledu                           8   2468   308.6   3.331
s_data1$gender:s_data1$prepcourse                          1     80    79.7   0.860
s_data1$race:s_data1$prepcourse                            4    670   167.6   1.809
s_data1$Parentaledu:s_data1$prepcourse                     2     42    21.2   0.229
s_data1$gender:s_data1$race:s_data1$Parentaledu            8   3786   473.2   5.108
s_data1$gender:s_data1$race:s_data1$prepcourse             4    997   249.2   2.690
s_data1$gender:s_data1$Parentaledu:s_data1$prepcourse      2    780   390.1   4.211
s_data1$race:s_data1$Parentaledu:s_data1$prepcourse        8   1654   206.7   2.231
Residuals                                                  8    741    92.6
                                                          Pr(>F)
s_data1$gender                                           0.65004
s_data1$race                                             0.21742
s_data1$Parentaledu                                      0.17890
s_data1$prepcourse                                       0.00209 **
s_data1$gender:s_data1$race                              0.81901
s_data1$gender:s_data1$Parentaledu                       0.41937
s_data1$race:s_data1$Parentaledu                         0.05426 .
s_data1$gender:s_data1$prepcourse                        0.38089
s_data1$race:s_data1$prepcourse                          0.22044
s_data1$Parentaledu:s_data1$prepcourse                   0.80045
s_data1$gender:s_data1$race:s_data1$Parentaledu          0.01655 *
s_data1$gender:s_data1$race:s_data1$prepcourse           0.10892
s_data1$gender:s_data1$Parentaledu:s_data1$prepcourse    0.05632 .
s_data1$race:s_data1$Parentaledu:s_data1$prepcourse      0.13865
Residuals
```

Significant factors at 5% level of significance:

i. prepcourse
ii. Interaction among gender,race,Parentaledu

i. Since prepcourse is a significant factor for average scores, we find its overall best level.

Table 4.1

```
> fisher_lsd_test4 <- LSD.test(anova_result4,"s_data1$prepcourse", alpha = 0.05)
> print(fisher_lsd_test4)
$statistics
   MSerror Df     Mean       CV  t.value      LSD
  92.63585  8 69.98556 13.75248 2.306004 5.730652

$parameters
        test p.ajusted             name.t ntr alpha
  Fisher-LSD      none s_data1$prepcourse   2  0.05

$means
          s_data1$AvgScore      std  r       se      LCL      UCL      Min
completed         75.53778 16.25346 30 1.757231 71.48560 79.58996 42.33333
none              64.43333 13.09358 30 1.757231 60.38115 68.48552 39.00000
               Max      Q25      Q50      Q75
completed 97.66667 68.91667 74.33333 88.16667
none      86.66667 58.33333 61.50000 74.75000

$comparison
NULL

$groups
          s_data1$AvgScore groups
completed         75.53778      a
none              64.43333      b

attr(,"class")
[1] "group"
```

For average scores, 'completed' is the overall best level for prepcourse.

Thus we can conclude that individuals who had completed the preparatory course ended up scoring higher average marks in the examination than those who did not take any preparatory course.

ii. Three way interaction present among gender,race and parental education

Two way layouts with parental education and gender for each level of race :

- Group A: Table 4.2.1

```
> summary(anova_result41)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data11$Parentaledu               2  146.1   73.03   0.313  0.742
s_data11$gender                    1   19.1   19.08   0.082  0.784
s_data11$Parentaledu:s_data11$gender  2  218.9  109.47   0.470  0.646
Residuals                          6 1398.9  233.15
```

Parental education, gender or their interaction is not a significant factor for average scores of individuals in race group A.

- Group B: Table 4.2.2

```
> summary(anova_result42)
                                    Df Sum Sq Mean Sq F value Pr(>F)
s_data12$Parentaledu                 2 2405.4  1202.7   7.181 0.0256 *
s_data12$gender                      1    0.0     0.0   0.000 0.9886
s_data12$Parentaledu:s_data12$gender 2  216.1   108.0   0.645 0.5575
Residuals                            6 1004.9   167.5
```

Parentaledu is a significant factor for average scores of individuals belonging to race group B.

We find the best level of Parentaledu for average scores of individuals in race group B:

Table 4.2.2.1

```
> fisher_lsd_test42 <- LSD.test(anova_result42,"s_data12$Parentaledu", alpha = 0.05)
> print(fisher_lsd_test42)
$statistics
   MSerror Df     Mean       CV  t.value      LSD
  167.4815  6 68.38889 18.92334 2.446912 22.39168

$parameters
        test p.ajusted              name.t ntr alpha
  Fisher-LSD      none s_data12$Parentaledu   3  0.05

$means
                 s_data12$AvgScore       std r       se      LCL      UCL
Baseline College          80.33333 14.155224 4 6.470732 64.50002 96.16664
Quality College           76.33333 12.268599 4 6.470732 60.50002 92.16664
School                    48.50000  7.490735 4 6.470732 32.66669 64.33331
                      Min      Max      Q25      Q50      Q75
Baseline College 60.00000 92.66667 77.25000 84.33333 87.41667
Quality College  60.33333 88.33333 70.33333 78.33333 84.33333
School           42.33333 58.33333 42.83333 46.66667 52.33333

$comparison
NULL

$groups
                 s_data12$AvgScore groups
Baseline College          80.33333      a
Quality College           76.33333      a
School                    48.50000      b

attr(,"class")
[1] "group"
```

Hence, 'Baseline College' and 'Quality College' are jointly the best levels of parental education for individuals belonging to race group B.

- Group C: Table 4.2.3

```
> summary(anova_result43)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data13$Parentaledu               2  135.2    67.6   0.481 0.6401
s_data13$gender                    1   25.0    25.0   0.178 0.6877
s_data13$Parentaledu:s_data13$gender  2 1853.5   926.7   6.593 0.0306 *
Residuals                          6  843.3   140.6
```

The interaction between Parentaledu and gender is a significant factor for average scores in individuals belonging to race group C.

Table 4.2.3.1

```
> #for males:
> anova_result431 <- aov(s_data131$AvgScore ~ s_data131$Parentaledu
+                        ,data=s_data131)
> summary(anova_result431)
                     Df Sum Sq Mean Sq F value Pr(>F)
s_data131$Parentaledu  2 1438.4   719.2   5.087  0.109
Residuals              3  424.2   141.4
> #for females:
> anova_result432 <- aov(s_data132$AvgScore ~ s_data132$Parentaledu
+                        ,data=s_data132)
> summary(anova_result432)
                     Df Sum Sq Mean Sq F value Pr(>F)
s_data132$Parentaledu  2  550.3   275.2   1.969  0.284
Residuals              3  419.2   139.7
```

Table 4.2.3.2

We find that parental education is not a significant factor for average scores for any level of gender. So, we do the similar one way ANOVA layout for each level of Parentaledu to find the cause of significant interaction between Parentaledu and gender in race group C .

- For Parentaledu level 'School' :

Table 4.2.3.3

```
> summary(anova_result433)
                   Df Sum Sq Mean Sq F value Pr(>F)
s_data1310$gender   1 1034.7  1034.7   12.25 0.0728 .
Residuals           2  168.9    84.5
```

Gender is a significant factor for average marks of individuals belonging to race group C and with parental education level 'School'.

So we find it's best level.

Table 4.2.3.3.1

```
> fisher_lsd_test433 <- LSD.test(anova_result433,"s_data1310$gender", alpha = 0.1)
> print(fisher_lsd_test433)
$statistics
   MSerror Df     Mean       CV  t.value      LSD
  84.47222  2 68.58333 13.40104 2.919986 26.83723

$parameters
        test p.ajusted            name.t ntr alpha
  Fisher-LSD      none s_data1310$gender   2   0.1

$means
       s_data1310$AvgScore       std r       se      LCL       UCL      Min
female            84.66667  0.942809 2 6.498932 65.68988 103.64345 84.00000
male              52.50000 12.963624 2 6.498932 33.52321  71.47679 43.33333
           Max      Q25      Q50      Q75
female 85.33333 84.33333 84.66667 85.00000
male   61.66667 47.91667 52.50000 57.08333

$comparison
NULL

$groups
       s_data1310$AvgScore groups
female            84.66667      a
male              52.50000      b

attr(,"class")
[1] "group"
```

Thus,females in race group C,with parental education level 'School' performed better on average than males in the same group.

- For Parentaledu level 'Baseline College' :

Table 4.2.3.4

```
> summary(anova_result434)
                   Df Sum Sq Mean Sq F value Pr(>F)
s_data1310$gender   1  330.0   330.0    2.97  0.227
Residuals           2  222.3   111.1
```

Gender is not a significant factor for average marks of individuals belonging to race group C , with parental education level 'Baseline College'.

- For Parentaledu level 'Quality College':

Table 4.2.3.5

```
> summary(anova_result435)
                   Df Sum Sq Mean Sq F value Pr(>F)
s_data1310$gender   1  513.8   513.8   2.273  0.271
Residuals           2  452.1   226.1
```

Gender is not a significant factor for average marks of individuals belonging to race group C ,with parental education level 'Quality College'.

- Group D: Table 4.2.4

```
> summary(anova_result44)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data14$Parentaledu               2   70.1   35.06   0.114  0.894
s_data14$gender                    1    0.0    0.04   0.000  0.992
s_data14$Parentaledu:s_data14$gender  2  559.8  279.90   0.910  0.452
Residuals                          6 1845.1  307.52
```

Parental education,gender or their interaction is not a significant factor of average scores for individuals in race group D.

- Group E: Table 4.2.5

```
> summary(anova_result45)
                                  Df Sum Sq Mean Sq F value Pr(>F)
s_data15$Parentaledu               2  110.1    55.0   0.192  0.830
s_data15$gender                    1  116.1   116.1   0.405  0.548
s_data15$Parentaledu:s_data15$gender  2 1117.1   558.6   1.947  0.223
Residuals                          6 1721.4   286.9
```

Parental education, gender or their interaction is not a significant factor of average scores for individuals in race group E.

7. CONCLUSION

➢ Response: Math score

We can conclude that –

1. Individuals having completed the preparatory course were better performers at math than those who had not taken the preparatory course.
2. Parental education, gender or their interaction is not a significant factor for Math scores of individuals in race Group A.
3. Parental Education is a significant factor for math scores of individuals in race Group B.

3.1. Joint best levels of parental education for math scores of individuals in race group B is Baseline College and Quality College.

4. Only Interaction between Parental education and gender is a significant factor in race group C.

4.1. For males in group C, Parental education is a significant factor for Math scores. Joint best levels of parental education for Math scores of male individuals in group C are Quality College and Baseline College.

4.2. Parental Education is not a significant factor for Math scores of female individuals in race Group C.

5. Parental education, gender or their interaction is not a significant factor for Math scores of individuals in race Group D.

6. Parental education, gender or their interaction is not a significant factor for Math scores of individuals in race Group E.

Response: Math score (Tabular representation)

<table>
<tr><td rowspan="2">Gender
Race</td><td colspan="3">Male</td><td colspan="3">Female</td></tr>
<tr><td>Parental Education</td><td>School</td><td>Baseline College</td><td>Quality College</td><td>School</td><td>Baseline College</td><td>Quality College</td></tr>
<tr><td>A</td><td colspan="6">Students in race group A have performed equally irrespective of their gender or parental education levels.</td></tr>
<tr><td>B</td><td></td><td colspan="2">Male wards of parents with education level ‘Baseline College’ & ‘Quality College’ performed comparatively better.</td><td></td><td colspan="2">Female wards of parents with education level ‘Baseline College’ & ‘Quality College’ comparatively performed better.</td></tr>
<tr><td>C</td><td></td><td colspan="2">Male wards of parents with education level ‘Baseline College’ and ‘ Quality College’ performed comparatively better.</td><td colspan="3">Female wards of parents with all education levels performed equally.</td></tr>
<tr><td>D</td><td colspan="6">Students in race group D have performed equally well irrespective of their gender or parental education levels.</td></tr>
<tr><td>E</td><td colspan="6">Students in race group E have performed equally well irrespective of their gender or parental education levels.</td></tr>
</table>

- ***Individuals having completed the preparatory course were better performers at math than those who had not taken the preparatory course.***

➢ Response: Reading score

We can conclude that-

1. Individuals who completed the preparatory course ended up scoring higher in the reading examination than those who did not take any preparatory course.
2. Parental education, gender or their interaction is not a significant factor for Reading scores of individuals in race group A.
3. Parental education is a significant factor for Reading scores of individuals in Race group B. Hence, we find the best level of parental education now.

    3.1. Joint best levels of parental education for reading scores of individuals in race group B are Baseline College and Quality College.

4. Parental education, gender or their interaction is not a significant factor for Reading scores of individuals in race group C.
5. Parental education, gender or their interaction is not a significant factor for Reading scores of individuals in race group D.
6. Parental education, gender or their interaction is not a significant factor for Reading scores of individuals in race group E.

Response: Reading score (Tabular representation)

Gender / Race	Male			Female		
Parental Education	School	Baseline College	Quality College	School	Baseline College	Quality College
A	*Students in race group A have performed equally irrespective of their gender or parental education levels.*					
B		*Male wards of parents with education level 'Baseline College' & 'Quality College' performed comparatively better.*			*Female wards of parents with education level 'Baseline College' & 'Quality College' performed comparatively better.*	
C	*Students in race group C have performed equally well irrespective of their gender or parental education levels.*					
D	*Students in race group D have performed equally well irrespective of their gender or parental education levels.*					
E	*Students in race group E have performed equally well irrespective of their gender or parental education levels.*					

- ***Individuals having completed the preparatory course were better performers at reading than those who had not taken the preparatory course.***

➢ Response: Writing score

We can conclude that-

1. Individuals who had completed the preparatory course ended up scoring higher in the writing examination than those who did not take any preparatory course.
2. Parental education, gender or their interaction is not a significant factor for Writing scores of individuals in race group A.
3. Parental education is a significant factor for Writing scores of individuals in Race group B.

   3.1. 'Baseline College' and 'Quality College' are jointly the best levels of parental education for race group B in accordance with reading scores obtained.
4. Only interaction between parental education and gender is a significant factor for individuals in race group C.

   4.1. Female individuals having parental education category as 'School' belonging to race group C are the better performers in the writing exam than male individuals belonging to the same category.

   4.2. Gender is not a significant factor for individuals with parental education as 'Baseline College' in race group C.

   4.3. Gender is not a significant factor for individuals with parental education as 'Quality College' in race group C.
5. Parental education, gender or their interaction is not a significant factor for Writing scores of individuals in race group D.
6. Parental education, gender or their interaction is not a significant factor for Writing scores of individuals in race group E.

Response: Writing score (Tabular representation)

<table>
<tr><td>Gender
Race</td><td colspan="3">Male</td><td colspan="3">Female</td></tr>
<tr><td>Parental Education</td><td>School</td><td>Baseline College</td><td>Quality College</td><td>School</td><td>Baseline College</td><td>Quality College</td></tr>
<tr><td>A</td><td colspan="6">Students in race group A have performed equally irrespective of their gender or parental education levels.</td></tr>
<tr><td>B</td><td></td><td colspan="2">Male wards of parents with education level ‘Baseline College’ & ‘Quality College’ have comparatively performed better.</td><td></td><td colspan="2">Female wards of parents with education level ‘Baseline College’ & ‘Quality College’ have comparatively performed better.</td></tr>
<tr><td>C</td><td></td><td></td><td></td><td>Female wards of parents with education level ‘School’ have comparatively performed better than their male counterparts.</td><td></td><td></td></tr>
<tr><td>D</td><td colspan="6">Students in race group D have performed equally well irrespective of their gender or parental education levels.</td></tr>
<tr><td>E</td><td colspan="6">Students in race group E have performed equally well irrespective of their gender or parental education levels.</td></tr>
</table>

- ***Individuals having completed the preparatory course were better performers at writing than those who had not taken the preparatory course.***

➢ Response: Average score

We can conclude that-

1. Individuals who had completed the preparatory course ended up scoring higher average marks in the examination than those who did not take any preparatory course.
2. Parental education, gender or their interaction is not a significant factor for average scores of individuals in race group A.
3. Parental education is a significant factor for average scores of individuals belonging to race group B.
   3.1. 'Baseline College' and 'Quality College' are jointly the best levels of parental education for average scores of individuals belonging to race group B.
4. Only interaction between Parentaledu and gender is a significant factor for average scores of individuals belonging to race group C.
   4.1. Gender is a significant factor for average marks of individuals belonging to race group C and with parental education level 'School'.
      1. Females in race group C,with parental education level 'School' performed better on average than males in the same group.
   4.2. Gender is not a significant factor for average marks of individuals belonging to race group C , with parental education level 'Baseline College'.
   4.3. Gender is not a significant factor for average marks of individuals belonging to race group C ,with parental education level 'Quality College'.
5. Parental education, gender or their interaction is not a significant factor of average scores for individuals in race group D.
6. Parental education, gender or their interaction is not a significant factor of average scores for individuals in race group E.

Response: Average score (Tabular representation)

<table>
<tr><th>Parental Education
Race</th><th colspan="2">School</th><th colspan="2">Baseline College</th><th colspan="2">Quality College</th></tr>
<tr><th>Gender</th><td>Male</td><td>Female</td><td>Male</td><td>Female</td><td>Male</td><td>Female</td></tr>
<tr><td>A</td><td colspan="6">Students in race group A have performed equally irrespective of their gender or parental education levels.</td></tr>
<tr><td>B</td><td colspan="2"></td><td colspan="4">Wards of parents with education level ‘Baseline College’ & ‘Quality College’ have comparatively performed better.</td></tr>
<tr><td>C</td><td colspan="2">Female wards of parents with education level ‘School’ have comparatively performed better than their male counterparts.</td><td colspan="4">Wards of parents with education level ‘Baseline College’ & ‘Quality College’ performed equally irrespective of gender.</td></tr>
<tr><td>D</td><td colspan="6">Students in race group D have performed equally well irrespective of their gender or parental education levels.</td></tr>
<tr><td>E</td><td colspan="6">Students in race group E have performed equally well irrespective of their gender or parental education levels.</td></tr>
</table>

- ***Individuals having completed the preparatory course were better performers than those who had not taken the preparatory course in terms of average score.***

8. Code Repository

R codes :

[The initial section presents R code snippets for exploratory data analysis conducted at the beginning. To streamline the presentation and prevent redundancy, only one example each of a histogram, scatter plot, heatmap, and two boxplots are provided. Similar graphical plots are generated analogously for the remaining data visualization tasks.]

Figure 1.4 (Histogram of Average Scores) :

```
data=StudentPerformance
data$AvgScore=(data$mathS+data$readingS+data$writingS)/3
hist(data$AvgScore,prob=TRUE,col='sky blue',xlab='Average Scores',
     main='Histogram of Average Scores')
curve(dnorm(x,mean(data$AvgScore),sd(data$AvgScore)),add=TRUE,lwd=2)
```

Figure 2.1 (Boxplot of Math Scores by levels of parental education) :

```
library(ggplot2)
# Create boxplot of initial vs final levels of Parental education math scores
boxplotf1 <- ggplot(data, aes(x = Parentaledu, y = mathS, fill = parentaledu)) +
  geom_boxplot() +
  labs(x = "Parental Education Initial and Final levels",
       y = "Math Score", fill = "Parental Education Initial levels") +
  ggtitle("Boxplot of Math Scores by levels of Parental Education") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
# Display the boxplot
print(boxplotf1)
```

Figure 3.3 (Scatter plot of Math Scores by Writing Scores) :

```
scatter_plot3 <- ggplot(data, aes(x = mathS, y = writingS, color = gender)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "black",lwd=1) +
  labs(x = "Math Score", y = "Writing Score", color = "Gender") +
  ggtitle("Scatter Plot of Math vs. Writing Scores by Gender") +
  theme_dark()
# Display the scatter plot
print(scatter_plot3)
```

Figure 3.4 (Heatmap of correlation matrix of Math,Reading and Writing Scores) :

```
# Compute correlation matrix
correlation_matrix <- cor(data[, c("mathS", "readingS","writingS")])
colnames(correlation_matrix) <- c("Math Scores", "Writing Scores", "Reading Scores")
rownames(correlation_matrix) <- c("Math Scores", "Writing Scores", "Reading Scores")
# Plot correlation heatmap
library(corrplot)
heatmap_plot <- corrplot(correlation_matrix, method = "color", type = "upper",
                         order = "hclust", addrect = 4)
# Display the correlation heatmap
print(heatmap_plot)
```

Figure 4.4 (Boxplot of average score by levels of prepcourse and gender):

```
#boxplot of average scores with prep course and gender
boxplot1 <- ggplot(data, aes(x = prepcourse, y = AvgScore, fill = gender)) +
  geom_boxplot() +
  labs(x = "Preparatory Course", y = "Average Score", fill = "Gender") +
  ggtitle("Boxplot of Average Scores by Preparatory Course") +
  theme_minimal()
# Display the boxplot
print(boxplot1)
```

[The subsequent section contains the R codes for the analyses considering average scores or 'AvgScore' as the response variable, to avoid redundancy. The analyses for math, reading and writing scores as response variables are done in the exact same way, only changing the response variable as and when required.]

Table 4 :

```
data$Parentaledu <- ifelse(data$parentaledu %in% c("some high school","high school" ), "School",
                           ifelse(data$parentaledu %in% c("associate's degree","some college"), "Baseline College", "Quality College"))
data$gender <- factor(data$gender, levels = c("male", "female"))
data$race <- factor(data$race, levels = c("group A", "group B", "group C", "group D", "group E"))
data$Parentaledu<-factor(data$Parentaledu,levels=c("School","Baseline College","Quality College"))
data$prepcourse<-factor(data$prepcourse,levels=c("none","completed"))
library(tidyverse)
library(magrittr)
library(dplyr)
library(conflicted)
library(agricolae)
set.seed(105)
s_data1<- data %>%
  group_by(gender, Parentaledu,race,prepcourse) %>%
  sample_n(1) %>%
  ungroup()
#Four-way Anova using gender,Parentaledu,race,prep course with response as average scores
anova_result4 <- aov(s_data1$AvgScore ~ s_data1$gender*s_data1$race*s_data1$Parentaledu*s_data1$prepcourse
                     -s_data1$gender:s_data1$race:s_data1$Parentaledu:s_data1$prepcourse
                     ,data = s_data1)
summary(anova_result4)
```

Table 4.1:

```
fisher_lsd_test4 <- LSD.test(anova_result4,
                      "s_data1$prepcourse", alpha = 0.05)
print(fisher_lsd_test4)
```

Table 4.2.1:

```
#keeping average scores as response variables,2 way layout on gender and parental education for each level of race
#group A:
data1=s_data1[s_data1$race=='group A',]
s_data11<- data1 %>%
  group_by(Parentaledu,gender) %>%
  sample_n(2) %>%
  ungroup()
anova_result41 <- aov(s_data11$AvgScore ~ s_data11$Parentaledu*s_data11$gender,data=s_data11)
summary(anova_result41)
```

Table 4.2.2:

```
#for group B:
data2=s_data1[s_data1$race=='group B',]
s_data12<- data2 %>%
  group_by(Parentaledu,gender) %>%
  sample_n(2) %>%
  ungroup()
anova_result42 <- aov(s_data12$AvgScore ~ s_data12$Parentaledu*s_data12$gender
                      ,data=s_data12)
summary(anova_result42)
```

Table 4.2.2.1:

```
fisher_lsd_test42 <- LSD.test(anova_result42,"s_data12$Parentaledu", alpha = 0.05)
print(fisher_lsd_test42)
```

Table 4.2.3:

```
#for group C:
data3=s_data1[s_data1$race=='group C',]
s_data13<- data3 %>%
  group_by(Parentaledu,gender) %>%
  sample_n(2) %>%
  ungroup()
anova_result43 <- aov(s_data13$AvgScore ~ s_data13$Parentaledu*s_data13$gender
                      ,data=s_data13)
summary(anova_result43)
```

Table 4.2.4:

```
#group D
data4=s_data1[s_data1$race=='group D',]
s_data14<- data4 %>%
  group_by(Parentaledu,gender) %>%
  sample_n(2) %>%
  ungroup()
anova_result44 <- aov(s_data14$AvgScore ~ s_data14$Parentaledu*s_data14$gender
                      ,data=s_data14)
summary(anova_result44)
```

Table 4.2.5:

```
#group E
data5=s_data1[s_data1$race=='group E',]
s_data15<- data5 %>%
  group_by(Parentaledu,gender) %>%
  sample_n(2) %>%
  ungroup()
anova_result45 <- aov(s_data15$AvgScore ~ s_data15$Parentaledu*s_data15$gender
                      ,data=s_data15)
summary(anova_result45)
```

Table 4.2.3.1 - 4.2.3.5 :

```
#for males:
data31=data3[data3$gender=='male',]
s_data131<- data31 %>%
  group_by(Parentaledu) %>%
  sample_n(2) %>%
  ungroup()
anova_result431 <- aov(s_data131$AvgScore ~ s_data131$Parentaledu,data=s_data131)
summary(anova_result431)
#for females:
data32=data3[data3$gender=='female',]
s_data132<- data32 %>%
  group_by(Parentaledu) %>%
  sample_n(2) %>%
  ungroup()
anova_result432 <- aov(s_data132$AvgScore ~ s_data132$Parentaledu,data=s_data132)
summary(anova_result432)
```

```
#for School parentaledu:
data310=data3[data3$Parentaledu=='School',]
s_data1310<- data310 %>%
  group_by(gender) %>%sample_n(2) %>%ungroup()
anova_result433 <- aov(s_data1310$AvgScore ~ s_data1310$gender,data=s_data1310)
summary(anova_result433)
fisher_lsd_test433 <- LSD.test(anova_result433,"s_data1310$gender", alpha = 0.1)
print(fisher_lsd_test433)
#for Baseline college parentaledu:
data310=data3[data3$Parentaledu=='Baseline College',]
s_data1310<- data310 %>%group_by(gender) %>%sample_n(2) %>%ungroup()
anova_result434 <- aov(s_data1310$AvgScore ~ s_data1310$gender,data=s_data1310)
summary(anova_result434)
#for Quality college parentaledu:
data310=data3[data3$Parentaledu=='Quality College',]
s_data1310<- data310 %>%group_by(gender) %>%sample_n(2) %>%ungroup()
anova_result435<- aov(s_data1310$AvgScore ~ s_data1310$gender,data=s_data1310)
summary(anova_result435)
```

Table 4.3.1 - 4.3.3:

```
#Studying three way interaction among Parentaledu,gender and prepcourse
#for those with Parentaledu level 'School'
data1=s_data1[s_data1$Parentaledu=='School',]
set.seed(105)
s_data11<- data1 %>%group_by(gender,prepcourse) %>%sample_n(2) %>%ungroup()
anova_result46 <- aov(s_data11$AvgScore ~ s_data11$gender*s_data11$prepcourse,data=s_data11)
summary(anova_result46)
#for those with Parentaledu level 'Baseline College'
data1=s_data1[s_data1$Parentaledu=='Baseline College',]
set.seed(105)
s_data11<- data1 %>%group_by(gender,prepcourse) %>%sample_n(2) %>%ungroup()
anova_result47 <- aov(s_data11$AvgScore ~ s_data11$gender*s_data11$prepcourse,data=s_data11)
summary(anova_result47)
fisher_lsd_test471 <- LSD.test(anova_result47,"s_data11$prepcourse", alpha = 0.05)
print(fisher_lsd_test471)
fisher_lsd_test472 <- LSD.test(anova_result47,"s_data11$gender", alpha = 0.1)
print(fisher_lsd_test472)
#for those with Parentaledu level 'Quality College'
data1=s_data1[s_data1$Parentaledu=='Quality College',]
set.seed(105)
s_data11<- data1 %>%group_by(gender,prepcourse) %>%sample_n(2) %>%ungroup()
anova_result48 <- aov(s_data11$AvgScore ~ s_data11$gender*s_data11$prepcourse,data=s_data11)
summary(anova_result48)
```

9. Future Prospects

Moving forward, potential avenues for further exploration include a deeper investigation into gender disparities in academic performance to uncover underlying factors and assess gender-specific interventions. Additionally, delving into the nuanced role of parental education in shaping academic outcomes could yield valuable insights, particularly in groups where it exerts a significant influence. Evaluating the effectiveness of preparatory courses and examining differential effects across demographic groups could inform targeted interventions for the school concerned. Further exploration of interaction effects, such as the significant interaction between gender and parental education observed in race group C, may elucidate underlying mechanisms. Longitudinal studies tracking academic progress over time and translating findings into actionable recommendations for educational interventions and policies could significantly contribute to promoting academic equity and improving performances of the pre-adolescent students, paving way for future academic excellence.

Exploring the effectiveness of the same offered preparatory course in improving academic performance across various demographic groups could shed light on tailored interventions. Additionally, longitudinal studies tracking academic trajectories over time could reveal critical periods for intervention and uncover factors influencing long-term academic success. Moreover, advocating for policy changes to address the specific needs of underrepresented or disadvantaged demographic groups identified in the study could lead to impactful systemic improvements in educational practices and outcomes of the academic records of the pre-adolescents admitted in the school under study.

Unfortunately, the 1001 observations available for our study were collected randomly from the students and happened to be in such a way that for the four factors considered gender, parental education, race and preparatory course (with 2,3,5 & 2 levels respectively), a maximum of only one observation could be chosen randomly and allocated to each of the 60 cells for further analyses. For statistical considerations, if the experiment could be designed in such a way that the initial 4-way ANOVA layout could have more than one (equal) number of observations in each cell, the four factor interaction could be studied in the model shedding light to more intricate, complicated relationships among the four factors under consideration. Additionally, the F-tests with right tailed critical region for testing the respective hypotheses in each case while performing ANOVA would be more powerful if we performed the 4-way layout ANOVA with more than one observation per cell.

10. Acknowledgement

I am deeply grateful for the invaluable support and guidance provided by my mentor, Sri Palas Pal, from the Department of Statistics, Ramakrishna Mission Residential College (Autonomous), Narendrapur, throughout the development of this project. His unwavering academic and mental support, available at every hour of the day, enabled me to meet deadlines without compromising on quality. I am forever thankful for his continuous enthusiasm in assisting me whenever I faced challenges during this project.

I also extend my profound gratitude to Dr. Parthasarathi Chakraborty, our esteemed Head of the Department of Statistics, for his guidance and support throughout my undergraduate program. His mentorship has not only enhanced my academic and professional growth but has also imparted invaluable life lessons alongside statistical knowledge.

Additionally, I extend my gratitude to all the teachers in our department for their exceptional teaching and dedication. Their unique teaching styles and passion for the subject have greatly influenced my understanding and sparked my active interest in Statistics throughout my academic journey. I am grateful for the opportunity to undertake this project during my final semester of the Bachelor of Science in Statistics (Honours) course.

11. Bibliography

1. Analysis of Variance – Scheffe
2. Outline (Volume Two) – A.M. Gun, M.K. Gupta, B. Dasgupta
3. Analysis of Variance – Guenther
4. Fundamentals of Statistics (Volume Two) - A.M. Gun, M.K. Gupta, B. Dasgupta
5. Design and Analysis of Experiments - Douglas C. Montgomery
6. An R Companion to Applied Regression - John Fox and Sanford Weisberg
7. Discovering Statistics Using R - Andy Field, Jeremy Miles, and Zoe Field
8. Scheffe Methods for Analysis of Variance - J. K. Lindsey