**Semester Project**

**Detailed analysis of the influence of various factors on students’ performance in exams.**

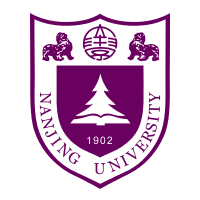
**By**

**MG21145010**

**Name: Muhammad Taha Nasir**

**Subject: Quantitative Information Analysis**

**Assignment submitted to the faculty of Information Resources Management at Nanjing University**



**Department of Information Resources Management**

**Nanjing University, China**

**31st July, 2022**

**Table of Contents**

[1) Introduction 1](#_Toc110197881)

[1.1 Background and Context: 1](#_Toc110197882)

[1.2 Problem Statement 1](#_Toc110197883)

[1.3 Research Questions 2](#_Toc110197884)

[1.4 Relevance and importance of research 2](#_Toc110197885)

[2) Literature review 2](#_Toc110197886)

[2.1 Factors affecting academic performance 2](#_Toc110197887)

[2.2 Central role of family 3](#_Toc110197888)

[2.3 Techniques for performance prediction 4](#_Toc110197889)

[3) Research Methodology 5](#_Toc110197890)

[3.1 Data Collection 5](#_Toc110197891)

[3.1.1 Dataset description: 5](#_Toc110197892)

[3.2 Hypothesis formulation 5](#_Toc110197893)

[3.3 Tools and techniques used 6](#_Toc110197894)

[3.3.1 Descriptive statistics 6](#_Toc110197895)

[3.3.2 T-test 6](#_Toc110197896)

[3.3.3 Chi-squared test 6](#_Toc110197897)

[3.3.4 Multiple Linear Regression 7](#_Toc110197898)

[4) Analysis, Results and Explanation 7](#_Toc110197899)

[4.1 Loading libraries 7](#_Toc110197900)

[4.2 Data import 7](#_Toc110197901)

[4.3 Data manipulation 8](#_Toc110197902)

[4.4 Descriptive statistics 8](#_Toc110197903)

[4.5 Data Visualization 9](#_Toc110197904)

[4.6 t-test 14](#_Toc110197905)

[4.7 Chi-squared test: 16](#_Toc110197906)

[4.8 Correlation 17](#_Toc110197907)

[4.9 Multiple Linear Regression Analysis 18](#_Toc110197908)

[5) Discussion 19](#_Toc110197909)

[5.1 Contributions 19](#_Toc110197910)

[5.2 Discussion about findings: 20](#_Toc110197911)

[5.3 Policy implications 21](#_Toc110197912)

[5.4 Limitations 21](#_Toc110197913)

[6) Conclusion 21](#_Toc110197914)

[7) References 23](#_Toc110197915)

**List of Tables**

[Table 1: Gender frequency 10](#_Toc110197940)

[Table 2: Race frequency 11](#_Toc110197941)

[Table 3: Test prep frequency 14](#_Toc110197942)

[Table 4: Correlation matrix 20](#_Toc110197943)

**List of Figures**

[Figure 1: Side-by\_side histograms of numerical variables 10](#_Toc110197947)

[Figure 2: Frequency of gender 10](#_Toc110197948)

[Figure 3: Frequency of race 11](#_Toc110197949)

[Figure 4: Frequency of parent\_education 12](#_Toc110197950)

[Figure 5: Frequency of lunch 13](#_Toc110197951)

[Figure 6: Frequency of test\_prep 13](#_Toc110197952)

[Figure 7: Bar plot of gender grouped by ethnicities 14](file:///D:\MS%20IRM\Sem%202\QIA\Final%20Project\Final%20Project_Muhammad%20Taha%20Nasir.docx#_Toc110197953)

[Figure 8: boxplot of parent education vs math score 17](#_Toc110197954)

[Figure 9: Correlation plot 17](#_Toc110197955)

[Figure 10: Correlation chart 21](#_Toc110197956)

# Introduction

## Background and Context:

Education is an essential element for the development of a country. Lack of knowledge in higher educational system could prevent system management to achieve quality in education. It is well known that children’s academic performances are affected by both their family backgrounds and contextual or structural factors such as the urban–rural difference and regional variation. This article evaluates the relative importance of family background versus structural factors in determining children’s academic achievements.[12]

An increasing interest has arisen during the past decade to identify the most important factors influencing students’ performance in higher education, especially by using data mining methods and techniques. This field of research is usually identified as educational data mining (EDM). The application of such research in assisting in the early identification of low performing students to overcome their learning challenges and improve their learning outcomes, which in turn serves the institutional goals of providing high-quality educational ecosystems, is credited with motivating this interest. EDM is also quickly evolving into a significant area of study thanks to its capacity to draw fresh insights from vast amounts of student data. This issue interests this paper in equal measure. Data mining methodology can help associating this knowledge gaps in higher education system. [13]

A better student model yields better instruction, which leads to improved learning. More accurate skill diagnosis leads to better prediction of what a student knows which provides better assessment. Better assessment leads to more efficient learning overall. The main objectives of data mining in practice tend to be prediction and description [1]. Predicting performance involves variables like attendance, IAT marks and assignment grades etc. in the student database to predict the unknown values. Data mining is the core process of knowledge discovery in databases. It is the process of extracting of useful patterns from the large database. In order to analyse large amount of information, the area of Knowledge Discovery in Databases (KDD) provides techniques by which the interesting patterns are extracted. Therefore, KDD utilizes methods at the cross point of machine learning, statistics and database systems.[2] Data mining is the application of efficient algorithms to detect the desired patterns contained within the given data.

## Problem Statement

After a detailed literature review on adverse selection review and having an in-depth study of the prior research and studies in ‘academic performance’, the research problem: “**Detailed analysis of the influence of various factors on students’ performance in exams”** has been selected.

## Research Questions

The research questions for this particular exercise are:

* Are there any differences between the exam scores of female and male students?
* Is completing the test preparation course related to gender?
* Is lunch eating status related to gender?
* Which variables explain the students score in maths, reading and writing?

These questions will be answered after running the analysis in RStudio.

## Relevance and importance of research

* Our study aims to analyse the academic performance of the students by studying the social and personal factors that influence student’s performance.
* Through the information generated through the experiment, the potential of data mining and regression techniques to predict and analyse student performance will be accessed which, in turn, will lead to developing education strategies and serving the community.
* The analysis of the influence of parents level of education, student’s race/ethnicity and age will give insights to the educational institutions (public schools) and parents on what to focus on if they want to improve the academic performance of students.

# Literature review

The following are the research findings based on the literature review:

## Factors affecting academic performance

All of the research reviews support the hypothesis that student performance depends on different socio-economic, psychological, environmental factors. The findings of research studies focused that student performance is affected by different factors such as learning abilities because new paradigm about learning assumes that all students can and should learn at higher levels but it should not be considered as constraint because there are other factors like race, gender, sex that can affect student’s performance. (Hansen, Joe B.2000). Some of the researchers even tried to explain the link between students’ achievements, economic circumstances and the risk of becoming a drop-out that proved to be positive (Goldman, N., Haney, W., and Koffler, S., 1988, Pallas, A., Natriello, G., McDill, E., 1989, Levin, H., 1986) B.A Chansarkar and A. Mishaeloudis (2001), explained the effects of age, qualification distance from learning place etc. on student performance. The performance of students on the module is not affected by such factors as age, sex and place of residence but is associated with qualification in quantitative subjects.[11]

It is also found that those who live near the university perform better than other students. Yvonne Beaumont Walters, Kola Soyibo (1998) further elaborated that student performance is very much dependent on SEB (socio economic back ground) as per their statement, “High school students’ level of performance is with statistically significant differences, linked to their gender, grade level, school location, school type, student type and socio-economic background (SEB).” Kirby, Winston et al. (2002) focused on student’s impatience (his time-discount behaviour) that influences his own academic performance. Goethe found out that weak students do better when grouped with other weak students.[10] (As implied by Zajonc’s analysis of older siblings (1976) it shows that students’ performance improves if they are with the students of their own kind. There are often different results by gender, as in Hoxby’s K-12 results (2000); Sacerdote (2001) finds that grades are higher when students have unusually academically strong roommates [20]. The results of Zimmerman (1999, 2001) were somewhat contradictory to Goethe results but again it proved that students’ performance depends on number of different factors, it says that weak peers might reduce the grades of middling or strong students. (Alexander, Gur et al. 1974; Fraser, Beamn et al. 1977) explained that some of the practices adopted by college administration in higher education like residential colleges or organized study groups also help to increases performance. Keeping in view all of the variables discussed by different researchers we have chosen only those variables that are recognizable in Pakistani setting.[15]

## Central role of family

The central role of the family in affecting children’s educational attainment is well documented. This was shown, for example, in the classic Blau–Duncan model of status attainment (Blau and Duncan, 1967). A growing number of later studies confirmed that family background, especially in early childhood, exerts strong influences on children’s educational outcomes, with children from higher SES families academically outperforming those from families with a lower SES (Carneiro and Heckman, 2003; Duncan et al., 1994; Duncan et al., 1998; Duncan et al., 2010). Moreover, early childhood educational inequality is predictive of inequalities in other domains in later life. How does family SES actually affect children’s outcomes? One perspective emphasizes family economic resources. A family’s economic condition determines how much parents can invest in their children’s education and development (Becker, 1991; Blau, 1999; Brooks-Gunn and Duncan, 1997; Coleman, 1988; Dahl and Lochner, 2012; Duncan et al., 1994; Duncan et al., 1998; Kaushal et al., 2011). Families with higher levels of income can provide material advantages, such as more learning opportunities and resources, that is, high-quality private tutoring (Zhang and Xie, 2015). Another perspective emphasizes families’ non-monetary resources, such as parenting attitudes and practices, and family environments.[4] Parents with a higher SES tend to have higher expectations of their child and to foster their child’s talents by incorporating organized activities. These class-based cultural and social factors could be viewed as a family’s social and cultural capital (Coleman, 1988). Thus, it has been observed that many early child development programs such as the Early Head Start program1 and the Nurse–Family Partnership2 not only provide children with direct interventions but also give their parents training in parenting skills (e.g., Gertler et al., 2014).[7]

## Techniques for performance prediction

A multitude of researches concerning not only the varied factors that influence the performance of students like personal, social, psychological and other environmental variables but also the techniques that have been used for the performance prediction is available in the literature. A number of selected studies are mentioned below for reference. Tismy et al. [13] collected data like group action, class test, seminar and assignment marks from the students' previous data to predict the performance at the beginning of the semester. They employed the classification to predict the students' division on the premise of prepared information. They used Naïve theorem and web-based application as a proposed system made to use the Naïve Bayesian mining technique to extract useful information. Naïve Bayesian algorithm provides more accuracy over other methods like regression, decision tree, and neural networks. Durairaj and Vijitha [15] reported that they developed a trusted model using DM techniques, which mines required information to predict student's performance in educational environments. The presented education system was proposed as a strategic management tool. Student details have been taken as vital information for analysis. The K-means test was used to choose the best cluster centre, which is to be treated as the centroid.[21] A model with five clusters was produced by the clustering method. Naïve Bayes algorithm was also applied to analyze data. In the research, the parameters used to evaluate the performance of the classifiers were TP rate, FP rate, and Precision, Recall F-Measure, and ROC area. Ahmed and Elaraby [5] applied one of the classification methods that are decision tree on student's database obtained from the Information System department of an educational institution to predict the student's performance based on previously recorded students' behaviour and activities. Their results showed that the study was able to predict- to a certain extent- the students' final grades in the selected course program. Abu Saa [8] conducted research on a group of 270 students enrolled in different colleges in Ajman University of Science and Technology in the United Arab Emirates. The researcher used multiple classification techniques (four decision tree algorithms and Naïve Bayes algorithm) for predicting the students' grade at the end of the semester. Aher and Lobo [16] collected two datasets for final year students from one college to predict student's final mark in the early phase of a particular course. The DM techniques were applied (classification and clustering) using WEKA DM software. Kong et al. [9] introduced a method of DM, which combines the concepts of contrast sets mining with association rules. They provided quantitative analysis for the similarity and difference of association rules obtained from the academic records datasets of multiple grades. Association rules were identified by generating positive association rules from frequent item sets. Two indicators have been selected (support and confidence) and negative association rules were generated. The analysis method combining contrast set mining and association rules were applied to pre-processed data. Mhetre and Nagar [17] proposed a classification-based predictive model to classify students: slow, average and fast based on the student's overall performance. Four classifiers were performed in the classification stage: Naïve Bayes, J48, ZeroR, and Random Tree to choose the most accurate one. Learners were classified based on various combinations of students' details such as GBA and assignments marks to get results that are based on the overall performance of the student. Experiments and results proved that Random Tree is potentially effective and an efficient classifier algorithm.[18]

# Research Methodology

## Data Collection

The dataset has been taken from Kaggle. Basically, this dataset contains the values of the categorical and numerical variables based on survey of a public school in the US. The initial source of this dataset is roycekimmons.com. The dataset contains 1000 records and 8 variables. It contains the marks obtained by the students in various subjects. The data is normally distributed.

### Dataset description:

The dataset contains the following 8 variables:

1. gender (2) race/ethnicity (3) parental level of education (4) lunch (5) test preparation course

(6) math score (7) reading score (8) writing score

The **categorical variables** are as follows:

* gender, race/ethnicity, parental level of education, lunch, test preparation course

The **numerical variables** are as follows:

* math score, reading score, writing score

The dataset has been precleaned and has no missing values.

## Hypothesis formulation

In the next step, hypothesis for the research questions have been formulated. There are three research questions, so three main hypothesis (H1) and three null hypotheses (H0) have been formulated. Two tests and one chi-squared test have been run and on the basis of the results, it has been decided during the analysis whether to accept or reject the null hypothesis. The H1 and H0 hypothesis and the results of the t-tests and chi-squared test run on them have been discussed in the analysis portion of the report.

## Tools and techniques used

Firstly, excel is used to open and preview the csv file. It has been concluded that there is no need for pre-processing and cleaning of the data. Then, the csv file has been imported into RStudio. R programming language has been used for the analysis and visualization.

The various techniques used in the analysis are as follows:

### Descriptive statistics

Descriptive statistics summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population.[26]

In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

In this particular project, mean, median, standard deviation and IQR (inter-quartile range has been calculated. Data visualization has been used for the visual representation of the data.

### T-test

t-test has been performed on the first two hypothesis to find out whether to accept or reject the hypothesis. A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another. [30]

The t-test is a parametric test of difference, meaning that it makes the same assumptions about the data as other parametric tests. The t-test assumes that the data:

* is independent.
* is (approximately) normally distributed.
* has a similar amount of variance within each group being compared (a.k.a. homogeneity of variance).

### Chi-squared test

Chi-squared test has been performed on the third hypothesis to accept or reject the null hypothesis and answer the research question. [29]

A chi-squared test (also chi-square or χ2 test) is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis, specifically Pearson's chi-squared test and variants thereof. Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

### Multiple Linear Regression

Linear regression model has been run to find out the weightage of the influence of ‘gender’, ‘test preparation’ and ‘reading score’ on math score and the weightage of the influence of ‘gender’ and ‘test preparation course’ on the writing score. [27]

Multiple linear regression is an extension of simple linear regression used to predict an outcome variable (y) on the basis of multiple distinct predictor variables (x). With three predictor variables (x), the prediction of y is expressed by the following equation:

***y = b0 + b1\*x1 + b2\*x2 + b3\*x3***

The “b” values are called the regression weights (or beta coefficients). They measure the association between the predictor variable and the outcome. “b\_j” can be interpreted as the average effect on y of a one unit increase in “x\_j”, holding all other predictors fixed.

# Analysis, Results and Explanation

In this section, analysis will be performed on the StudentsPerformance.csv file. The necessary code, plots and explanation will be given in this section.

## Loading libraries

In the first step, the libraries needed for the analysis are installed. If they’re pre-installed, then they’re loaded in RStudio using the library() function. require() function has also been used. R Markdown has been used to knit all the data into the word file. The following libraries have been loaded:

library(tidyr) library(ggplot2) library(dplyr) library(fBasics) library(kableExtra) library(funModeling) library(plotly) library(MASS) library(corrplot)

## Data import

The pre-processed data in excel has been imported into Rstudio using read.csv function.Column names have been defined. The code and dataset is as follows:

performance <- read.csv("StudentsPerformance.csv")  
  
names\_columns <- c("gender","race","parent\_education","lunch","test\_prep","math\_score","reading\_score","writing\_score")  
colnames(performance) <- names\_columns  
head(performance)

## gender race parent\_education lunch test\_prep math\_score  
## 1 female group B bachelor's degree standard none 72  
## 2 female group C some college standard completed 69  
## 3 female group B master's degree standard none 90  
## 4 male group A associate's degree free/reduced none 47  
## 5 male group C some college standard none 76  
## 6 female group B associate's degree standard none 71  
## reading\_score writing\_score  
## 1 72 74  
## 2 90 88  
## 3 95 93  
## 4 57 44  
## 5 78 75  
## 6 83 78

## Data manipulation

dplyr library has been used to manipulate the data. We have converted the variables with type=character to type=factor. Then the str() function has been used to analyse the structure of the performance data frame.

performance$gender<-factor(performance$gender)  
performance$race<-factor(performance$race)  
performance$parent\_education<-factor(performance$parent\_education)  
performance$lunch<-factor(performance$lunch)  
performance$test\_prep<-factor(performance$test\_prep)  
str(performance)

## 'data.frame': 1000 obs. of 8 variables:  
## $ gender : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 2 2 1 ...  
## $ race : Factor w/ 5 levels "group A","group B",..: 2 3 2 1 3 2 2 2 4 2 ...  
## $ parent\_education: Factor w/ 6 levels "associate's degree",..: 2 5 4 1 5 1 5 5 3 3 ...  
## $ lunch : Factor w/ 2 levels "free/reduced",..: 2 2 2 1 2 2 2 1 1 1 ...  
## $ test\_prep : Factor w/ 2 levels "completed","none": 2 1 2 2 2 2 1 2 1 2 ...  
## $ math\_score : int 72 69 90 47 76 71 88 40 64 38 ...  
## $ reading\_score : int 72 90 95 57 78 83 95 43 64 60 ...  
## $ writing\_score : int 74 88 93 44 75 78 92 39 67 50 ...

## Descriptive statistics

The next step in the analysis is descriptive statistics. Summary statistics has been used to display the no. of various levels of the categorical variables and the mean, median, mode,min, max, 1st quartile and 3rd quartile of the numerical variables. IQR function will be used to determine the inter-quartile range of the numerical data. The use of summary() function and its output has been shown below:

summary(performance)

## gender race parent\_education lunch   
## female:518 group A: 89 associate's degree:222 free/reduced:355   
## male :482 group B:190 bachelor's degree :118 standard :645   
## group C:319 high school :196   
## group D:262 master's degree : 59   
## group E:140 some college :226   
## some high school :179   
## test\_prep math\_score reading\_score writing\_score   
## completed:358 Min. : 0.00 Min. : 17.00 Min. : 10.00   
## none :642 1st Qu.: 57.00 1st Qu.: 59.00 1st Qu.: 57.75   
## Median : 66.00 Median : 70.00 Median : 69.00   
## Mean : 66.09 Mean : 69.17 Mean : 68.05   
## 3rd Qu.: 77.00 3rd Qu.: 79.00 3rd Qu.: 79.00   
## Max. :100.00 Max. :100.00 Max. :100.00

Then, profiling\_num() function has been used to show the detailed descriptive statistics of the numerical variables.

profiling\_num(performance)

## variable mean std\_dev variation\_coef p\_01 p\_05 p\_25 p\_50 p\_75  
## 1 math\_score 66.089 15.16308 0.2294342 27.99 40.95 57.00 66 77  
## 2 reading\_score 69.169 14.60019 0.2110800 31.99 44.00 59.00 70 79  
## 3 writing\_score 68.054 15.19566 0.2232882 31.98 42.95 57.75 69 79  
## p\_95 p\_99 skewness kurtosis iqr range\_98 range\_80  
## 1 90.05 98.01 -0.2785166 3.267597 20.00 [27.99, 98.01] [47, 86]  
## 2 92.00 100.00 -0.2587157 2.926081 20.00 [31.99, 100] [51, 87.1]  
## 3 92.00 100.00 -0.2890096 2.960808 21.25 [31.98, 100] [48, 87]

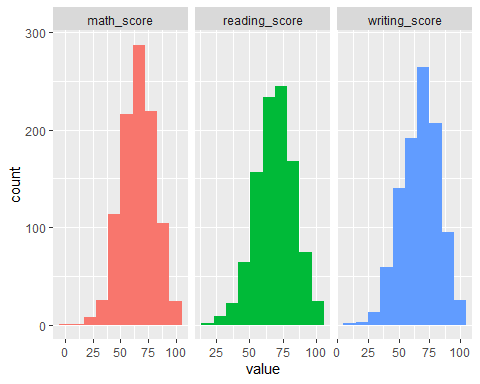
We can see that Mathematics exam score average of students is 66.08. Reading exam score average of students is 69.16. Writing exam score average of students is 68.05.

## Data Visualization

For data visualization, three side-by-side histograms of the numerical variables will be plotted using the plot\_num() function as follows:

plot\_num(performance)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.

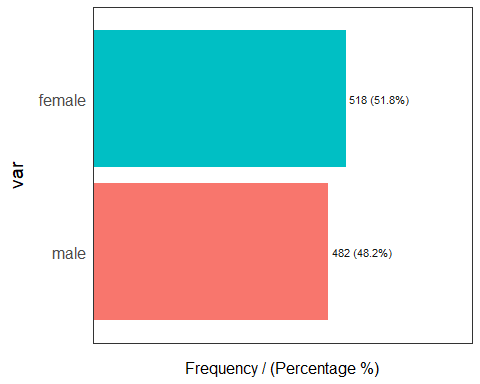


**Figure 1: Side-by-side histograms of numerical variables**

The freq() function has been used to plot a table and a bar graph showing the frequency, percentage and cumulative frequency of the gender variable.

freq(performance$gender)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.



**Figure 2: Frequency of gender**

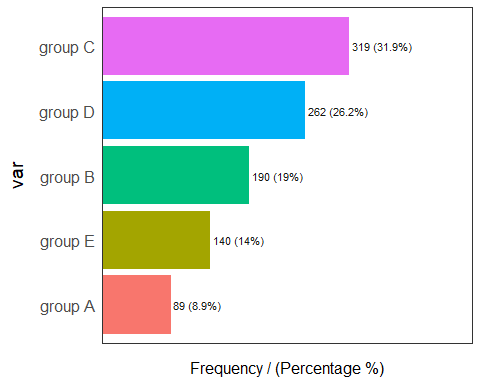
Table 1: Gender frequency

|  |
| --- |
| var frequency percentage cumulative\_perc  1 female 518 51.8 51.8  2 male 482 48.2 100.0 |

The bar graph shows that the percentage of female students is higher that the male students in the dataset. Since the dataset contains the data of all the students of the public school in US. Hence, it can be concluded that the number of female students is higher than the number of male students in the school. Similarly freq() function has also been run on the race variable as follows:

freq(performance$race)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.



**Figure 3: Frequency of race**

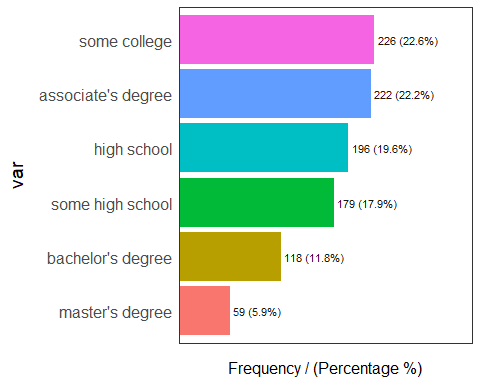
Table 2: Race frequency

|  |
| --- |
| var frequency percentage cumulative\_perc  1 group C 319 31.9 31.9  2 group D 262 26.2 58.1  3 group B 190 19.0 77.1  4 group E 140 14.0 91.1  5 group A 89 8.9 100.0 |

From the bar graph of race vs frequency/percentage, we can conclude that the most common race ethnicity is Group C, while the least common is Group A. The parent education level variable of the students consists of 6 categories.

freq(performance$parent\_education)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.



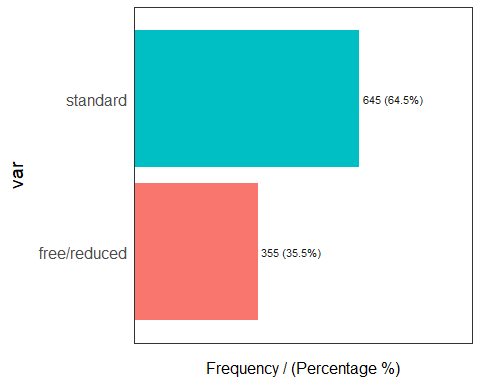
**Figure 4: Frequency of parent\_education**

## var frequency percentage cumulative\_perc  
## 1 some college 226 22.6 22.6  
## 2 associate's degree 222 22.2 44.8  
## 3 high school 196 19.6 64.4  
## 4 some high school 179 17.9 82.3  
## 5 bachelor's degree 118 11.8 94.1  
## 6 master's degree 59 5.9 100.0

The most common education in the parent\_education variable is ‘some college’ while the least common is ‘master’s degree’. The detailed statistics and bar plot of the lunch variable is as follows:

freq(performance$lunch)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.



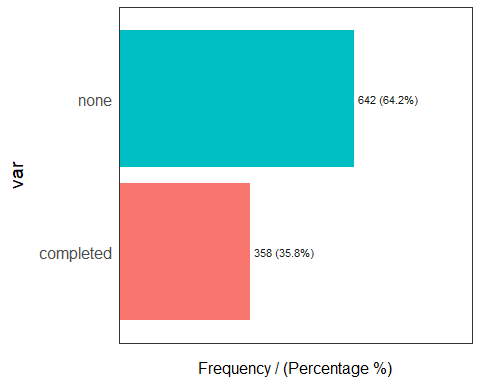
**Figure 5: Frequency of lunch**

|  |
| --- |
| var frequency percentage cumulative\_perc  1 standard 645 64.5 64.5  2 free/reduced 355 35.5 100.0 |

64.5% of the students eat lunch at standard price and 35.5% eat free/reduced lunch.

freq(performance$test\_prep)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.

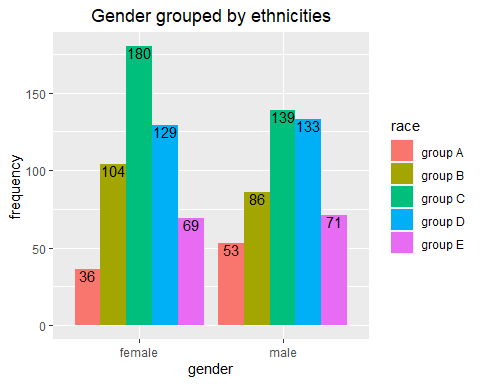


**Figure 6: Frequency of test\_prep**

Table 3: Test prep frequency

|  |
| --- |
| ## var frequency percentage cumulative\_perc ## 1 none 642 64.2 64.2 ## 2 completed 358 35.8 100.0 |

From the table and bar plot, we can conclude that while 642 of the students did not complete the exam preparation course, 358 students did.

In the next step, the distribution of gender grouped by race will be visualized using a bar plot. The code has been hidden by using echo=FALSE argument. Only the plot will be shown. Both the geom\_bar() and geom\_text()functions have been used to generate the plot.

**Figure 7: Bar plot of gender grouped by ethnicities**

## t-test

**1st Research Question**:

The 1st research question of the analysis is as follows:

**Are there any differences between the exam scores of female and male students?**

To answer this question, three null and main hypothesis have been formulated. The first null hypothesis for this research question is:

**H0** : There is no difference between the mathematics scores of male and female students. The first main hypothesis for this research question is:

**H1** : There is a difference between the mathematics scores of male and female students. Now, the t-test has been conducted after the formulation of the hypothesis.

attach(performance)  
t.test(math\_score~gender, var.equal=TRUE)

##   
## Two Sample t-test  
##   
## data: math\_score by gender  
## t = -5.3832, df = 998, p-value = 9.12e-08  
## alternative hypothesis: true difference in means between group female and group male is not equal to 0  
## 95 percent confidence interval:  
## -6.952285 -3.237737  
## sample estimates:  
## mean in group female mean in group male   
## 63.63320 68.72822

***t=-5.398*** and ***p value < 0.05***

Hence, H0 is rejected. There is a difference between gender and math test score. The second null hypothesis for this research question is:

**H0**: There is no difference between the reading scores of male and female students. The second main hypothesis for this research question is:

**H1**: There is a difference between the reading scores of male and female students. The t-test has been conducted after the formulation of the second hypothesis.

t.test(reading\_score~gender, var.equal=TRUE, data=performance)

##   
## Two Sample t-test  
##   
## data: reading\_score by gender  
## t = 7.9593, df = 998, p-value = 4.681e-15  
## alternative hypothesis: true difference in means between group female and group male is not equal to 0  
## 95 percent confidence interval:  
## 5.375946 8.894212  
## sample estimates:  
## mean in group female mean in group male   
## 72.60811 65.47303

***t=7.9593*** and ***p value <0.05***, hence H0 is rejected. There is a difference between gender and reading test score. The third null and main hypothesis are as follows:

**H0**: There is no difference between the writing scores of male and female students.

**H1**: There is a difference between the writing scores of male and female students.

The t-test has been conducted after the formulation of the third hypothesis as well.

t.test(writing\_score~gender,var.equal=TRUE )

##   
## Two Sample t-test  
##   
## data: writing\_score by gender  
## t = 9.9796, df = 998, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group female and group male is not equal to 0  
## 95 percent confidence interval:  
## 7.35558 10.95638  
## sample estimates:  
## mean in group female mean in group male   
## 72.46718 63.31120

***t=9.9796*** and ***p value <0.05***, hence H0 is rejected i.e it proves the statement “There is a difference between gender and writing test score”.

## Chi-squared test:

The H0 and H1 are as follows:

**H0**: Completing the test preparation course is not related to gender.

**H1**: Completing the test preparation course is related to gender.

chi\_tab1<- table(gender, test\_prep)  
  
chisq.test(chi\_tab1)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: chi\_tab1  
## X-squared = 0.015529, df = 1, p-value = 0.9008

As p value =0.9>0.05, hence H0 is accepted. This means that gender is not all related to test preparation course variable. The H0 and H1 for the next chi-squared test are as follows:

**H0**: Lunch eating status is not related to gender.

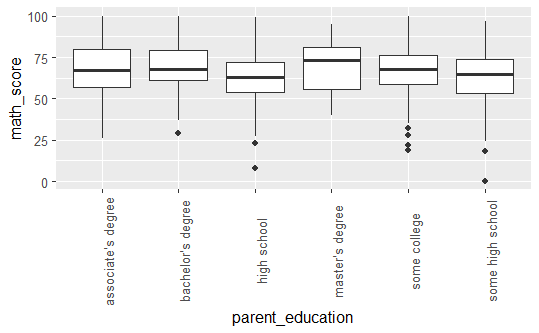
**H1**: Lunch eating status is related to gender.

chi\_tab2<- table(gender, lunch)  
  
chisq.test(chi\_tab2)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: chi\_tab2  
## X-squared = 0.37174, df = 1, p-value = 0.5421

As p value=0.5421>0.05, H0 is accepted.Thus means that lunch eating status is not related to gender.

For identifying the relationship between parents’ education and student scores, a boxplot has been created in R. The plot is shown below:



**Figure 8: boxplot of parent education vs math score**

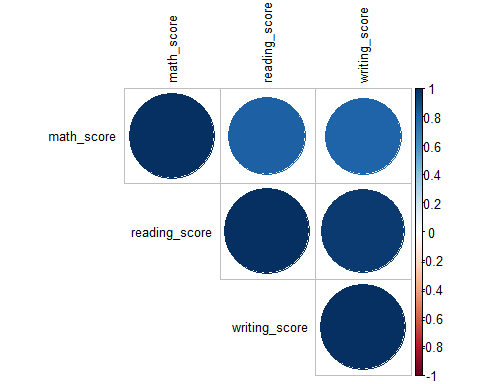
As we can see in the figure, the median of math scores of students is higher for parents who have either done masters or bachelor or studied in college for some time. There isn’t any definite pattern so we cannot infer a causal relationship. Generally, parents with better education have children with the least chance of failure in math exams as there is no outlier of the master’s degree boxplot.

## Correlation

There is a strong positive correlation between the variables as shown by the correlation matrix.

num\_data <- performance[, c( "math\_score", "reading\_score", "writing\_score")]  
m<-cor(num\_data)  
m

The correlation plot is shown below:



**Figure 9: Correlation plot**

The correlation plot shows that there is a strong correlation between math score, writing score and reading score.

## Multiple Linear Regression Analysis

First of all, linear regression is used to determine the extent to which the mathematics score can be explained by the gender and reading score.

reg1<- lm(math\_score~ gender+reading\_score, data=performance)  
summary(reg1)

##   
## Call:  
## lm(formula = math\_score ~ gender + reading\_score, data = performance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.7447 -4.3918 -0.0747 4.1590 18.1338   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.2230 1.1056 -4.724 2.64e-06 \*\*\*  
## gendermale 11.8614 0.4292 27.634 < 2e-16 \*\*\*  
## reading\_score 0.9483 0.0147 64.523 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.577 on 997 degrees of freedom  
## Multiple R-squared: 0.8122, Adjusted R-squared: 0.8119   
## F-statistic: 2157 on 2 and 997 DF, p-value: < 2.2e-16

***Equation: math.score=−5.2230+11.8614gender+0.9483reading***

R-squared = 0.8122, this implies that 81.2% of the mathematics score can be explained by the variables of gender and reading test score. The second regression analysis is used to determine the extent to which the mathematics score can be explained by the variables gender, test\_prep and reading\_score.

reg2<- lm(math\_score~ gender+test\_prep+reading\_score, data=performance)  
summary(reg2)

##   
## Call:  
## lm(formula = math\_score ~ gender + test\_prep + reading\_score,   
## data = performance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.1746 -4.2620 -0.1302 4.1333 17.5549   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.1570 1.2351 -5.795 9.19e-09 \*\*\*  
## gendermale 11.9632 0.4279 27.955 < 2e-16 \*\*\*  
## test\_prepnone 1.5327 0.4457 3.438 0.000609 \*\*\*  
## reading\_score 0.9614 0.0151 63.661 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.541 on 996 degrees of freedom  
## Multiple R-squared: 0.8144, Adjusted R-squared: 0.8139   
## F-statistic: 1457 on 3 and 996 DF, p-value: < 2.2e-16

***Equation: math.score=−7.1570+11.9632gender+1.5327course+0.9614reading***

R-squared = 0.8144,this implies that 81.4% of the mathematics score can be explained by the variables of gender,reading test scores and test preparation course. The third regression analysis is run to determine the extent to which gender and test\_prep variables are used to explain the writing score.

reg3<-lm(writing\_score~ gender+test\_prep, data=performance)  
summary(reg3)

##   
## Call:  
## lm(formula = writing\_score ~ gender + test\_prep, data = performance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -58.925 -8.711 0.317 9.296 32.289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 78.8971 0.8372 94.23 <2e-16 \*\*\*  
## gendermale -9.2137 0.8665 -10.63 <2e-16 \*\*\*  
## test\_prepnone -9.9722 0.9031 -11.04 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.69 on 997 degrees of freedom  
## Multiple R-squared: 0.1898, Adjusted R-squared: 0.1882   
## F-statistic: 116.8 on 2 and 997 DF, p-value: < 2.2e-16

***Equation: writing.score=78.8971−9.2137gender−9.9722course***

R-squared = 0.1898, this implies that 19% of the writing score can be explained by the variables of gender and test preparation course.

# Discussion

## Contributions

This research has contributed in identifying the relationship between various variables and student performance. We have found out whether there are any differences between the exam scores of female and male students or not. We have also identified the relationship between the gender and test preparation course and gender and lunch eating status using linear regression.

This study can contribute to find out the factors, which are responsible for student’s inelastic behavior towards study along with identifying those factors, which help a student to make progress in his studies. We have also contributed in finding the relationship between parent education and student scores. This research also highlights the role of regression analysis in the research of ‘factors affecting student performance’.

## Discussion about findings:

A lot of results have already been discussed in the results and analysis section but further clarification will be given in this section.

The bar plot of gender grouped by race shows that the most common race for females is group C (180 students) and the most common one for males is group C as well (139 students). Moreover, several hypotheses have been created to answer the 1st research question: **Are there any differences between the exam scores of female and male students?** T-tests have been conducted to answer these research questions. For the first t-test, **t***=-5.398* and *p value < 0.05.* Hence, H0 was rejectedwhich means that **there is a difference between gender and math test score.** For the second test, t=7.9593 and p value <0.05, hence H0 was rejected which implied that **there is a difference between gender and reading test score**.For the third t-test, t=9.9796 and p value <0.05, hence H0 is rejected i.e it proves the statement “There is a difference between gender and writing test score”. Hence, we concluded that there is a difference between the exam scores of female and male students.

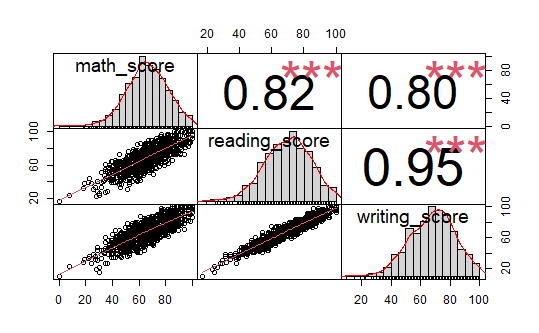
For the 1st chi-square test, p value =0.9>0.05, hence H0 was accepted. This means that gender is not all related to test preparation course variable. For the 2nd chi-square test, p value =0.9>0.05, hence H0 is accepted. This means that gender was not related to test preparation course variable. The correlation matrix shown below shows that there is a strong positive correlation among all the numerical variables:

Table 4: Correlation matrix

|  |
| --- |
| math\_score reading\_score writing\_score  math\_score 1.0000000 0.8175797 0.8026420  reading\_score 0.8175797 1.0000000 0.9545981  writing\_score 0.8026420 0.9545981 1.0000000 |

Linear regression has been conducted three times. The results show that 81.2% of the mathematics score can be explained by the variables of gender and reading test score. Similarly, 81.4% of the mathematics score can be explained by the variables of gender, reading test scores and test preparation course. (2nd regression). Lastly, 19% of the writing score can be explained by the variables of gender and test preparation course (3rd regression).

The correlation chart below has been formulated using the chart.Correlation() function of the ‘PerformanceAnalytics’ library of RStudio. In addition to scatter plots depicting the relationship between the numerical variables, the histogram depicts the distribution of the each numerical variable.



**Figure 10: Correlation chart**

## Policy implications

On the basis of findings of study, the following are the policy implications:

* may be given proper examination training before getting into final examination, in order to avoid overconfidence as well as exam phobia.
* Internal environment of examination may be peaceful and conducive to the students
* Difficulty level of questions in question paper may be moderate i.e., neither too easy nor too difficult
* paper evaluator may pay more concentration while marking answer sheet.

## Limitations

Further research is needed to understand how causal processes operate differently across different social contexts, including the roles of schools, teachers, and peers. Although our results are consistent with implications of authors who did research similar to our topic, we do not have a direct way to test this hypothesis. In other words, although we have uncovered different distribution patterns of academic achievement in a public school in the USA, we will wait for future research to confirm whether or not this relationship between various variables is causal or not. PSM and DID can be used to determine this causal relationship.

# Conclusion

In conclusion, after completing data collection, hypothesis for the research have been formulated. The data has been imported into RStudio and the required libraries have been loaded. Then, data has been manipulated to be ready for the analysis. Descriptive statistics is used to determine the mean, median, mode, IQR, standard deviation and variance of the numerical data.

In the next step, data has been visualized for the visual explanation of the results. To accept or reject the null hypothesis t-tests and chi-squared tests have been run. Then, multiple linear regression has been done. It has been concluded that:

* There is a difference between gender and reading test score.
* Test preparation course is related to gender.
* Lunch eating status is not related to gender.
* 81.2% of the mathematics score can be explained by the variables of gender and reading test score. 81.4% of the mathematics score can be explained by the variables of gender, reading test scores and test preparation course. 19% of the writing score can be explained by the variables of gender and test preparation course (3rd regression).

# References

**[1]** Han J., Kamber M., and Pie J. (2006). Data Mining Concepts and Techniques. 2nd edition, Morgan

Kaufmann Publishers.

**[2]** Al-Radaideh Q., Al-Shawakfa Emad M., and Al-Najjar Mustafa I. (2006). Mining Student Data Using Decision Trees. The 2006 International Arab Conference on Information Technology.

**[3]** Nguyen N., Paul J., and Peter H. (2007). A Comparative Analysis of Techniques for Predicting Academic Performance. In Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference. pp. 7-12.

**[4]** Sunita B Aher, Mr. LOBO L.M.R.J. (2011). Data Mining in Educational System using WEKA, International Conference on Emerging Technology Trends (ICETT). Proceedings Published by International Journal of Computer Applications.

**[5]** Kabra R. R., Bichkar R.S. (2011). Performance Prediction of Engineering Students using Decision Trees, International Journal of Computer Applications (0975-8887), Vol-36-No.11.

**[6]** Kumar S. Anupama and Dr. Vijayalakshmi M.N. (2011). Efficiency of Decision Trees in Predicting Students Academic Performance. Computer Science & Information Technology 02, pp. 335–343.

**[7]** Baradwaj Brijesh Kumar and Pal Saurabh (2011). Mining Educational Data to Analyze Student

Performance. International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6.

**[8]** Sembiring Sajadin,. Zarlis M , Hartama Dedy ,Ramliana S,Wani Elvi (2011). Prediction of Student Academic Performance by an application of data mining techniques.International Conference on Management and Artificial Intelligence IPEDR vol.6 (2011) © (2011) IACSIT Press, Bali, Indonesia

**[9]** Al-Radaideh Q., Al-Ananbeh Ahmad.,Al-Shawakfa Emad M . (2011). A Classification Model for Predicting the Study Track for Suitable Student. [www.arpapress.com/Volumes/Vol8Issue2/IJRRAS\_8\_2\_](http://www.arpapress.com/Volumes/Vol8Issue2/IJRRAS_8_2_) 15.pdf

**[10]** Yadav Surjeet Kumar, Pal Saurabh Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification (2012). World of Computer Science and Information Technology Journal (WCSIT) .ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012.

**[11]** Kabakchieva Dorina (2012). Student Performance Prediction by Using Data Mining Classification

Algorithms. International Journal of Computer Science and Management Research .ISSN 2278-733X: Vol 1 Issue 4 November

**[12]** Shovon Md. Hedayetul Islam, Haque Mahfuza (2012).An Approach of Improving Student’s Academic Performance by using K-means clustering algorithm and Decision tree . (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.3,

No. 8, 2012

**[13]** Caplan NS, Choy MH and Whitmore JK (1991) Children of the Boat People: A Study of Educational Success. Ann Arbor: University of Michigan Press.

**[14]** Card D and Krueger AB (1992) Does school quality matter? Returns to education and

the characteristics of public schools in the United States. Journal of Political Economy

100(1): 1–40.

**[15]**Card D and Payne AA (2002) School finance reform, the distribution of school spending,

and the distribution of student test scores. Journal of Public Economics 83(1): 49–82.

**[16]**Carneiro P and Heckman JJ (2003) Human capital policy. In: Heckman JJ, Krueger AB and

Friedman BM (eds) Inequality in America: What Role for Human Capital Policies?

Cambridge, MA: MIT Press, 77–239.

**[17]**Carnoy M and Loeb S (2002) Does external accountability affect student outcomes? A crossstate analysis. Educational Evaluation and Policy Analysis 24(4): 305–331.

**[18]**Cheadle JE and Amato PR (2011) A quantitative assessment of Lareau’s qualitative conclusions about class, race, and parenting. Journal of Family Issues 32(5): 679–706.

**[19]**Chen X, Liu M and Li D (2000) Parental warmth, control, and indulgence and their relations to adjustment in Chinese children: A longitudinal study. Journal of Family

Psychology 14(3): 401–419.

**[20]**Chen X, Yi H, Zhang L, et al. (2014) Do poor students benefit from China’s merger program? Transfer path and educational performance. Asia Pacific Journal of Education 34(1): 15–35.

**[21]**Chu JH, Loyalka P, Chu J, et al. (2015) The impact of teacher credentials on student achievement in China. China Economic Review 36: 14–24.

**[22]**Coleman JS (1966) Equality of Educational Opportunity. Ann Arbor: Inter-university

Consortium for Political and Social Research.

**[23]**Coleman JS (1988) Social capital in the creation of human capital. American Journal of

Sociology 94: S95–S120.

**[24]**C. Ricketts, S. J. Wilks, (2002), ”Improving Student Performance Through Computer-based Assessment: insights from recent research “ Assessment & Evaluation in Higher Education, Volume 27, Number 5, 1, pp. 475-479(5)

**[25]**George R. Goethals (October, 2001) “Peer effects, gender, and intellectual performance among students at a highly selective college: a social comparison of abilities analysis” Discussion Paper-61

**[26]**Gordon c. Winston, David j. Zimmerman, (January 2003) “Peer Effects In Higher Education”

NBER Working Paper No. 9501, JEL no I21

**[27]** Hansen, Joe B. (April 2000) “Student Performance and Student Growth as measure of success: A evaluator’s perspective” Paper presented at annual meeting of the American Educational Research Association New Orleans, Louisiana, April 25, 2000

**[28]** Jessica Hoel, Jeffrey Parker, and Jon Rivenburg (August 2005) “A Test for Classmate Peer Effects in Higher Education” Institutional Research at Reed College.

**[29]**Patrick J. McEwan, Kristen A. Soderberg (March 2004) “Roommate effects on grades: Evidence from first-year housing assignments ”Social Science Summer Research Program at Wellesley College.

**[30]** Russell, M., Haney, W. (January 1997) Testing Writing on Computers: An Experiment Comparing Student Performance on Tests Conducted via Computer and via Paper-and-Pencil, Volume 5, Number 3, ISSN 1068-2341.