Contents

Chapter 1

Objectives

1. To analyse all the different columns and rows.
2. To analyse different factors affecting marks such as gender, ethnicity, parent's qualification.
3. To perform Exploratory data analysis on the given data frame
4. To perform both uni variate and bi variate analysis on the given data frame
5. To draw proper conclusions regarding the same using R programming language

Chapter 2

Introduction

2.1 Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.Data mining is a particular data analysis technique that focuses on statistical modeling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing mainly on business information. In statistical applications, data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data while CDA focuses on confirming or falsifying existing hypotheses. Predictive analytic focuses on the application of statistical models for predictive forecasting or classification, while text analytic applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All of the above are varieties of data analysis.Data integration is a precursor to data analysis, and data analysis is closely linked to data visualization and data dissemination.

2.1.1 Process of Data Analysis

1. Data collection: The first stage in data analysis involves collecting data from various sources, such as surveys, experiments, databases, or web scraping. It is essential to ensure that the data collected is accurate, relevant, and complete.

2. Data cleaning: Once the data has been collected, it is essential to clean it to eliminate any errors, inconsistencies, or missing values that may affect the quality of the analysis. This process involves identifying and correcting errors, removing outliers, and filling in missing values.

3. Data exploration: Data exploration involves visualizing and summarizing the data to gain an understanding of its distribution, patterns, and relationships between variables. This stage helps to identify any interesting insights that can guide the analysis.

4. Data modeling: Once the data has been explored, a suitable model can be chosen to analyze it. The model chosen depends on the type of data and the research question. Some common modeling techniques include regression analysis, decision trees, and clustering.

5. Data interpretation: After the model has been applied to the data, the results need to be interpreted to draw meaningful conclusions. This involves analyzing the output from the model and assessing its validity.

6. Presentation: The final stage involves presenting the findings in a clear and concise manner. This can be done through reports, charts, graphs, or presentations.

2.1.2 Uni variate Analysis

Univariate analysis is a type of data analysis that focuses on examining one variable at a time. It involves analyzing and summarizing the characteristics of a single variable, such as its frequency, distribution, central tendency, and variability. Univariate analysis is commonly used to gain an understanding of the data before conducting more complex analyses, such as multivariate analysis. In univariate analysis, the data is typically represented using descriptive statistics, such as histograms, box plots, and summary statistics. These statistics provide information about the distribution of the variable, such as its range, median, mean, standard deviation, and skewness. Univariate analysis is used in many fields, such as business, finance, social sciences, and healthcare, among others. For example, in finance, univariate analysis may be used to analyze the performance of a single stock or bond, while in healthcare, it may be used to examine the distribution of a particular disease or health outcome in a population. Overall, univariate analysis is a powerful tool for gaining insights into the characteristics of a single variable and can be used to inform decision-making in a wide range of fields.

2.1.3 Bi variate Analysis

Bivariate analysis is a type of statistical analysis that examines the relationship between two variables. In other words, it analyzes how one variable (the independent variable) affects or influences the other variable (the dependent variable). Bivariate analysis typically involves measuring the strength and direction of the relationship between the two variables. This is commonly done using correlation coefficients, such as Pearson's correlation coefficient or Spearman's rank correlation coefficient. These coefficients measure the degree of association between the two variables, with values ranging from -1 to +1. A correlation coefficient of +1 indicates a perfect positive relationship, while a correlation coefficient of -1 indicates a perfect negative relationship. A coefficient of 0 indicates no relationship. Bivariate analysis is commonly used in many fields, such as economics, social sciences, and healthcare, among others. For example, in economics, bivariate analysis may be used to analyze the relationship between a company's revenue and its advertising expenditure, while in healthcare, it may be used to examine the relationship between smoking and lung cancer. Overall, bivariate analysis is a powerful tool for understanding the relationship between two variables and can be used to inform decision-making in a wide range of fields.

2.2 Importance of Data Analysis

Data analysis is an essential process that provides valuable insights into the data, and its importance can be summarized as follows:

1. Improved decision-making: Data analysis can help in making informed decisions by providing insights into patterns, trends, and relationships within the data. It can also identify key drivers and factors that are contributing to specific outcomes or events.

2. Increased efficiency and productivity: Data analysis can help to identify areas where efficiencies can be gained, enabling organizations to reduce costs, streamline operations, and optimize resources.

3. Improved customer experience: Data analysis can help organizations understand customer behavior and preferences, allowing them to tailor their products and services to meet the needs of their customers better.

4. Competitive advantage: Data analysis can help organizations gain a competitive advantage by identifying trends and patterns that can inform strategic decisions and enable them to stay ahead of the competition.

5. Better risk management: Data analysis can help organizations to identify and manage risks by providing insights into potential risks and their impact, allowing them to take proactive measures to mitigate or avoid them.

6. Improved quality and innovation: Data analysis can help organizations identify areas for improvement in products and services, enabling them to innovate and deliver higher quality offerings.

Overall, data analysis is a critical process that can help organizations to make better-informed decisions, increase efficiency and productivity, and gain a competitive advantage in today's data-driven world.

Chapter 3

Report

3.1 Data Set

The data set is under the Education domain discuss about different features that contribute to a student scoring marks in a test like gender, parent's level of education, the ethnicity they belong to, the status of test preparation score and their marks.

The data set contains the following:

1. Gender : Male or Female
2. Ethnicity or Race : Group A,B,C,D,E
3. Parental Level of Education : Associate's Degree,Bachelor's Degree,High school,Master's Degree,Some College,Some High School
4. Lunch Status : Standard or free/reduced
5. Test Preparation Status : Completed or None
6. Mathematics Score
7. Reading Score
8. Writing Score

3.2 Uni variate Analysis

In this data set, Uni variate analysis deals with single variables such as Gender, composition of Ethnicity, composition of Education and composition of lunch status in graphical methods and basic descriptive statistical tools for numerical methods.

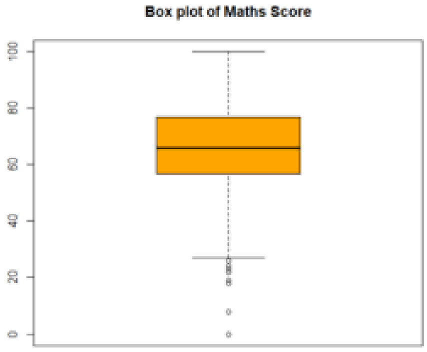
3.2.1 Source code

```
#reading of the CSV File
Students_Performance <-read.csv("C:\Users\HP\Downloads\
StudentsPerformance.csv",dec = ",")

#display of the CSV File
Students_Performance
#uni variate analysis

#summary
summary(Students_Performance)

#graphical representation
#boxplot of marks
boxplot(Students_Performance$math.score ,main =
"Box␣plot␣of␣Maths␣Score",col ="orange")
boxplot(Students_Performance$reading.score ,main =
"Box␣plot␣of␣Reading␣Score",col = "red")
boxplot(Students_Performance$writing.score ,main =
"Box␣plot␣of␣Writing␣Score",col = "magenta")

#barplot of different categories
barplot(table(Students_Performance$gender),xlab = "Gender",
col = c("light␣green","dark␣green"),ylab="Count",
main="Gender␣Wise␣count")
legend("topright",legend = c("Female","Male"),
fill=c("light␣green","dark␣green"))

#pie chart of different categories
pie(table(Students_Performance$lunch), col = c("orange","blue"),
main = "Pie␣Chart␣representing␣the␣different␣lunches␣student's␣opt")

pie(table(Students_Performance$parental.level.of.education),
col = c("pink","brown","green","lightblue","purple","orange"),
main = "Pie␣Chart␣representing␣the␣different␣Educational␣Status")

#grouping function
grouped_reading <-cut(Students_Performance$reading.score,
breaks = c(0,40,60,80,100),labels = c
("Fail","Meritorious","Outstanding","Exceptional"))
table(grouped_reading)
```

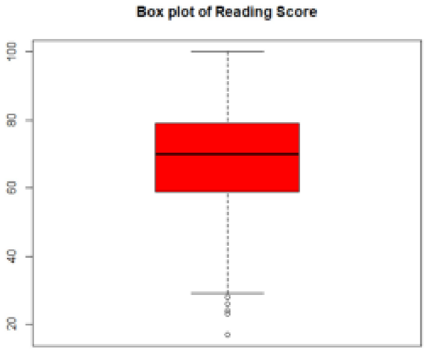
```
barplot(table(grouped_math),main=”Students␣Performance␣in␣Maths”,
xlab=”Grade”,ylab=”Count”,col = c(
”red”,”green”,”lightblue”,”purple”))
\\
grouped_writing <-cut(Students_Performance$writing.score,
breaks = c(0,40,60,80,100),
labels = c(”Fail”,”Meritorious”,”Outstanding”,”Exceptional”))
table(grouped_writing)
barplot(table(grouped_reading),main=”Students␣Performance␣in␣Reading”,
xlab=”Grade”,ylab=”Count”,col = c(”red”,”green”,”lightblue”,”purple”))
\\
grouped_math <-cut(Students_Performance$math.score,
breaks = c(0,40,60,80,100),
labels = c(”Fail”,”Meritorious”,”Outstanding”,”Exceptional”))
table(grouped_math)
barplot(table(grouped_math),main=”Students␣Performance␣in␣Writing”,
xlab=”Grade”,ylab=”Count”,col =
c(”red”,”green”,”lightblue”,”purple”))
```

3.2.2 Output

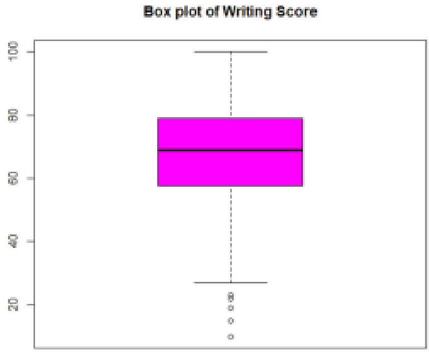
```
gender: Length:1000 Class :character Mode :character

race.ethnicity: Length:1000 Class :character Mode :character

parental.level.of.education: Length:1000 Class :character Mode :character

lunch: Length:1000 Class :character Mode :character

test.preparation.course : Length:1000 Class :character Mode :character

math.score:
min. 1st Qu. Median Mean 3rd Qu. Max.
0.00 57.00 66.00 66.09 77.00 100.00

reading.score:
Min. 1st Qu. Median Mean 3rd Qu. Max.
17.00 59.00 70.00 69.17 79.00 100.00

writing.score
Min. 1st Qu. Median Mean 3rd Qu. Max.
10.00 57.75 69.00 68.05 79.00 100.00

Average.Score.Rounded
Min. 1st Qu. Median Mean 3rd Qu. Max.
9.00 58.00 68.00 67.76 78.00 100.00
```



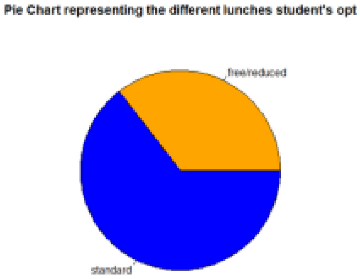
Box plot of Maths Score



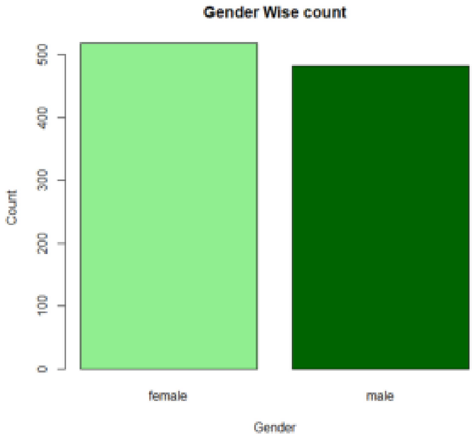
Box plot of Reading Score



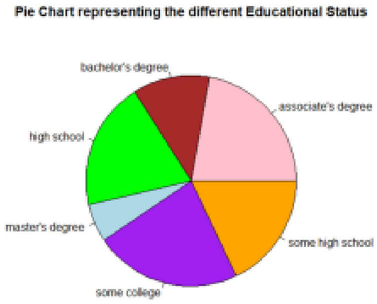
Box plot of Writing Score



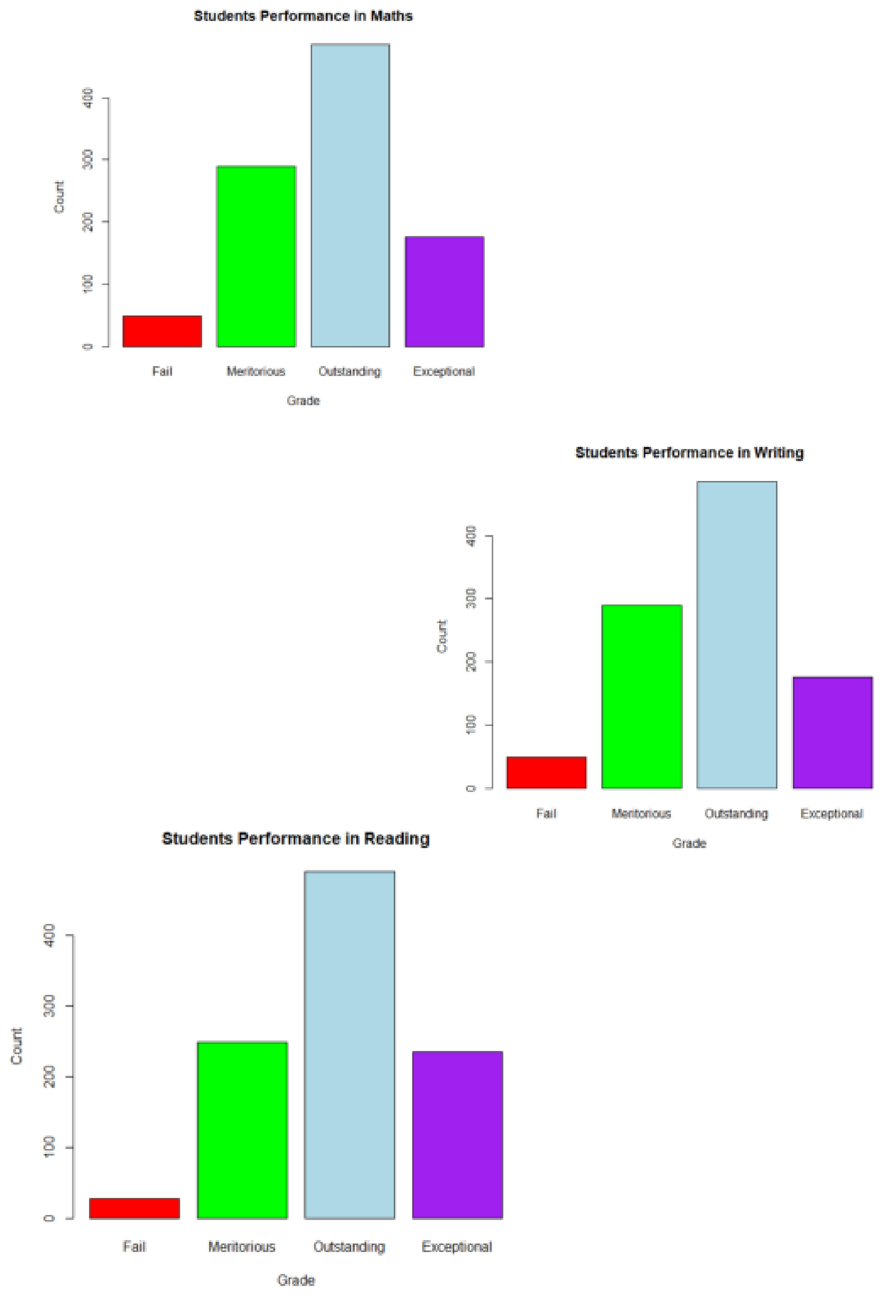
Pie Chart representing the different lunches student's opt



Gender Wise count



Pie Chart representing the different Educational Status


Students Performance in Maths
Count
0
100
200
300
400
Fail
Meritorious
Outstanding
Exceptional
Grade
Students Performance in Writing
Count
0
100
200
300
400
Fail
Meritorious
Outstanding
Exceptional
Grade
Students Performance in Reading
Count
0
100
200
300
400
Fail
Meritorious
Outstanding
Exceptional
Grade

3.3 Bi variate Analysis

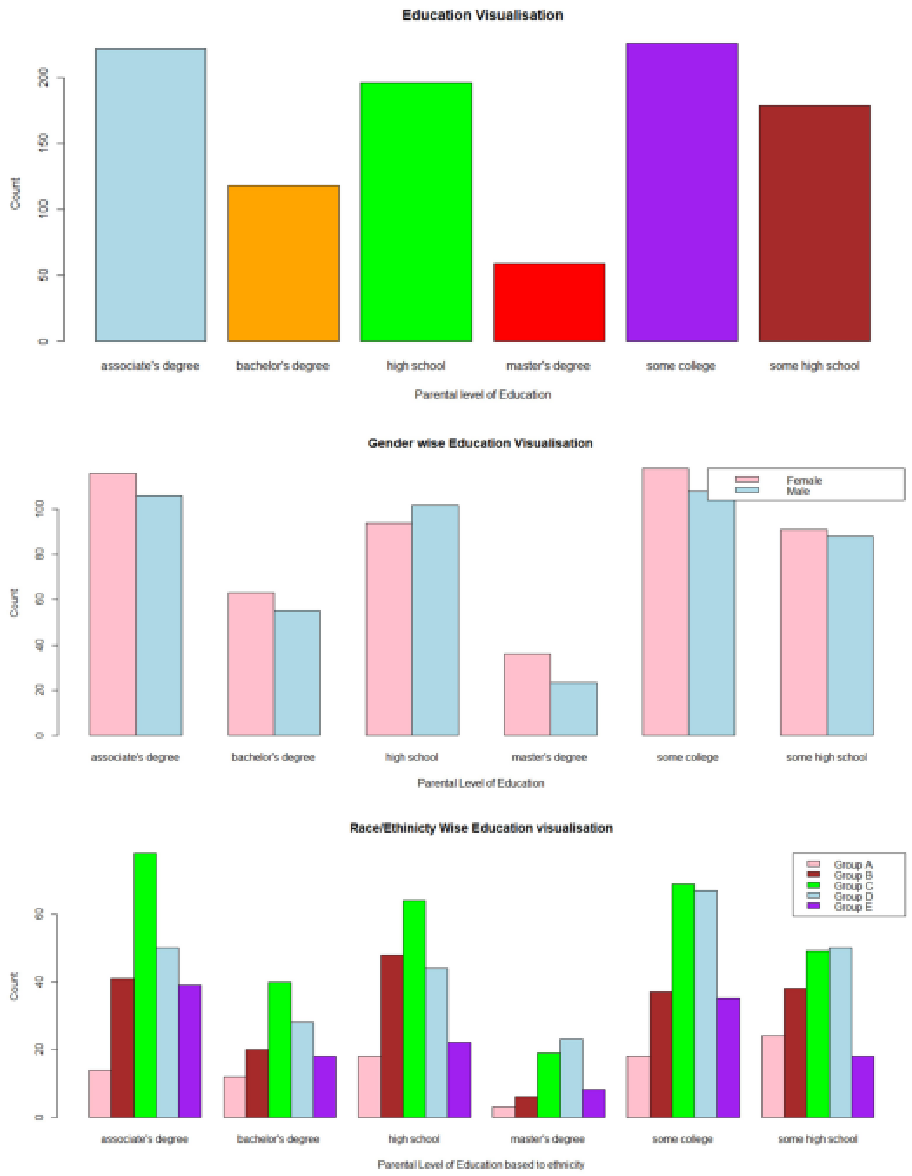
In this data set, bi variate analysis deals with 2 or more variables with regards to Gender, composition of Ethnicity, composition of Education, composition of lunch status in graphical methods and marks scored.

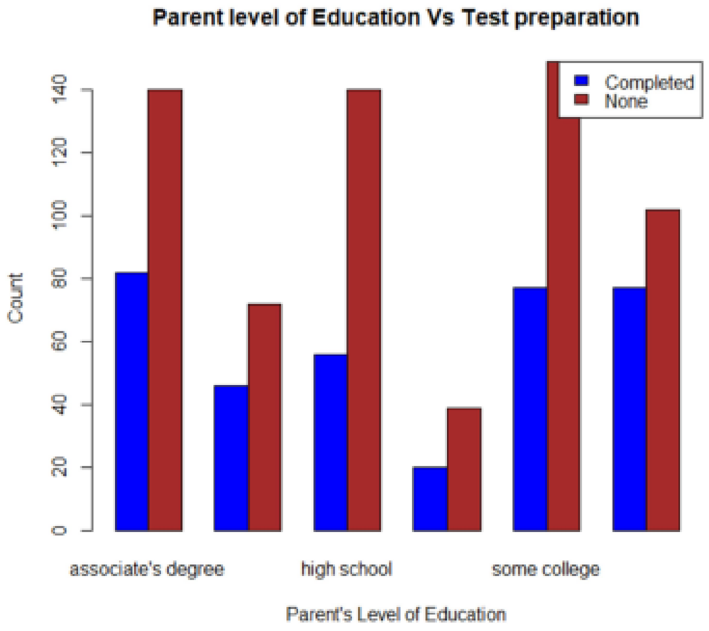
3.3.1 Source code

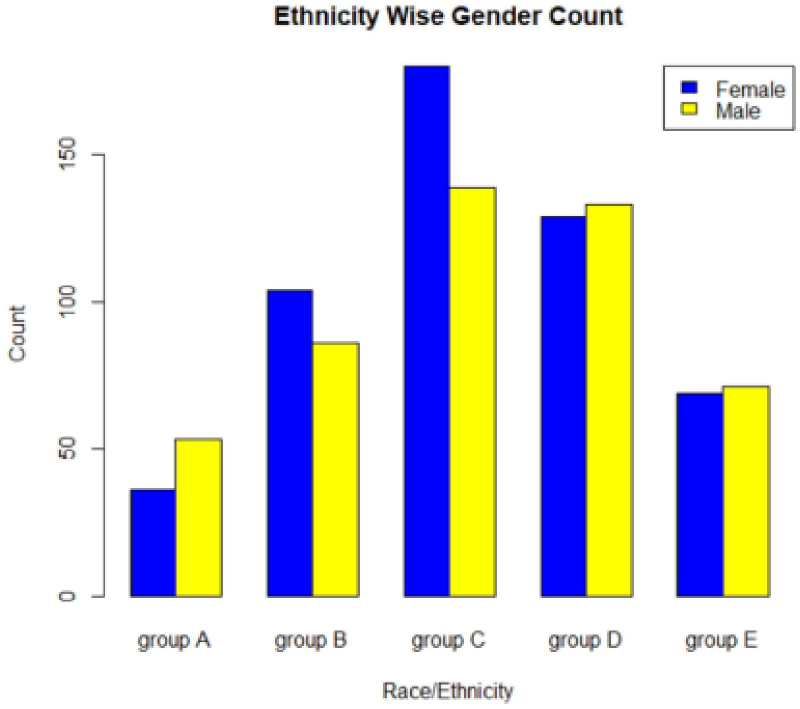
```
#bivariate analysis

#categorical vs categorical
barplot(table(Students_Performance$parental.level.of.education),
xlab="Parental␣level␣of␣Education",ylab = "Count", main=
"Education␣Visualisation",col=
c("lightblue","orange","green","red","purple","brown"))
legend("topright",legend = c("Associate's␣Degree","Bachelor's␣Degree",
"High␣school","Master's␣Degree","Some␣College","Some␣High
School"), fill=c("lightblue","orange","green","red","purple","brown"))

barplot(table(Students_Performance$gender, Students_Performance$
parental.level.of.education), beside = TRUE, xlab="Parental␣Level␣of
Education",ylab="Count",main="Gender␣wise␣Education
Visualisation",col=c("pink","lightblue"))
legend("topright",legend = c("Female","Male"), fill = c("pink",
"lightblue"))

barplot(table(Students_Performance$race.ethnicity,students_performance$
parental.level.of.education), beside = TRUE, xlab="Parental␣Level␣of␣Education
based␣to␣ethnicity",ylab = "Count",main = "Race/Ethinicty␣Wise␣Education
visualisation",col = c("pink","brown","green","lightblue","purple"))
legend("topright",legend = c("Group␣A","Group␣B","Group␣C","Group␣D","Group
E"), fill = c("pink","brown","green","lightblue","purple"))

barplot(table(Students_Performance$gender, Students_Performance
$race.ethnicity), main="Ethnicity␣Wise␣Gender␣Count",
xlab="Race/Ethnicity", ylab = "Count",col =
c("blue","yellow"),beside=TRUE)
legend("topright",legend = c("Female","Male"),
fill = c("blue","yellow"))

barplot(table(Students_Performance$test.preparation.course,
Students_Performance$parental.level.of.education), beside = TRUE, xlab="Parent's
Level␣of␣Education",ylab="Count",main="Parent␣level␣of␣Education␣Vs␣Test
preparation",col = c("blue","brown"))
legend("topright",legend = c("Completed","None"), fill = c("blue","brown"))

grouped_average <-cut(Students_Performance$Average.Score.Rounded, breaks =
c(0,40,60,80,100), labels = c("Fail","Meritorious","Outstanding","Exceptional"))
```

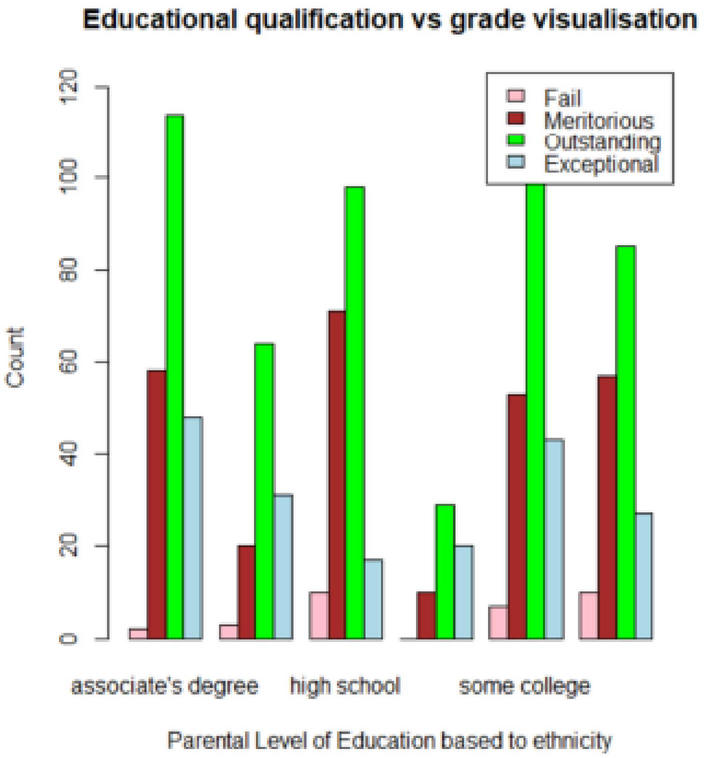
```
grouped_average
barplot(table(students_performance$lunch, grouped_average), beside=TRUE,
col=c("pink","lightblue"),main = "Lunch␣type␣Vs␣Average␣Marks",xlab =
"Grade",ylab = "count")
legend("topright",legend = c("Standard","free/reduced"), fill =
c("pink","lightblue"))

barplot(table(grouped_average, students_performance$parental.level.of.education),
beside = TRUE, xlab="Parental␣Level␣of␣Education␣based␣to␣ethnicity",
ylab = "Count",main = "Race/Ethinicty␣Wise␣Education␣visualisation",
col = c("pink","brown","green","lightblue"))
legend("topright",legend = c("Fail","Meritorious","Outstanding","Exceptional"),
fill = c("pink","brown","green","lightblue"))

# numerical vs numerical
plot(Students_Performance$reading.score, Students_Performance$
math.score,main = "Reading␣Vs␣Math␣Score",xlab="Reading␣Score",
ylab="Math␣Score",pch = 16,col = "lightgreen")
plot(Students_Performance$writing.score, Students_Performance$
math.score,main = "Writing␣Vs␣Math␣Score",xlab="Writing␣Score",
ylab="Math␣Score",pch = 16,col = "pink")
plot(Students_Performance$writing.score, Students_Performance$
reading.score,main = "Reading␣Vs␣Writing␣Score",xlab="Reading␣Score",
ylab="Writing␣Score",pch = 16,col = "purple")

barplot(table(students_performance$test.preparation.course, grouped_reading),
beside = TRUE, ylab = "Count",xlab="Marks␣in␣Maths",main="Maths␣Scores␣Vs␣Test
Preparation␣Status",col = c("pink","lightblue"))
legend("topright",legend = c("Completed","None"), fill=c("pink","lightblue"))

barplot(table(students_performance$test.preparation.course, grouped_writing),
beside = TRUE, ylab = "Count",xlab="Marks␣in␣Writing",main="Writing␣Scores␣Vs
Test␣Preparation␣Status",col = c("pink","lightblue"))
legend("topright",legend = c("Completed","None"), fill=c("pink","lightblue"))

barplot(table(students_performance$test.preparation.course, grouped_math),
beside = TRUE, ylab = "Count",xlab="Marks␣in␣Reading",main="Reading␣Scores␣Vs
Test␣Preparation␣Status",col = c("pink","lightblue"))
legend("topright",legend = c("Completed","None"), fill=c("pink","lightblue"))
```

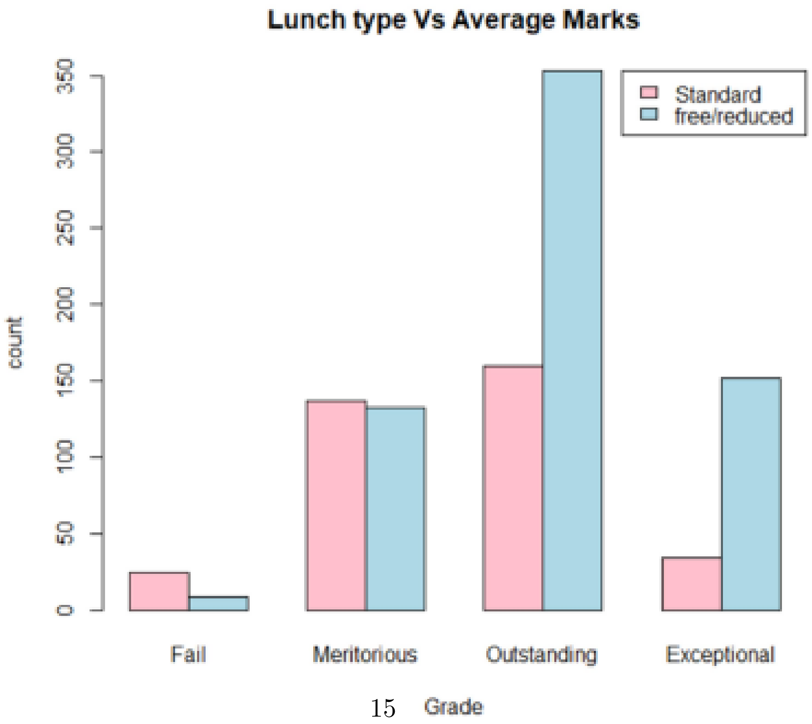
3.3.2 Output

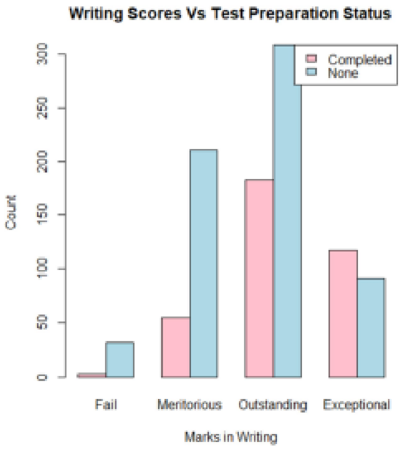

Education Visualisation
Count
0
50
100
150
200
associate's degree
bachelor's degree
high school
master's degree
some college
some high school
Parental level of Education
Gender wise Education Visualisation
Female
Male
Count
0
20
40
60
80
100
associate's degree
bachelor's degree
high school
master's degree
some college
some high school
Parental Level of Education
Race/Ethinicty Wise Education visualisation
Group A
Group B
Group C
Group D
Group E
Count
0
20
40
60
associate's degree
bachelor's degree
high school
master's degree
some college
some high school
Parental Level of Education based to ethnicity

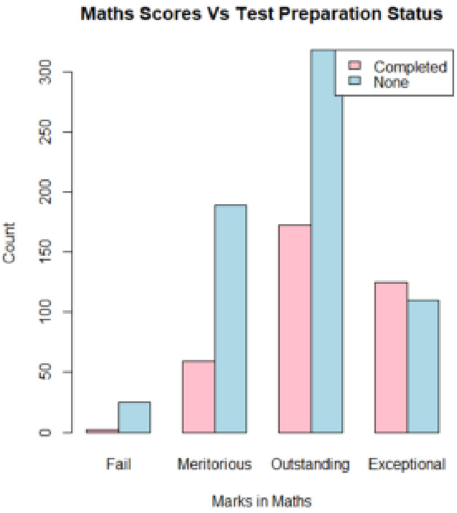

Parent level of Education Vs Test preparation
Completed
None
Count
0
20
40
60
80
100
120
140
associate's degree
high school
some college
Parent's Level of Education

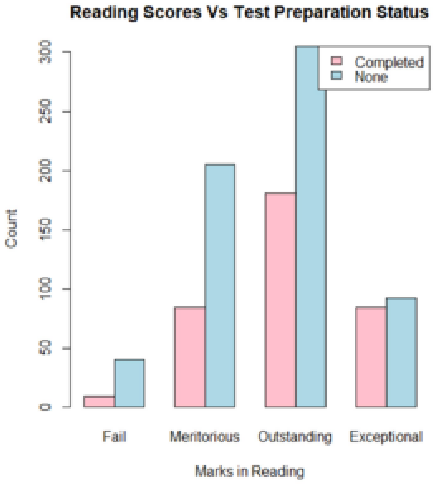

Ethnicity Wise Gender Count
Female
Male
Count
0
50
100
150
group A
group B
group C
group D
group E
Race/Ethnicity

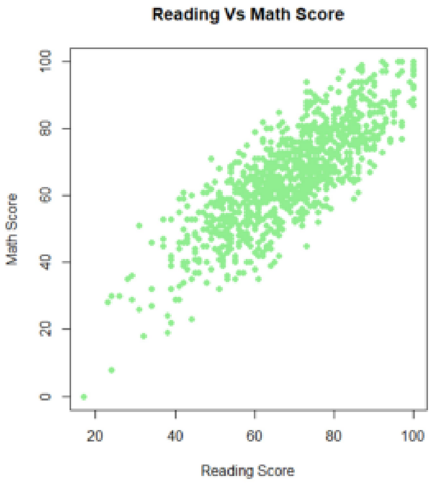

Educational qualification vs grade visualisation
Fail
Meritorious
Outstanding
Exceptional
Count
0
20
40
60
80
100
120
associate's degree
high school
some college
Parental Level of Education based to ethnicity


Lunch type Vs Average Marks
Standard
free/reduced
count
0
50
100
150
200
250
300
350
Fail
Meritorious
Outstanding
Exceptional
Grade

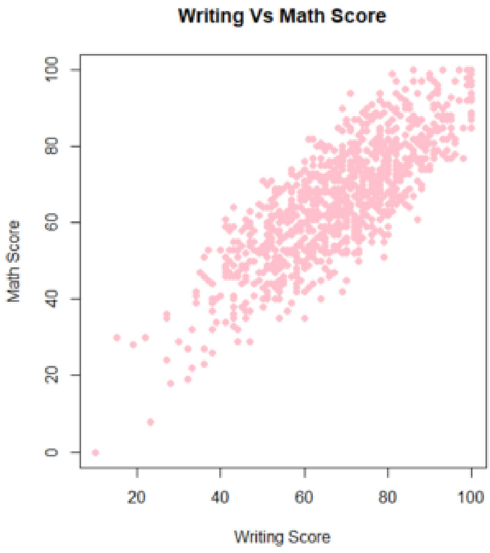

Writing Scores Vs Test Preparation Status
Completed
None
Count
0
50
100
150
200
250
300
Fail
Meritorious
Outstanding
Exceptional
Marks in Writing


Maths Scores Vs Test Preparation Status
Completed
None
Count
0
50
100
150
200
250
300
Fail
Meritorious
Outstanding
Exceptional
Marks in Maths

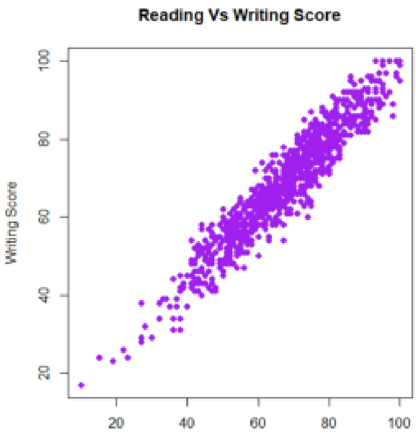

Reading Scores Vs Test Preparation Status
Completed
None
Count
0
50
100
150
200
250
300
Fail
Meritorious
Outstanding
Exceptional
Marks in Reading



Reading Vs Math Score



Writing Vs Math Score



Reading Vs Writing Score

Chapter 4

Conclusion

1. The data set has 1000 rows and 10 columns and no missing data, hence it is suitable for EDA.
2. The max. marks of all the 3 subjects stands at 100, whereas the min. marks vary for all the 3 subject.
3. Yet the other measures of descriptive statistics remain the same for all the 3 subjects
4. All the 3 subjects data frame is not symmetrical i.e., it is a skewed data set.
5. The marks are concentrated between 60 - 100 and there are very less people who have scored below 25 in all the 3 subjects
6. The data set has a slightly high population of females i.e., 518 F for 482 M.
7. Out of 1000 students 2/3rd opt for Standard Lunch and 1/3rd opt for free or reduced lunch.
8. The Educational qualifications of the parents are composite. Associate's degree: 222; Bachelor's degree: 118; High school: 196; Master's degree: 59; some college: 226; Some High school: 179.
9. With regards to ethnicity Group A is the minority with 89 students and Group C with the majority of 319 students.
10. Out of the 1000 students in this data frame more than 50 percentage of students have not done any test preparation course
11. The following are the categories that marks are converted into for grading purposes.
    (a) 00 to 40 marks :- Fail
    (b) 40 to 60 marks :- Meritorious
    (c) 60 to 80 marks :- Outstanding
    (d) 80 to 100 marks:- Outstanding
12. The following table shows the different categories of marks and subjects:

Subject	Fail	Meritorious	Outstanding	Exceptional
Reading	27	248	490	235
Writing	35	266	491	208
Mathematics	49	289	485	176
Total	32	269	513	186

13. Test preparation course and its relationship with students scoring marks.
    (a) Mathematics : Though the course does not have much affect of marks, but the course helps in students scoring more than 80 marks
    (b) Writing : Though the course does not have much affect of marks, but the course helps in students scoring more than 80 marks. The non-completion of the course helps people score more.
    (c) Reading : Though the course does not have much affect of marks, but the course helps in students scoring more than 80 marks
    (d) Total : Though the course does not have much affect of marks, but the course helps in students scoring more than 80 marks.
14. Females students have better educational qualification in their parents, hence we could conclude that females have better educational .
15. there are a very few people who go on to do a master's degree. Most of them own an associate degree, All ethnicity have more or less the same educational opportunities
16. The data has some gender disparity, yet all ethnic groups have the same composition of males and females.
17. Educational Qualification has no effect on the choice of test preparation course, while having a higher educational qualification means a lot of students taking the course, there is no exact relationship
18. We can see that students who opt for reduced/free lunch perform better than students who opt for Standard lunches.
19. Better educational qualification of parents do have an effect on the grade of students, meaning IQ levels are hereditary
20. Relationship between the marks of 3 Subjects:
    (a) Reading and Maths : Reading and Maths score have a strong linear correlation of 0.8.
    (b) Writing and Maths : Writing and Maths score have a strong linear correlation of 0.8.
    (c) Reading and Writing : Reading and Writing score have a very strong linear correlation of 0.95.