Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6010 - Probability Theory and Introductory Statistics**

**Assignment:**

MODULE 2 PRACTICE ASSIGNMENT 2

**Submitted on:**

November 13, 2022

**Submitted to:**  **Submitted by:**

Professor: BEHZAD AHMADI NIKSHITA RANGANATHAN

**ABSTRACT**

Students’ academic performance improvement is a task that involves many factors to monitor and analyze for the academic community of higher learning. Analyzing the student test data gives meaning to the information collected and is essential to appropriately utilize data interpretation and use it to determine an effective approach for driving individual performance and communicating the test results.

Student test data analysis provides a means to look at student performance to provide evidence about student learning in the curriculum of various subjects, provide information about subject strengths and weaknesses, and help us in decision-making. The importance of data analytics is growing across all industries including service sectors like educational institutions generating vast amounts of knowledge acumen that can offer insightful information.

Student data analytics discusses the quantitative and qualitative methods used to boost productivity, data is extracted and then categorized to identify and explore the student study behavioral data and patterns. We study data to make decisions based on the outcome of the test which enables the practitioner to understand at which level the students are in the learning and helps to set the goals and learning intention for the next steps in the learning process and plan the learning program.

**INTRODUCTION**

This project aims to provide a detailed exploratory analysis of the student marks secured by the students in high school students from the United States. The test data that is used within the project is sourced from [*Kaggle*](https://www.kaggle.com/datasets/spscientist/students-performance-in-exams) by the data owner, Jakki Seshapanpu.

**About this Dataset:** The dataset considered is based on student performance. This dataset consists of 1000 records of high school student marks in 3 Subjects Math, Reading & writing subjects. It comprises of the student details of 8 attributes of both male & female students in the US. The dataset has information such as race/ethnicity, gender of the student, level of education of parents, lunch type, and test preparation course.

The data from the dataset is imported into RStudio and analyzed to understand several insights to understand the influence of several factors on the student’s performance.

Below are the data descriptions of each variable of the data that briefly describe the contents of the data set. The features of the dataset are as follows:

|  |  |  |
| --- | --- | --- |
| **No** | **Feature** | **Dictionary** |
| 1 | gender | Gender of the student (Male/Female) |
| 2 | race/ethnicity | Ethnicity/race of the student (group A/B/C/D/E) |
| 3 | parental level of education | Highest educational qualification of the parents of each student |
| 4 | lunch | Type of the lunch included in the package selected for the student (standard/reduced) |
| 5 | test preparation course | Test preparation course was completed by the student or not |
| 6 | math score | Score in Math Subject |
| 7 | reading score | Score in Reading Subject |
| 8 | writing score | Score in Writing Subject |

*Table 1: Dictionary with the Features of the Student Performance Dataset*

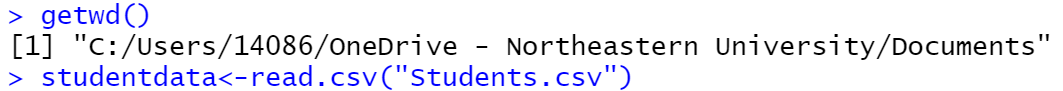
The purpose of this project is to provide an exploratory analysis of the test data of students. The following are the questions derived to do the data analysis on student performance which is presented in a variety of data visualizations:

1. In which subjects’ majority of the students master?
2. Which subjects did the students struggle with?
3. How are the relationships between 3 subjects on test scores?
4. Whether student scores impacted by the effectiveness of the test preparation course?
5. What effect does parent education have on their children's performance?
6. Is there relevance of student race with respect to student performance?
7. Which student gender shows the best performance in the subjects?
8. Which type of lunch opted for by students influences student performance?

**DATA CLEANING AND MANIPULATION**

* Importing the Students CSV file

<studentdata> vector contains the details of the student performance



**Figure 1-read.csv()**

* Manipulating the dataset
* Revising the column names

For ease of understanding, new column names are provided.

Text

Description automatically generated

**Figure 2-Changing column names**

* Visualization and checking of <NA> values

gg\_miss\_which is a function under the naniar package and is used to check the column that has the blanks.

miss\_var\_summary() summarizes the missingness under each variable.

There are no NA values in the studentdata dataset.

**Graphical user interface, text, application

Description automatically generatedTable

Description automatically generated with medium confidence**

A picture containing logo

Description automatically generated

**Figure 3-<NA> values**

* Checking for duplication

I checked for identical rows in the dataset We can confirm using anyDuplicated() that there are no duplicate rows in the student performance dataset.

Table

Description automatically generated

A picture containing logo

Description automatically generated

**Figure 4 – duplicated()and anyDuplicated()**

* Changing datatypes

as.factor() changes the data type of the columns from one datatype to factor.

Text

Description automatically generated

**Figure 5 – as.factor()**

**DATA ANALYSIS**

* Analyzing the studentdata dataset

The str() function displays the various datatypes for the variables

A picture containing text

Description automatically generated

**Figure 6-str()**

dim() views the number of columns and rows of a dataset. We have 1000 observations and 8 features in this case.

A picture containing text

Description automatically generated

**Figure 7-dim()**

summary() - It can be observed that the variables of the dataset are of datatype character and integer.

A screenshot of a computer

Description automatically generated with medium confidence

**Figure 8-summary()**

glimpse() is part of the dplyr package, and we can see the preview of columns of the dataset with the help of this function.

Text, letter

Description automatically generated

**Figure 9-glimpse()**

skim() is useful for getting a statistical summary of the features

Table

Description automatically generated with medium confidence

**Figure 10-skim()**

headTail() - This function is a part of the psych package. We can see the first 4 and last 4 records of the studentdata below.

Text

Description automatically generated

**Figure 11-headTail()**

psych::describe() – This command gives details of descriptive statistics for studentdata data frame in one glance

The terms "mean" and "standard deviation" refers to the average of the numbers and the variability around the mean, respectively.

The majority of the features have kurtosis values below 3, which denotes a flat and less peaks distribution (Platykurtic).Skew values between -0.5 to 0.5 show symmetric distribution whereas less than -0.5 shows negatively skewed distribution.

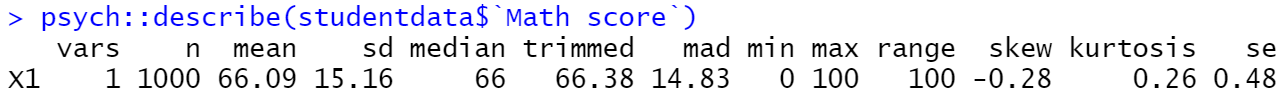
Table

Description automatically generated

**Figure 12-describe()**

* Descriptive statistics of the scores

Math Scores



**Figure 13-describe() – Math scores**

Text

Description automatically generated

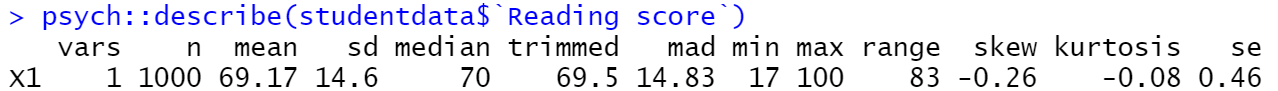
**Figure 14-IQR() and t.test() – Math scores**

Text

Description automatically generated

**Figure 15-var() and quantile() – Math scores**

Reading Scores



**Figure 16-describe() – Reading scores**

Text

Description automatically generated

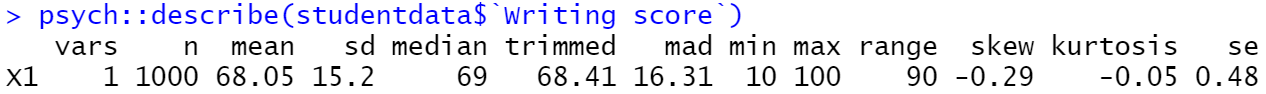
**Figure 17-IQR() and t.test() – Reading scores**

Text

Description automatically generated

**Figure 18-var() and quantile() – Reading scores**

Writing Scores



**Figure 19-describe() – Writing scores**

Text

Description automatically generated

**Figure 20-IQR() and t.test() – Writing scores**

Text

Description automatically generated

**Figure 21-var() and quantile() – Writing scores**

* Descriptive statistics by groups

What is a three-line table format that is commonly used in white papers?

Three-line tables consist of 3 lines (The first one is above the column titles, the second one is under the column titles and the third one is at the end of the table). There are no vertical lines present in them.

I used print\_table to display the descriptive statistical summary using three-line tables for each group.

By Gender

A screenshot of a computer

Description automatically generated with low confidence

**A screenshot of a computer

Description automatically generated with low confidence**

**Figure 22-describeBy() Gender**

**Table

Description automatically generated**Below three-line tables have details like mean, median, standard deviation, variance, skew, etc based on gender grouping(female and male)

**Figure 23-Three Line table (Female group)**

**Table

Description automatically generated**

**Figure 24-Three Line table (Male group)**

By Ethnicity

A screenshot of a computer

Description automatically generated with low confidence

**A screenshot of a computer

Description automatically generated with low confidence**

**Calendar

Description automatically generated with medium confidence**

**Figure 25- describeBy() Ethnicity**

By Test Prep

**Table

Description automatically generated**

**Figure 26- describeBy() Test Prep**

Calendar

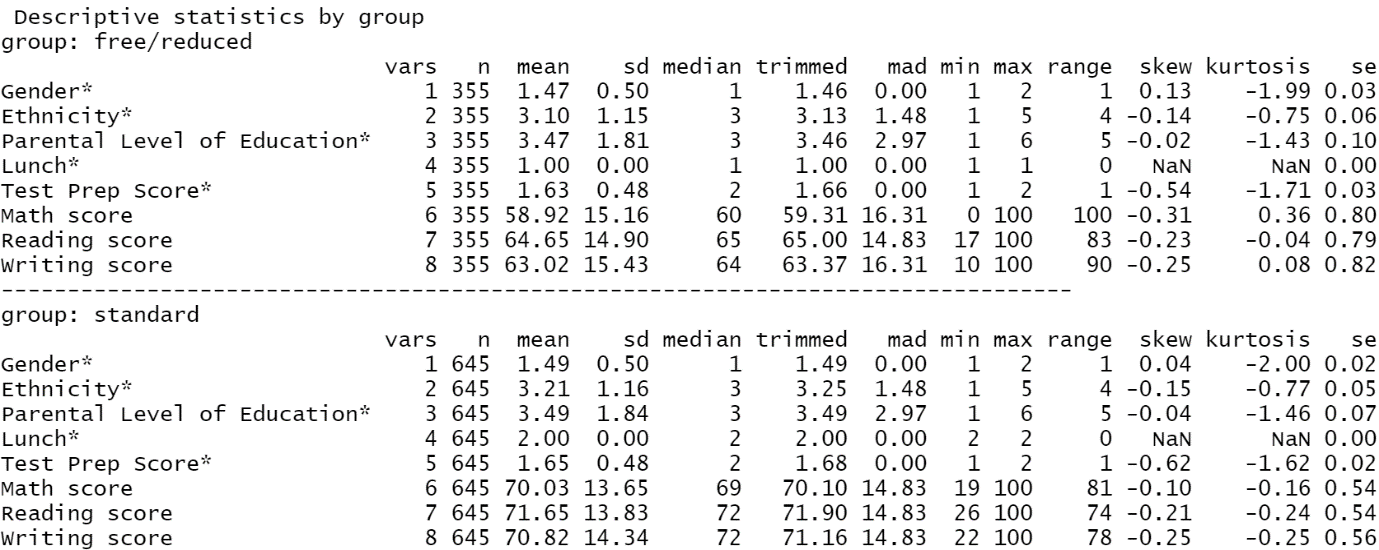
Description automatically generated

**Calendar

Description automatically generated****Figure 27-Three Line table (Test Prep None group)**

**Figure 28-Three Line table (Test Prep Completed group)**

By Lunch



**Figure 29-- describeBy() Lunch**

By Parental Level of Education

**A screenshot of a computer

Description automatically generated with low confidence**

**A picture containing text, receipt

Description automatically generated**

**A screenshot of a computer

Description automatically generated with low confidence**

**Figure 30-- describeBy() Parental Level of Education**

**DATA VISUALIZATION**

* When do you use a jitter chart?

A jitter is an irregular motion or variation. A jitter plot is similar to a scatterplot (representation in the form of dots) but is easier to see how measurement and categorical variables relate to one another. When fewer data points are needed to illustrate the data, jitter charts are more insightful.

* Boxplot

A box plot's first quartile (Q1) and third quartile (Q3) correspond to the 25th and 75th percentiles, respectively. Between the Q1 and Q3 quartiles is the interquartile range (IQR). The values outside the extreme limits (Maximum and minimum) of the boxplots are known as outliers.

Maximum = Q3+1.5 \* (IQR), Minimum = Q1-1.5 \* (IQR)

* Chart, histogram

  Description automatically generatedPlot 1

**Figure 31-- Histograms**

All three subjects, Math, Reading & writing scores in the histograms are negatively skewed. 27.5% of students have scored between 60-70 in Math. 22.5% of students have scored above 80 in reading compared to writing which is 20% in writing. In math, 17.5% of students approximately scored above 80.

There is no significant difference in scores are evident between Reading and writing scores when we considered the students’ scores of 60-80. In math subject, 45 students have not even scored an average of 50% score which shows that the students were struggling.

Overall, the reading score of the students was best compared to other subjects.

* Chart

  Description automatically generatedPlot 2

**Figure 32- Correlation Matrix**

Reading and writing scores have a correlation of 0.95 which shows a strong positive linear relationship. This shows that as one variable (subject) declines, the other variable also declines, or when one variable rises while the other rises. because the two variables' directions are similar. Writing and Math have a correlation of 0.8 and Reading and Math have a correlation of 0.82 which represents a fairly strong association.

* Chart, box and whisker chart

  Description automatically generatedChart, box and whisker chart

  Description automatically generated**Chart, box and whisker chart

  Description automatically generated**Plot 3

**Figure 33- Boxplots**

In Math subject, male students, who have completed the test prep course have scored well. The male students who completed the course have reached 60-80 scores whereas the students who did not complete the course have attained 50-75 scores.

Similarly, reading & writing is dominated by female students irrespective of the course studied. Students who have done the course have scored 60-90 whereas the score is 55-80 range for that course not completed.

* Diagram, schematic

  Description automatically generatedPlot 4

Diagram

Description automatically generated

**Diagram

Description automatically generated**

**Figure 34- Boxplots with Jitter**

In all 3 subjects of Math, Reading & writing, it is evident that parents who have the level of education being “Masters” then their children have reached above average as 75 score which is higher compared to others.

Similarly, parents who have the level of education being “high school”, then those children have attained average of approximately 65 score which is low.

This shows that if level of education of parents is higher that resulted in students scoring higher.

* Chart, scatter chart

  Description automatically generatedPlot 5

**Figure 35- Scatterplots**

There is a strong relationship between all 3 subjects so there is an impact of increase or decrease in 1 subject reflecting on the other. If one student does better in Math, then his score was seen to improve and better in Reading and writing as well.

* **Chart, scatter chart

  Description automatically generated**Chart, scatter chart

  Description automatically generatedPlot 6

**Chart, scatter chart

Description automatically generated**

**Figure 36- Scatterplots with Boxplots**

Male students have performed better in Math subject, while female students have done better in Reading and writing.

* A picture containing diagram

  Description automatically generatedPlot 7

**Figure 37- Jitter plot**

Students who had opted for “standard lunch” got a better score and performed well compared to students who had “Free/reduced lunch”.

This shows that healthier meals have played a vital role in increasing the student score that raised student achievement on average.”

* **Chart

  Description automatically generatedChart, histogram

  Description automatically generated**Plot 8

**Chart

Description automatically generated**

**Figure 38- Density plot**

The data confirms that there is a significant connection between race and academic achievement. The student score of group “E” is best compared to any other group. Also, it is evident that Groups “C” & “D” have similar student performance as they have reached the same range of scores.

**CONCLUSION**

Numerous researchers have investigated the various aspects of what influences students' academic achievement in their research. The academic achievement of a student is significantly influenced by both internal and external classroom conditions. In the dataset taken for analysis, we see that most of the data considered are external such as parent education, lunch, and test preparation course. Studies have shown that factors such as learning environments, gender variances, and age disparities, among others, can have an impact on students’ performance. (Hansen, Joe B., 2000).

**Key takeaway points:**

1. Female students are more in overall to take up the test compared to males.
2. More strugglers were in Math subjects compared to reading & writing.
3. If a student scores good score in Math, then due to a strong linear correlation, he tends to do better in the other two subjects as well.
4. Students who have completed the test preparation course have done better in all 3 subjects, compared to students who have not opted for the course. Also, reading & writing is dominated by female students irrespective of the course studied.
5. Parents who have a higher level of education, than their children have scored higher compared to those parents who have less education.
6. There is a significant difference between race & student score.
7. Male students have performed better in Math subject, while female students have done better in Reading and writing.
8. Healthier meals have played a vital role in increasing the student score that raised student achievement on average.

**REFERENCES**

Hansen, J. B. (2000, April 25). *ERIC - ED443876 - Student Performance and Student Growth as Measures of Success: An Evaluator’s Perspective., 2000-Apr-25.* <https://eric.ed.gov/?id=ED443876>

Seshapanpu, J. (2019, November). Students Performance in Exams. Kaggle. Retrieved November 10, 2022, from https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

Low-code data app development. Plotly. (n.d.). Retrieved October 28, 2022, from https://plotly.com/

**APPENDIX**

#---------------------- Week\_2\_Module\_2 R Script ----------------------#

print("Author : Nikshita Ranganathan")

print("Week 2 Assignment - Module 2 R Practice")

print("Course Name - ALY6010: Probability Theory and Introductory Statistics")

# Loading the packages

library(dplyr)

library(psych)

library(ggplot2)

library(naniar)

library(skimr)

library(moments)

library(magrittr)

# Importing the dataset of student performance

# Check the current working directory

getwd()

studentdata<-read.csv("Students.csv")

#------------------- Data Cleaning -------------------#

# Changing the column names

names(studentdata)

names(studentdata)<-c('Gender','Ethnicity','Parental Level of Education','Lunch','Test Prep Score','Math score','Reading score','Writing score')

# Checking NA values

gg\_miss\_which(studentdata)

miss\_var\_summary(studentdata)

sum(is.na(studentdata))

# Checking for duplication

duplicated(studentdata)

anyDuplicated(studentdata)

# Changing datatypes to factors

studentdata$Gender=as.factor(studentdata$Gender)

studentdata$Lunch=as.factor(studentdata$Lunch)

studentdata$Ethnicity=as.factor(studentdata$Ethnicity)

studentdata$`Parental Level of Education`=as.factor(studentdata$`Parental Level of Education`)

studentdata$`Test Prep Score`=as.factor(studentdata$`Test Prep Score`)

#------------------- Exploratory Data Analysis -------------------#

# Analyzing studentdata dataset

str(studentdata)

dim(studentdata)

summary(studentdata)

glimpse(studentdata)

skim(studentdata)

headTail(studentdata)

psych::describe(studentdata)

# Descriptive Statistics (describe and other functions)

psych::describe(studentdata$`Math score`)

IQR(studentdata$`Math score`)

t.test(studentdata$`Math score`)

var(studentdata$`Math score`)

quantile(studentdata$`Math score`)

scale(studentdata$`Math score`)

psych::describe(studentdata$`Reading score`)

IQR(studentdata$`Reading score`)

t.test(studentdata$`Reading score`)

var(studentdata$`Reading score`)

quantile(studentdata$`Reading score`)

scale(studentdata$`Reading score`)

psych::describe(studentdata$`Writing score`)

IQR(studentdata$`Writing score`)

t.test(studentdata$`Writing score`)

var(studentdata$`Writing score`)

quantile(studentdata$`Writing score`)

scale(studentdata$`Writing score`)

install.packages("bruceR")

library(bruceR)

print\_table(studentdata)

# Descriptive Statistics by Group (Using describeby)

# BY GENDER

describeBy(studentdata,group = studentdata$Gender)

female<-filter(studentdata,studentdata$Gender=='female')

f<-describe(female)

print\_table(f,digits=1)

male<-filter(studentdata,studentdata$Gender=='male')

m<-describe(male)

print\_table(m,digits=1)

# BY ETHNICITY

describeBy(studentdata,group = studentdata$Ethnicity)

# BY TEST PREP

describeBy(studentdata,group = studentdata$`Test Prep Score`)

none<-filter(studentdata,studentdata$`Test Prep Score`=='none')

n<-describe(none)

print\_table(n,digits=1)

completed<-filter(studentdata,studentdata$`Test Prep Score`=='completed')

c<-describe(completed)

print\_table(c,digits=1)

# BY LUNCH

describeBy(studentdata,group = studentdata$`Lunch`)

# BY LEVEL OF EDUCATION

describeBy(studentdata,group = studentdata$`Parental Level of Education`)

#------------------- Data Visualizations -------------------#

# Plot 1

par( mfrow= c(1,3) )

hist(studentdata$`Math score`,ylim = c(0,300),xlim=c(0,100),main="Histogram A",xlab="Math Scores",col="#E1DEFC")

hist(studentdata$`Reading score`,ylim = c(0,300),xlim=c(0,100),main="Histogram B",xlab="Reading Scores",col="lightsalmon")

hist(studentdata$`Writing score`,ylim = c(0,300),xlim=c(0,100),main="Histogram C",xlab="Writing Scores",col="yellowgreen")

# Plot 2 - Correlation Matrix

scores<-studentdata %>% select(`Math score`,`Reading score`,`Writing score`)

install.packages("ggcorrplot")

library(ggcorrplot)

ggcorrplot::ggcorrplot(cor(scores),colors = c("#6D9EC1", "white", "#E46726"),lab = TRUE,title = "Correlation Matrix")

# Plot 3

library(wesanderson)

ggplot(studentdata,aes(x=`Test Prep Score`, y=`Math score`, fill=`Gender`)) +geom\_boxplot() +facet\_wrap(~`Test Prep Score`,scale="free")+ scale\_fill\_manual (values=wes\_palette(n=2, name="BottleRocket2"))

ggplot(studentdata,aes(x=`Test Prep Score`, y=`Reading score`, fill=`Gender`)) + geom\_boxplot()+facet\_wrap(~`Test Prep Score`, scale="free") + scale\_fill\_manual (values=wes\_palette(n=2, name="Darjeeling2"))

ggplot(studentdata,aes(x=`Test Prep Score`, y=`Writing score`, fill=`Gender`))+ geom\_boxplot() + facet\_wrap(~`Test Prep Score`, scale="free") + scale\_fill\_manual (values=wes\_palette(n=2, name="GrandBudapest2"))

# Plot 4 -

ggplot(studentdata,aes(x=`Parental Level of Education`,y=`Math score`,fill=`Parental Level of Education`))+geom\_boxplot(alpha=0.4,notch=TRUE,outlier.colour = "red",outlier.size = 2)+geom\_jitter(color="black", size=0.4, alpha=0.9)+scale\_color\_brewer(palette="Dark2")+ggtitle("Boxplot (A) with jitter")

ggplot(studentdata,aes(x=`Parental Level of Education`,y=`Reading score`,fill=`Parental Level of Education`))+geom\_boxplot(alpha=0.4,notch=TRUE,outlier.colour = "red",outlier.size = 2)+geom\_jitter(color="black", size=0.4, alpha=0.9)+scale\_color\_brewer(palette="Dark2")+ggtitle("Boxplot(B) with jitter")

ggplot(studentdata,aes(x=`Parental Level of Education`,y=`Writing score`,fill=`Parental Level of Education`))+geom\_boxplot(alpha=0.4,notch=TRUE,outlier.colour = "red",outlier.size = 2)+geom\_jitter(color="black", size=0.4, alpha=0.9)+scale\_color\_brewer(palette="Dark2")+ggtitle("Boxplot(C) with jitter")

# Plot 5 - Scatterplots

Reading=studentdata$`Reading score`

Math=studentdata$`Math score`

Writing=studentdata$`Writing score`

par( mfrow= c(1,3) )

plot(Math, Reading, main = "Scatterplot A",xlab = "Math scores", ylab = "Reading scores",pch = 16, frame = FALSE,col="sandybrown")

abline(lm(Reading ~ Math, data = studentdata), col = "black",lty=2,lwd=1.5)

plot(Reading, Writing, main = "Scatterplot B",xlab = "Reading scores", ylab = "Writing scores",pch = 16, frame = FALSE,col="#56B4E9")

abline(lm(Writing ~ Reading, data = studentdata), col = "black",lty=2,lwd=1.5)

plot(Writing, Math, main = "Scatterplot C",xlab = "Writing scores", ylab = "Math scores",pch = 16, frame = FALSE,col="#00A087FF")

abline(lm(Math ~ Writing, data = studentdata), col = "black",lty=2,lwd=1.5)

# Plot 6

library(ggplot2)

install.packages("ggExtra")

library(ggExtra)

fig<-ggplot(studentdata, aes(x=Math, y=Reading,color=`Gender`)) +geom\_point() +scale\_color\_manual(values=wes\_palette(n=2, name="Chevalier1"))+theme(legend.position="top")

ggMarginal(fig, type="boxplot")

fig2<-ggplot(studentdata, aes(x=Reading, y=Writing,color=`Gender`)) +geom\_point() +scale\_color\_manual(values=wes\_palette(n=2, name="Royal1"))+theme(legend.position="top")

ggMarginal(fig2, type="boxplot")

fig3<-ggplot(studentdata, aes(x=Writing, y=Math,color=`Gender`)) +geom\_point() +scale\_color\_manual(values=wes\_palette(n=2, name="Moonrise3"))+theme(legend.position="top")

ggMarginal(fig3, type="boxplot")

# Plot 7

install.packages("gridExtra")

library(gridExtra)

fig4<-ggplot(studentdata,aes(x=`Lunch`,y=`Math score`,col=`Lunch`))+geom\_jitter()+coord\_flip()+scale\_color\_brewer(palette="Accent")

fig5<-ggplot(studentdata,aes(x=`Lunch`,y=`Reading score`,col=`Lunch`))+geom\_jitter()+coord\_flip()+scale\_color\_brewer(palette="Accent")

fig6<-ggplot(studentdata,aes(x=`Lunch`,y=`Writing score`,col=`Lunch`))+geom\_jitter()+coord\_flip()+scale\_color\_brewer(palette="Accent")

grid.arrange(fig4, fig5, fig6)

# Plot 8

library(dplyr)

library(hrbrthemes)

library(tidyr)

library(viridis)

ggplot(studentdata,aes(x=`Math score`,group=`Ethnicity`,fill=`Ethnicity`))+geom\_density(alpha=0.3)+theme\_ipsum()+ggtitle("Multiple Density Chart A")

ggplot(studentdata,aes(x=`Reading score`,group=`Ethnicity`,fill=`Ethnicity`))+geom\_density(alpha=0.3)+theme\_ipsum()+ggtitle("Multiple Density Chart B")

ggplot(studentdata,aes(x=`Writing score`,group=`Ethnicity`,fill=`Ethnicity`))+geom\_density(alpha=0.3)+theme\_ipsum()+ggtitle("Multiple Density Chart C")