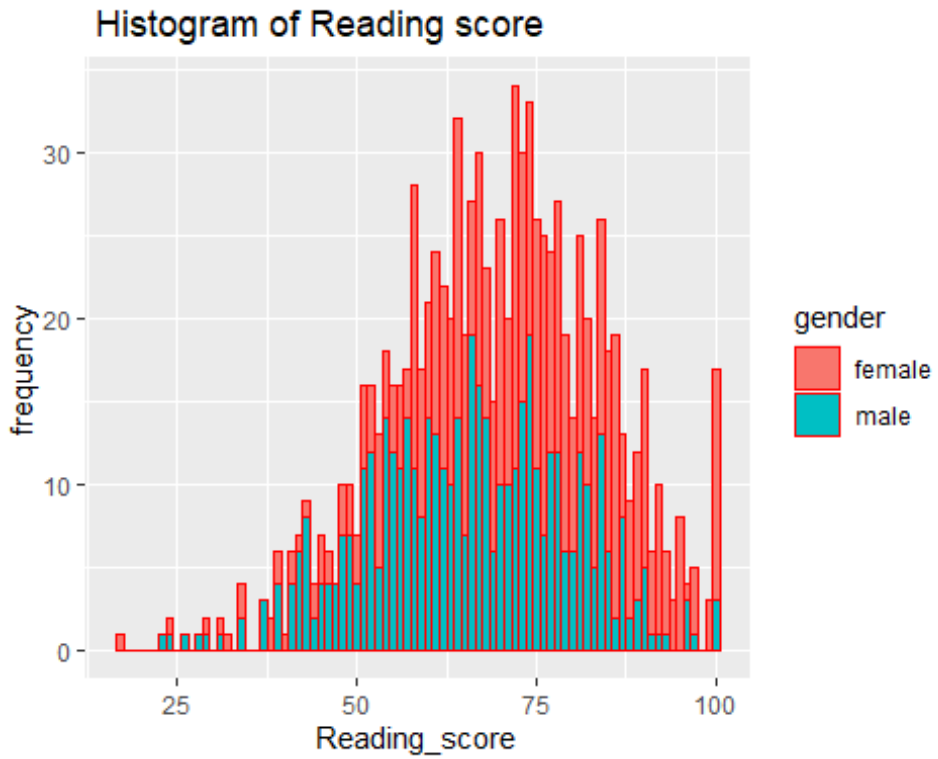
Visualization.pdf

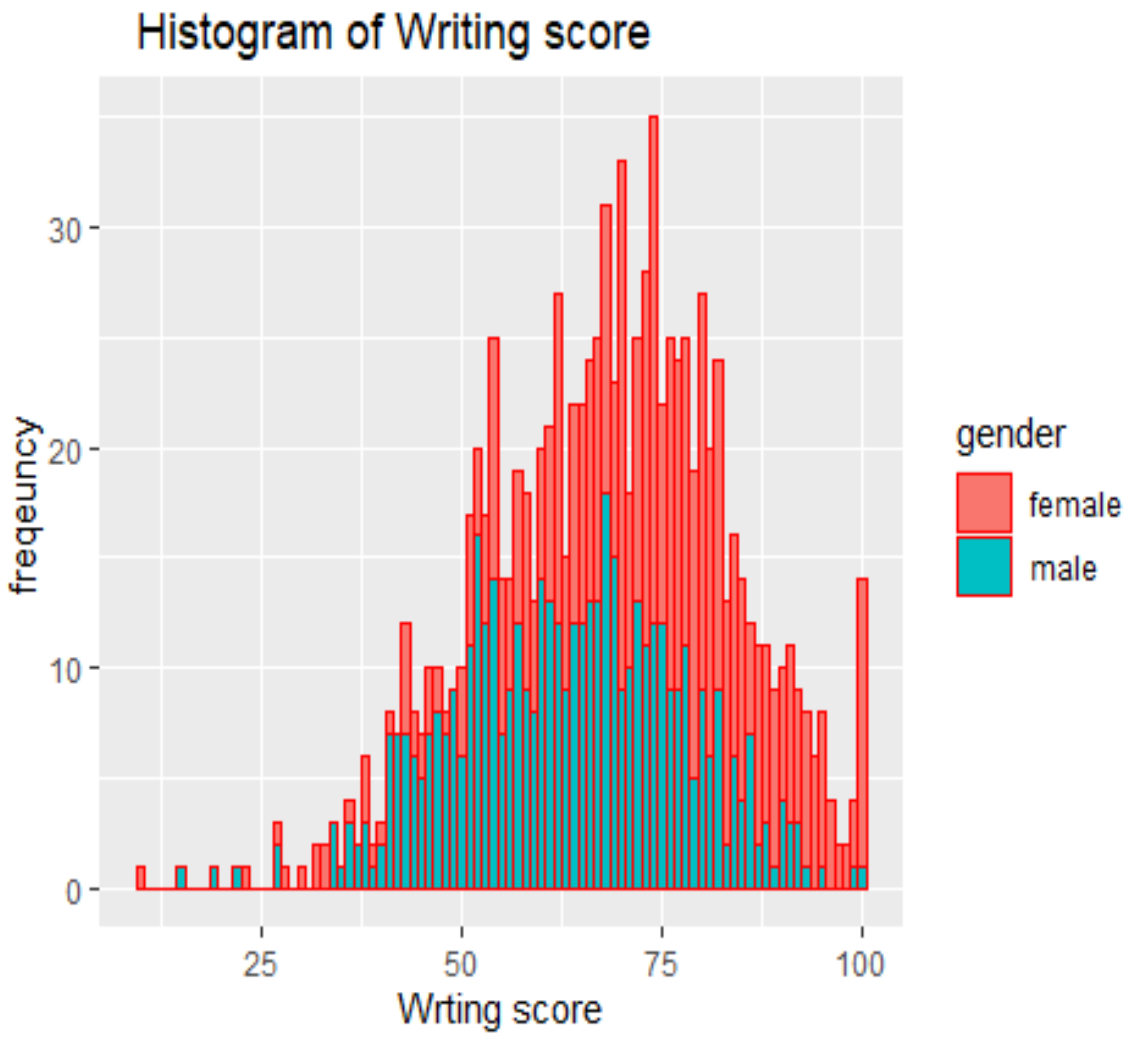
```
library(readr)
sp= read_csv("StudentsPerformance.csv")

##
## -- Column specification --------------------------------------------------
------
## cols(
##   gender = col_character(),
##   `race/ethnicity` = col_character(),
##   `parental level of education` = col_character(),
##   lunch = col_character(),
##   `test preparation course` = col_character(),
##   `math score` = col_double(),
##   `reading score` = col_double(),
##   `writing score` = col_double()
## )

View(sp)

# creating data frame
data1=data.frame(sp)
View(data1)

#Finding unique vales in each column
for (i in seq(1,ncol(data1)-3,1)){
  print(unique(data1[i]))
}

##   gender
## 1 female
## 4   male
##    race.ethnicity
## 1         group B
## 2         group C
## 4         group A
## 9         group D
## 33        group E
##    parental.level.of.education
## 1            bachelor's degree
## 2                 some college
## 3              master's degree
## 4           associate's degree
## 9                  high school
## 16           some high school
##          lunch
## 1     standard
```

```
## 4 free/reduced
##   test.preparation.course
## 1                    none
## 2               completed

# cleaning data/missing values
clean_data=complete.cases(data1)
a=data1[clean_data,]
View(a)
any(is.na(data1))

## [1] FALSE

# Summarizing all the columns
summary(data1)

##     gender          race.ethnicity     parental.level.of.education
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     lunch           test.preparation.course   math.score     reading.score
##  Length:1000        Length:1000             Min.   :  0.00   Min.   :
17.00
##  Class :character   Class :character        1st Qu.: 57.00   1st Qu.:
59.00
##  Mode  :character   Mode  :character        Median : 66.00   Median :
70.00
##                                             Mean   : 66.09   Mean   :
69.17
##                                             3rd Qu.: 77.00   3rd Qu.:
79.00
##                                             Max.   :100.00   Max.
:100.00
##  writing.score
##  Min.   : 10.00
##  1st Qu.: 57.75
##  Median : 69.00
##  Mean   : 68.05
##  3rd Qu.: 79.00
##  Max.   :100.00
```

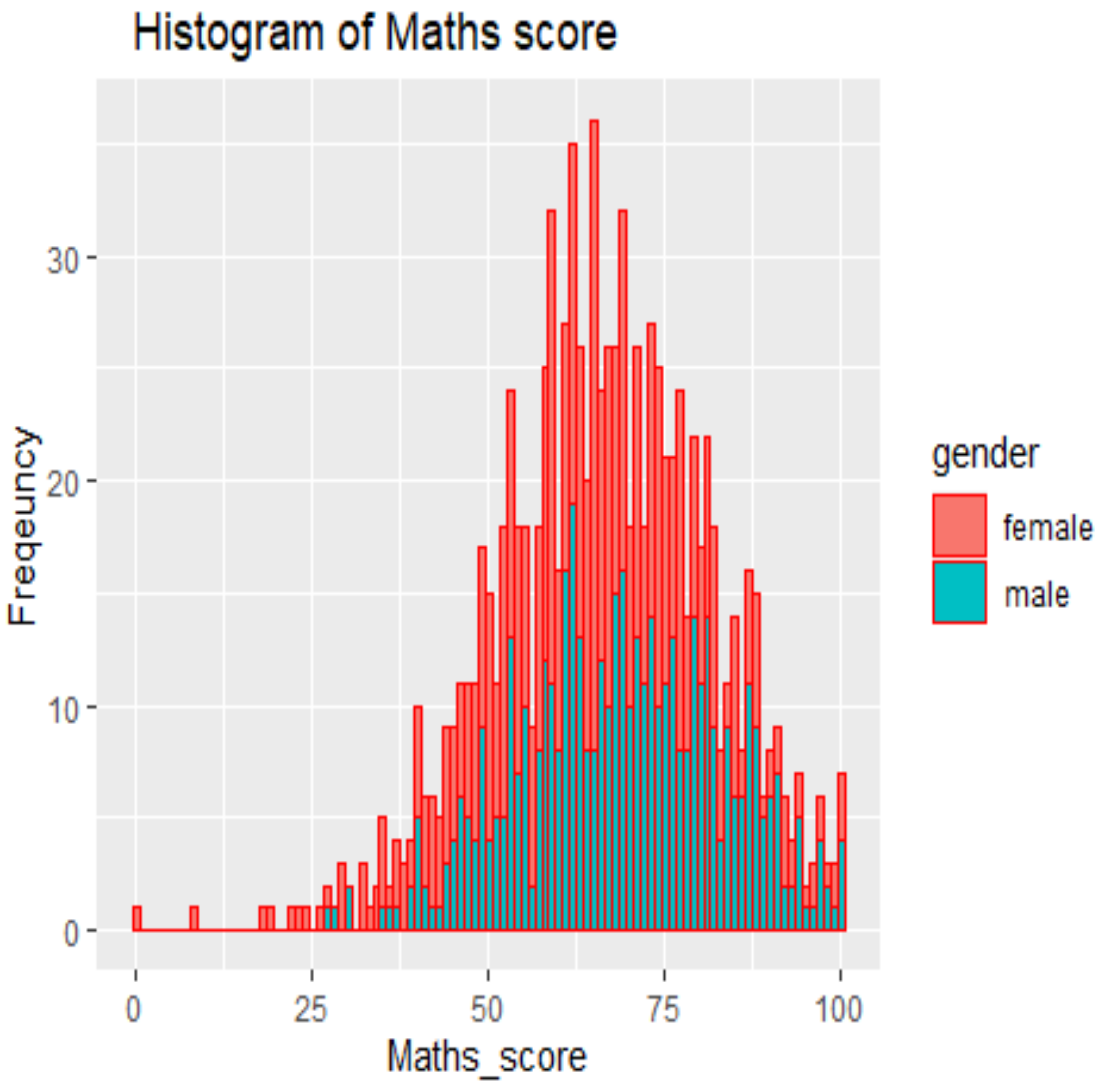
```r
#Data Visualization

library(ggplot2)
#Frequency of Reading score in terms of Gender
ggplot(data=data1,aes(x=reading.score,fill=gender))+
  geom_histogram(col="red",binwidth =1)+ylab("frequency")+
  ggtitle(" Histogram of Reading score")+theme(text=element_text(size=11))+
  xlab("Reading_score")
```




```r
# Frequency of Writing score in terms of Gender
ggplot(data=data1,aes(x=writing.score,fill=gender))+
  geom_histogram(col="red",binwidth =1)+
  ggtitle("   Histogram of Writing score")+
  theme(text=element_text(size=11))+xlab("Wrting score")+
  ylab("freqeuncy")
```



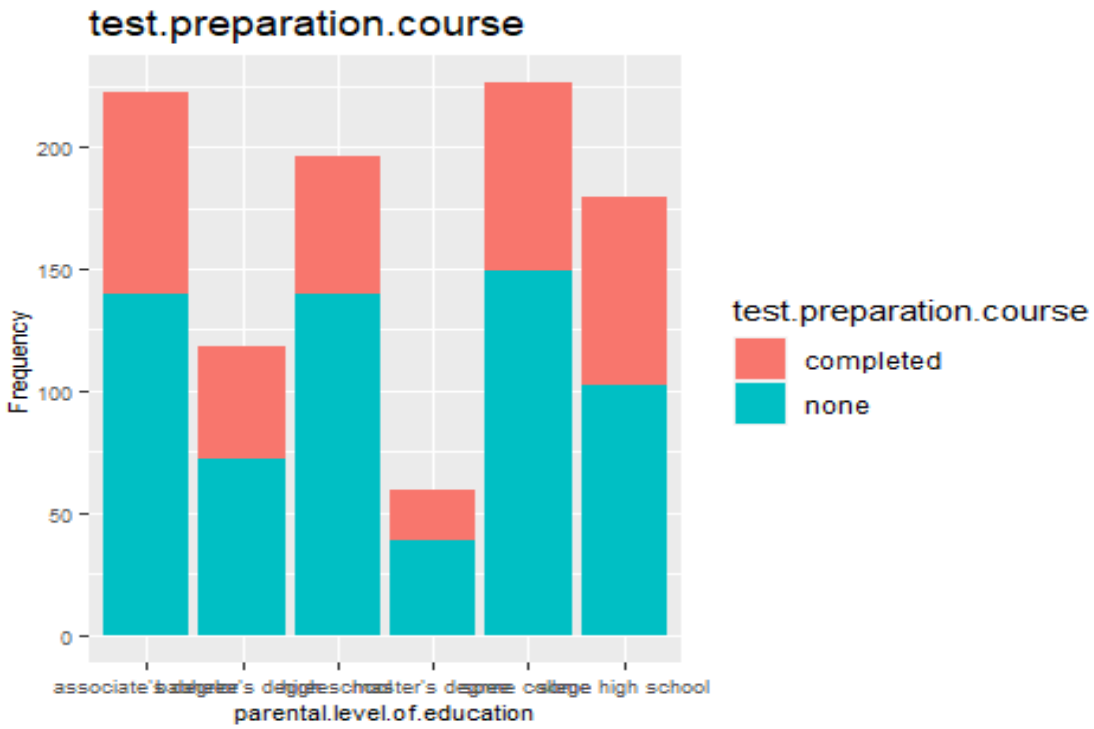

```
#Frequency of Maths Score in terms of Genders
ggplot(data=data1,aes(x=math.score,fill=gender))+
  geom_histogram(col="red",binwidth =1)+ ylab("Freqeuncy")+
  ggtitle("   Histogram of Maths score")+
  theme(text=element_text(size=11))+xlab("Maths_score")
```



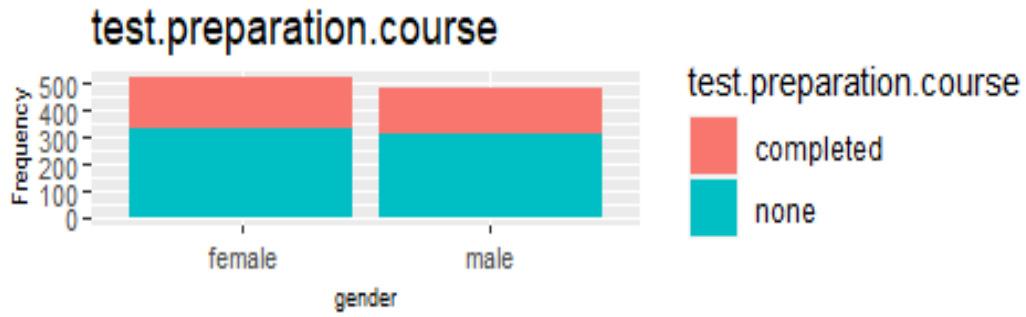

```
# Frequency of Genders Quantity in terms of Test Prep Course
y1=ggplot(data=data1,aes(x=gender,fill=test.preparation.course))+
  geom_bar()+ylab("Frequency")+ggtitle("test.preparation.course")+
  theme(axis.text = element_text(size=8),axis.title = element_text(size=7))
print(y1)
```

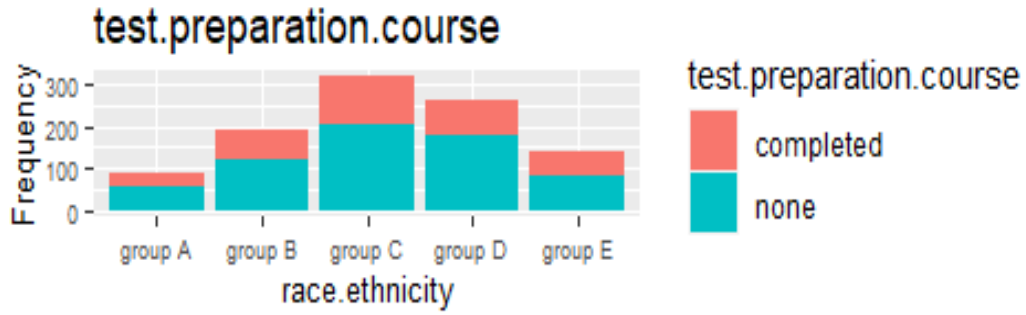

```
# Frequency Race/ethics in terms of Test Prep Course
y2=ggplot(data=data1,aes(x=race.ethnicity,fill=test.preparation.course))+
  geom_bar()+ylab("Frequency")+ggtitle("test.preparation.course")+
  theme(axis.title = element_text(size=10),axis.text = element_text(size=7))
print(y2)
```

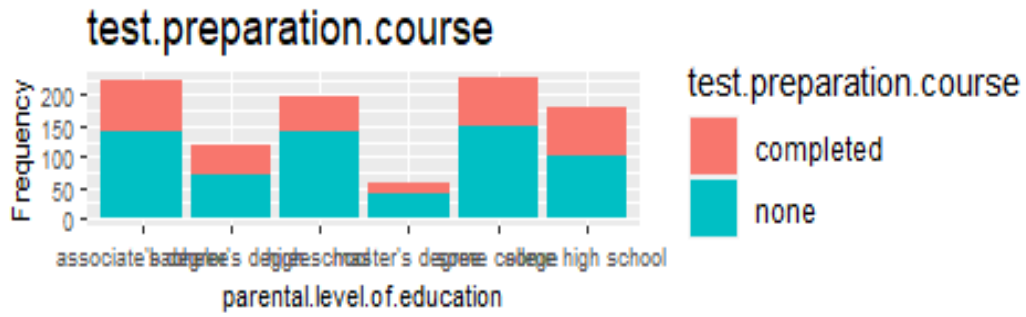

```
# Counting Prep level of education in terms of Test Prep Course
y3=ggplot(data=data1,aes(x=parental.level.of.education,fill=test.preparation.
course))+
  geom_bar()+ggtitle("test.preparation.course")+
  theme(axis.title = element_text(size = 8),axis.text = element_text(size =
7))+
  ylab("Frequency")
print(y3)
```



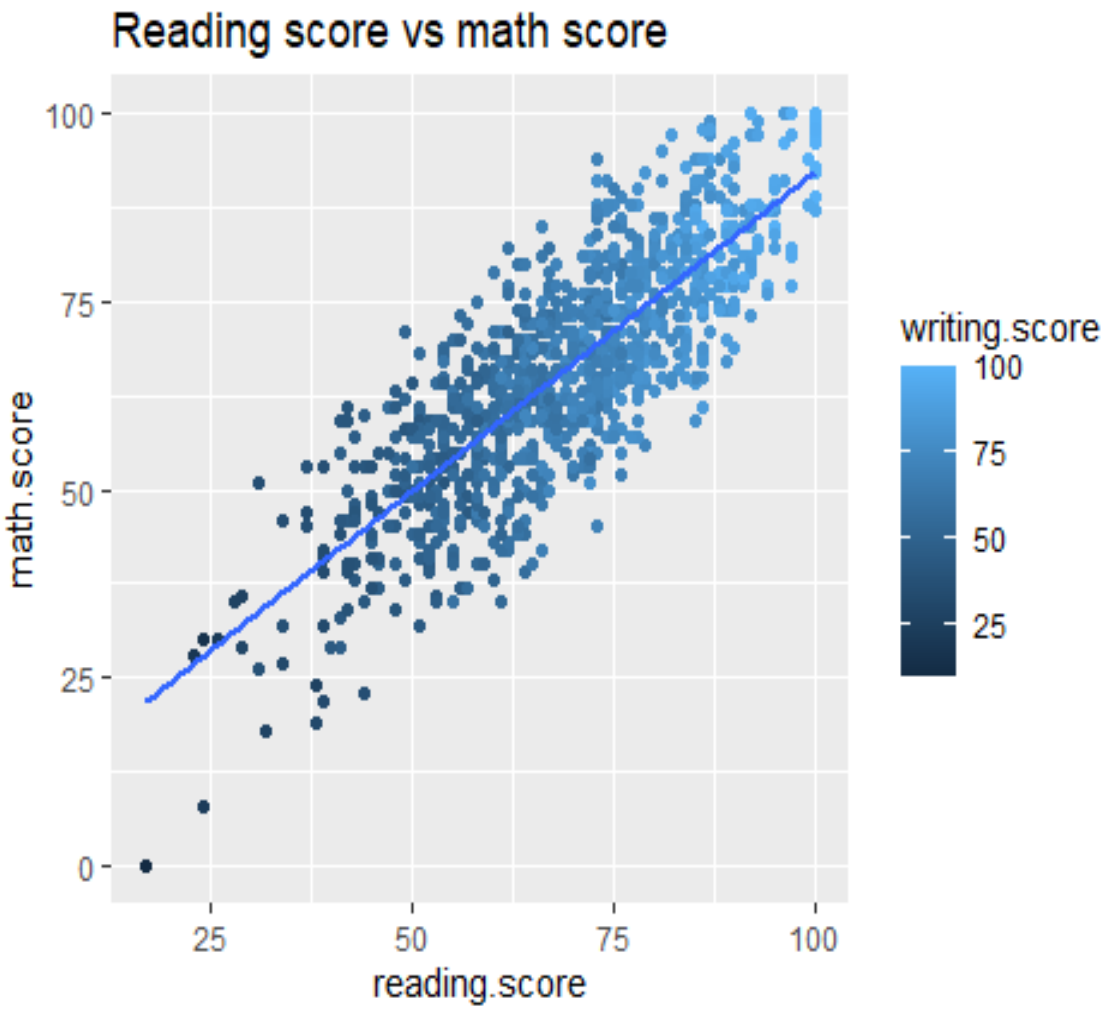

```
library(ggpubr)
ggarrange(y1,y2,y3,ncol=1,nrow=3)
```



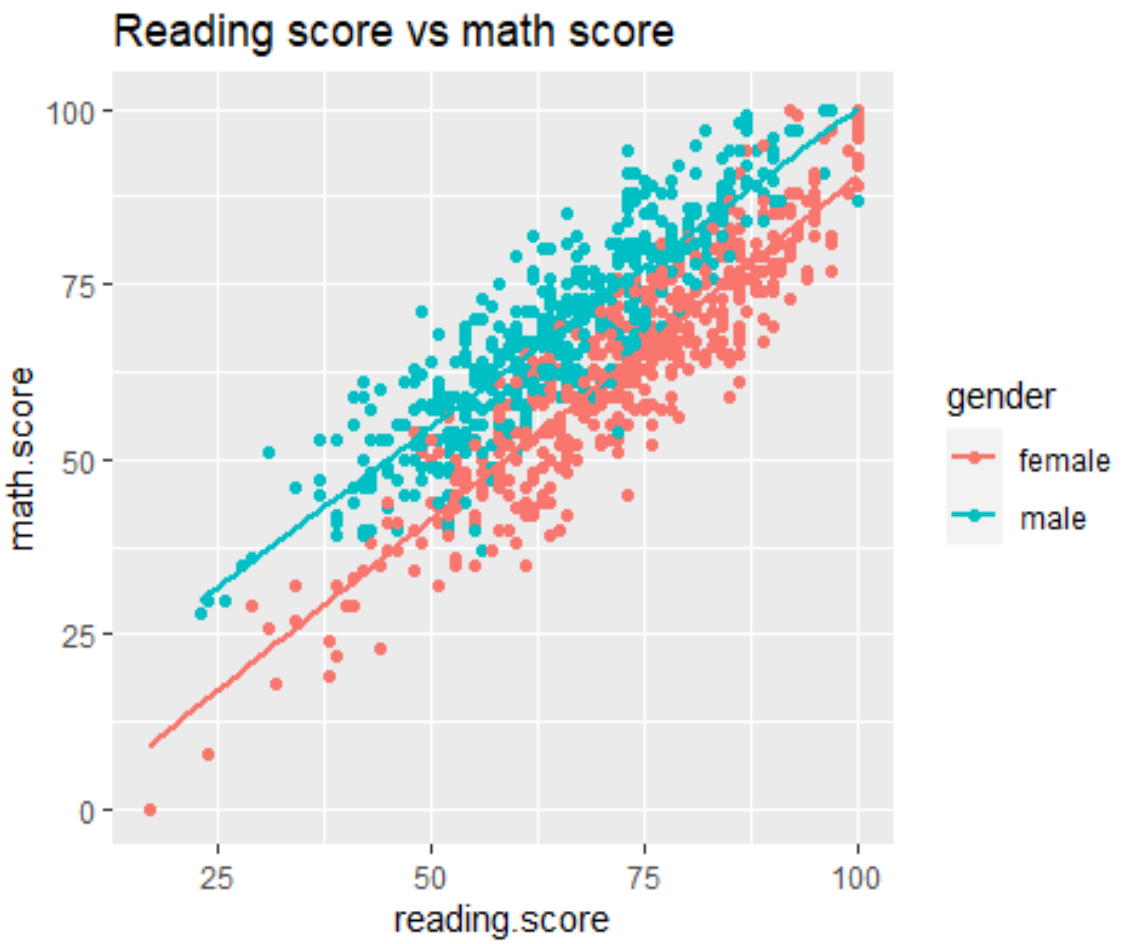




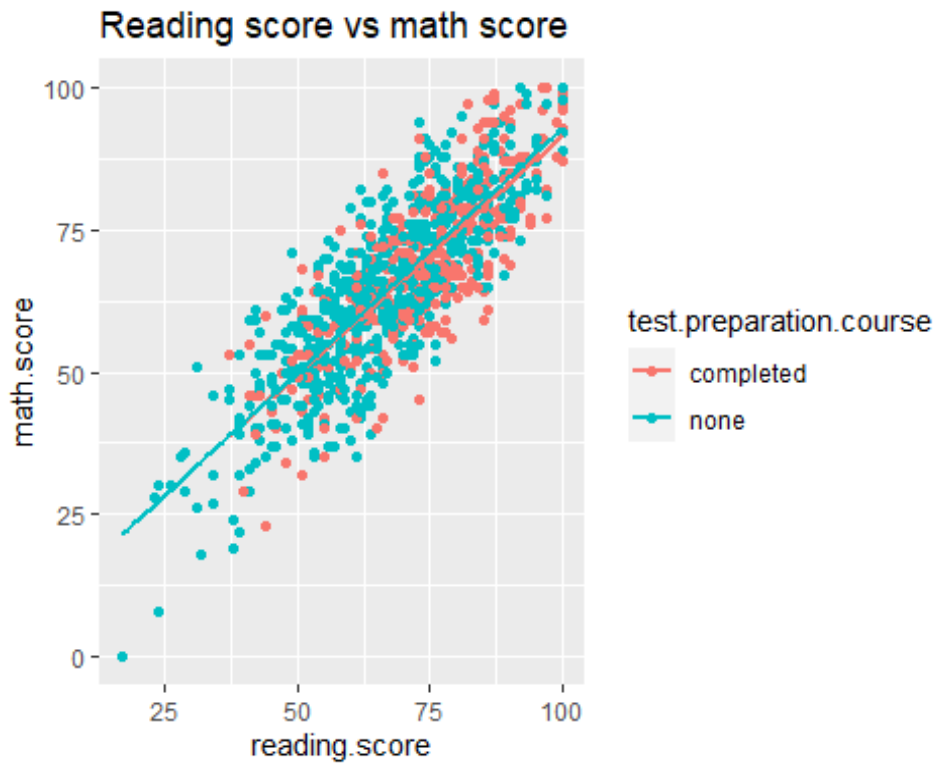

```
# Relationship between Reading Score and Maths Score in terms of Writing
Score
ggplot(data=data1,aes(x=reading.score,y=math.score,col=writing.score))+
  geom_point()+ggtitle("Reading score vs math score")+
  geom_smooth(method = 'lm',se=FALSE)

## `geom_smooth()` using formula 'y ~ x'
```



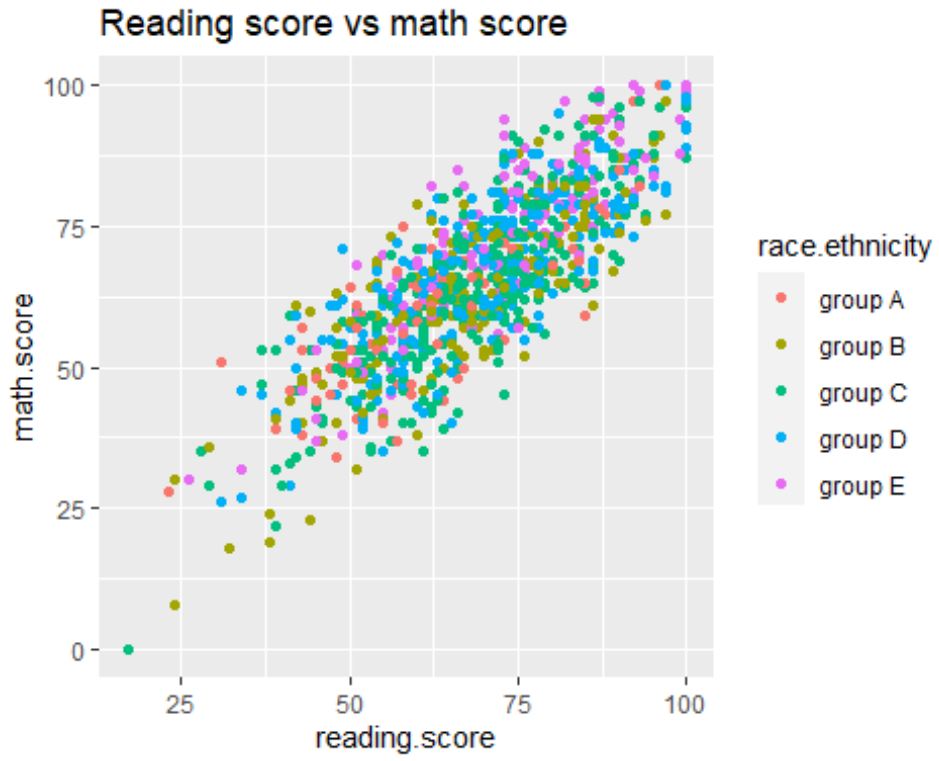

```
# Relationship between Reading Score and Maths Score in terms of Gender
ggplot(data=data1,aes(x=reading.score,y=math.score,col=gender))+geom_point()+
  geom_smooth(method='lm',se=FALSE)+ggtitle("Reading score vs math score")

## `geom_smooth()` using formula 'y ~ x'
```



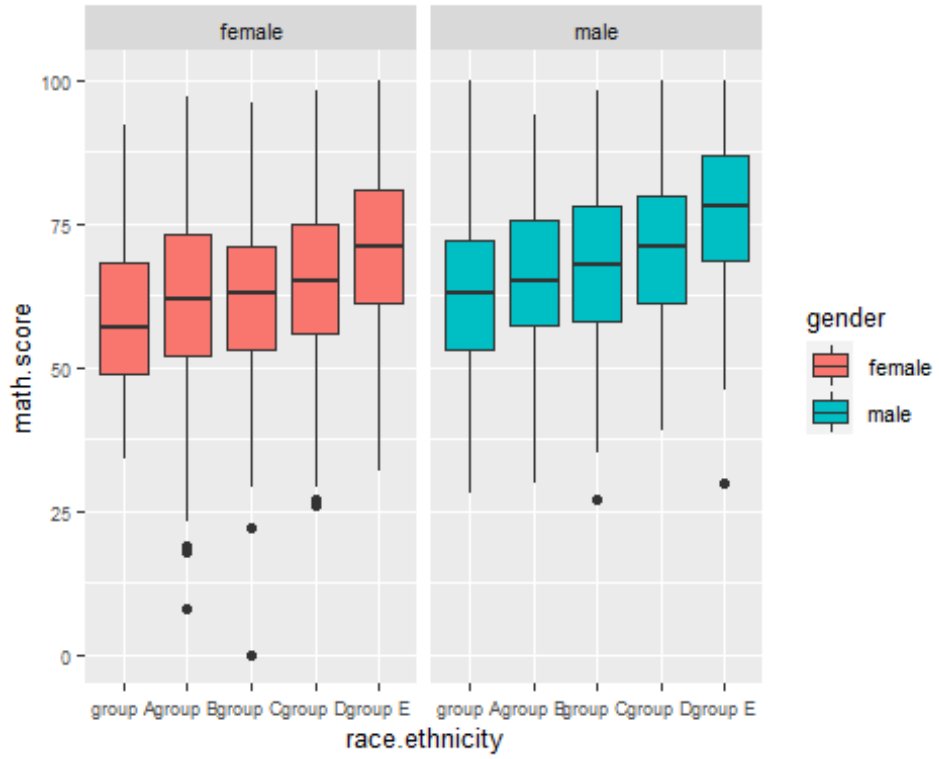

```
# Relationship between Reading Score and Maths Score in terms of Test prep
ggplot(data=data1,aes(x=reading.score,y=math.score,col=test.preparation.cours
e))+
  geom_point()+ ggtitle("Reading score vs math score")+
  geom_smooth(method='lm',se=FALSE)

## `geom_smooth()` using formula 'y ~ x'
```



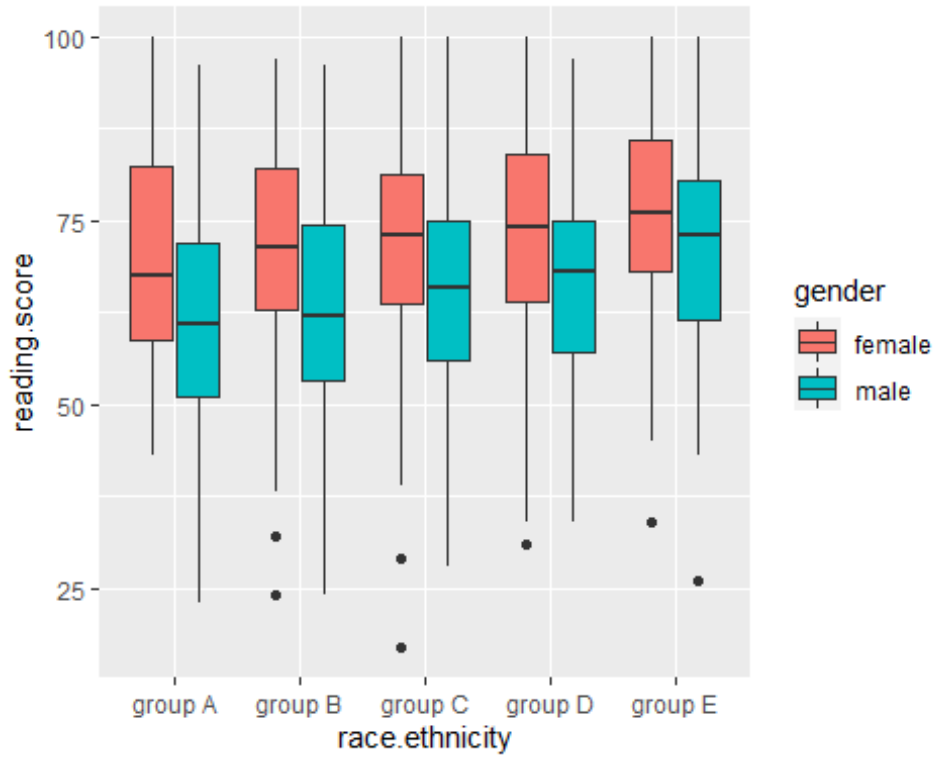

```
# Relationship between Reading Score and Maths Score in terms of Race/Ethics
ggplot(data=data1,aes(x=reading.score,y=math.score,col=race.ethnicity))+
  geom_point()+ ggtitle("Reading score vs math score")
```



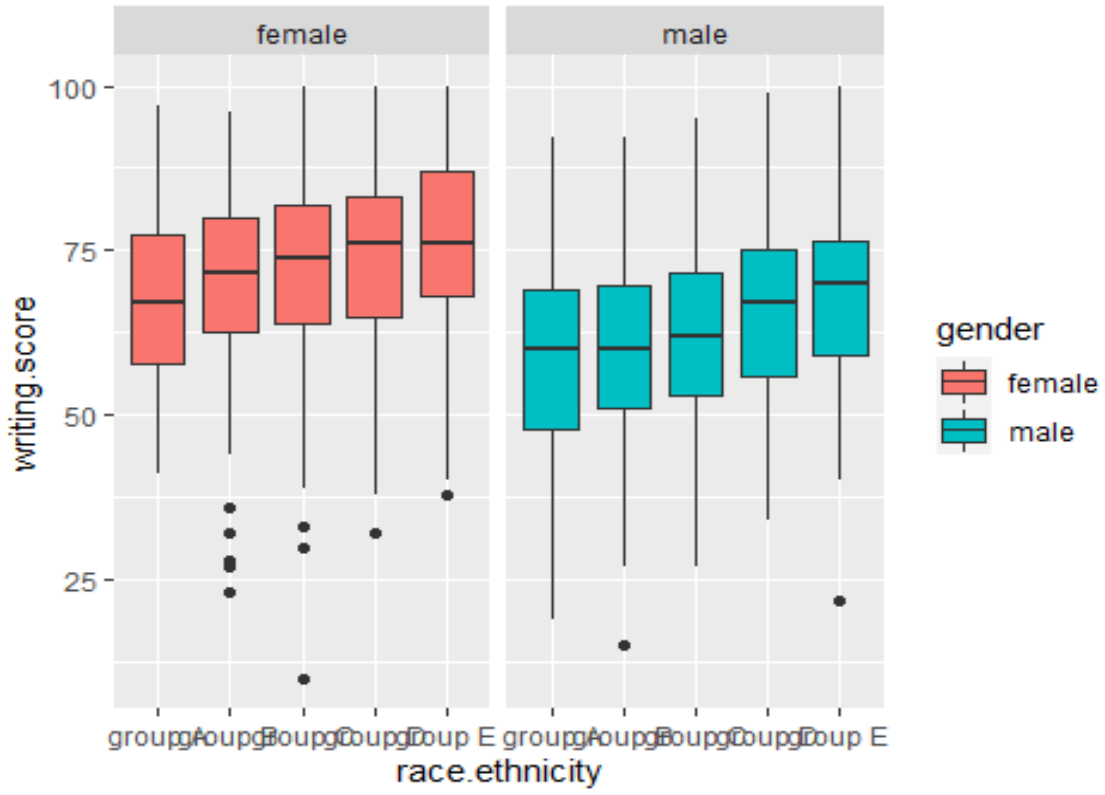
```r
# Correlation between Maths Score and Reading Score
cor.test(data1$math.score,data1$reading.score)

##
##  Pearson's product-moment correlation
##
## data:  data1$math.score and data1$reading.score
## t = 44.855, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7959276 0.8371428
## sample estimates:
##       cor
## 0.8175797

#boxplot
#Summary of Maths score for Race/ethics in terms of Gender
ggplot(data=data1,aes(x=race.ethnicity,y=math.score,fill=gender))+
  geom_boxplot()+facet_grid(~gender)+
  theme(text = element_text(size = 10),axis.text = element_text(size = 7) )
```




```r
#Summary of Reading score for Race/ethics in terms of Gender
#reading score;
ggplot(data=data1,aes(x=race.ethnicity,y=reading.score,fill=gender))+
  geom_boxplot()
```

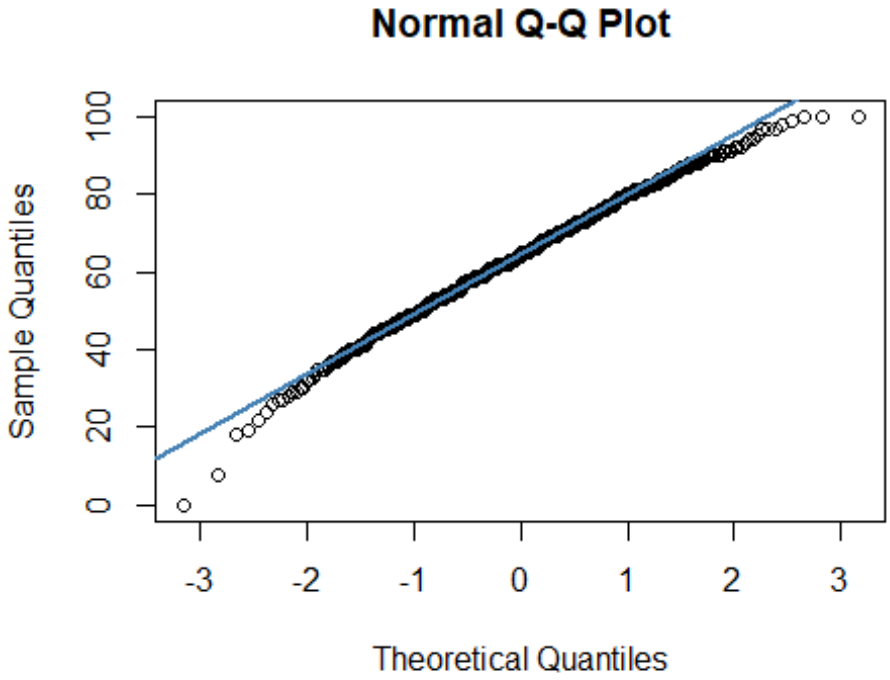



```
# Summary of writing score for Race/ethics in terms of Gender
#writing score;
ggplot(data=data1,aes(x=race.ethnicity,y=writing.score,fill=gender))+
  geom_boxplot()+facet_grid(~gender)
```



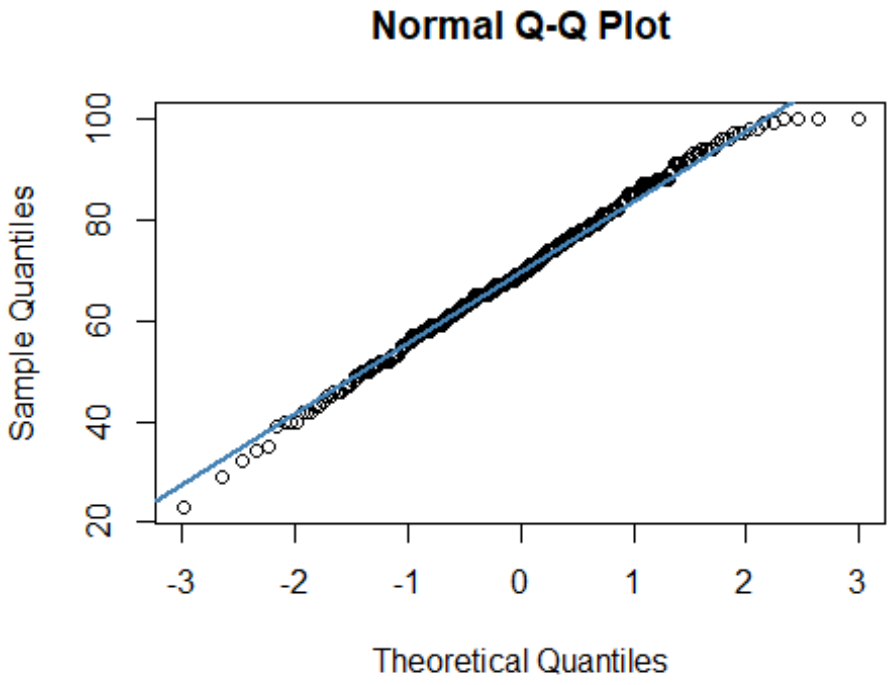

```
# Now applying hypothesis Test:

i)
Is there a difference in Maths mean score among students who's test
preparation is "None" or "complete"

#Null hypothesis: There is no difference in Maths mean score among students
#who's test preparation is "None" or "complete"

#Alternate hypothesis: There is difference in Maths mean score among students
#who's test preparation is "None" or "complete"


library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(moments)
# making Two separate data frames for maths score in terms of test
preparation
a=data1%>%select("test.preparation.course","math.score")%>%
  filter(data1$test.preparation.course=="none")
View(a)
b=data1%>%select("test.preparation.course","math.score")%>%
  filter(data1$test.preparation.course=="completed")
View(b)


#normality test to check Normal distribution;
# i)visualization test for normality for maths score with complete Tests prep
qqnorm(a$math.score)
qqline(a$math.score, col = "steelblue",lwd = 2)
```




```r
# taken alpha =5%

# ii)Statistic test for normality
shapiro.test(a$math.score)

##
##  Shapiro-Wilk normality test
##
## data:  a$math.score
## W = 0.99212, p-value = 0.001754
```

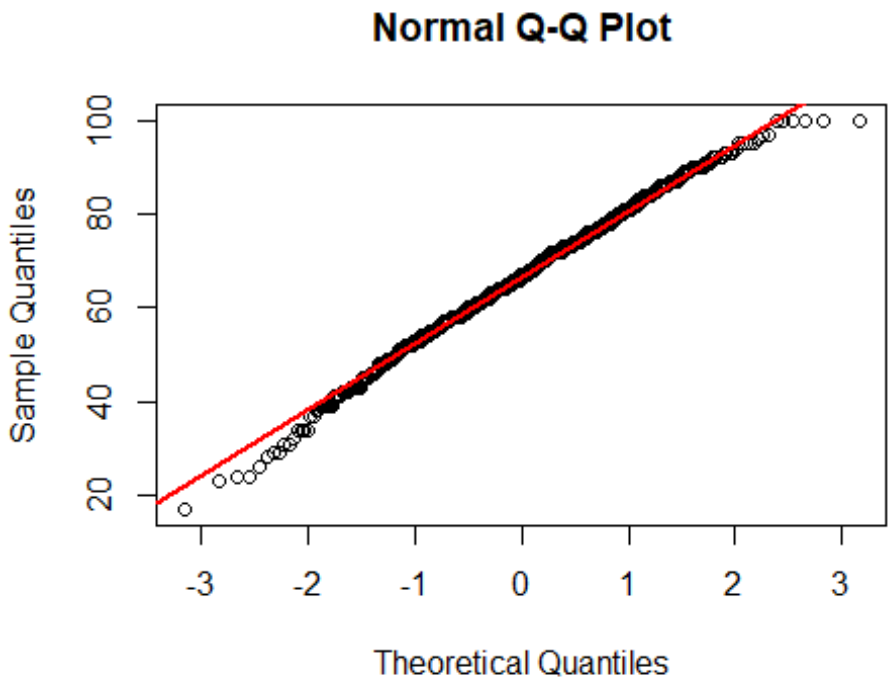
```r
# since P-value < alpha value so it is not normal distribution

agostino.test(a$math.score)

##
##  D'Agostino skewness test
##
## data:  a$math.score
## skew = -0.32796, z = -3.34757, p-value = 0.0008152
## alternative hypothesis: data have a skewness

# since P-value < alpha value so it is not normal distribution

#visualization test for normality for maths score with None Tests prep
qqnorm(b$math.score) # qq-plot
qqline(b$math.score, col = "steelblue",lwd = 2)
```




```r
#Statistic test for normality
shapiro.test(b$math.score)

##
##  Shapiro-Wilk normality test
##
## data:  b$math.score
## W = 0.99366, p-value = 0.1393
```

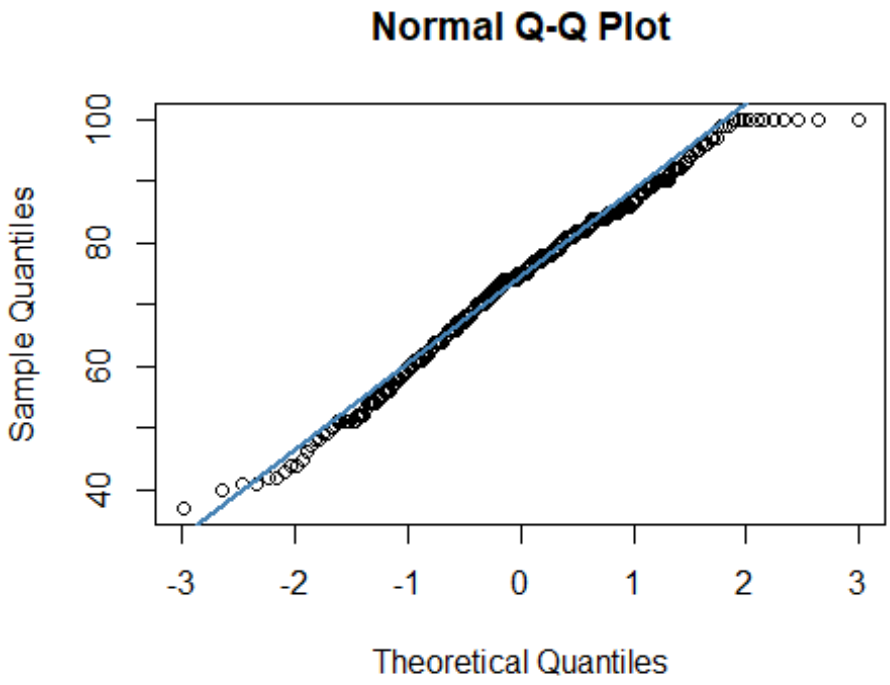
```r
# since P-value > alpha value so it is normal distribution

agostino.test(b$math.score)

##
##  D'Agostino skewness test
##
## data:  b$math.score
## skew = -0.1469, z = -1.1516, p-value = 0.2495
## alternative hypothesis: data have a skewness

# since P-value > alpha value so it is normal distribution

# checking the difference between the two data frames of maths score
wilcox.test(a$math.score,b$math.score)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  a$math.score and b$math.score
## W = 91424, p-value = 8.015e-08
## alternative hypothesis: true location shift is not equal to 0

# hence it is proves that mean of both data frames are different as
#alternate hypothesis is seen as a result







#ii)
Is there a difference in Reading mean score among students who's test
#preparation is "None" or "complete"

Null hypothesis: There is no difference in Reading mean score among students
who's test preparation is "None" or "complete"

Alternate hypothesis: There is difference in Reading mean score among
students
who's test preparation is "None" or "complete"



# making Two separate data frames for maths score in terms of test
preparation
a=data1%>%select("test.preparation.course","reading.score")%>%
  filter(data1$test.preparation.course=="none")
View(a)
b=data1%>%select("test.preparation.course","reading.score")%>%
```

```
  filter(data1$test.preparation.course=="completed")
View(b)


#normality test to check Normal distribution;
# i)visualization test for normality for maths score with complete Tests prep
qqnorm(a$reading.score)
qqline(a$reading.score, col = "red",lwd = 2)
```



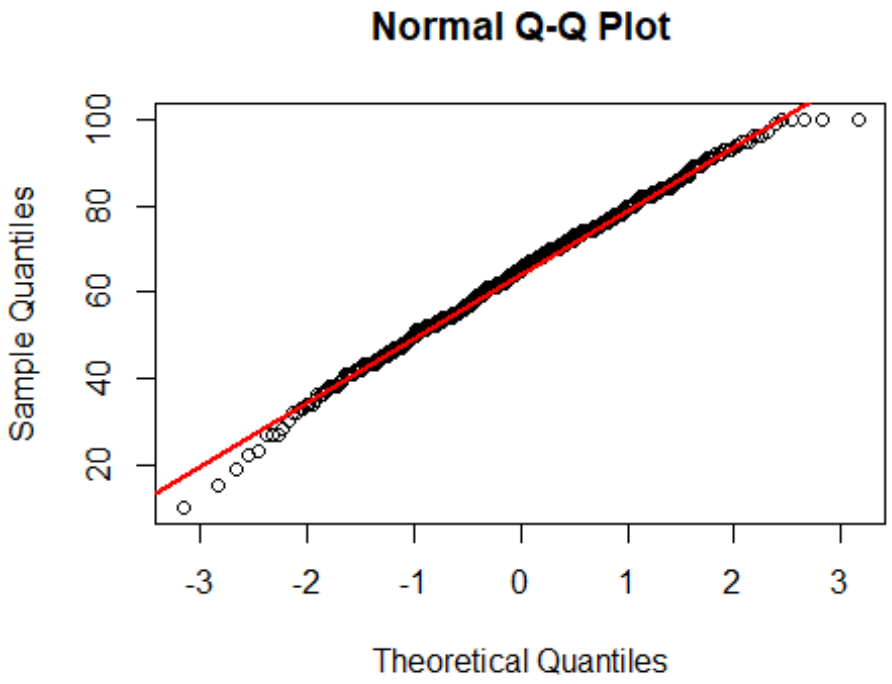

```
# taken alpha =5%

# ii)Statistic test for normality
shapiro.test(a$reading.score)

##
##  Shapiro-Wilk normality test
##
## data:  a$reading.score
## W = 0.99433, p-value = 0.017

# since P-value < alpha value so it is not normal distribution

agostino.test(a$reading.score)

##
##  D'Agostino skewness test
##
```

```
## data:  a$reading.score
## skew = -0.2331, z = -2.4077, p-value = 0.01605
## alternative hypothesis: data have a skewness

# since P-value < alpha value so it is not normal distribution

#visualization test for normality for maths score with None Tests prep
qqnorm(b$reading.score) # qq-plot
qqline(b$reading.score, col = "steelblue",lwd = 2)
```

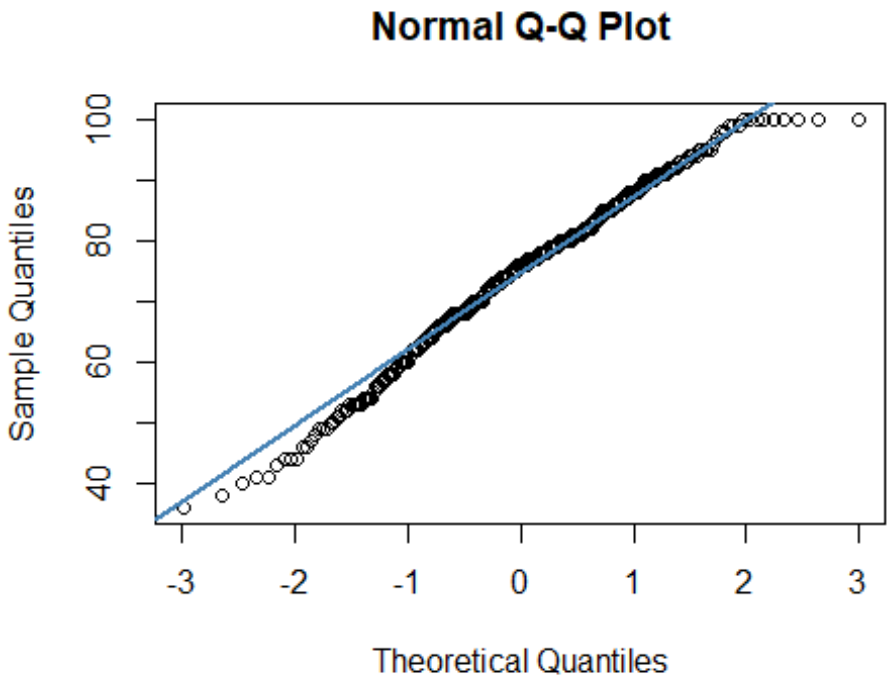



```
#Statistic test for normality
shapiro.test(b$reading.score)

##
##  Shapiro-Wilk normality test
##
## data:  b$reading.score
## W = 0.98563, p-value = 0.001264

# since P-value < alpha value so it is not normal distribution

agostino.test(b$reading.score)

##
##  D'Agostino skewness test
##
## data:  b$reading.score
```

```
## skew = -0.28696, z = -2.21927, p-value = 0.02647
## alternative hypothesis: data have a skewness

# since P-value < alpha value so it is not normal distribution

# checking the difference between the two data frames of maths score
wilcox.test(a$reading.score,b$reading.score)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  a$reading.score and b$reading.score
## W = 81339, p-value = 1.712e-14
## alternative hypothesis: true location shift is not equal to 0

# hence it is proves that mean of both data frames are different as
#alternate hypothesis is seen as a result



#iii)
Is there a difference in Writing mean score among students who's test
preparation is "None" or "complete"

Null hypothesis: There is no difference in writing mean score among students
who's test preparation is "None" or "complete"

Alternate hypothesis: There is difference in writing mean score among
students
who's test preparation is "None" or "complete"



# making Two separate data frames for maths score in terms of test
preparation
a=data1%>%select("test.preparation.course","writing.score")%>%
  filter(data1$test.preparation.course=="none")
View(a)
b=data1%>%select("test.preparation.course","writing.score")%>%
  filter(data1$test.preparation.course=="completed")
View(b)


#normality test to check Normal distribution;
# i)visualization test for normality for maths score with complete Tests prep
qqnorm(a$writing.score)
qqline(a$writing.score, col = "red",lwd = 2)
```




```
# taken alpha =5%

# ii)Statistic test for normality
shapiro.test(a$writing.score)

##
##  Shapiro-Wilk normality test
##
## data:  a$writing.score
## W = 0.99517, p-value = 0.04211

# since P-value < alpha value so it is not normal distribution

agostino.test(a$writing.score)

##
##  D'Agostino skewness test
##
## data:  a$writing.score
## skew = -0.21476, z = -2.22243, p-value = 0.02625
## alternative hypothesis: data have a skewness

# since P-value < alpha value so it is not normal distribution

#visualization test for normality for maths score with None Tests prep
qqnorm(b$writing.score) # qq-plot
qqline(b$writing.score, col = "steelblue",lwd = 2)
```




```
#Statistic test for normality
shapiro.test(b$writing.score)

##
##  Shapiro-Wilk normality test
##
## data:  b$writing.score
## W = 0.98552, p-value = 0.001186

# since P-value < alpha value so it is not normal distribution

agostino.test(b$writing.score)

##
##  D'Agostino skewness test
##
## data:  b$writing.score
## skew = -0.34683, z = -2.66064, p-value = 0.007799
## alternative hypothesis: data have a skewness

# since P-value < alpha value so it is not normal distribution

# checking the difference between the two data frames of maths score
wilcox.test(a$writing.score,b$writing.score)

##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  a$writing.score and b$writing.score
## W = 71027, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0

# hence it is proves that mean of  both data frames are different as
#alternate hypothesis is seen as a result
```