Statistical Programming

Question 3

Student Name: Mai Do

Student ID: 19132761

Word count: 1623

1. Overview

The Student Performance dataset is downloaded from Kaggle Student Performance in Exams (https://www.kaggle.com/spscientist/students-performance-in-exams). It consists of 1000 observations on 8 separate variables. 5 columns are categorical, and 3 columns are numeric variables. The variables include:

- Gender Ethnicity
- Parental level of Education
- Lunch
- Test preparation course
- Math score
- Reading score
- Writing score

The inspiration is to understand the influence of the student's background and determine the features which play an essential role in affecting academic performance. The background has to do with their ethnicity, education level of their parents and the type of lunch they have. Other innate characteristics deemed relevant for the students' performance is the gender, whether they had completed a preparatory course before taking the tests. The students' performance is gauged on their scores obtained in the reading, writing and math tests.

2. The Dataset

```
# Read the dataset
data <- read.csv (file.choose(), header = T)
```

Library

```
library(dplyr)

library(ggplot2)

library(gridExtra)
```

Combining the three scores into the average score as a measure of student performance. Further, we display the overview of data using the first six observations, the new avg.score included. The data contains no missing observations. From the numerical summaries below the reading.scores, writing scores and avg.scores shows left skewness as their means were lesser than the medians while the math.score is slightly right-skewed.

```
# Average score in 3 subjects
avg.score <- rowSums (data[ , 6:8])/3
data <- cbind (data, avg.score)

# Preview of the dataset
head(data)

##   gender race.ethnicity parental.level.of.education      lunch
## 1 female        group B           bachelor's degree    standard
## 2 female        group C                some college     standard
## 3 female        group B            master's degree     standard
## 4   male        group A         associate's degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B         associate's degree     standard
##   test.preparation.course math.score reading.score writing.score avg.score
## 1                    none         72            72            74  72.66667
## 2               completed         69            90            88  82.33333
## 3                    none         90            95            93  92.66667
## 4                    none         47            57            44  49.33333
## 5                    none         76            78            75  76.33333
## 6                    none         71            83            78  77.33333
```

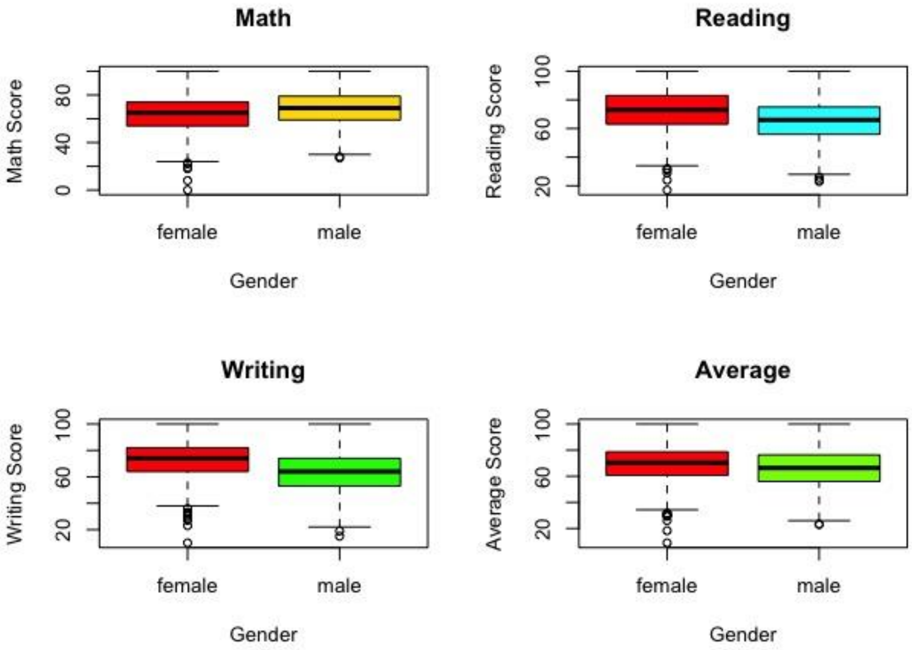
```
summary(data)
```

```
##     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      avg.score
##  Min.   : 10.00   Min.   :  9.00
##  1st Qu.: 57.75   1st Qu.: 58.33
##  Median : 69.00   Median : 68.33
##  Mean   : 68.05   Mean   : 67.77
##  3rd Qu.: 79.00   3rd Qu.: 77.67
##  Max.   :100.00   Max.   :100.00
```

3. Visualisation

The graphical summaries of the data give us the first glimpse of how the variables relate t each other before we proceed to the numerical summaries. The plots give a prior knowledge of the useful elements in our data.

Comparison of Gender attributes to the Marks.



From the plot, it is evident that females tend to have a higher reading and writing score in comparison to males. In contrast, the male has a higher math score than female.

However, we need to use Hypothesis tests are designed to detect whether an "effect" is systematic or is the result of random variation.

```
# Comparing the means of Math by Gender – the unpaired samples t-test
var.test(data$math.score~data$gender)
```

```
##  F test to compare two variances
##
## data:  data$math.score by data$gender
## F = 1.1644, num df = 517, denom df = 481, p-value = 0.09016
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9764071 1.3877941
## sample estimates:
## ratio of variances
##           1.164396
```

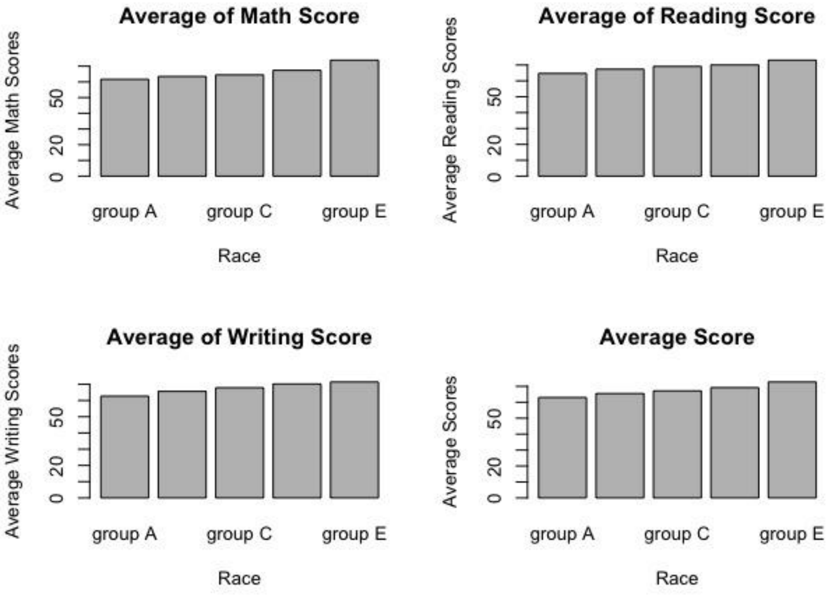
The p-value is greater than 0.05, so do not reject the null hypothesis that the variances are equal. We can thus proceed to perform t-test, assuming the equality of variance.

```
# Perform an unpaired samples t-test using
t.test(data$math.score~data$gender, var.equal=TRUE)
```

```
##  Two Sample t-test
## data:  data$math.score by data$gender
## t = -5.3832, df = 998, p-value = 9.12e-08
## alternative hypothesis: true difference in means 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
```

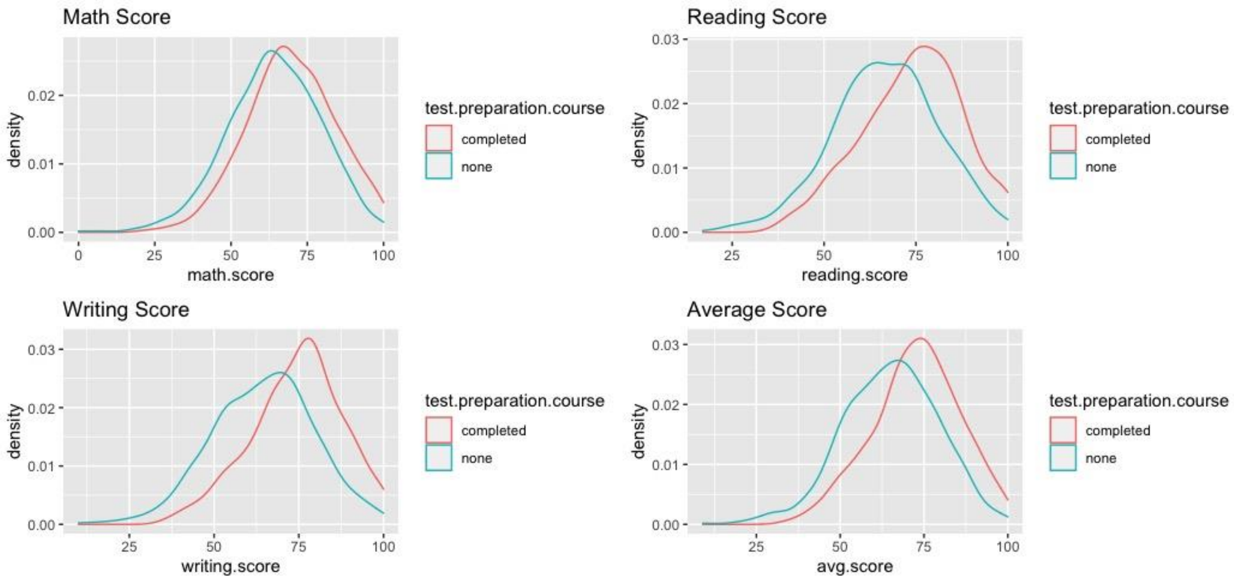
From the t-test, the p-value is (9.12e-08) found to be p-value < 0.001. The p-value is lesser than the significance level(0.05), and the 95% CI of the mean difference does not capture H0 = 0. This indicates that mean male score percent is higher mean female score percent, and we reject the null hypothesis (H0) and test is statistically significant.

Comparison of Race Ethnicity attributes to the Marks.



A boxplot visualisation shows that Group E students scored well compared to all other races, E > D > C > B > A in perspective.

Comparison of Test Preparation attributes to the Marks.



We infer from the above plots that the students who completed the test preparation course had higher scores in all the 3 subjects and the average score compared to the students who had not taken any test preparation course.

Hypothesis Testing

```
# Comparing the means of Reading by Pre-Test Course
var.test(data$reading.score~data$test.preparation.course)

##
##  F test to compare two variances
##
## data:  data$reading.score by data$test.preparation.course
## F = 0.88911, num df = 357, denom df = 641, p-value = 0.2144
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7420038 1.0705553
## sample estimates:
## ratio of variances
##          0.8891108
```

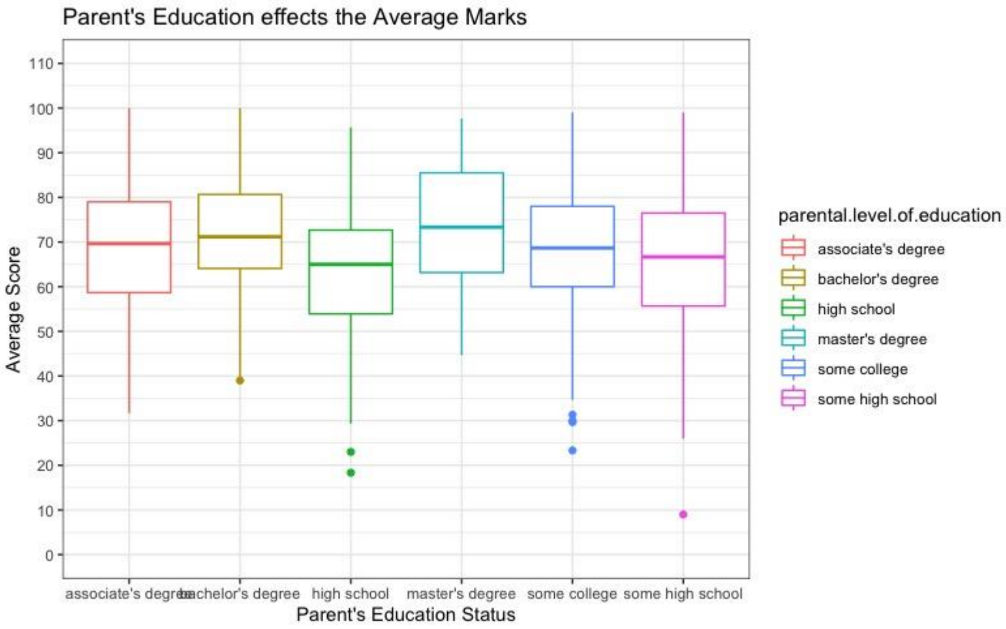
The p-value is greater than 0.05, so do not reject the null hypothesis that the variances are equal. Hence we can proceed to t-test:

```
##
##  Two Sample t-test
##
## data:  data$reading.score by data$test.preparation.course
## t = 7.8717, df = 998, p-value = 9.082e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.524900 9.194274
## sample estimates:
## mean in group completed      mean in group none
##              73.89385             66.53427
```

From the t-test, the p-value is found to be 9.082e-15. The p-value is lesser than the significance level (0.05), and the 95% CI of the mean difference does not capture $H_0 = 0$.

Thus, there is statistical evidence that the mean percentage of students who have taken the pre-test course is higher than the mean percentage of students who have not taken the pre-test course. The test is statistically significant and thus taking a pre-test necessary for improving reading scores.

Comparison of Parent's Education attributes to the Average Marks.



It is easy to see that students who have highly educated parents (masters degree), also have a higher average score.

The F-test below confirms that the (F=10.75 p-value =4.38e-10) above results. The p-value is less than the level of significance 0.05 hence reject the null hypothesis; hence there is a difference in the means of avg.score of students with parents of various learning levels.

```
# ANOVA test to compare parents education and average marks
ANOVA<- aov(data$avg.score ~ data$parental.level.of.education)
summary(ANOVA)
```

```
##                                   Df Sum Sq Mean Sq F value   Pr(>F)
## data$parental.level.of.education   5  10420  2084.1   10.75 4.38e-10 ***
## Residuals                        994 192648   193.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Descriptive Statistics

Summary on Pre-Test Course determining the average score

```
## # A tibble: 2 x 10
##   test.preparation.cou...   Min    Q1 Median    Q3   Max  Mean    SD     n Missing
##   <chr>                   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
## 1 completed                34.3  65     73.5  82.2   100  72.7  13.0   358       0
## 2 none                      9    55.4   65.3  75     100  65.0  14.2   642       0
```

The students who took the pre-test had higher average score (72.7) compared to those who took no pre-test course (65.0) as inferred from the boxplots in the above section.

Summary on Gender determining the average score

```
## # A tibble: 2 x 10
##   gender   Min    Q1 Median    Q3   Max  Mean    SD     n Missing
##   <chr>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
## 1 female     9  60.7   70.3  78.7   100  69.6  14.5   518       0
## 2 male      23  56     66.3  76.2   100  65.8  13.7   482       0
```

Female performed well on the average score (69.6) compared to males (65.8).

Summary on Race Ethnicity determining the average score

```
## # A tibble: 2 x 10
##   race.ethnicity   Min    Q1 Median    Q3   Max  Mean    SD     n Missing
##   <chr>          <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
## 1 group A      23.3   52      61.3     73   96.3    63.0    14.4   89     0
## 2 group B      18.3   56.7     65      76.8 96.7    65.5    14.7   190    0
## 3 group C       9     57.7    68.3     77   98.7    67.1    13.9   319    0
## 4 group D      31     60.3     70      78.6   99    69.2    13.3   262    0
## 5 group E      26     64.7     73.5.   82.4. 100    72.8    14.6   140    0
```

The group E does the best in the average score (72.8) while the group A performs the poorest compared to all the other groups (63).

5. Conclusion

We can thus conclude from the data visualisation, t-tests and ANOVA test that Gender, Race/Ethnicity, Test Preparing and Parental level of Education have significant roles to Student Performance. The females perform averagely higher than males in writing, reading and average performance while men outshine them in math. Group E of the ethnicity is ranked top in performance in all the test and average scores. The analysis also shows that the pre-test course is advantageous for improving student performance. Lastly, the students' performance improves with depending on the level parents learning, the students whose parents hold masters degrees had the highest scores.