Final Project : Student Performance in Exams

MATH60604A - Statistical Modelling

Submitted to : Prof. Juliana Schulz

Presented by:

Zihang Cai - 11319479

Helen Ma - 11313446

Jiahua Shang - 11319456

Danni Weng - 11304053

Introduction

Consider the “exams.csv” data (from https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams (https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams)). The dataset includes test scores from 1000 kids in reading, writing, and math, as well as some additional data about the students. The following variables are included in the file:

Variable	Description
gender	Female or Male
race.ethnicity	Race/Ethnicity group A, B, C, D, E
parental.level.of.education	Parents’ education level: some high school, some college, high school, associate’s degree, bachelor’s degree, master’s degree
lunch	Student’s meal plan: reduced/free or standard
test.preparation.course	Indicator if the student took the preparation course prior to the test (completed or none)
math.score	Test score in math
reading.score	Test score in reading
writing.score	Test score in writing

```
##     gender    race.ethnicity     parental.level.of.education
##  female:483   group A: 79    associate's degree:203
##  male  :517   group B:205    bachelor's degree :112
##               group C:323    high school       :202
##               group D:262    master's degree   : 70
##               group E:131    some college      :222
##                              some high school  :191
##           lunch     test.preparation.course   math.score    reading.score
##  free/reduced:348   completed:335           Min.   : 13.0   Min.   : 27
##  standard    :652   none     :665           1st Qu.: 56.0   1st Qu.: 60
##                                             Median : 66.5   Median : 70
##                                             Mean   : 66.4   Mean   : 69
##                                             3rd Qu.: 77.0   3rd Qu.: 79
##                                             Max.   :100.0   Max.   :100
##  writing.score
##  Min.   : 23.00
##  1st Qu.: 58.00
##  Median : 68.00
##  Mean   : 67.74
##  3rd Qu.: 79.00
##  Max.   :100.00
```

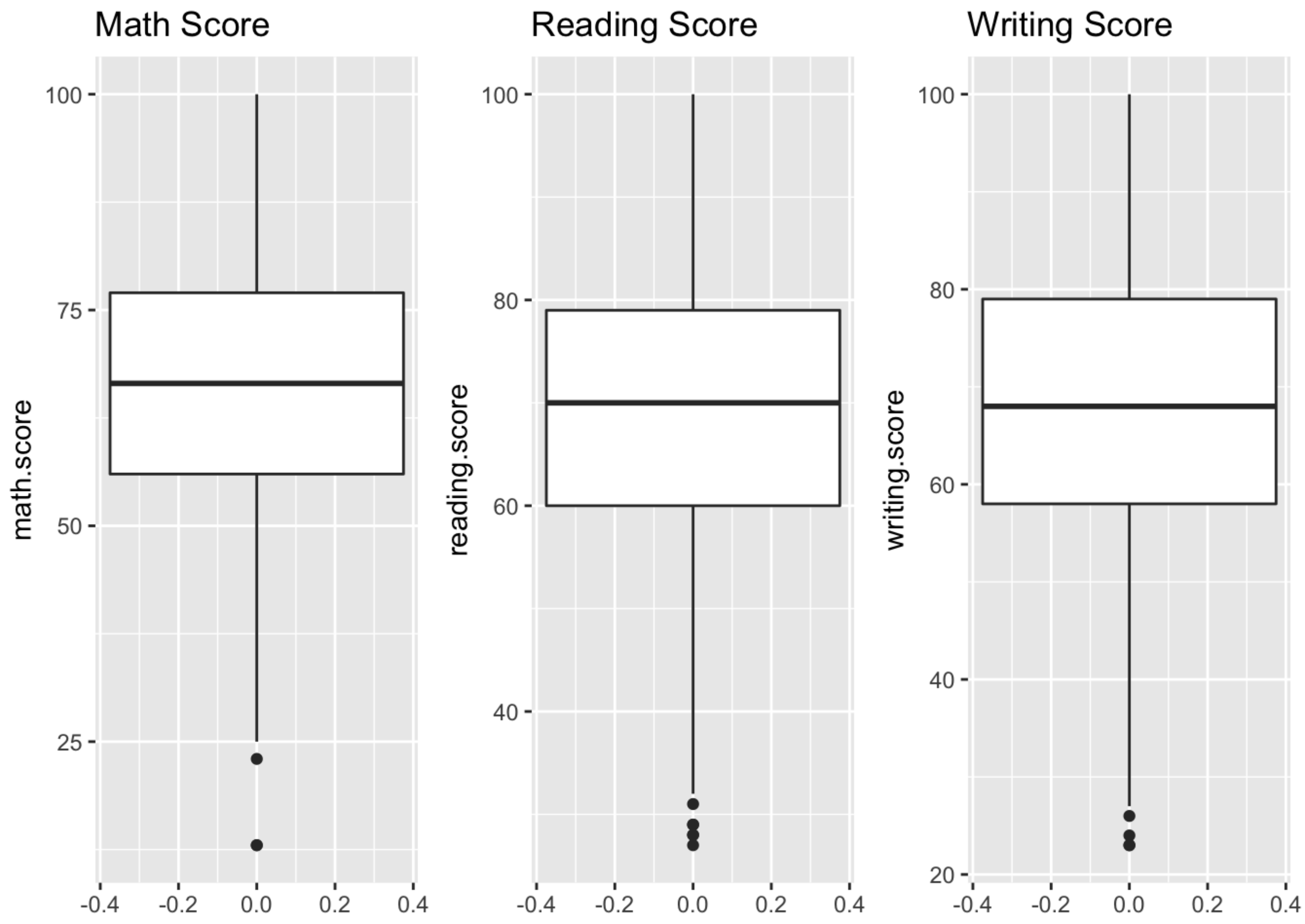
Data Exploration

From the summary above, the dataset is fairly structured and does not contain any NULL rows at a first glance. We will now explore each variable more in details.

Score

From the summary, we can see that the data contains three numeric variables of scores for each subject : `math.score` , `reading.score` and `writing.score` . All variables are capped at a maximum score of 100% and this seem coherent. The mean score in math, reading and writing are 66.4%, 69% and 67.74% respectively and for the purpose of this analysis, we will be creating a new variable `avg_score` , which is the average of the three score of three scores as the response variable.

```
##    math.score    reading.score writing.score
##  Min.   : 13.0   Min.   : 27   Min.   : 23.00
##  1st Qu.: 56.0   1st Qu.: 60   1st Qu.: 58.00
##  Median : 66.5   Median : 70   Median : 68.00
##  Mean   : 66.4   Mean   : 69   Mean   : 67.74
##  3rd Qu.: 77.0   3rd Qu.: 79   3rd Qu.: 79.00
##  Max.   :100.0   Max.   :100   Max.   :100.00
```

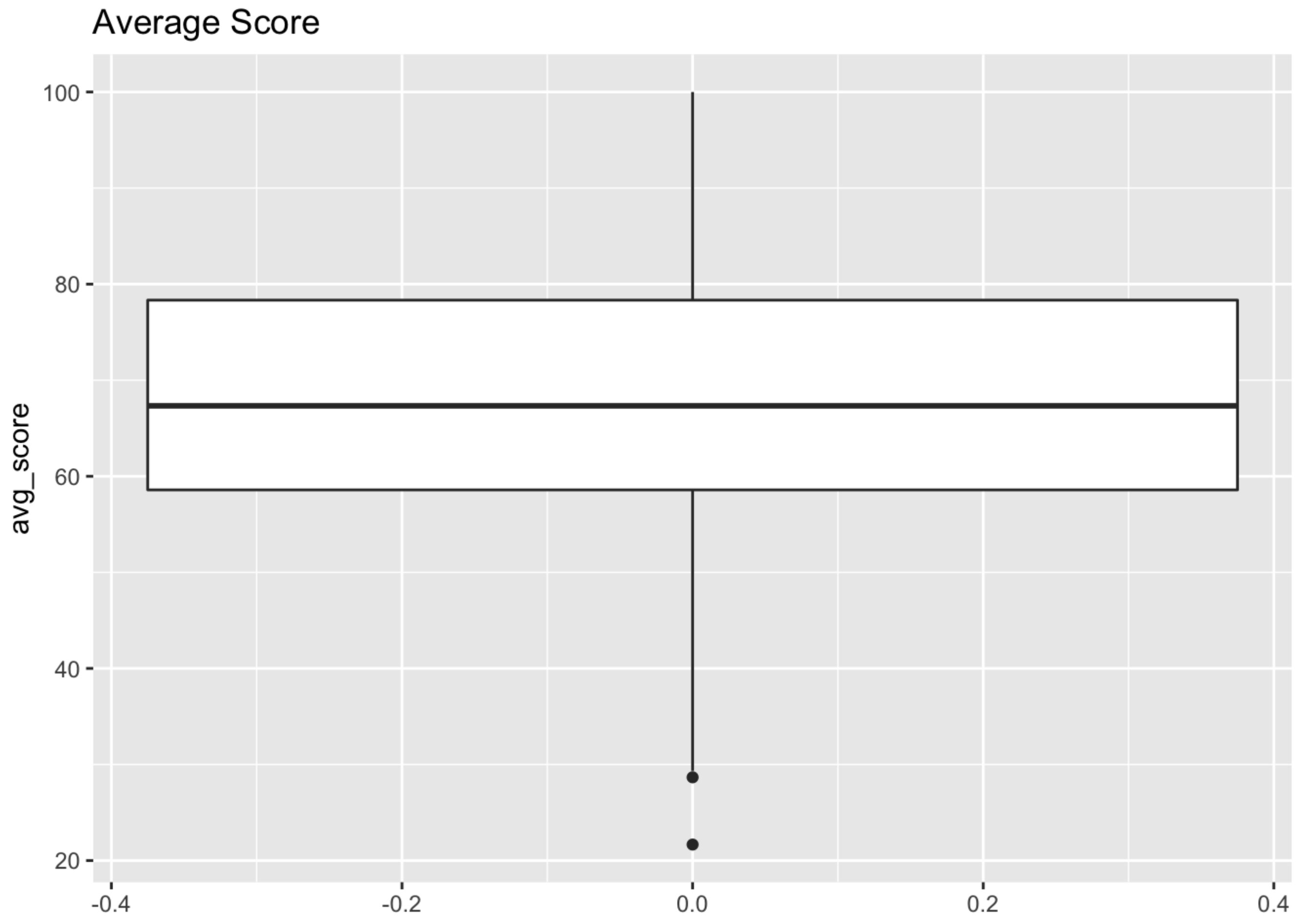


The code to create the new variable `avg_score` is shown below:

```
data<-raw_data[,1:5]
#create new variable Avg Score
data$avg_score<-(raw_data$math.score+raw_data$reading.score+raw_data$writing.score)/3
```

The variable `avg_score` ranges from a minimum score of 21.67% to maximum score of 100%. From boxplot, the data points seem to be distributed nicely in this range as the median is fairly close to the mean value. Moreover, the upper quartile at 58.6% and lower quartile at 78.3% seem to be roughly symmetric distance from the median, which seems like a plausible distribution for students tests scores. There are a few points that may be further away from the center (scores of 20%-30%), but they do not seem too problematic.

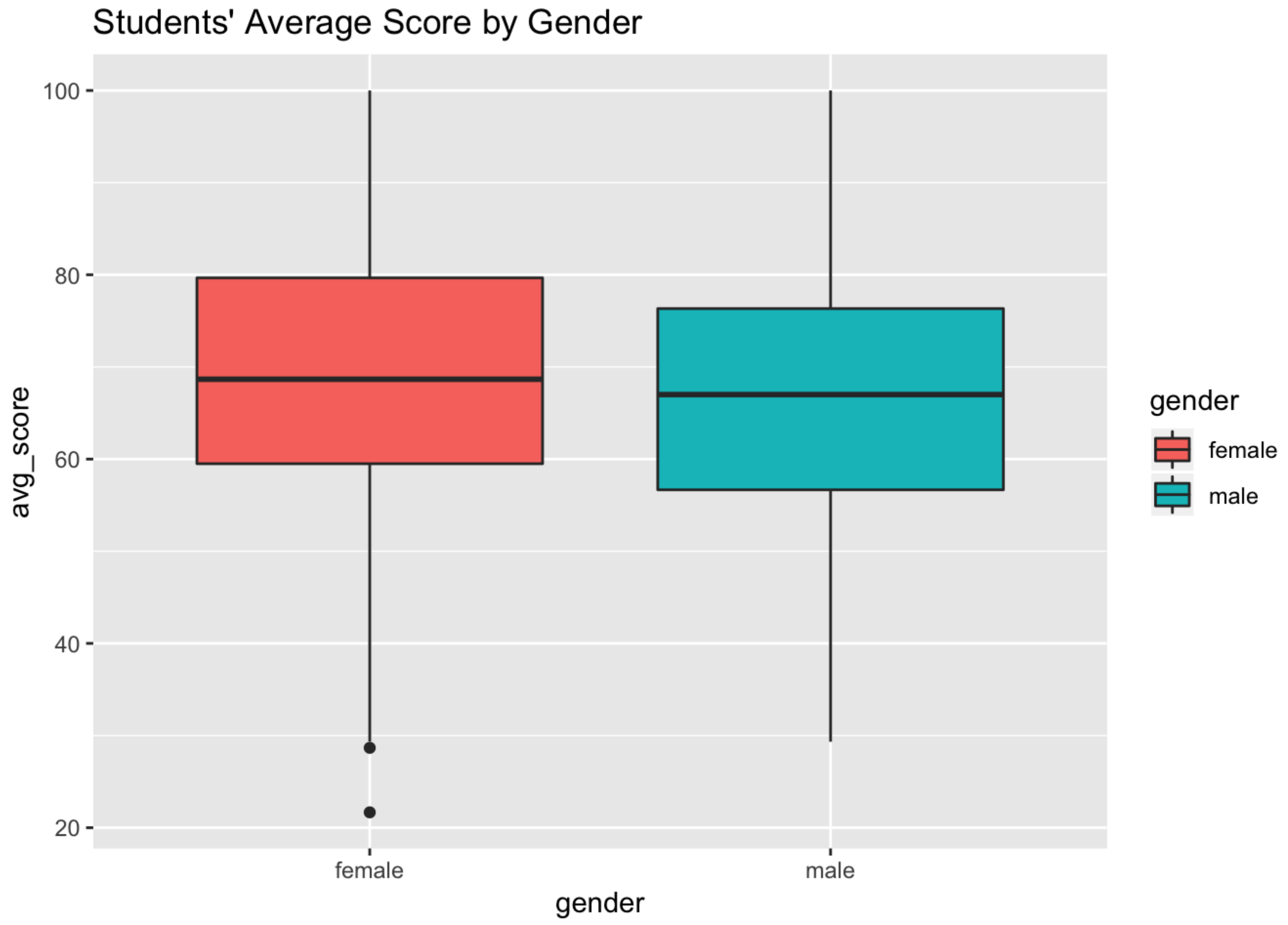
```
##     gender    race.ethnicity     parental.level.of.education
##  female:483   group A: 79    associate's degree:203        
##  male  :517   group B:205    bachelor's degree :112        
##               group C:323    high school       :202        
##               group D:262    master's degree   : 70        
##               group E:131    some college      :222        
##                              some high school  :191        
##           lunch     test.preparation.course   avg_score     
##  free/reduced:348   completed:335           Min.   : 21.67  
##  standard    :652   none     :665           1st Qu.: 58.58  
##                                             Median : 67.33  
##                                             Mean   : 67.71  
##                                             3rd Qu.: 78.33  
##                                             Max.   :100.00  
```



Gender

The variable `Gender` is a categorical variable with two levels : Female or Male. There are 483 observations of female students and 517 observations of male students in the dataset. This distribution seem to be balanced between both genders. When analyzing the average scores by gender, we can observe that the score distribution seem slightly higher for female students, but the differences are slight. Moreover, the lowest scores noticed earlier seem to belong to female students.

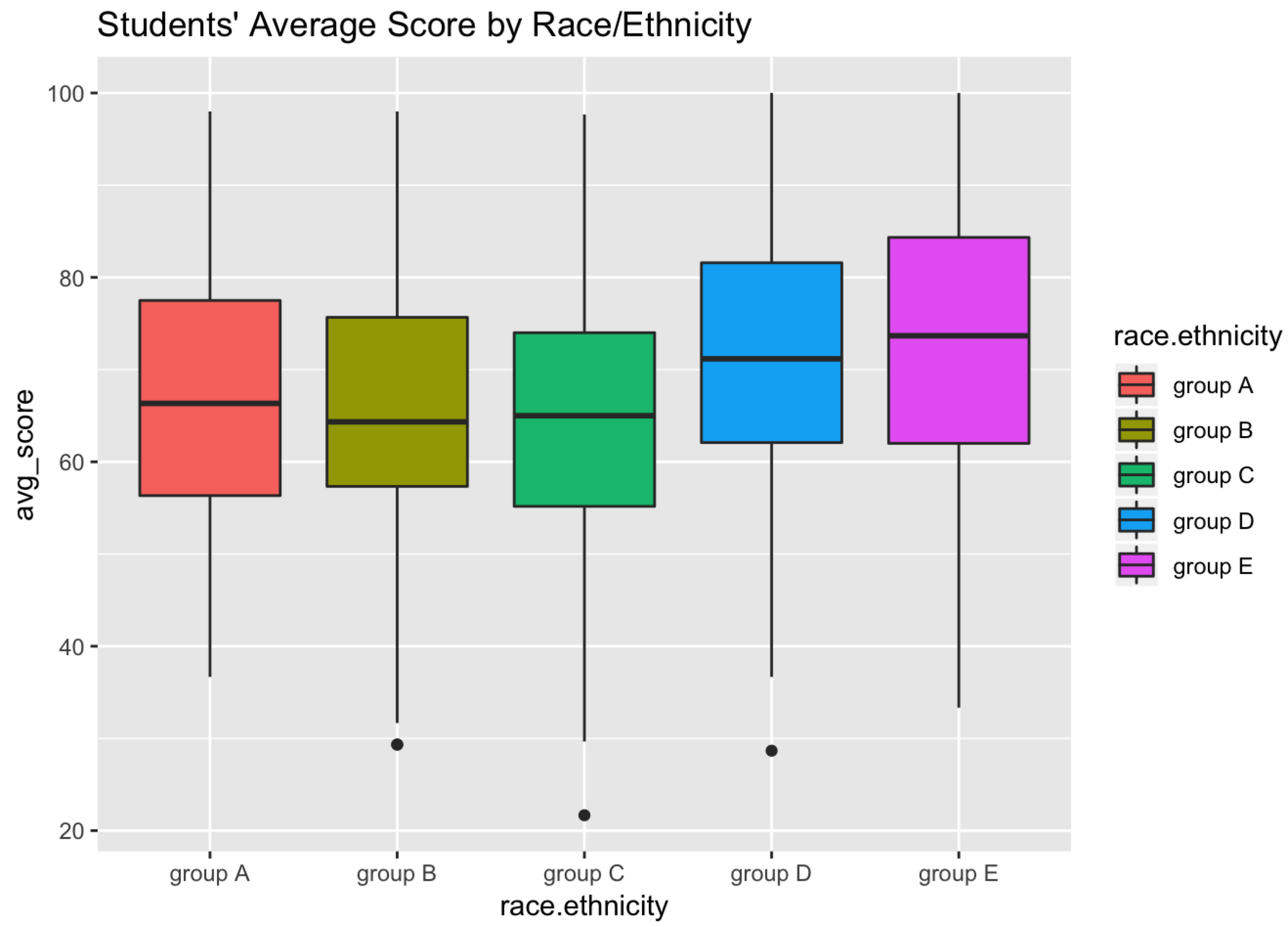
```
##        frequencies percentage cumulativepercentage
## female         483       48.3                 48.3
## male           517       51.7                100.0
## Totals        1000      100.0                100.0
```



Race/Ethnicity

The variable `Race/Ethnicity` is a categorical variable grouped into five levels : group A, B, C, D and E. From the frequency table, we can observe that there's an uneven distribution of observations among ethnicity group: Group C represents the largest proportion at 32.3% and group C only represents 7.9% of total observations. Given that we know the actual distribution of students population data, it could be possible that this dataset has an under-representation of Group A and over-representation of group C, which we have to keep in mind when drawing any conclusions or interpretations later.

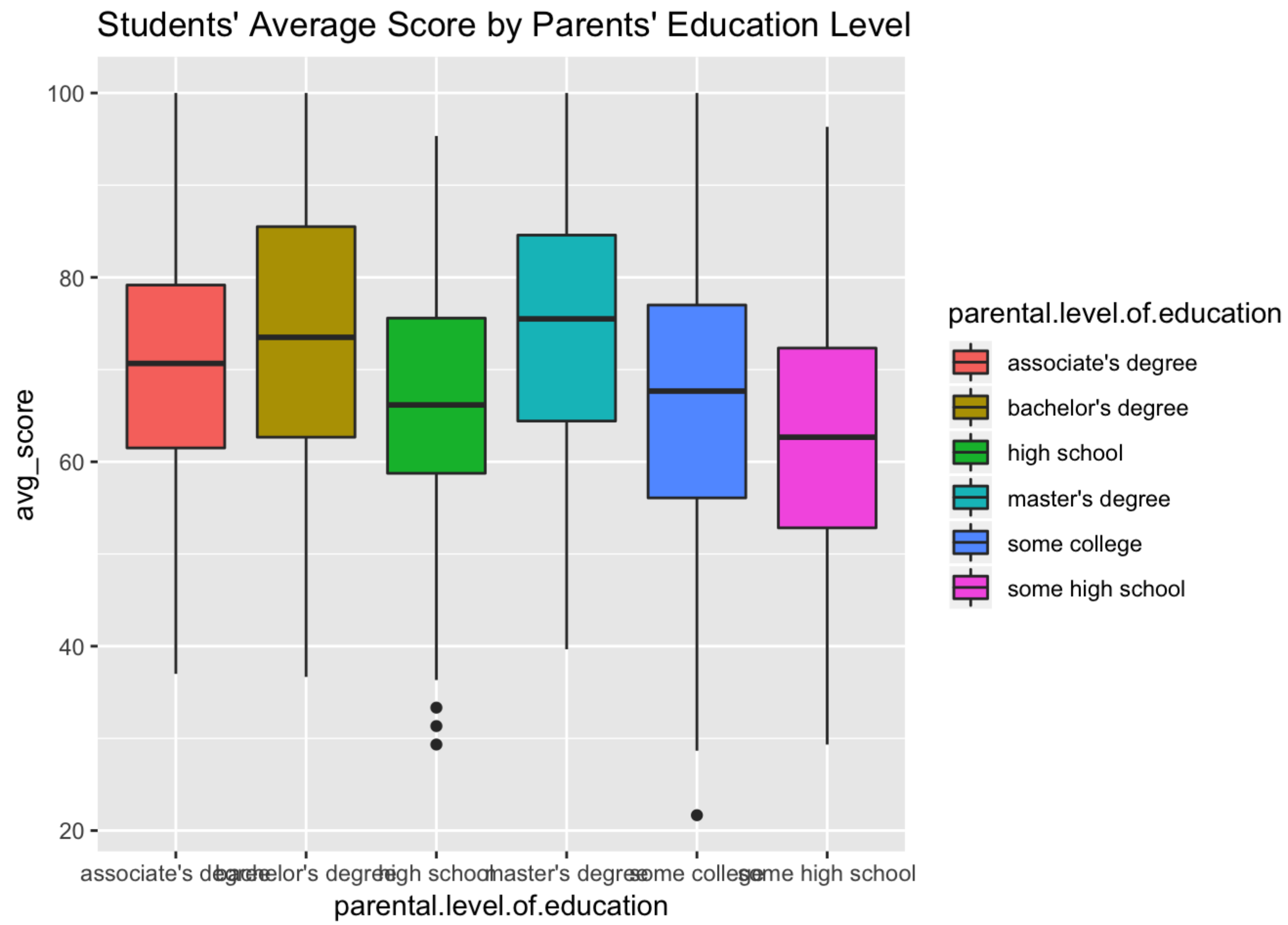
```
##         frequencies percentage cumulativepercentage
## group A          79        7.9                  7.9
## group B         205       20.5                 28.4
## group C         323       32.3                 60.7
## group D         262       26.2                 86.9
## group E         131       13.1                100.0
## Totals         1000      100.0                100.0
```



Parent’s education level

The variable `Parental.level.of.education` is a categorical variable grouped into six levels : some high school, high school, some college, associate’s degree, bachelor’s degree and master’s degree. From the frequency table, we can observe that the distributions of observations are roughly similar for each group, except for the higher levels of education (bachelor and master’s degree). Given that this variable has somewhat a ordinal structure (for example, a high school diploma is lower education level than a college degree, which is lower than a master degree), we decide to group adjacent levels of educations levels together and condensed them into three levels.

```
##                      frequencies percentage cumulativepercentage
## associate's degree           203       20.3                 20.3
## bachelor's degree            112       11.2                 31.5
## high school                  202       20.2                 51.7
## master's degree               70        7.0                 58.7
## some college                 222       22.2                 80.9
## some high school             191       19.1                100.0
## Totals                      1000      100.0                100.0
```

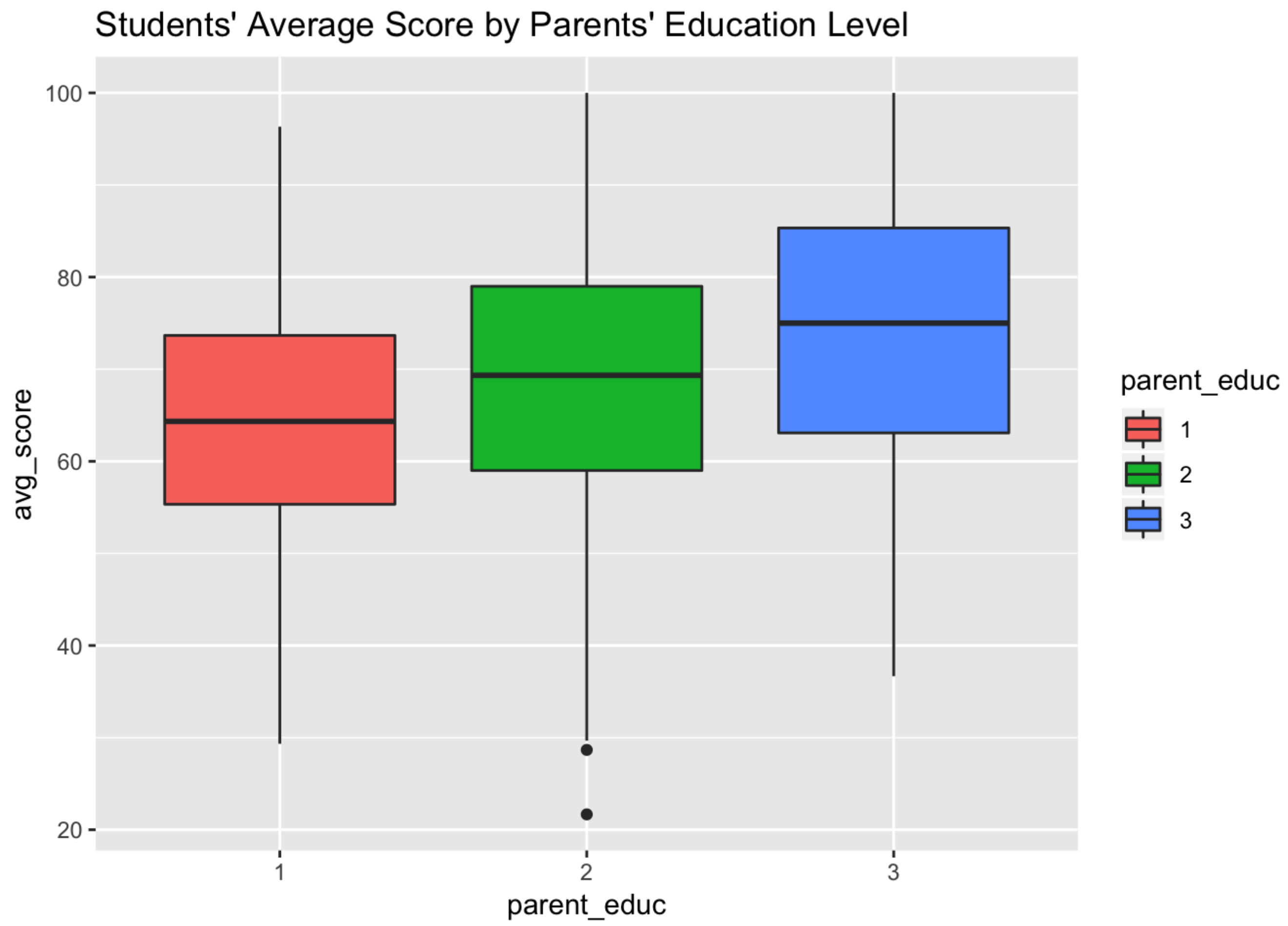


Therefore, a new variable `parent_educ` will be created with three levels:
1: high school and some high school
2: some college and associate’s degree
3: bachelor’s and master’s degree

```
data$parent_educ <- data$parental.level.of.education
levels(data$parent_educ) <- list("1" = c("high school", "some high school"),
                                  "2" = c("some college", "associate's degree"),
                                  "3" = c("bachelor's degree", "master's degree"))
data$parent_educ<-relevel(as.factor(data$parent_educ),ref="1")
```

The new categories are distributed as follows:

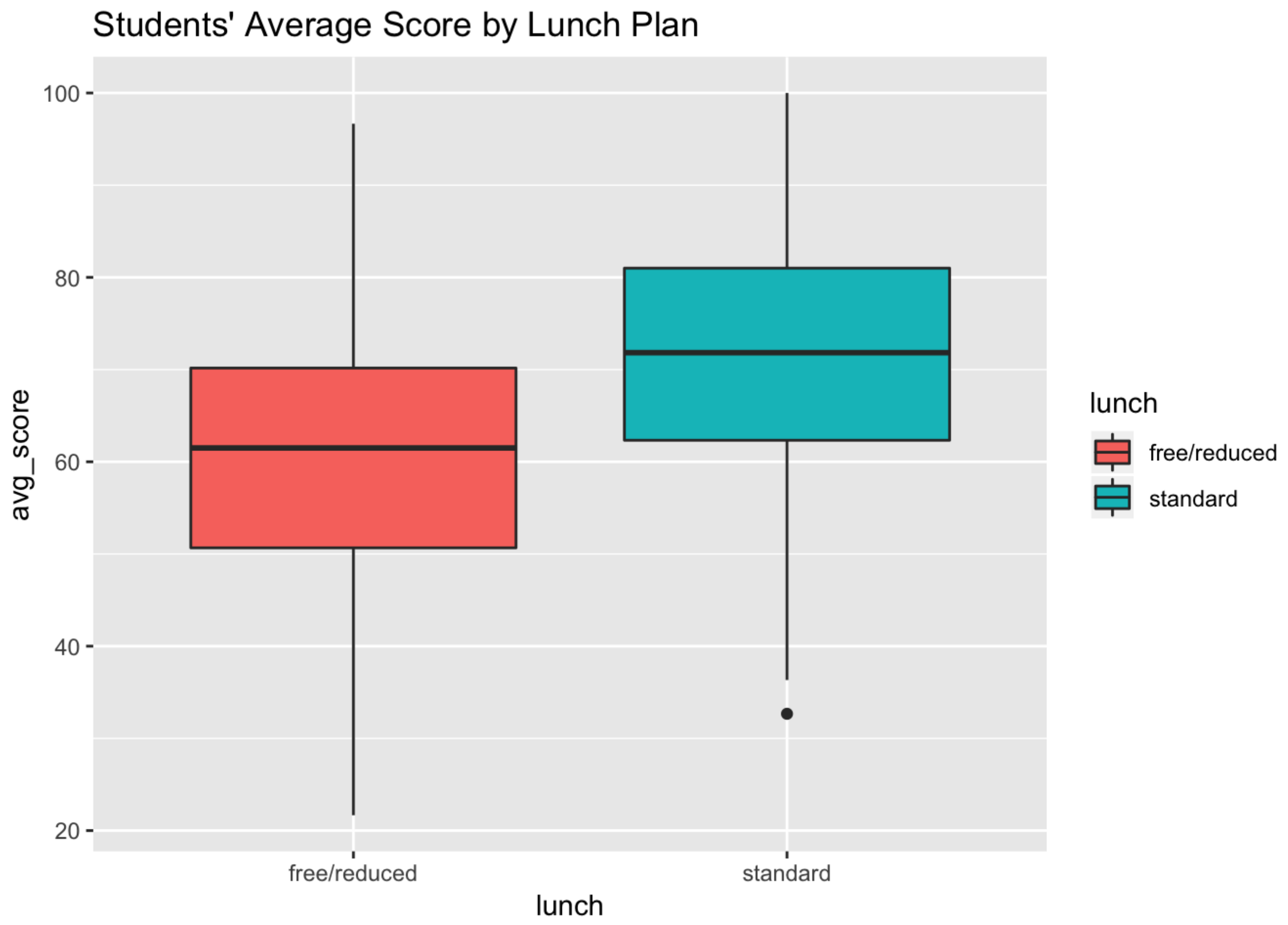
```
##         frequencies percentage cumulativepercentage
## 1               393       39.3                 39.3
## 2               425       42.5                 81.8
## 3               182       18.2                100.0
## Totals         1000      100.0                100.0
```



Lunch Meal Plan

The variable `Lunch` is a categorical variable with two levels describing the option of lunch plan the student has selected at school. The first option is the free or reduced meal plan and the other alternative is the standard meal plan. Looking at the distribution of observations, it consists of 34.8% of the “free/reduced” plan and 65.2% observations of standard lunch plan. Again, the standard plan seem to be slightly over-represented in this data set. When looking at the plot of average scores by lunch plan selected, it seems to suggest that students who selected the standard lunch plan tend to score higher than students who selected the reduced plan, but additional investigation is required to assess whether this difference is statistically significant.

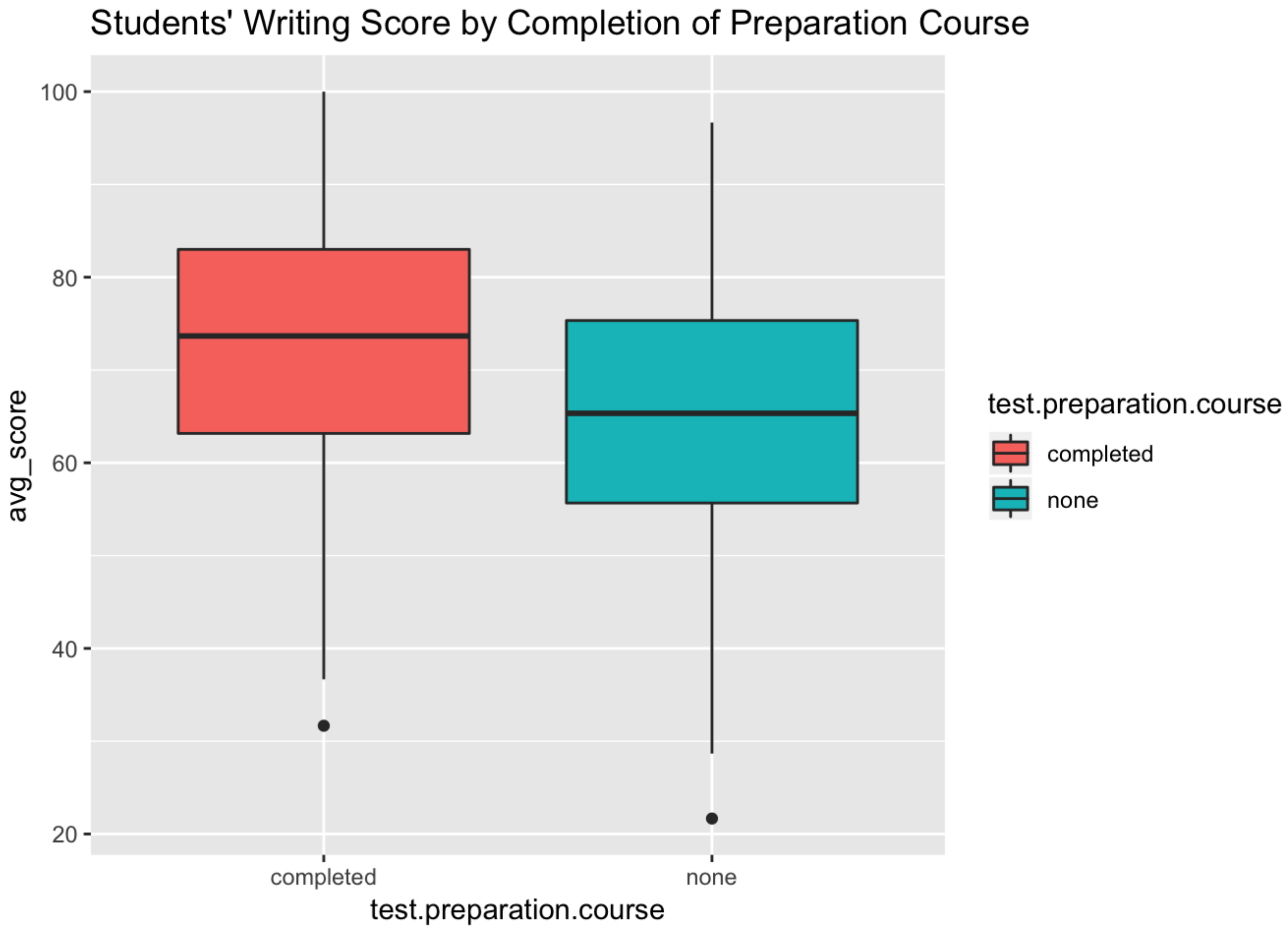
```
##              frequencies percentage cumulativepercentage
## free/reduced         348       34.8                 34.8
## standard             652       65.2                100.0
## Totals              1000      100.0                100.0
```



Test preparation course

The last variable is an indicator variable flagging whether or not the student has completed the test preparation course prior to the exam. Looking at the distribution of data, it seems that only 33.5% of students in the dataset have completed the preparation course. We will recode this variable to take value 1 if the student has completed the preparation course and value 0 otherwise.

```
##           frequencies percentage cumulativepercentage
## completed         335       33.5                 33.5
## none              665       66.5                100.0
## Totals           1000      100.0                100.0
```



```
#Changing indicator variable for test preparation course, if completed = 1, otherwise 0
data$prep_course<-as.factor(ifelse(data$test.preparation.course=="completed", 1,0))
```

Final data

Following all the modification to the raw data above, the final dataset contains the following variables.

```
mydata<-data[,c(6,1,2,7,4,8)]
summary(mydata)
```

```
##    avg_score         gender    race.ethnicity parent_educ
##  Min.   : 21.67   female:483   group A: 79    1:393
##  1st Qu.: 58.58   male  :517   group B:205    2:425
##  Median : 67.33                group C:323    3:182
##  Mean   : 67.71                group D:262
##  3rd Qu.: 78.33                group E:131
##  Max.   :100.00
##            lunch     prep_course
##  free/reduced:348   0:665
##  standard    :652   1:335
##
##
##
##
```

```
#write.csv(mydata,"Data/exam_final.csv")
```

Variable	Description
avg_score	Average score in math, reading and writing
gender	Female or Male
race.ethnicity	Race/Ethnicity group A, B, C, D, E
parent_educ	Parents’ education level: 1: some high school/high school, 2:some college/associate’s degree, 3:bachelor’s degree/master’s degree
lunch	Student’s meal plan: reduced/free or standard
prep course	Indicator if the student took the preparation course prior to the test (1 if completed, 0 otherwise)

Question 1

We would like to investigate the factors influencing a student’s score. Beginning by fitting a linear regression with all variables to the response variable `avg_score` .

a. Provide the fitted model (group A as reference level of race.ethnicity; some high school as reference level of parent_educ), and check whether ‘avg_score’ is well explained by this model.
b. Is race.ethnicity globally significant? Compare group C to group A and interpret the differences found?
c. Fit a new linear model including an interaction between*** `gender` and `lunch` and all other variables as well and provide the fitted model. Then justify whether the lunch type influences the effect of gender on average score significantly.
d. Formally test if the interpretations to above questions are valid. Carry a residual analysis of the model in part a) and comment on the results.

Question 2

Let's say we decide that a student passes the "exam" if their combined average across all three subjects is higher than 60%. For the sake of simplicity, we will refer to the combination of all three subjects as the "exam." Create an indicator variable `Pass` , which takes value 1 if the student's average score is above 60% and value 0 if below (indicating that he failed). We are now interested to examine the chances that the student will pass the exam.

a. Begin by fit a logistic regression that includes only the variable*** `Prep_Course` ***. Use students who have not completed the preparation course as the reference level.

    i. Provide the fitted model on the log-odds scale and probability scale.

    ii. Interpret all of the regression coefficients on an appropriate scale. What is the estimated probability of passing the exam when a student has not completed the preparation course? What about for a student who have completed the preparation course?

    iii. A student is unsure if enrolling in the preparation course will actually improve his chances of passing the test. Based on the results, what would you recommend? Support your explanation with a statistical test.

b. Suppose that now we are interested to investigate the number of subject that a student passes, using the same threshold of 60%. Create an indicator variable `total_pass` , which counts the number of subjects in which he receives a grade of 60% or more. We are now interested to examine the number of school subjects each student passes.

    i. Fit a Poisson regression model using the variables `Gender` , `Race.Ethnicity` , `Parent_Educ` , `Lunch` and `Prep_course` . Provide the fitted model on the mean scale.

    ii. Give an interpretation of the intercept in model.

    iii. Assess the global significance of the variable `Parent_Educ` in the model.

    iv. Discuss in a few sentences the main assumptions differences between a Poisson and Quasi-Poisson model. What are the benefits or drawbacks of using a Quasi-Poisson model instead of the Poisson model.

c. Repeat b) with a Negative Binomial regression model this time.

    i. Compare the Poisson model to the Negative Binomial using a proper statistical test. Would the Poisson model be an adequate simplification of the Negative Binomial model?

    ii. Use the AIC and BIC criterion to compare the Poisson model and Negative Binomial now. Which model is selected by each criterion?

Question 3

Consider the distribution Weibull $f(x) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} exp(-(\frac{x}{\lambda})^k), where\ x \geq 0; \lambda, k > 0$

a. Write an expression for the likelihood function.

b. Write an expression for the log-likelihood function.

c. What is the maximum likelihood estimator for*** k, λ ?