BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TPHCM

KHOA CÔNG NGHỆ THÔNG TIN

MÔN: LẬP TRÌNH R CHO PHÂN TÍCH

BÁO CÁO TIẾN ĐỘ

ĐỀ TÀI:

PHÂN TÍCH CÁC YẾU TỐ NGOẠI CẢNH ẢNH HƯỞNG ĐẾN ĐIỂM SỐ CỦA HỌC SINH

GVHD: QUÁCH ĐÌNH HOÀNG

SVTH: TRẦN VĂN DUY 19133016

LÊ PHƯƠNG NAM 19133036

TRẦN CÔNG TRƯỜNG 19133062

CAO ANH VĂN 19133067

TP. Hồ Chí Minh, ngày 27 tháng 11 năm 2021

1. Giới thiệu

Trong thời đại 4.0 việc vận dụng công nghệ vào các lĩnh vực trong trong cuộc sống có vai trò rất quan trọng, đặc biệt là vận dụng các các công nghệ vào việc phân tích dữ liệu giúp ta dự đoán được các kết quả xảy ra trong tương lai, tìm ra giải pháp để đưa ra một kết quả tốt nhất. Vì vậy, ở bài phân tích này, chúng em sẽ thực hiện áp dụng ngôn ngữ R để phân tích về thành tích học tập của học sinh qua các kỳ thi, áp dụng các kiến thức đã học về phân tích dữ liệu bằng ngôn ngữ R. Từ các phân tích đó có thể đưa ra dự đoán về hiệu suất thi cử sau này của học sinh

Nguồn dữ liệu

Nguồn dữ liệu của nhóm được được lấy từ một vài trường trung học cơ sở ở Mỹ và được tổng hợp bởi trang kaggle.com:
https://www.kaggle.com/spscientist/studentsperformance-in-exams

2. Thông tin về bộ dữ liệu

Dữ liệu bao gồm 8 cột : Tập dữ liệu này bao gồm điểm số từ ba kỳ thi và nhiều yếu tố cá nhân, xã hội và kinh tế có tác động tương tác lên chúng.

Sơ lược về tập dữ liệu

• Vài dòng đầu của dữ liệu:

```
##   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
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

• Các thuộc tính trong bảng dữ liệu:

```
colnames(data)

## [1] "gender"                      "race.ethnicity"
## [3] "parental.level.of.education" "lunch"
## [5] "test.preparation.course"     "math.score"
## [7] "reading.score"               "writing.score"
```

• Giải thích ý nghĩa của các thuộc tính

- gender: giới tính.
- race/ethnicity: chủng tộc / dân tộc của mỗi sinh viên trong tập dữ liệu.
- parental level of education: trình độ học vấn của cha mẹ.
- luch: chất lượng của bữa ăn trưa.
- test preparation course: học sinh đã hoàn thành khóa luyện thi hay chưa.
- math.score: điểm toán.
- reading.score: điểm đọc.
- writting.score: điểm viết.

3. Câu hỏi nghiên cứu

1. Yếu tố nào ảnh hưởng đến điểm số nhất
2. Dự đoán điểm trung bình(là điểm trung bình của math.score, reading.core, writing.core) theo các biến giới tính, dân tộc, trình độ học vấn của bố mẹ, khóa luyện thi và bữa ăn trưa.

Biến kết quả: điểm trung bình

Biến dự đoán: giới tính, dân tộc, trình độ học vấn của bố mẹ, khóa luyện thi và bữa ăn trưa.

4. Các phần đã làm được

- Kiểm tra dữ liệu
- Phân tích EDA.
- Thử nghiệm dự đoán điểm trung bình sử dụng thuật toán RANDOM FOREST MODEL

4.1 Kiểm tra dữ liệu

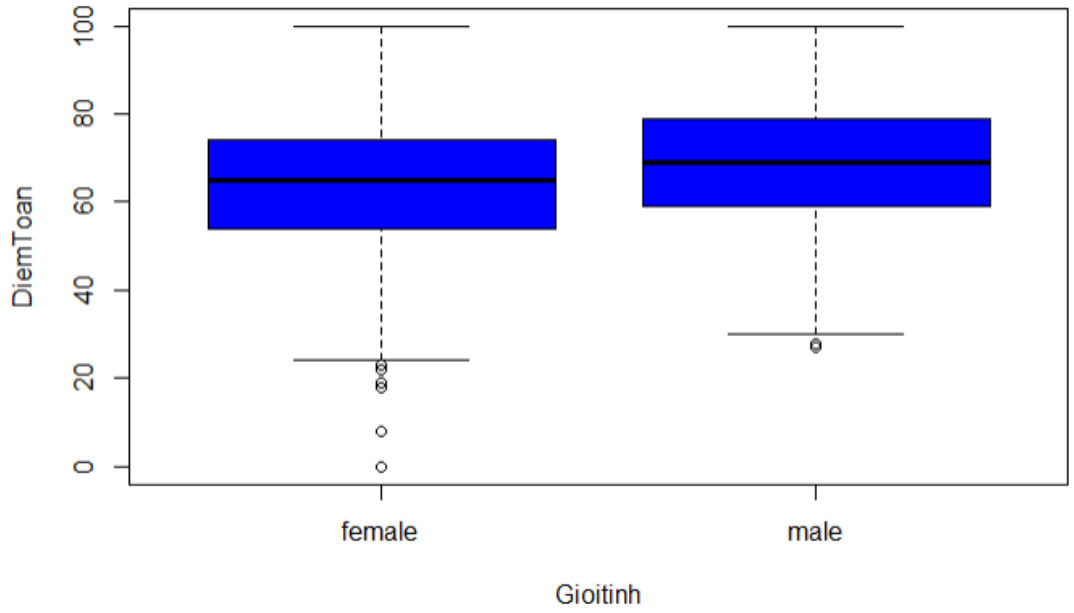
Kiểm tra dữ liệu có giá trị NA hay không.

```
> sapply(data, function(x)sum(is.na(x)))
                   gender             race.ethnicity parental.level.of.education                      lunch
                        0                          0                          0                          0
   test.preparation.course                 math.score              reading.score              writing.score
                        0                          0                          0                          0
```

=> Dữ liệu không có giá trị NA nào.

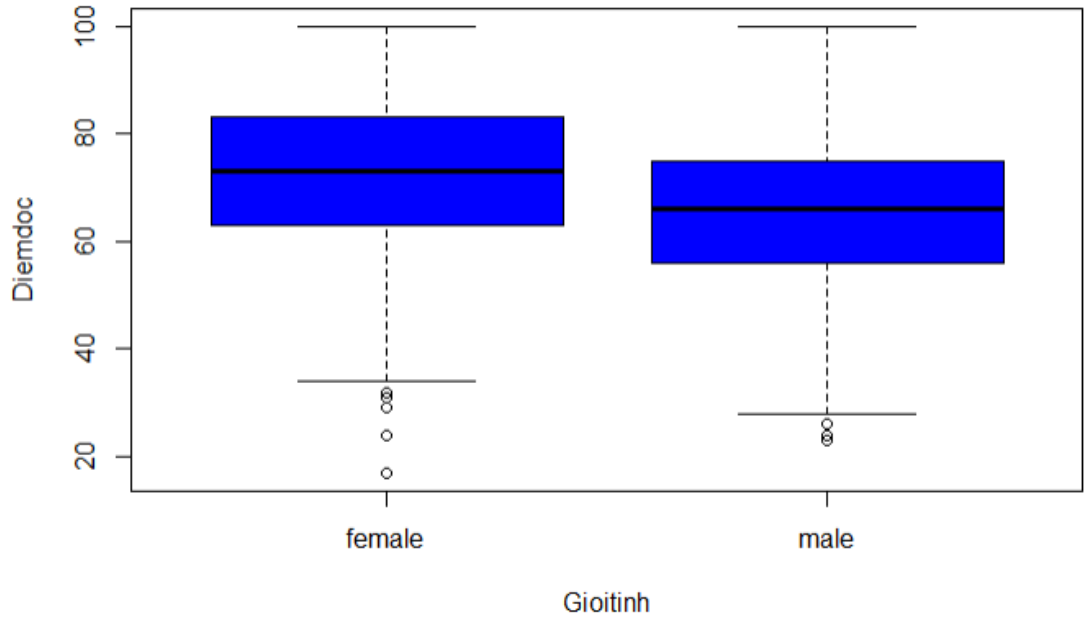
4.2 Phân tích EDA

4.2.1 Mối tương quan giữa giới tính và điểm toán



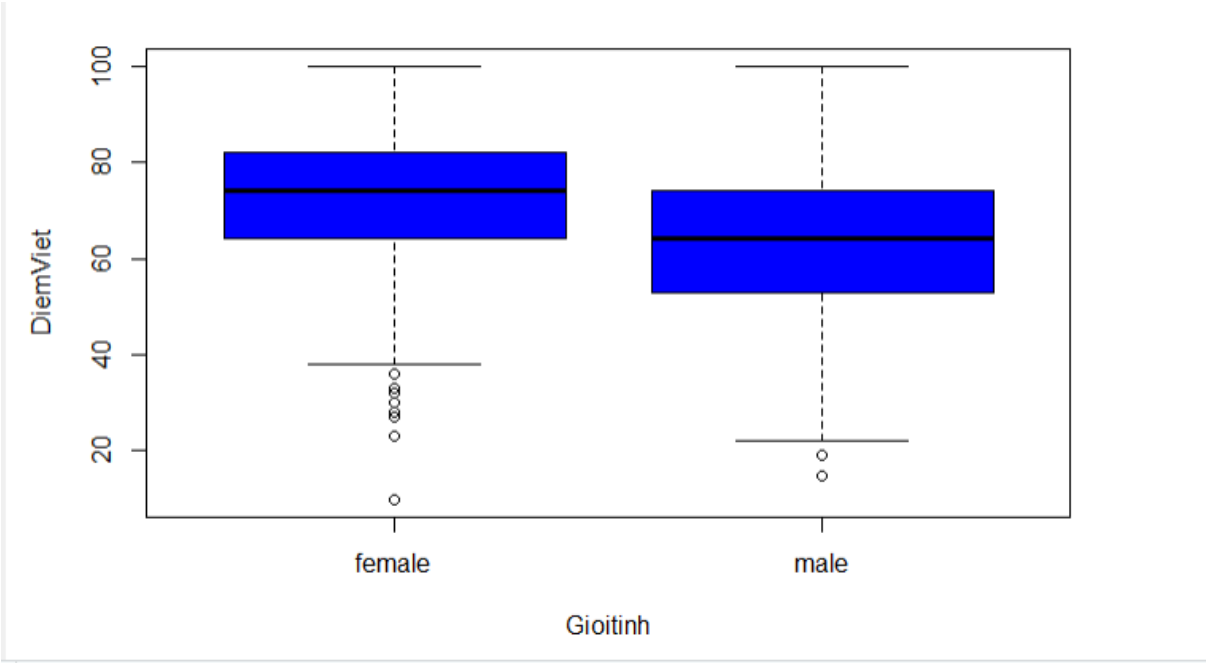
*Nhận xét: Điểm toán của nam cao hơn của nữ

4.2.2 Mối tương quan giữa giới tính và điểm đọc



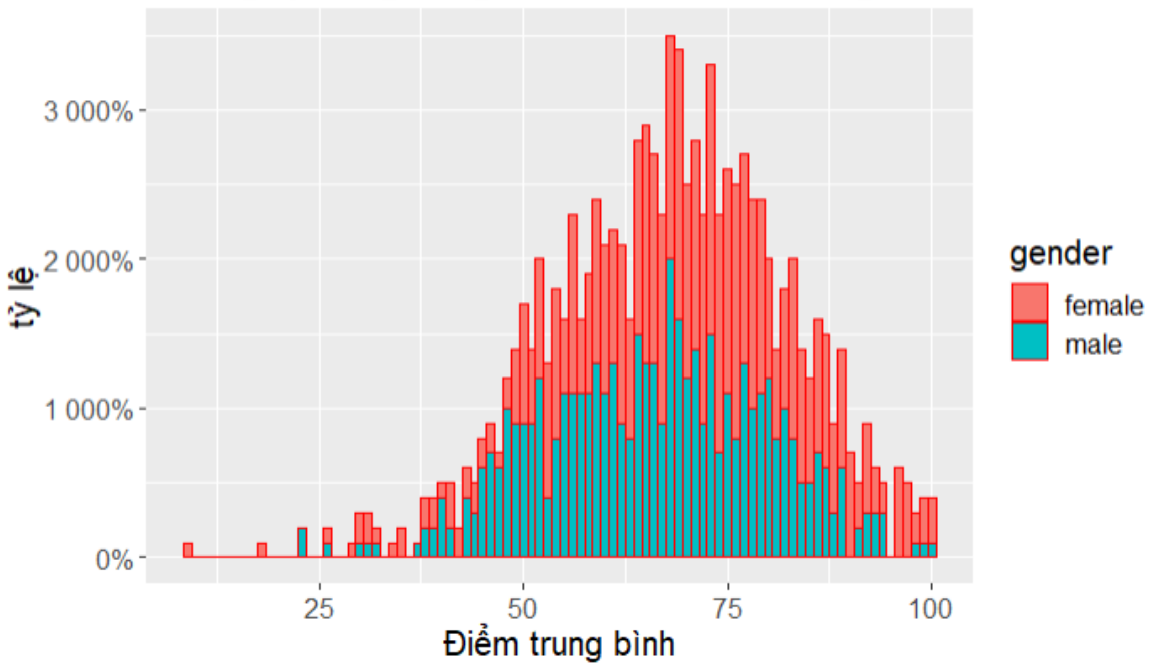
*Nhận xét: Điểm đọc của nữ cao hơn của nam

4.2.3 Mối tương quan giữa giới tính và điểm viết



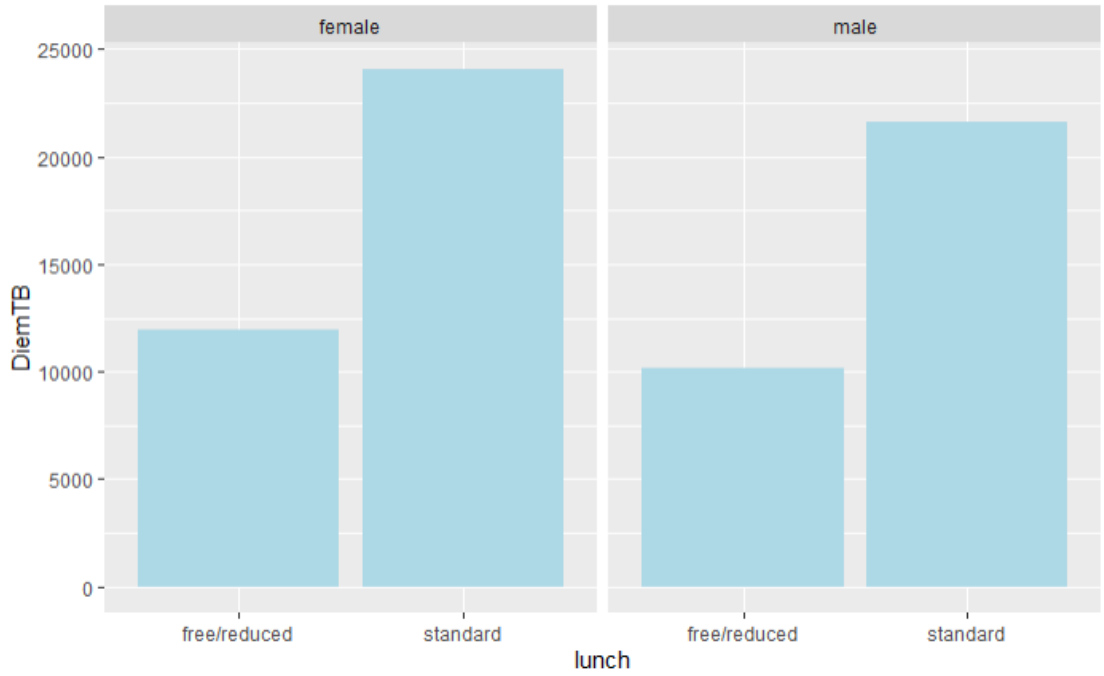
*Nhận xét: Điểm viết của nữ cao hơn điểm viết của nam

4.2.4 Mối tương quan giữa giới tính và điểm trung bình



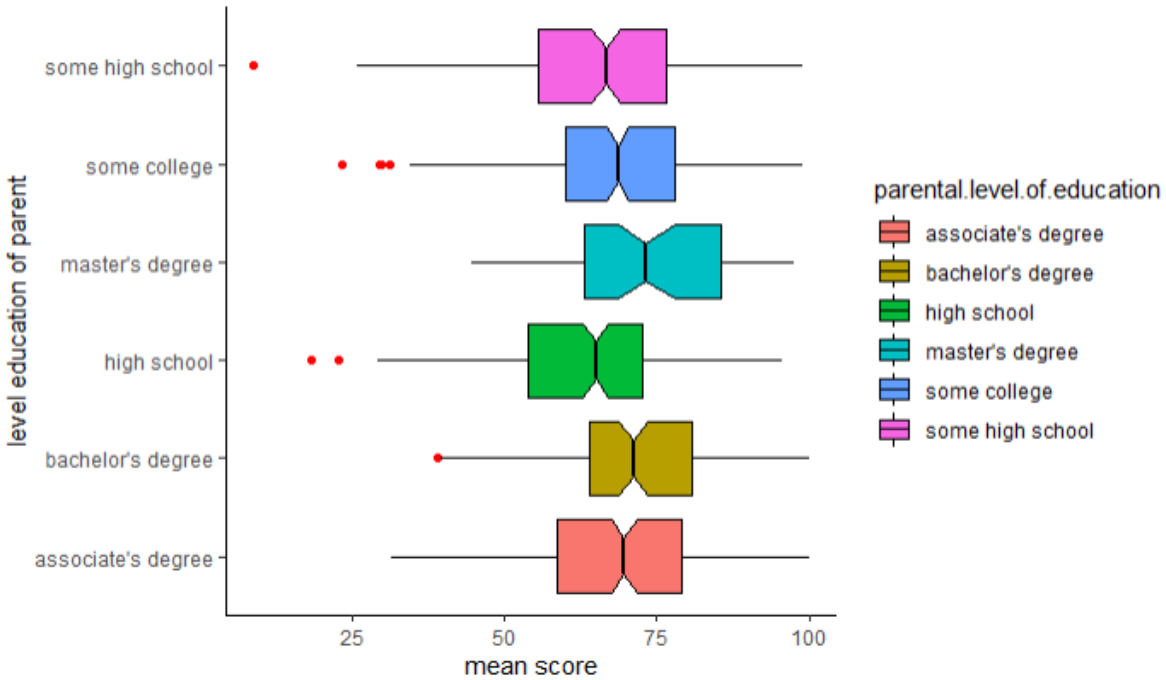
*Nhận xét: Tỷ lệ nữ giới có điểm trung bình cao hơn so với nam giới

4.2.5 Mối tương quan giữa giới tính và điểm trung bình chất lượng bữa ăn trưa



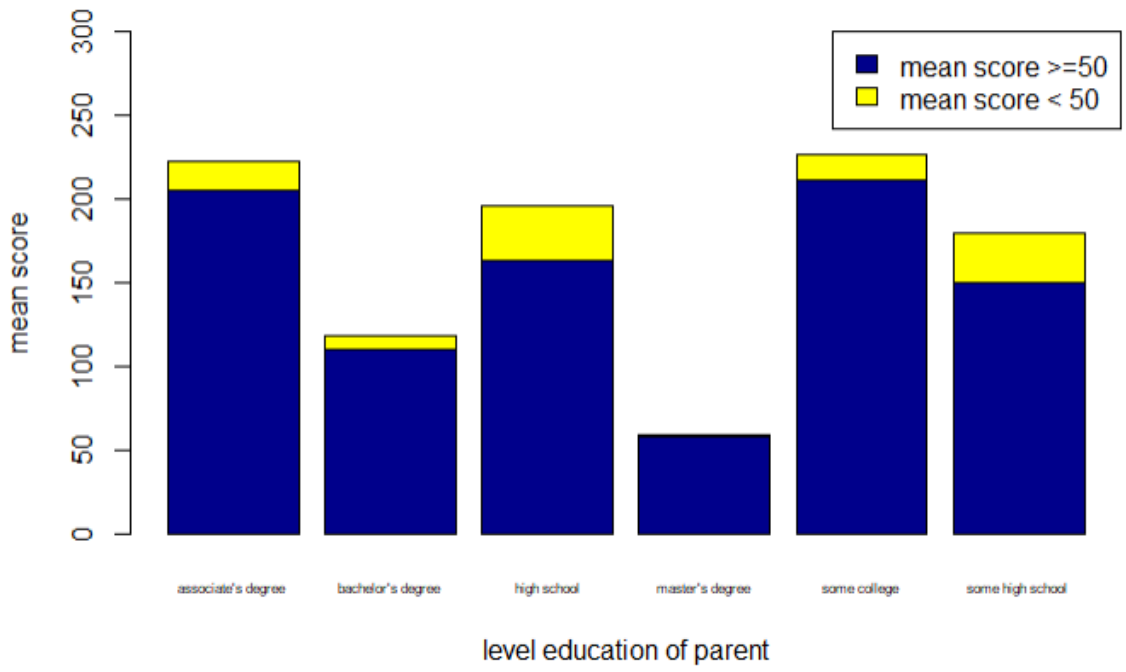
* Nhận xét: Điểm trung bình ở cả nam và nữ có bữa ăn trưa là Standard sẽ cao hơn so với bữa ăn theo free/reduced

4.2.6 Mối tương quan giữa trình độ học vấn của bố mẹ và điểm trung bình.



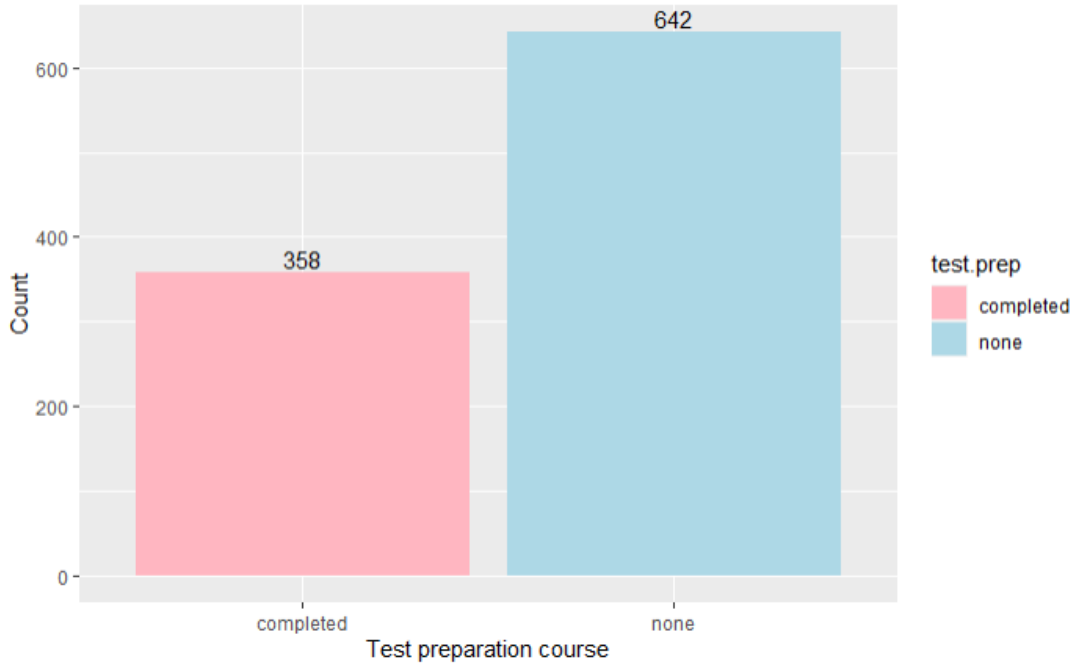
*Từ biểu đồ hộp trên ta có nhận xét: nhóm học sinh có cha mẹ đạt trình độ master's degree có median của điểm trung bình cao nhất không có các outlier với số điểm quá thấp. Nhóm học sinh có cha mẹ ở trình độ high school có median của điểm trung bình thấp nhất, còn lại các nhóm học sinh với cha mẹ có trình độ khác nhau đều gần ngang nhau.

- Sau đó chia điểm trung bình ra làm 2 loại là: >= 50 và < 50 và vẽ biểu đồ như sau:



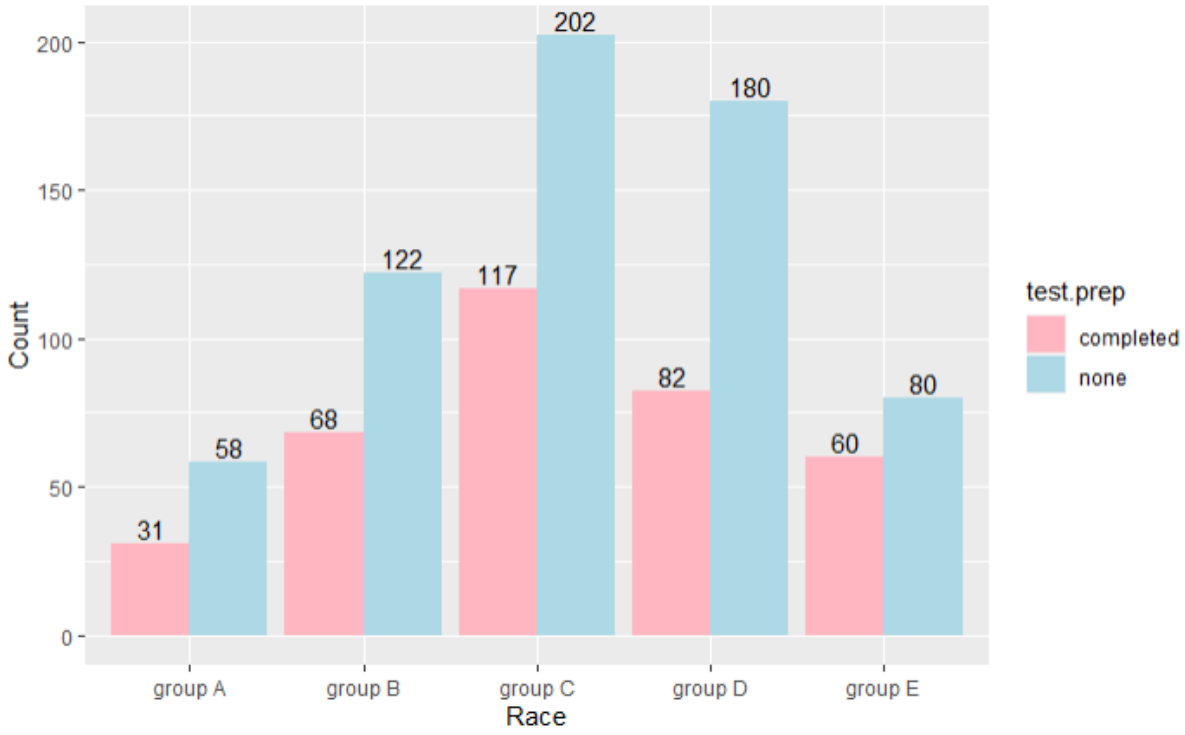
* Từ biểu đồ này, ta có thể thấy nhóm học sinh có cha mẹ với trình độ master's degree có tất cả học sinh đều có điểm trung bình 3 môn lớn hơn mức 50, còn lại tất cả các nhóm học sinh khác đều có một phần học sinh có điểm trung bình dưới 50. Với nhóm học sinh có cha mẹ có trình độ high school và some high school có lượng học sinh dưới 50 điểm chiếm tỉ trọng cao hơn các nhóm còn lại.

4.2.7 Biểu đồ thể hiện số lượng học sinh hoàn thành khóa luyện thi



*Nhận xét: có khoảng 2/3 học sinh không thực hiện khóa luyện thi.

4.2.8 Biểu đồ thể hiện số lượng học sinh hoàn thành khóa luyện thi theo dân tộc.

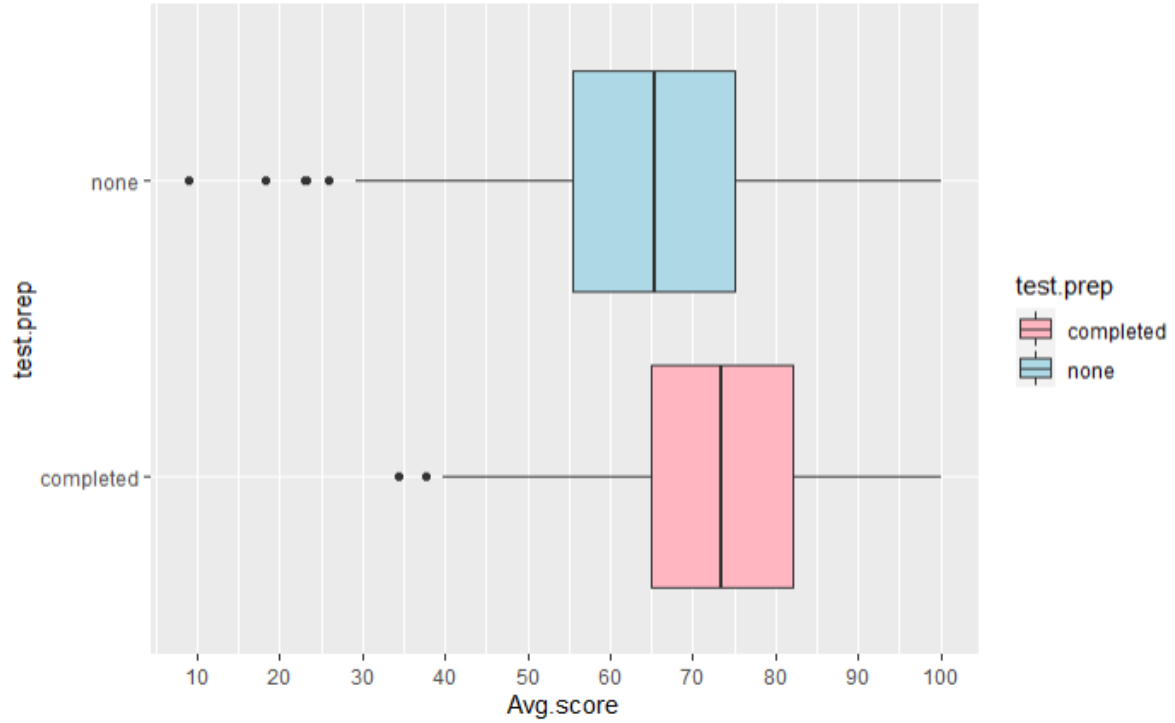


*Nhận xét:

- Ở group A có khoảng 35% học sinh hoàn thành khóa luyện thi.
- Ở group B có khoảng 36% học sinh hoàn thành khóa luyện thi.
- Ở group C có khoảng 37% học sinh hoàn thành khóa luyện thi.
- Ở group D có khoảng 31% học sinh hoàn thành khóa luyện thi.
- Ở group E có khoảng 42% học sinh hoàn thành khóa luyện thi.

Kết luận: Theo biểu đồ trên ta có thể thấy số lượng học sinh hoàn thành khóa luyện thi ở group C là nhiều nhất và group A có số lượng hoàn thành khóa luyện thi ít nhất. Tỉ lệ học sinh hoàn thành khóa luyện thi ở group E là cao nhất còn ở group D là thấp nhất.

4.2.9 Biểu đồ phân bố điểm trung bình của học sinh khi hoàn thành và không hoàn thành khóa luyện thi.



*Nhận xét: Từ biểu đồ trên cho ta biết

- Khi không hoàn thành khóa luyện thi thì:

+ Điểm thấp nhất(trừ các outliner) là 28, cao nhất là 100.

+ Ít nhất 75% học sinh đạt được 55 điểm trở lên.

+ ít nhất một nửa số học sinh đạt được 65 điểm trở lên.

- Khi hoàn thành khóa luyện thi thì:

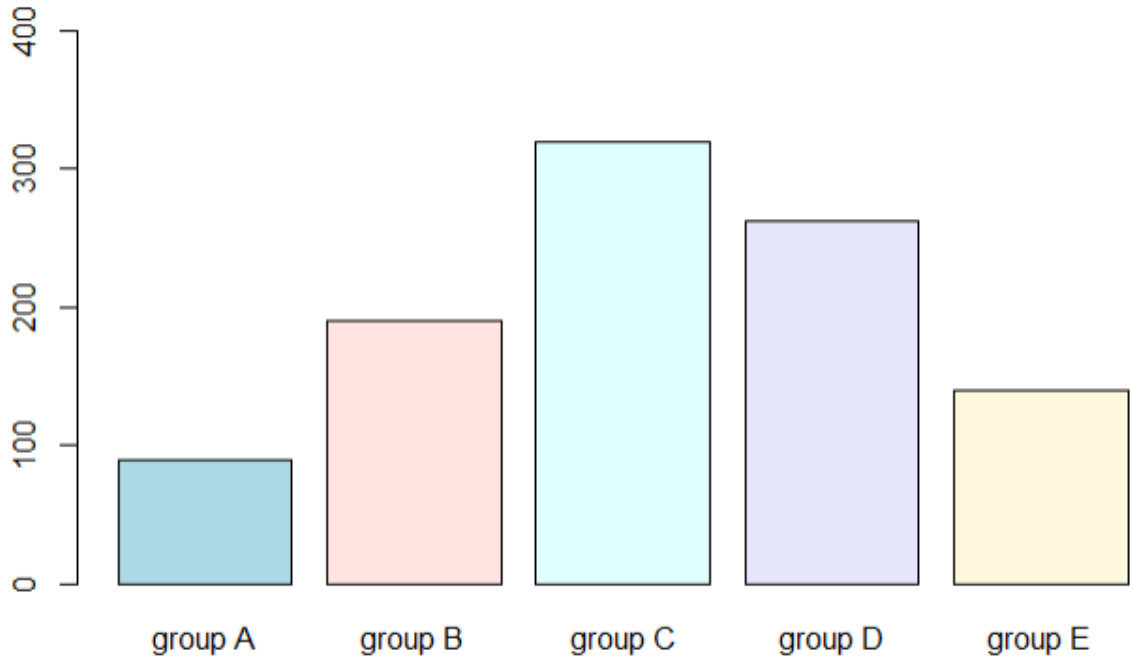
+ Điểm thấp nhất(trừ các outliner) là 39, cao nhất là 100.

+ Ít nhất 75% học sinh đạt được 65 điểm trở lên.

+ ít nhất một nửa số học sinh đạt được 73 điểm trở lên.

Kết luận: Dựa theo số liệu từ biểu đồ trên ta có thể thấy khi hoàn thành khóa luyện thi thì điểm số có cao hơn so với khi không hoàn thành khóa luyện thi(cụ thể thì ít nhất 75% học sinh hoàn thành khóa luyện thi tăng được khoảng 18,18% so với khi không hoàn thành và một nửa số học sinh hoàn thành tăng khoảng 12,31% so với khi không hoàn thành).

4.2.10 Biểu đồ số lượng học sinh theo sắc tộc.

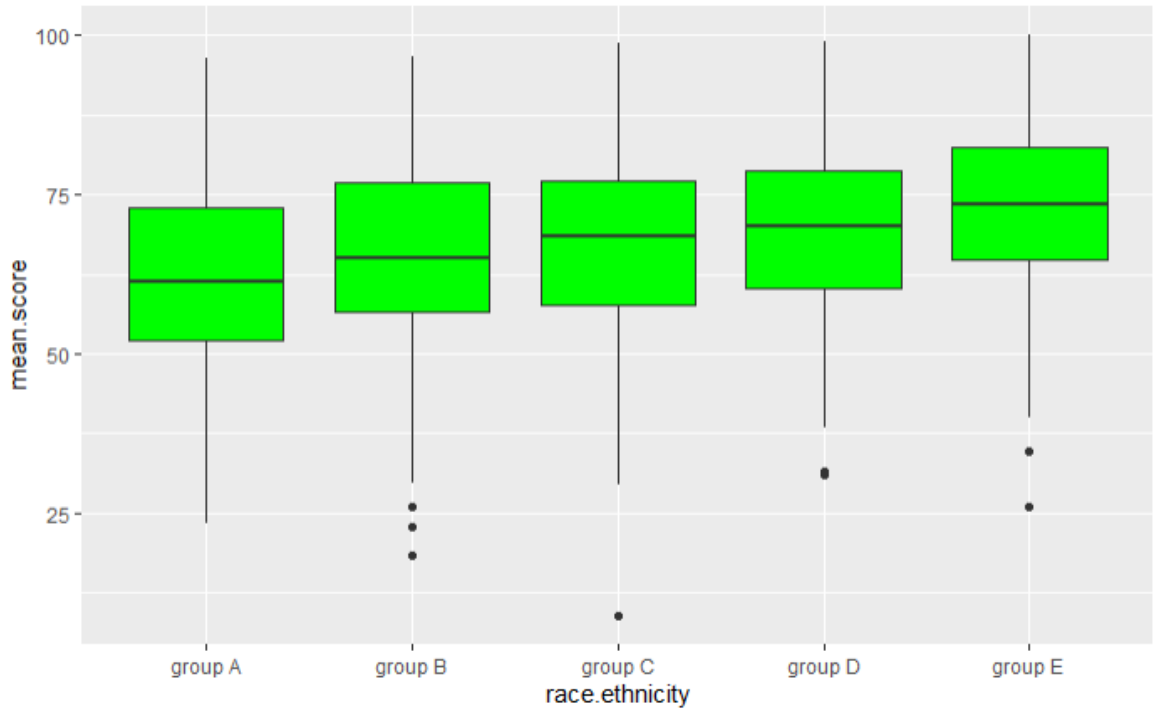


*Nhận xét: Ta thấy số lượng học sinh có thuộc tính race.ethnicity là group C nhiều nhất và ít nhất là group A.

4.2.11 Biểu đồ thể hiện điểm trung bình theo sắc tộc.

Tính trung bình của mean.score theo từng biến phân loại race.ethnicity

```
 group A  group B  group C  group D  group E
62.99251 65.46842 67.13166 69.17939 72.75238
```



*Nhận xét: Ta thấy điểm trung bình của sinh viên có thuộc tính race.ethnicity là group C là lớn nhất và thấp nhất là group A.

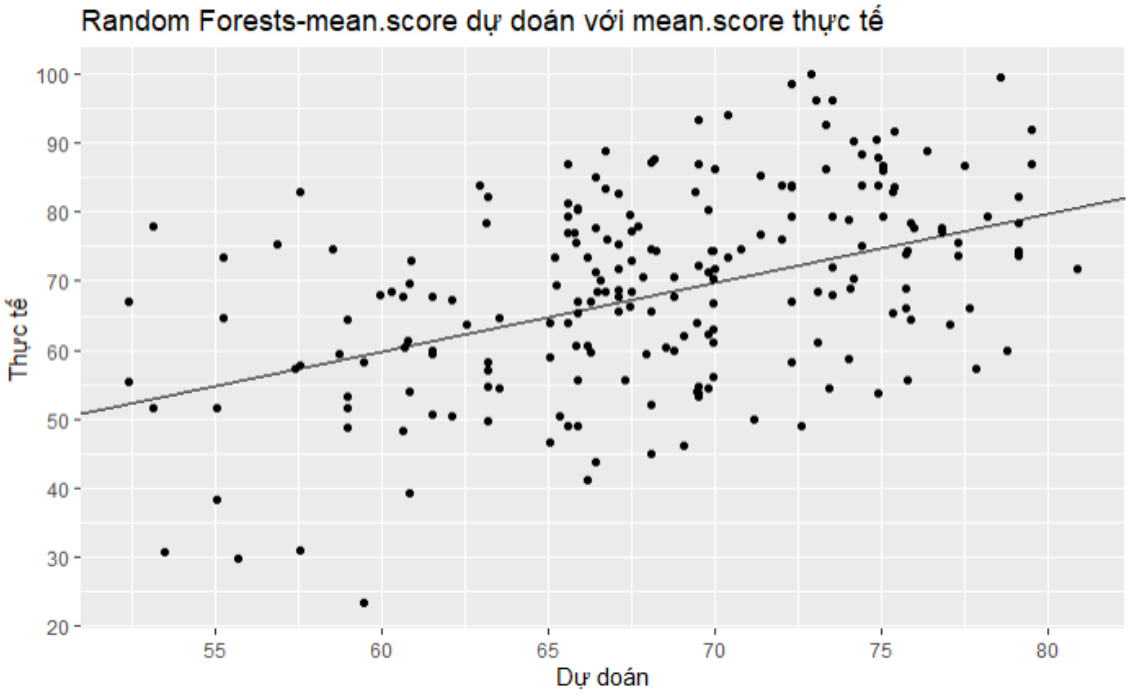
4.2.12 Kết luận.

Sau khi phân tích EDA tổng quan từ những biểu đồ trên sẽ có những kết luận được rút ra từ tập dữ liệu đó là:

- Tỷ lệ nữ giới có điểm trung bình cao hơn so với nam giới.
- Điểm trung bình ở cả nam và nữ có bữa ăn trưa là Standard sẽ cao hơn so với bữa ăn theo free/reduced.
- Có khoảng 2/3 học sinh không thực hiện khóa luyện thi.
- Số lượng học sinh hoàn thành khóa luyện thi ở group C là nhiều nhất và group A có số lượng hoàn thành khóa luyện thi ít nhất.
- Khi hoàn thành khóa luyện thi thì điểm số có cao hơn so với khi không hoàn thành khóa luyện thi.
- Số lượng học sinh có thuộc tính race.ethnicity là group C nhiều nhất và ít nhất là group A.
- Điểm trung bình của học sinh có thuộc tính race.ethnicity là group C là lớn nhất và thấp nhất là group A.

4.3 Thử nghiệm dự đoán điểm trung bình sử dụng thuật toán RANDOM FOREST MODEL.

- Đầu tiên chia bộ dữ liệu làm 2 phần train và test. Tập train lấy ngẫu nhiên 80% của tập dữ liệu, 20% còn lại làm tập test.
- Sau đó xây dựng mô hình RF trên tập dữ liệu.
- Và cuối cùng là dự đoán với mô hình RF mới xây dựng ở trên và tính RMSE(là thước đo mức độ hiệu quả của mô hình) dưới đây là biểu đồ và RMSE tính được:

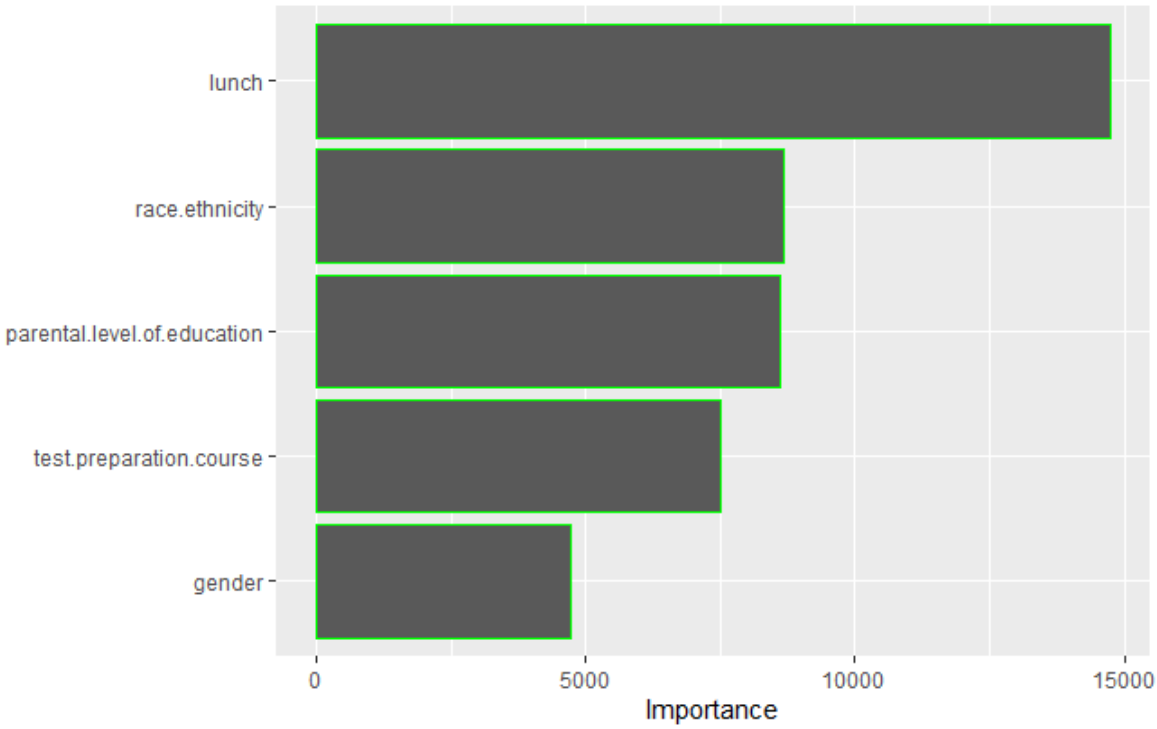


$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{n}}$$

rmse
<dbl>
12.52889

*Nhận xét: nhìn vào biểu đồ thì ta thấy được mức độ dự đoán chính xác của mô hình không được cao và RMSE lớn nên độ hiệu quả của mô hình thấp

Dựa trên tập train kiểm tra mức độ ảnh hưởng của các biến đến kết quả ta được biểu đồ như sau:



*Kết luận : Ta thấy biến luch sẽ có ảnh hưởng nhiều nhất đến điểm thi của học sinh.

5. Kế hoạch tiếp theo

Tìm hiểu, xây dựng các mô hình khác hoặc cải thiện mô hình cũ để độ hiểu quả của dự đoán cao hơn