ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC BÁCH KHOA

BÀI TẬP LỚN XÁC SUẤT – THỐNG KÊ

NHÓM: 06 – CHỦ ĐỀ: 06

LỚP: L15 - HK221

GVHD: Ts. Nguyễn Thị Mộng Ngọc

STT	MSSV	HỌ VÀ TÊN	Phân công	Chú thích
1	2114988	Nguyễn Đại Tiến	Hoạt động 2	Nhóm trưởng
2	2114913	Nguyễn Trương Phước Thọ	Hoạt động 1	
3	2110160	Trịnh Văn Hân	Hoạt động 2	
4	2114332	Lê Thị Yến Nhung	Hoạt động 1	
5	2114908	Phan Viết Thoan	Hoạt động 1	

Tp. Hồ Chí Minh, tháng 12 năm 2022

BẢNG PHÂN CÔNG CÔNG VIỆC

STT	Mã số SV	Họ và tên	Nhiệm vụ được phân công	Đóng góp
1	2114988	Nguyễn Đại Tiến	Hoạt động 2, tổng hợp word, thuyết trình	100 %
2	2114913	Nguyễn Trương Phước Thọ	Hoạt động 1, thuyết trình	100 %
3	2110160	Trịnh Văn Hân	Hoạt động 2	100 %
4	2114332	Lê Thị Yến Nhung	Hoạt động 1, hiệu chỉnh word	100 %
5	2114908	Phan Viết Thoan	Hoạt động 1	100 %

MỤC LỤC

DANH MỤC HÌNH ẢNH

I. CƠ SỞ LÍ THUYẾT

Ở bài tập lớn lần này, chúng xem sẽ tập trung thảo luận về mô hình hồi quy tuyến tính bội

1. Mô hình hồi quy tuyến tính bội

Phương trình hồi quy tổng thể với k biến độc lập có dạng như sau

$$Y = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

Trong đó:

- β_o: là hệ số tung độ góc
- β_1: là hệ dốc của Y theo biến X_1 và giữa các biến $X_2, X_3, \ldots, X_k$ không đổi.
- β_k: là hệ dốc của Y theo biến X_k và giữa các biến $X_1, X_2, X_3, \ldots, X_k$ không đổi.
- ϵ_i : là thành phần ngẫu nhiên (yếu tố nhiễu), có kì vọng bằng 0 và phương sai không đổi σ^2

Giả sử có một mẫu quan sát với giá trị thực tế là ($Y_i, X_{2i}, \ldots, X_{ki}$) với (i=1,2, 3,...k). Ta sẽ sự dụng thông tin từ mẫu để xây dựng các ước lượng cho các hệ số β_j (với j=1,2, 3,…k). Từ các giá trị ước lượng này có thể viết thành hàm hồi quy mẫu như sau:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2} X_{2i}, \ldots \widehat{\beta_k} X_{ki}$$

Trong đó $\widehat{Y_1}$ là giá trị ước lượng cho Y_1 và sai lệch giữa hai giá trị này được gọi là phần dư

2. Đánh giá sự phù hợp của mô hình

- ***Tính toán hệ số xác định bội***

– Khi có nhiều biến độc lập trong mô hình thì R^2 vẫn được sử dụng để xác định phần biến thiên trong biến phụ thuộc và tất cả các biến độc lập trong mô hình, tuy nhiên lúc này R^2 được gọi là hệ số xác định bội, công thức tính toán hệ số xác định bội như sau:

$$R^2 = \frac{SSR}{SST}$$

– Cụ thể trong trường hợp khi R^2 thì ta có thể kết luận rằng 82% biến thiên trong giá trị của biến phụ thuộc có thể được giải thích bởi mối liên hệ tuyến tính giữa biến phụ thuộc với các biến độc lập trong mô hình, tuy nhiên chú ý rằng không phải tất cả các biến độc lập này đều có tầm quan trọng ngang nhau đối với khả năng giải thích cho biến thiên trong biến phụ thuộc của mô hình.

- ***Hệ số xác định hiệu chỉnh***

– Hệ số xác định hiệu chỉnh ký hiệu R^2_{adj} là một cách khác để đo lường tỷ lệ phần trăm của biến thiên được giải thích trong biến phụ thuộc mà có tính đến mối liên hệ giữa cỡ mẫu và số biến độc lập trong mô hình hồi quy bội, công thức của nó như sau:

$$R^2_{adj} = 1 - (1 - R^2)[\frac{n-k}{n-k-1}]$$

Trong đó n là cỡ mẫu, k là số biến độc lập trong mô hình.

– Sự gia tăng trong R^2 có thể không bù đắp được thiệt hại do mất thêm bậc tự do khi thêm biến, thế nhưng R^2_{adj} có tính đến chi phí này và điều chỉnh giá trị R^2_{adj} theo nó một cách phù hợp. Khi một biến độc lập được thêm vào không có đóng góp xứng đáng vào khả năng giải thích cho biến phụ thuộc thì R^2_{adj} sẽ luôn luôn giảm đi mặc dù R^2 tăng

– Điều đó cho thấy với mô hình hồi quy đa biến, nhất là khi số biến độc lập khá lớn trong tương quan với cỡ mẫu thì ta nên dùng R^2_{adj} để đánh giá khả năng giải thích của mô hình. Vì vậy thông thường khi đánh giá độ phù hợp của mô hình hồi quy bội,bên cạnh thông tin về R^2_{adj} người ta cũng dùng thêm thông tin về R^2_{adj} để tham khảo.

- ***Đánh giá ý nghĩa toàn diện của mô hình***

– Mô hình hồi quy mà chúng ta xây dựng là dựa trên dữ liệu của một mẫu lấy từ tổng thể vì vậy nó có thể bị ảnh hưởng của sai số lấy mẫu, vì thế chúng ta cần kiểm định ý nghĩa thống kê của toàn bộ mô hình.

– Chúng ta có thể dựng một giả thuyết như sau:

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$ (hay $R^2 = 0$)

$H_1 : \exists \beta_j \neq 0$ (hay $R^2 \neq 0$)

Nếu giả thuyết H_0 trên đúng nghĩa là tất cả các hệ số độ dốc đều đồng thời bằng 0 thì mô hình hồi quy đã xây dựng không hề có tác dụng trong việc dự đoán hay mô tả về biến phụ thuộc.

– Đại lượng F thống kê (trong bảng ANOVA) chính là con số thống kê được sử dụng để kiểm định giả thuyết về ý nghĩa toàn diện của mô hình hồi quy, công thức của đại lượng F được hình thành như sau:

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

Trong đó SSR là tổng bình phương hồi quy (Regression Sum of Squares), SSE là tổng bình phương sai số (Error Sum of Squares), *n* và *k* lần lượt là cỡ mẫu và biến độc lập.

Chú ý là để quyết định ta phải tra bảng thống kê F tìm giá trị tới hạn tương ứng với mức ý nghĩa ta chọn trước. Mà muốn tra bảng F ta phải có thêm thông tin về bậc tự do ở tử số và mẫu số, ta qui ước bậc tự do của tử số k và bậc tự do của mẫu số là (n = k = 1).

– Từ đây, ta có quy trình đánh giá ý nghĩa toàn diện của mô hình như sau:

Bước 1: Đặt giả thuyết:

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$

- $H_1 : \nexists_j \neq 0$

Bước 2: Chọn độ tin cậy cho kiểm định từ đó có mức ý nghĩa α

Bước 3: Với bậc tự do xác định như trên, tra bảng phân phối F ta được giá trị F tới hạn.

Bước 4: So sánh giá trị F kiểm định tính được theo công thức trên và giá trị F tới hạn.

Bước 5: Kết luận.

Nếu F kiểm định > F tới hạn, ta có thể kết luận rằng mô hình hồi quy bội với các biến độc lập ta đưa vào có thể giải thích một cách có ý nghĩa cho biến thiên giá trị của biến phụ thuộc.

***Tính toán sai số chuẩn ước lượng*:**

– Mục tiêu của việc xây dựng mô hình hồi quy là để có thể xác định được giá trị của biến phụ thuộc khi biết trước các giá trị cụ thể của biến độc lập. Một số thống kê cho thấy mô hình hồi quy thực hiện mục tiêu này tốt đến đâu là lệch chuẩn của mô hình hồi quy (còn gọi tên là Sai số chuẩn ước lượng). Giá trị ước lượng từ thông tin mẫu của độ lệch chuẩn của mô hình hồi quy (sai số chuẩn ước lượng) được tính toán như sau đây:

$$s_{Y/X} = \sqrt{\frac{SSE}{n-k-1}}$$

Trong đó n là cỡ mẫu, k là biến độc lập trong mô hình

– Sai số chuẩn ước lượng đo lường sự phân tán của các giá trị thực tế đo lường được của biến phụ thuộc quanh những giá trị của biến phụ thuộc được dự đoán bằng đường hồi quy.

***Đánh giá ý nghĩa của của từng biến độc lập riêng biệt*:**

Ở kiểm định F, giả sử H_1 được chấp nhận ta kết luận rằng mô hình toàn diện có ý nghĩa. Điều này có ý nghĩa là có ít nhất một biến độc lập trong mô hình có thể giải thích được một cách có ý nghĩa cho biến thiên phụ thuộc. Tuy nhiên điều này không có ý nghĩa là tất cả các biến độc lập đưa vào mô hình đều có ý nghĩa, để xác định biến độc lập nào có ý nghĩa chúng ta kiểm định giả thuyết sau:

- $H_0 : \beta_1 = 0$
- $H_1 : \nexists_j \neq 0$

Chúng ta có thể dùng kiểm định *t* để kiểm định nghĩa của mỗi hệ số hồi quy với độ tin cậy được chọn trước, *t* được xác định bằng công thức:

$$t = \frac{b_j - 0}{S_{bj}}$$

Trong đó b_j là hệ số dốc trong mô hình hồi quy mẫu cho biến độc lập thứ *j*, S_{bj} là sai số chuẩn ước lượng lượng của hệ số độ dốc của biến độc lập thứ *j*. Giá trị *t* tính toán được sẽ được so sánh với giá trị *t* tới hạn tra tùa bảng phân phối student với (n-k-1) bậc tự do và mức ý nghĩa $\frac{\alpha}{2}$

3. Phương pháp bình phương nhỏ nhất

– Phương pháp bình phương nhỏ nhất được đưa ra bởi nhà Toán học Đức Carl Friedrich Gauss- đây là một trong các phương pháp ước lượng hồi quy tuyến tính phổ biến nhất.

– Với tổng thể, sai số (error) kí hiệu là e, còn trong mẫu nghiên cứu lúc này được gọi là phần dư và được kí hiệu là ϵ. Biến thiên phần dư được tính bằng tổng bình phương tất cả các phần dư cộng lại.

– Nguyên tắc của phương pháp hồi quy OLS là làm cho biến thiên phần dư này trong phép hồi quy là nhỏ nhất. Khi biểu diễn trên mặt phẳng O_{xy}, đường hồi quy là đường thẳng đi qua đám đông các điểm dữ liệu mà ở đó, khoảng cách từ các điểm dữ liệu (tuyệt đối của ϵ đến đường hồi quy là ngắn nhất).

– Từ đồ thị **scatter** biểu diễn mối quan hệ giữa các biến độc lập và biến phụ thuộc, các điểm dữ liệu sẽ nằm phân tán nhưng có xu hướng chung tạo thành một đường thẳng. Chúng ta có thể rất nhiều đường thẳng hồi quy đi qua đám đông các điểm dữ liệu này chứ không phải chỉ một đường duy nhất, vấn đề là ta phải chọn ra đường thẳng nào mô tả sát nhất xu hướng dữ liệu. Bình phương nhỏ nhất OLS sẽ tìm ra đường thẳng đó dựa trên nguyên tác cực tiểu hóa khoảng cách từ các điểm dữ liệu đến đường thẳng. Trong hình ở trên đường màu đỏ là đường hồi quy OLS.

4. Các giả định của mô hình hồi quy

a) *Hàm hồi quy là tuyến tính theo các tham số.*

Điều này có nghĩa là quá trình thực hành hồi quy trên thực tế được miêu tả bởi mối quan hệ dưới dạng: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_k x_k + \epsilon$ hoặc mối quan hệ thực thế có thể được viết lại ví dụ như dưới dạng lấy loga cả hai vế.

b) *$E(\epsilon_i) = 0$: Kỳ vọng của các yếu tố ngẫu nhiên bằng 0.*

Trung bình tổng thể sai số là bằng 0. Điều này có nghĩa là có một số giá trị sai số mang dấu dương và một số mang dấu âm. Do hàm xem như là đường trung bình nên có thể giả định rằng các sai số ngẫu nhiên trên sẽ bị loại trừ nhau, ở mức trung bình trong tổng thể.

c) $Cov\ (\epsilon_i, \epsilon_j) = 0$ *: Không có sự tương quan giữa các ϵ_i*

Không có sự tương quan giữa các quan sát của yếu tố sai số. Nếu ta xem xét các chuỗi số liệu thời gian (dữ liệu được thu nhập từ một nguồn trong nhiều khoảng thời gian khác nhau), yếu tố sai

số ϵ_i trong khoảng thời gian này không có bất kỳ một tương quan nào với yếu tố sai số trong khoảng thời gian trước đó.

d) $Cov\ (\epsilon_i, \epsilon_j) = 0$ *: ϵ và X không có sự tương quan với nhau.*

Khi bất kỳ biến giải thích nào lớn hơn hay nhỏ đi thì yếu tố sai số sẽ không thay đổi theo nó.

e) $Var\ (\epsilon_i = \sigma^2)$ *: Phương sai bằng nhau và thuần nhất với mọi* ϵ_i

Tất cả giá trị ϵ_i được phân phối giống nhau với cùng σ^2 sao cho $Var\ (\epsilon_i) = E\ (\epsilon_i^2) = \sigma^2$

f) ϵ_i *phân phối chuẩn*

Điều này rất quan trọng khi phát sinh khoảng tin cậy và thực hiện kiểm định giả thuyết trong những phạm vi mẫu là nhỏ. Nhưng phạm vi mẫu lớn hơn, điều này trở nên không mấy quan trọng.

g) Giữa các $X_2, X_3, \dots, X_k$ *không có quan hệ tuyến tính*

Nếu $X_2, X_3, \dots, X_k$ có quan hệ tuyến tính thì người ta nói rằng có hiện tượng đa cộng tuyến. Hay không tồn tại $\lambda_1 \equiv 0 : \lambda_1 x_{1i} + \lambda_2 x_{2i} + \lambda_3 x_{3i} + \dots + \lambda_k x_{ki} + v_i = 0$

Kiểm tra giả định thông qua các biểu đồ :

- Đồ thị thứ 1 (Residuals vs Fitted) vẽ các giá trị dự báo với các giá trị thặng dư (sai số) tương ứng, dùng để kiểm tra tính tuyến tính của dữ liệu và tính đồng nhất của các phương sai sai số. Nếu như giả định về tính tuyến tính của dữ liệu KHÔNG thỏa, ta sẽ quan sát thấy rằng các điểm thặng dư (residuals) trên đồ thị sẽ phân bố theo một hình mẫu (pattern) đặc trưng nào đó (ví dụ parabol). Nếu đường màu đỏ trên đồ thị phân tán là đường thẳng nằm ngang mà không phải là đường cong, thì giả định tính tuyến tính của dữ liệu được thỏa mãn. Để kiểm tra giả định phương sai đồng nhất thì các điểm thặng dư phải phân tán đều nhau xung quanh đường thẳng y = 0.
- Đồ thị thứ 2 (Normal Q-Q) cho phép kiểm tra giả định về phân phối chuẩn của các sai số. Nếu các điểm thặng dư nằm trên cùng 1 đường thẳng thì điều kiện về phân phối chuẩn được thỏa.
- Đồ thị thứ 3 (Scale - Location) vẽ căn bậc hai của các giá trị thặng dư được chuẩn hóa với các giá trị dự báo, được dùng để kiểm tra giả định phương sai của các sai số là hằng số. Nếu như đường màu đỏ trên đồ thị là đường thẳng nằm ngang và các điểm thặng dư phân tán đều xung quanh đường thẳng này thì giả định này được thỏa. Nếu như đường màu đỏ có độ dốc (hoặc cong) hoặc các điểm thặng dư phân tán không đều xung quanh đường thẳng này, thì giả định bị vi phạm.
- Đồ thị thứ 4 (Residuals vs Leverage) cho phép xác định những điểm có ảnh hưởng cao (influential observations), nếu chúng có hiện diện trong bộ dữ liệu. Những điểm có ảnh hưởng cao này có thể là các điểm outliers, là những điểm có thể gây nhiều ảnh hưởng nhất khi phân tích dữ liệu. Nếu như ta quan sát thấy một đường thẳng màu đỏ đứt nét (Cook's distance), và có một số điểm vượt qua đường thẳng khoảng cách này, nghĩa là các điểm đó là các điểm có ảnh hưởng cao.

Nếu như ta chỉ quan sát thấy đường thẳng khoảng cách Cook ở góc của đồ thị và không có điểm nào vượt qua nó, nghĩa không có điểm nào thực sự có ảnh hưởng cao.

II. HOẠT ĐỘNG 1

1. Mô tả bài toán

Tập tin "diem_so.csv" chứa thông tin về điểm toán của các em học sinh trung học thuộc hai trường học ở Bồ Đào Nha. Các thuộc tính dữ liệu bao gồm điểm học sinh, nơi cư trú, và một số hoạt động xã hội khác. Dữ liệu được thu thập bằng cách sử dụng báo cáo của các trường và các kết quả khảo sát sinh viên. Dữ liệu gốc được cung cấp tại: https://archive.ics.uci.edu/ml/datasets/student+performance.

Các biến chính trong bộ dữ liệu:

- G1: Điểm thi học kì 1.
- G2: Điểm thi học kì 2.
- G3: Điểm cuối khoá.
- studytime: Thời gian tự học trên tuần (1 - ít hơn 2 giờ, 2 - từ 2 đê 5 giờ, 3 - từ 5-10 giờ, or 4 - lớn hơn 10 giờ).
- failures: số lần không qua môn (1,2,3, hoặc 4 chỉ nhiều hơn hoặc bằng 4 lần).
- absences: số lần nghỉ học.
- paid - Có tham gia các lớp học thêm môn Toán ngoài trường (có/không).
- sex: Giới tính của học sinh. (Nam/nữ).

2. Thực hiện

a. Đọc dữ liệu (Import data)

Đọc tệp tin " diem_so.csv " và gán với tên diem_so

```
diem_so <- read.csv("C:/Users/ADMIN/Downloads/diem_so.csv") #doc du lieu
head(diem_so,10)  # xem 10 dong dau tien
```

```
    X school sex age address famsize Pstatus Medu Fedu     Mjob      Fjob     reason guardian traveltime studytime failures
1   1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course   mother          2         2        0
2   2     GP   F  17       U     GT3       T    1    1  at_home    other     course   father          1         2        0
3   3     GP   F  15       U     LE3       T    1    1  at_home    other      other   mother          1         2        3
4   4     GP   F  15       U     GT3       T    4    2   health services       home   mother          1         3        0
5   5     GP   F  16       U     GT3       T    3    3    other    other       home   father          1         2        0
6   6     GP   M  16       U     LE3       T    4    3 services    other reputation   mother          1         2        0
7   7     GP   M  16       U     LE3       T    2    2    other    other       home   mother          1         2        0
8   8     GP   F  17       U     GT3       A    4    4    other  teacher       home   mother          2         2        0
9   9     GP   M  15       U     LE3       A    3    2 services    other       home   mother          1         2        0
10 10     GP   M  15       U     GT3       T    3    4    other    other       home   mother          1         2        0
   schoolsup famsup paid activities nursery higher internet romantic famrel freetime goout Dalc Walc health absences G1 G2 G3
1        yes     no   no         no     yes    yes       no       no      4        3     4    1    1      3        6  5  6  6
2         no    yes   no         no      no    yes      yes       no      5        3     3    1    1      3        4  5 NA  6
3        yes     no  yes         no     yes    yes      yes       no      4        3     2    2    3      3       10  7  8 10
4         no    yes  yes        yes     yes    yes      yes      yes      3        2     2    1    1      5        2 15 14 15
5         no    yes  yes         no     yes    yes       no       no      4        3     2    1    2      5        4  6 10 10
6         no    yes  yes        yes     yes    yes      yes       no      5        4     2    1    2      5       10 15 NA 15
7         no     no   no         no     yes    yes      yes       no      4        4     4    1    1      3        0 12 12 11
8        yes    yes   no         no     yes    yes       no       no      4        1     4    1    1      1        6  6  5  6
9         no    yes  yes         no     yes    yes      yes       no      4        2     2    1    1      1        0 16 NA 19
10        no    yes  yes        yes     yes    yes      yes       no      5        5     1    1    1      5        0 14 15 15
```

Hình 1: Kết quả khi xem 10 dòng đầu tiên của file "diem_so"

b. Làm sạch dữ liệu

❖ Từ dữ liệu trong diem_so, trích ra một dữ liệu con bao gồm các biến chính của đề bài và đặt tên là "new_DF".

```
new_DF <- diem_so [,c("G1","G2","G3","studytime","failures","absences","paid","sex")] # trich cac bien
head(new_DF,10)  # xem 10 dong dau tien
```

```
   G1 G2 G3 studytime failures absences paid sex
1   5  6  6         2        0        6   no   F
2   5 NA  6         2        0        4   no   F
3   7  8 10         2        3       10  yes   F
4  15 14 15         3        0        2  yes   F
5   6 10 10         2        0        4  yes   F
6  15 NA 15         2        0       10  yes   M
7  12 12 11         2        0        0   no   M
8   6  5  6         2        0        6   no   F
9  16 NA 19         2        0        0  yes   M
10 14 15 15         2        0        0  yes   M
```

Hình 2: Kết quả khi xem 10 dòng đầu tiên của file "new_DF"

❖ Kiểm tra dữ liệu khuyết trong tệp tin "new_DF".

```
apply(is.na(new_DF),2,which) # kiem tra va xuat vị tri du lieu khuyet (NA)
apply(is.na(new_DF),2,sum) # kiem tra va dem gia tri NA
apply(is.na(new_DF),2,mean) # ti le NA
```

```
$G1
integer(0)

$G2
[1]   2   6   9  80 100

$G3
integer(0)

$studytime
integer(0)

$failures
integer(0)

$absences
integer(0)

$paid
integer(0)

$sex
integer(0)
    G1     G2     G3 studytime failures absences   paid    sex
     0      5      0         0        0        0      0      0
    G1     G2     G3 studytime failures absences   paid    sex
0.00000 0.01266 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
```

Hình 3: Kết quả kiểm tra dữ liệu khuyết trong tệp tin "new_DF"

Nhận xét: Ta nhận thấy có 5 dữ liệu khuyết ở biến G2 (cụ thể là ở dòng 2, 6, 9, 80 và 100), số lượng dữ liệu khuyết chiếm 1.2658% so với tổng số quan sát của dữ liệu. Do dữ liệu bị khuyết là hoàn toàn ngẫu nhiên (Missing Not at Random) và tỉ lệ dữ liệu khuyết thấp. Nên trong bài này ta sẽ chọn cách xóa các quan sát có chứa dữ liệu khuyết.

❖ Xóa dữ liệu khuyết trong tệp tin "new_DF".

```
new_DF <- na.omit(new_DF) # xoa quan sat chua NA va luu lai vao tep tin
apply(is.na(new_DF),2,sum) #kiem tra lai NA
head(new_DF,10) #xem 10 dong dau tien
```

```
 G1  G2  G3 studytime failures absences paid sex
  0   0   0         0        0        0    0   0

   G1 G2 G3 studytime failures absences paid sex
1   5  6  6         2        0        6   no   F
3   7  8 10         2        3       10  yes   F
4  15 14 15         3        0        2  yes   F
5   6 10 10         2        0        4  yes   F
7  12 12 11         2        0        0   no   M
8   6  5  6         2        0        6   no   F
10 14 15 15         2        0        0  yes   M
11 10  8  9         2        0        0  yes   F
12 10 12 12         3        0        4   no   F
13 14 14 14         1        0        2  yes   M
```

Hình 4: Kết quả sau khi xoá dữ liệu khuyết trong tệp tin "new_DF"

c. Trực quan hóa dữ liệu

❖ Đối với các biến liên tục G1, G2, G3 thì thực hiện thống kê mô tả và xuất kết quả dưới dạng bảng.

```
mean <- apply(new_DF[,c("G1","G2","G3")],2,mean) #tinh trung binh mau
sd <- apply(new_DF[,c("G1","G2","G3")],2,sd) #tinh do lech chuan hieu chinh
Q1 <- apply(new_DF[,c("G1","G2","G3")],2,quantile,probs=0.25) #tinh diem phan vi 1
median <- apply(new_DF[,c("G1","G2","G3")],2,median) #tinh trung vi
Q3 <- apply(new_DF[,c("G1","G2","G3")],2,quantile,probs=0.75) #tinh diem phan vi 3
min <- apply(new_DF[,c("G1","G2","G3")],2,min) #tinh gia tri nho nhat
max <- apply(new_DF[,c("G1","G2","G3")],2,max) #tinh gia tri lon nhat
t(data.frame(mean,sd,Q1,median,Q3,min,max)) #tao bang voi cac bien
```

```
                G1        G2        G3
mean    10.925641 10.717949 10.412821
sd       3.290886  3.737868  4.568962
Q1       8.000000  9.000000  8.000000
median  11.000000 11.000000 11.000000
Q3      13.000000 13.000000 13.750000
min      3.000000  0.000000  0.000000
max     19.000000 19.000000 20.000000
```

Hình 5: Kết quả tính thống kê mô tả cho các biến G1, G2, G3

❖ Đối với các biến phân loại studytime, failures, paid, sex thì tiến hành lập bảng thống kê số lượng cho từng biến.

– Thống kê số lượng cho biến studytime.

```
table(new_DF$studytime) #lap bang thong ke so luong cho cac phan loai
```

```
  1   2   3   4
105 194  64  27
```

Hình 6: Kết quả thống kê số lượng cho biến studytime.

Nhận xét: Dựa vào bảng thống kê ta có thể biết được: Số học sinh có thời gian tự học trên tuần ít hơn 2 giờ là 105 học sinh; Số học sinh có thời gian tự học trên tuần từ 2 - 5 giờ là 194 học sinh; Số học sinh có thời gian tự học trên tuần từ 5 - 10 giờ là 64 học sinh; Số học sinh có thời gian tự học trên tuần lớn hơn 10 giờ là 27 học sinh.

- Thống kê số lượng cho biến failures.

```
table(new_DF$failures) #lap bang thong ke so luong cho cac phan loai
```

```
  0   1   2   3
307  50  17  16
```

Hình 7: Kết quả thống kê số lượng cho biến failures.

Nhận xét: Dựa vào bảng thống kê ta có thể biết được: Số học sinh có 1 lần không qua môn là 307 học sinh; Số học sinh có 2 lần không qua môn là 50 học sinh; Số học sinh có 3 lần không qua môn là 17 học sinh; Số học sinh có 4 lần hoặc hơn 4 lần không qua môn là 16 học sinh.

- Thống kê số lượng cho biến paid.

```
table(new_DF$paid) #lap bang thong ke so luong cho cac phan loai
```

```
 no yes
212 178
```

Hình 8: Kết quả thống kê số lượng cho biến paid

Nhận xét: Dựa vào bảng thống kê ta có thể biết được: Số học sinh không tham gia các lớp học thêm môn Toán ngoài trường là 212 học sinh; Số học sinh có tham gia các lớp học thêm môn Toán ngoài trường là 178 học sinh.

– Thống kế số lượng cho biến sex.

```
table(new_DF$sex) #lap bang thong ke so luong cho cac phan loai
```

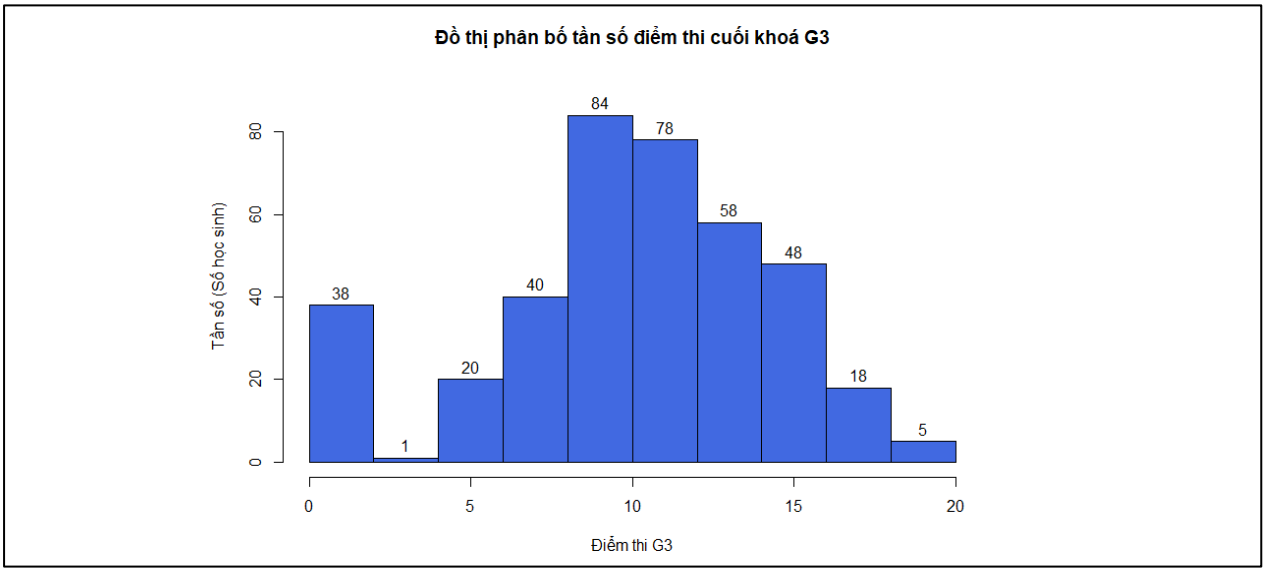
```
  F   M
205 185
```

Hình 9: Kết quả thống kê số lượng cho biến sex

Nhận xét: Dựa vào bảng thống kê ta có thể biết được: Số học sinh có giới tính nữ là 205 học sinh; Số học sinh có giới tính nam là 185 học sinh.

❖ Vẽ đồ thị Histogram cho biến G3.

```
hist(new_DF$G3,main ="Đồ thị phân bố tần số điểm thi cuối khoá G3",xlab ="Điểm thi G3",ylab ="Tần số (Số học sinh)",label =T,ylim=c(0,90),col ="royalblue")
```

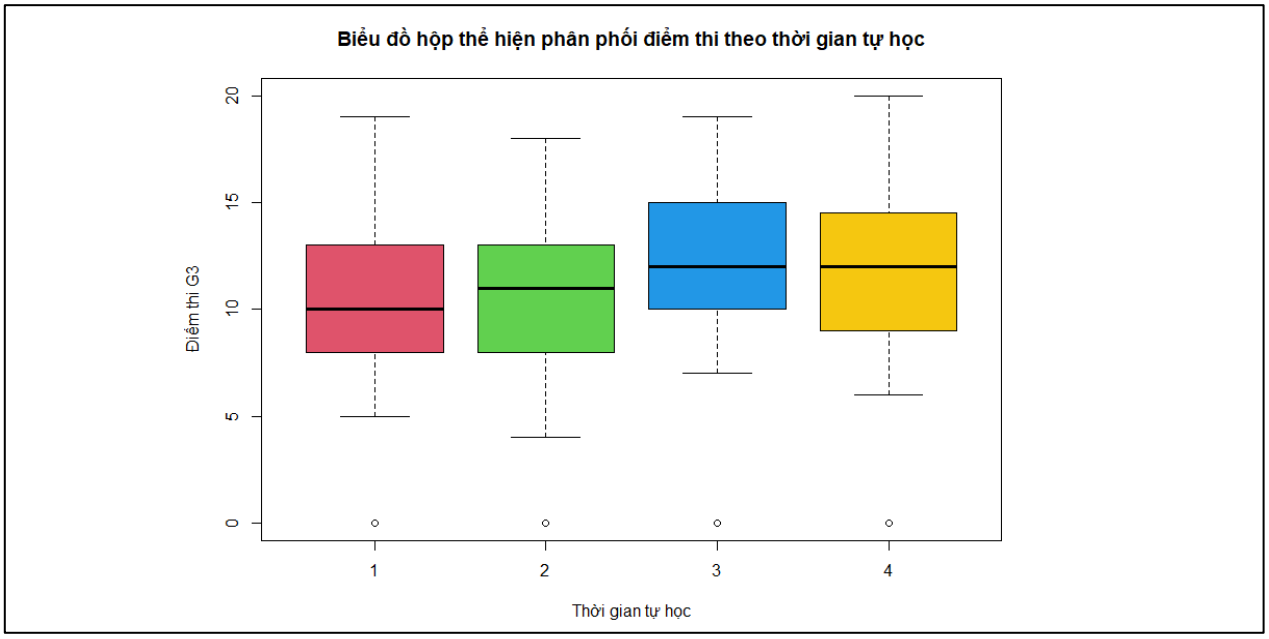


Hình 10: Đồ thị Histogram cho biến G3

Nhận xét: Dựa vào đồ thị, ta nhận thấy điểm cuối khoá của học sinh tập trung phần lớn ở mức từ khoảng 6 - 16 điểm, cao nhất ở mức 8 - 10 điểm (84 học sinh) và thấp nhất ở mức 2 - 4 điểm (1 học sinh). Điểm bất thường của đồ thị là số lượng học sinh ở mức 0 - 2 điểm chiếm số lượng khá lớn (38 học sinh), điều này gây ảnh hưởng không tốt đến mô hình hồi quy sắp xây dựng.

❖ Vẽ đồ thị Boxplot thể hiện phân phối của biến G3 theo phân loại của biến studytime.

```
boxplot(G3~studytime,data=new_DF,main="Biểu đồ hộp thể hiện phân phối điểm thi theo thời gian tự học",xlab="Thời gian tự học", ylab="Điểm thi G3", col=c(2,3,4,6))
```



Hình 11: Đồ thị Boxplot thể hiện phân phối của biến G3 theo phân loại của biến studytime

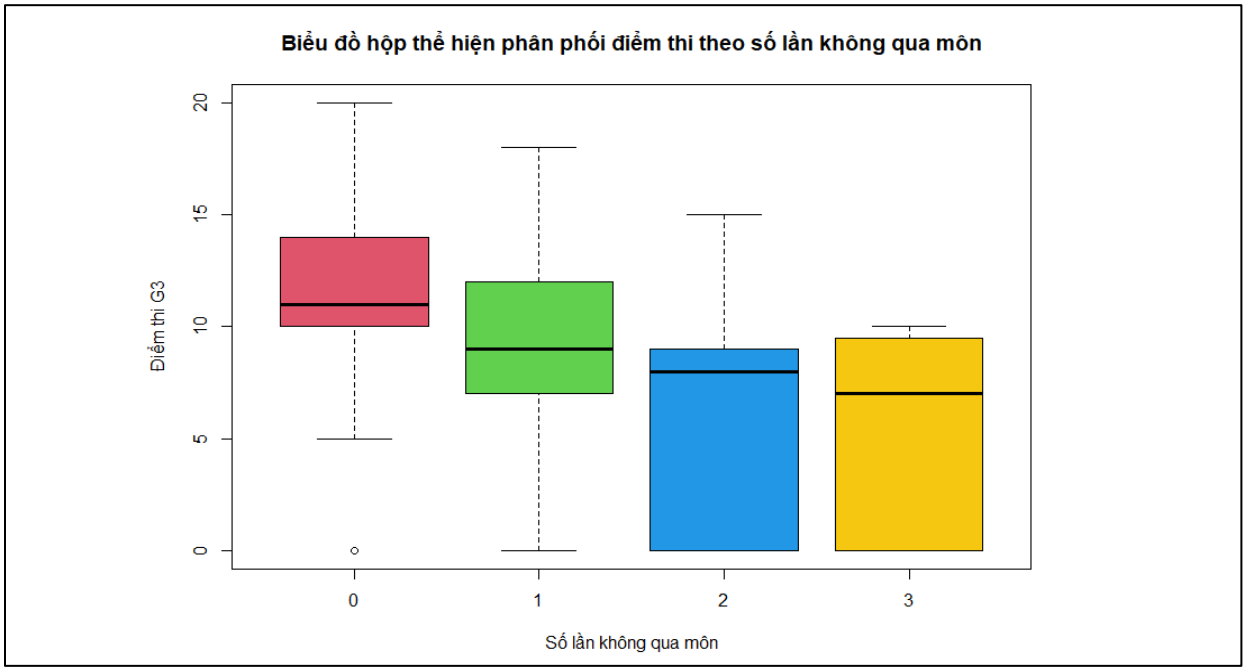
Nhận xét:

- Nhóm học sinh có thời gian tự học trên tuần ít hơn 2 giờ.
  - Có 25% học sinh có điểm cuối khóa từ 8 điểm trở xuống.
  - Có 50% học sinh có điểm cuối khóa từ 10 điểm trở xuống.
  - Có 75% học sinh có điểm cuối khóa từ 13 điểm trở xuống.
- Nhóm học sinh có thời gian tự học trên tuần từ 2 - 5 giờ.
  - Có 25% học sinh có điểm cuối khóa từ 8 điểm trở xuống.
  - Có 50% học sinh có điểm cuối khóa từ 11 điểm trở xuống.
  - Có 75% học sinh có điểm cuối khóa từ 13 điểm trở xuống.
- Nhóm học sinh có thời gian tự học trên tuần từ 5 - 10 giờ.
  - Có 25% học sinh có điểm cuối khóa từ 10 điểm trở xuống.
  - Có 50% học sinh có điểm cuối khóa từ 12 điểm trở xuống.
  - Có 75% học sinh có điểm cuối khóa từ 15 điểm trở xuống.
- Nhóm học sinh có thời gian tự học trên tuần lớn hơn 10 giờ.
  - Có 25% học sinh có điểm cuối khóa từ 9 điểm trở xuống.
  - Có 50% học sinh có điểm cuối khóa từ 12 điểm trở xuống.
  - Có 75% học sinh có điểm cuối khóa từ 14.5 điểm trở xuống.
- Qua các dữ liệu về điểm cuối khóa theo phân loại thời gian tự học trên tuần, có thể dự đoán được nhóm có thời gian tự học trên tuần ít hơn 2 giờ có kết quả thi kém hơn so với các nhóm còn lại

do có khoảng phân bố giá trị điểm thi thấp hơn. Nhóm có thời gian tự học trên tuần từ 5 - 10 giờ có kết quả thi tốt hơn so với các nhóm còn lại do có khoảng phân bố giá trị điểm thi cao hơn.

❖ Vẽ đồ thị Boxplot thể hiện phân phối của biến G3 theo phân loại của biến failures.

```
boxplot(G3~failures,data=new_DF,main="Biểu đồ hộp thể hiện phân phối điểm thi theo số lần không qua môn",xlab="Số lần không qua môn", ylab="Điểm thi G3",col=c(2,3,4,7))
```



Hình 12: Đồ thị Boxplot thể hiện phân phối của biến G3 theo phân loại của biến failures

Nhận xét:

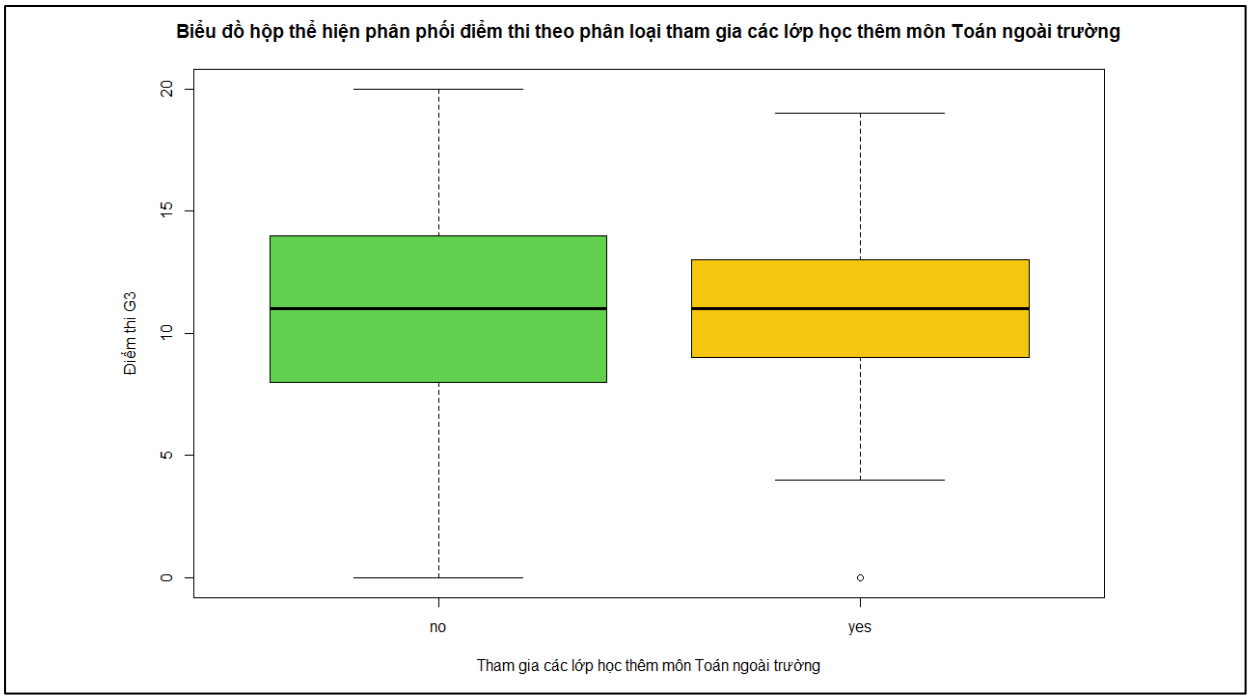
- Nhóm học sinh có 1 lần không qua môn.
  - Có 25% học sinh có điểm cuối khóa từ 10 điểm trở xuống.
  - Có 50% học sinh có điểm cuối khóa từ 11 điểm trở xuống.
  - Có 75% học sinh có điểm cuối khóa từ 14 điểm trở xuống.
- Nhóm học sinh có 2 lần không qua môn.
  - Có 25% học sinh có điểm cuối khóa từ 7 điểm trở xuống.
  - Có 50% học sinh có điểm cuối khóa từ 9 điểm trở xuống.
  - Có 75% học sinh có điểm cuối khóa từ 12 điểm trở xuống.
- Nhóm học sinh có 3 lần không qua môn.
  - Có 50% học sinh có điểm cuối khóa từ 8 điểm trở xuống.
  - Có 75% học sinh có điểm cuối khóa từ 9 điểm trở xuống.
- Nhóm học sinh có 4 lần hoặc hơn 4 lần không qua môn.
  - Có 50% học sinh có điểm cuối khóa từ 7 điểm trở xuống.

- Có 75% học sinh có điểm cuối khóa từ 10.5 điểm trở xuống.

- Qua các dữ liệu về điểm cuối khóa theo phân loại số lần không qua môn, có thể dự đoán được nhóm có 1 lần không qua môn có kết quả thi cao hơn so với các nhóm còn lại do có khoảng phân bố điểm thi cao hơn. Nhóm có 4 hoặc nhiều hơn 4 lần không qua môn có kết quả thi thấp hơn so với các nhóm còn lại do có khoảng phân bố điểm thi thấp hơn. Điều này cho thấy học sinh có số lần không qua môn càng nhiều thì điểm cuối khóa sẽ càng thấp.

❖ Vẽ đồ thị Boxplot thể hiện phân phối của biến G3 theo phân loại của biến paid.

```
boxplot(G3~paid,data=new_DF,main="Biểu đồ hộp thể hiện phân phối điểm thi theo phân loại tham gia các lớp học thêm môn Toán ngoài trường",xlab="Tham gia các lớp học thêm môn Toán ngoài trường", ylab="Điểm thi G3",col=c(3,7))
```



Hình 13: Đồ thị Boxplot thể hiện phân phối của biến G3 theo phân loại của biến paid

Nhận xét:

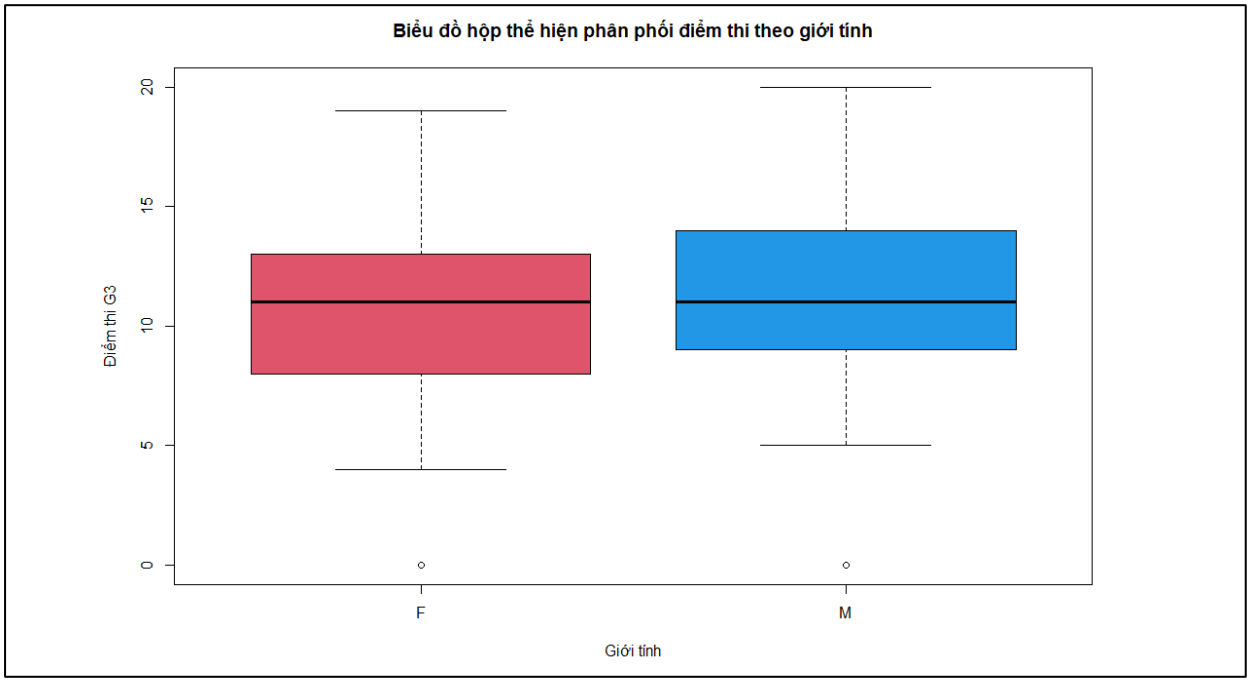
- Nhóm học sinh không tham gia các lớp học thêm môn Toán ngoài trường.
  - Có 25% học sinh có điểm cuối khóa từ 8 điểm trở xuống.
  - Có 50% học sinh có điểm cuối khóa từ 11 điểm trở xuống.
  - Có 75% học sinh có điểm cuối khóa từ 14 điểm trở xuống.
- Nhóm học sinh có tham gia các lớp học thêm môn Toán ngoài trường.
  - Có 25% học sinh có điểm cuối khóa từ 9 điểm trở xuống.
  - Có 50% học sinh có điểm cuối khóa từ 11 điểm trở xuống.

- Có 75% học sinh có điểm cuối khóa từ 13 điểm trở xuống.

- Qua các dữ liệu về điểm cuối khóa theo phân loại có hay không tham gia các lớp học thêm môn Toán ngoài trường, có thể nhận thấy khoảng phân bố điểm thi của hai nhóm nhau là gần như nhau. Tuy nhiên, nhóm có tham gia các lớp học thêm môn Toán ngoài trường có độ dao động điểm cuối khóa nhỏ hơn so với nhóm không tham gia. Điều này chứng tỏ học sinh có tham gia các lớp học môn Toán ngoài trường sẽ có điểm cuối khóa chệch lệch nhau không nhiều so với học sinh không tham gia các lớp học môn Toán ngoài trường.

❖ Vẽ đồ thị Boxplot thể hiện phân phối của biến G3 theo phân loại của biến sex.

```
boxplot(G3~sex,data=new_DF,main="Biểu đồ hộp thể hiện phân phối điểm thi theo giới tính",xlab="Giới tính", ylab="Điểm thi G3",col=c(2,4))
```



Hình 14: Đồ thị Boxplot thể hiện phân phối của biến G3 theo phân loại của biến sex

Nhận xét:

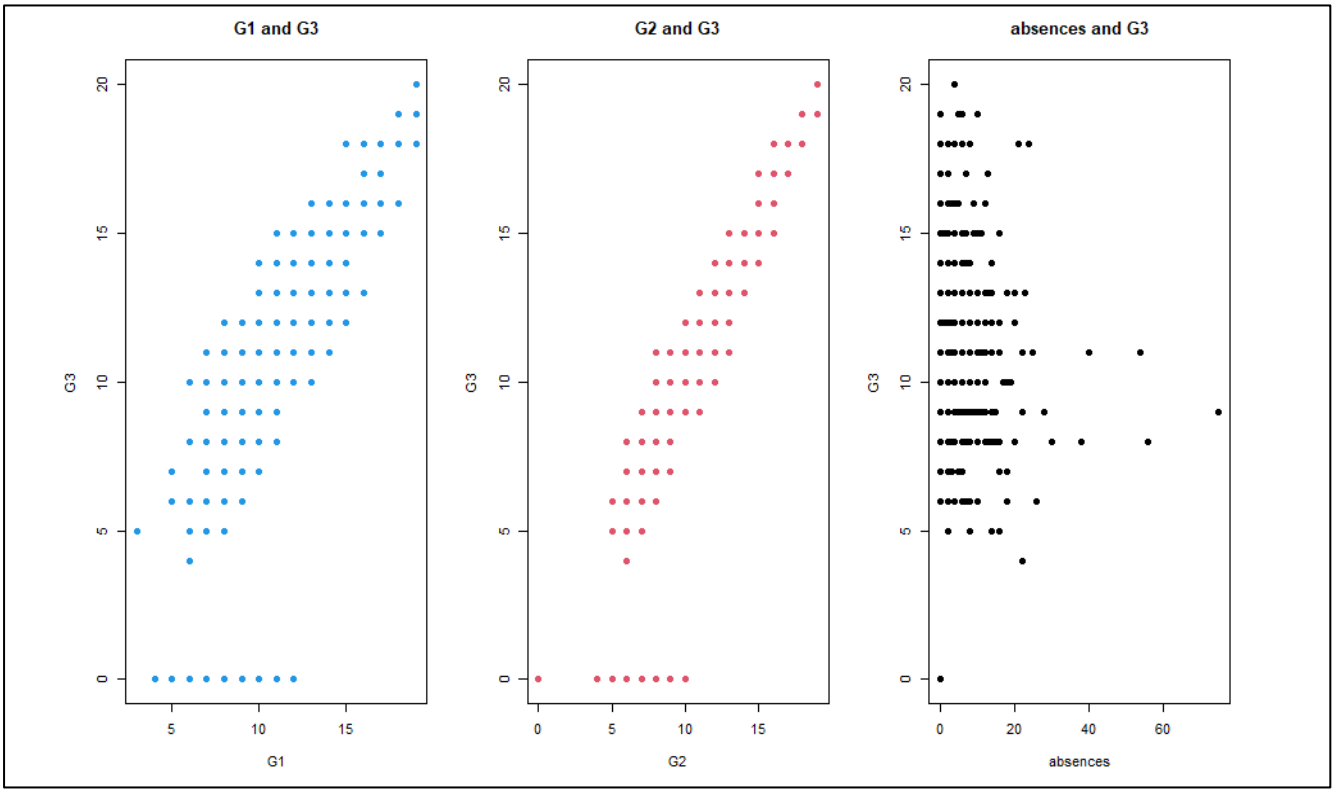
- Nhóm học sinh có giới tính nữ.
  - Có 25% học sinh có điểm cuối khóa từ 8 điểm trở xuống.
  - Có 50% học sinh có điểm cuối khóa từ 11 điểm trở xuống.
  - Có 75% học sinh có điểm cuối khóa từ 13 điểm trở xuống.
- Nhóm học sinh có giới tính nam.
  - Có 25% học sinh có điểm cuối khóa từ 9 điểm trở xuống.
  - Có 50% học sinh có điểm cuối khóa từ 11 điểm trở xuống.

- Có 75% học sinh có điểm cuối khóa từ 14 điểm trở xuống.

- Qua các dữ liệu về điểm cuối khóa theo phân loại giới tính, có thể nhận thấy khoảng phân bố điểm thi của nhóm học sinh nam cao hơn nhóm học sinh nữ. Nên ta có thể dự đoán được điểm thi cuối khóa của nhóm học sinh nam cao hơn nhóm học sinh nữ.

❖ Vẽ đồ thị phân tán thể hiện phân phối của biến G3 theo biến G1, G2, absences.

```
par(mfrow=c(1,3)) #Xep 3 bieu do vao 1 hang
plot(G3~G1,data=new_DF,main ="G1 and G3",col =4,pch=16) #Ve do thi phan tan cua G3 theo G1
plot(G3~G2,data=new_DF,main ="G2 and G3",col=2,pch=16) #Ve do thi phan tan cua G3 theo G2
plot(G3~absences,data=new_DF,main ="absences and G3",col=1,pch=16) #Ve do thi phan tan cua G3 theo absences
```



Hình 15: Đồ thị phân tán thể hiện phân phối của biến G3 theo biến G1, G2, absences

Nhận xét: Từ các đồ thị phân tán, ta có thể nhận xét rằng các biến G1, G2 có mối quan hệ tuyến tính với biến G3 (nói rõ hơn là mối quan hệ đồng biến vì G1, G2 tăng thì G3 cũng tăng theo), tuy nhiên với biến absences lại không có quan hệ tuyến tính với biến G3.

3. Xây dựng mô hình hồi quy tuyến tính để đánh giá các nhân tố có thể ảnh hưởng đến điểm thi cuối kỳ của học sinh

Ta xây dựng hình hồi quy bội (gọi là mô hình 1) bao gồm:

- Biến phụ thuộc: G3.

- Biến dư báo (biến độc lập): G1, G2, studytime, failures, absences, paid, sex.

Mô hình được biểu diễn như sau: $G3 = \beta_0 + \beta_1 \times G_1 + \beta_2 \times G_2 + \beta_3 \times studytime + \beta_4 \times failures + \beta_5 \times absences + \beta_6 \times paid + \beta_7 \times sex$

Ta thực hiện ước lượng các hệ số β_i, i = 0, ..., 7.

```
model_1 <- lm(G3~G1 + G2 + studytime + failures + absences +paid + sex,new_DF) #Xay dung mo hinh 1 va luu voi ten model_1
summary(model_1) #Ket qua mo hinh 1
```

```
Call:
lm(formula = G3 ~ G1 + G2 + studytime + failures + absences +
    paid + sex, data = new_DF)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2309 -0.3952  0.2652  0.9791  3.6101

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.68296    0.46944  -3.585 0.000381 ***
G1           0.15705    0.05706   2.752 0.006197 **
G2           0.97128    0.04995  19.443  < 2e-16 ***
studytime   -0.15211    0.12551  -1.212 0.226300
failures    -0.26358    0.14378  -1.833 0.067559 .
absences     0.03769    0.01220   3.088 0.002161 **
paidyes      0.12889    0.20324   0.634 0.526358
sexM         0.19834    0.20819   0.953 0.341355
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.919 on 382 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8236
F-statistic: 260.5 on 7 and 382 DF,  p-value: < 2.2e-16
```

Hình 16: Kết quả mô hình hồi quy tuyến tính model_1

Nhận xét: Từ kết quả phân tích, ta thu được:

$$\widehat{\beta_0} = -1.68296;\ \widehat{\beta_1} = 0.15705;\ \widehat{\beta_2} = 0.97128;\ \widehat{\beta_3} = -0.15211;\ \widehat{\beta_4} = -0.26358$$

$$\widehat{\beta_5} = 0.03769;\ \widehat{\beta_6} = 0.12889;\ \widehat{\beta_7} = 0.19834$$

Như vậy, đường thẳng hồi quy ước lượng cho bởi phương trình sau:

$$\widehat{G3} = -1.68296 + 0.15705 \times G1 + 0.97128 \times G2 - 0.15211 \times studytime - 0.26358 \times failures + 0.03769 \times absences + 0.12889 \times paidyes + 0.19834 \times sexM.$$

Kiểm định hệ số hồi quy:

- Giả thuyết H_0: Hệ số hồi quy không có ý nghĩa thống kê ($\beta_i = 0$)
- Giả thuyết H_1: Hệ số hồi quy có ý nghĩa thống kê ($\beta_i \neq 0$)

+ $Pr(>|t|)$ của các hệ số ứng với biến G1, G2, absences bé hơn mức ý nghĩa $\alpha = 0.05$ nên ta bác bỏ giả thuyết H_0, chấp nhận giả thuyết H_1. Do đó hệ số ứng với các biến này có ý nghĩa với mô hình hồi quy ta xây dựng.

+ $Pr(>|t|)$ của các hệ số ứng với biến studytime, failures, paidyes, sexM lớn hơn mức ý nghĩa α = 0.05 nên ta chưa đủ cơ sở để bác bỏ giả thuyết H_0. Do đó hệ số ứng với các biến này không có ý nghĩa với mô hình hồi quy ta xây dựng. Ta có thể cân nhắc việc loại bỏ các biến này ra khỏi mô hình.

Ta xây dựng thêm mô hình hồi quy tuyến tính.

Mô hình 2: Loại bỏ đi biến studytime, failures, paid, sex từ mô hình 1.

```
model_2 <- lm(G3 ~ G1 + G2 + absences,new_DF) #Xay dung mo hinh 2 da loai bo bien studytime, failures, paid, sex tu mo hinh 1 va luu voi ten model_2
summary(model_2) #Ket qua mo hinh 2
```

```
Call:
lm(formula = G3 ~ G1 + G2 + absences, data = new_DF)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3471 -0.3582  0.3133  0.9811  3.9465

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.12101    0.34769  -6.100 2.57e-09 ***
G1           0.15971    0.05615   2.844  0.00469 **
G2           0.98711    0.04943  19.968  < 2e-16 ***
absences     0.03660    0.01216   3.011  0.00277 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.925 on 386 degrees of freedom
Multiple R-squared:  0.8238,    Adjusted R-squared:  0.8224
F-statistic: 601.6 on 3 and 386 DF,  p-value: < 2.2e-16
```

Hình 17: Kết quả mô hình hồi quy tuyến tính model_2

So sánh giữa mô hình 1 và mô hình 2.

- Giả thuyết H_0: Mô hình 2 hiệu quả hơn.
- Giả thuyết H_1: Mô hình 1 hiệu quả hơn.

```
anova(model_1,model_2) #So sanh mo hinh 1 va mo hinh 2
```

```
Analysis of Variance Table

Model 1: G3 ~ G1 + G2 + studytime + failures + absences + paid + sex
Model 2: G3 ~ G1 + G2 + absences
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    382 1406.5
2    386 1430.7 -4   -24.217 1.6443 0.1624
```

Hình 18: Kết quả so sánh model_1 và model_2

Nhận xét: Vì *p_value* = 0.1624 lớn hơn mức ý nghĩa α = 0.05 nên chưa bác bỏ được giả thuyết H_0 => Mô hình 2 hiệu quả hơn do đó ta chọn mô hình 2.

Phân tích sự tác động của các nhân tố lên điểm thi cuối kỳ

Như vậy mô hình hồi quy tuyến tính về ảnh hưởng của các nhân tố lên điểm thi cuối kỳ G3 được cho bởi:

$$\widehat{G3} = -2.12101 + 0.15971 \times G1 + 0.98711 \times G2 + 0.03660 \times absences$$

Trước hết, ta thấy rằng p − value tương ứng với thống kê F bé hơn 2.2e − 16, có ý nghĩa rất cao. Điều này chỉ ra rằng, ít nhất một biến dự báo trong mô hình có ý nghĩa giải thích rất cao cho biến điểm thi cuối kỳ.

Để xét ảnh hưởng cụ thể của từng biến độc lập, ta xét trọng số (hệ số β_i) và p − value tương ứng. Ta thấy rằng p−value tương ứng với biến G2 bé hơn 2e−16, điều này nói lên rằng ảnh hưởng của G2 có ý nghĩa rất cao lên điểm thi cuối kỳ G3. Ngoài ra, biến G1, absence ít ảnh hưởng lên điểm thi cuối kỳ G3.

Mặt khác, hệ số hồi quy β_i của một biến dự báo cũng có thể được xem như ảnh hưởng trung bình lên biến phụ thuộc điểm thi cuối kỳ G3 khi tăng một đơn vị của biến dự báo đó, giả sử rằng các biến dự báo khác không đổi. Cụ thể, $\hat{\beta}_1$ = 0.15971 thì khi điểm thi học kỳ 1 tăng 1đ, ta có thể kỳ vọng điểm thi cuối kỳ sẽ tăng lên 0.15971đ về mặt trung bình (giả sử rằng các biến dự báo khác không đổi). Với $\hat{\beta}_2$ = 0.98711 thì khi điểm thi học kỳ 2 tăng 1đ, ta có thể kỳ vọng điểm thi cuối kỳ sẽ tăng lên 0.98711đ về mặt trung bình (giả sử rằng các biến dự báo khác không đổi). Tương tự, đối với các biến còn lại.

Hệ số R^2 hiệu chỉnh bằng 0.8224 nghĩa là 82.24% sự biến thiên trong điểm thi cuối kỳ được giải thích bởi các biến các biến độc lập.

Kiểm tra các giả định của mô hình

Nhắc lại các giả định của mô hình hồi quy: $Y_i = \beta_0 + \beta_1.X_1 + ...\beta_i .X_i + \epsilon_i$, i = 1, ...n.

+ Tính tuyến tính của dữ liệu: mối quan hệ giữa biến dự báo X và biến phụ thuộc Y được giả sử là tuyến tính.

+ Sai số có phân phối chuẩn

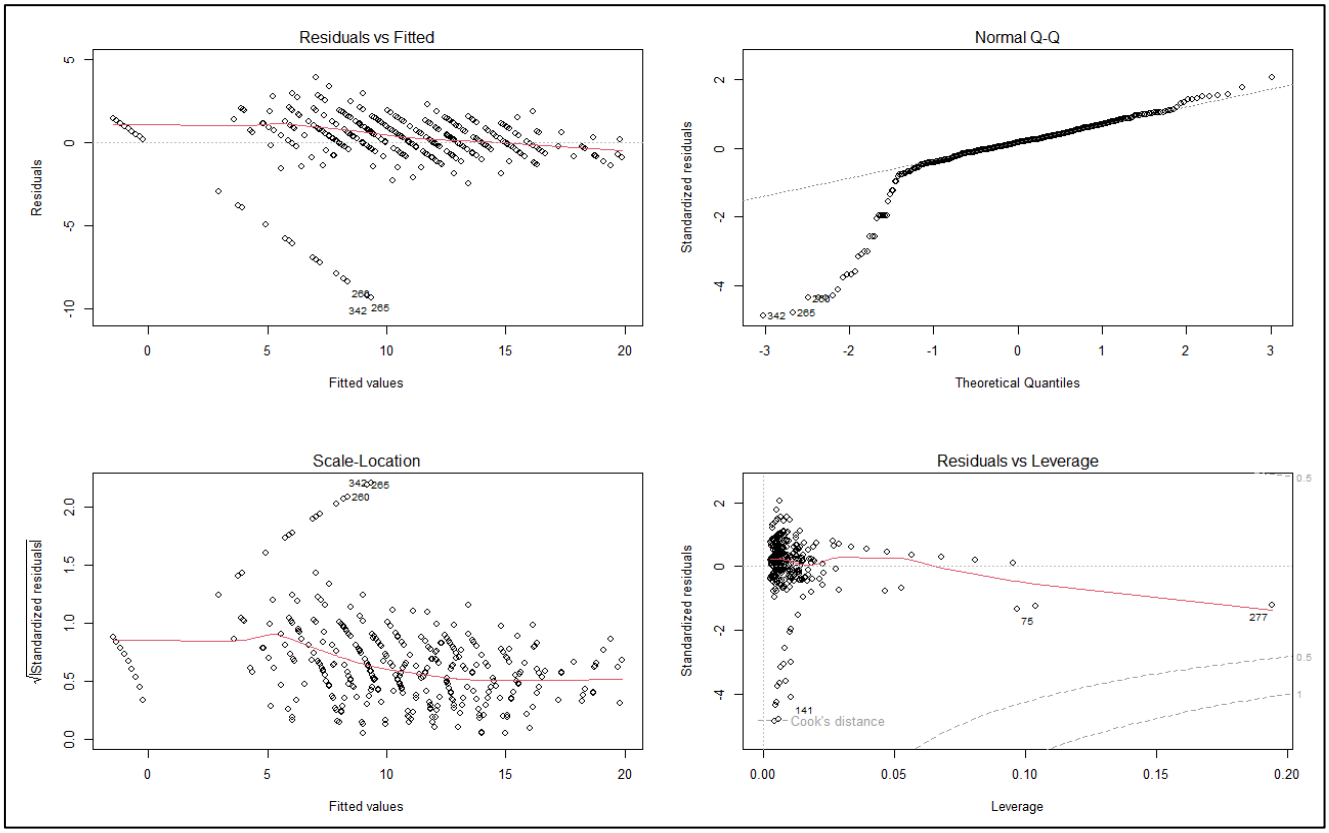
+ Phương sai của các sai số là hằng số.

+ Các sai số ϵ có kỳ vọng = 0.

+ Các sai số ϵ_1, ..., ϵ_n thì độc lập với nhau.

Ta thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình.

```
par(mfrow = c(2, 2)) #xep 4 bieu do thanh 2 hang 2 cot
plot(model_2) #ve do thi phan tich thang du
```



Hình 19: Kết quả khi vẽ các đồ thị phân tích thặng dư

Nhận xét:

- *Đồ thị 1 vẽ các giá trị sai số tương ứng với các giá trị dự báo, kiểm tra giả định tính tuyến tính của dữ liệu, giả định các sai số có kỳ vọng bằng 0 và giả định phương sai của các sai số là hằng số.*

  * Ta nhận thấy đường màu đỏ tương đối là đường thẳng nằm ngang nên giả định tính tuyến tính của dữ liệu thỏa mãn.

  * Ta nhận thấy đường màu đỏ nằm gần sát đường Residuals bằng 0 nên giả định các sai số có kỳ vọng bằng 0 thỏa mãn.

  * Các điểm sai số không phân tán ngẫu nhiên dọc theo đường màu đỏ nên giả định phương sai của các sai số là hằng số không thoả mãn.

- *Đồ thị 2 vẽ các sai số đã được chuẩn hóa kiểm tra giả định phân phối chuẩn.*

  Ta nhận thấy các sai số đã được chuẩn hóa chưa nằm trên một đường thẳng nên giả định này không thỏa mãn.

- *Đồ thị 3 vẽ căn bậc hai của các sai số đã được chuẩn hoá kiểm tra giả định phương sai của các sai số là hằng số.*

Ta nhận thấy đường màu đỏ không phải là đường thẳng nằm ngang và các sai số chưa phân tán ngẫu nhiên dọc theo đường màu đỏ nên giả định phương sai của các sai số là hằng số không thoả mãn.

- *Đồ thị 4 xác định các điểm có ảnh hưởng cao nếu nó hiện diện trong bộ dữ liệu.*
  Các điểm 141, 75, 277 là các điểm có khả năng gây ảnh hưởng cao trong bộ dữ liệu. Tuy nhiên các điểm này chưa vượt qua đường Cook's distance nên chúng chưa thật sự là những điểm có ảnh hưởng cao do đó không cần loại bỏ ra khỏi mô hình.

4. Thực hiện dự báo cho điểm toán của học sinh

Dự báo điểm thi cuối kỳ môn Toán nếu một học sinh có điểm thi học kỳ 1 là 15, điểm thi cuối học kỳ 2 là 12, học sinh này số lần nghỉ học là 2.

```
X = data.frame("G1"= 15,"G2"= 12,"absences"= 2)
predict(model_2,X,interval="confidence") #thuc hien du bao cho bien X bang model_2
```

```
       fit      lwr      upr
1 12.19306 11.78521 12.60091
```

Hình 20: Kết quả khi thực hiện dự báo

Nhận xét: Điểm thi cuối kỳ trung bình dự báo được là 12.19306đ, khoảng tin cậy cho giá trị dự báo (11.78521; 12.60091).

II. HOẠT ĐỘNG 2

Ở hoạt động này, nhóm chúng em quyết định khảo sát số lượng thuê xe công cộng trong tập dữ liệu Seoul Bike Sharing Demand

1. Mô tả bài toán

Hiện nay, dịch vụ thuê xe đạp công cộng đang được giới thiệu đến nhiều thành phố lớn trên thế giới để nâng cao sự thoải mái khi di chuyển. Điều quan trọng là phải cung cấp dịch vụ đúng thời điểm, đủ số lượng để giảm thời gian chờ. Vì vậy, việc cung cấp nguồn cung xe đạp cho thuê ổn định trở thành mối quan tâm chính ở các thành phố này. Ở hoạt động lần này, chúng ta sẽ dự đoán nhu cầu thuê xe đạp mỗi giờ để có nguồn cung xe đạp cho thuê ổn định.

Các biến có trong tập dữ liệu :

- Date : ngày/tháng/năm
- Rented Bike Count : số lượng thuê xe đạp mỗi giờ
- Hour : giờ trong ngày (giá trị là các số nguyên từ 0 đến 23)
- Temperature : nhiệt độ (°C)
- Humidity : độ ẩm (%)
- Windspeed : tốc độ gió (m/s)

- Visibility : tầm nhìn (10m)
- Dew point temperature : nhiệt độ điểm sương (°C)
- Solar radiation : bức xạ mặt trời (MJ/m^2)
- Rainfall : lượng mưa (mm)
- Snowfall : lượng tuyết (mm)
- Seasons : mùa trong năm (Spring. Summer, Autumn, Winter)
- Holiday : ngày lễ (Holiday và No holiday)
- Functioning day : ngày hoạt động (Yes và No)

2. Trực quan hóa dữ liệu

- Ta cùng đọc dữ liệu từ file SeoulBikeData.csv và xem qua 10 dòng đầu tiên của tập dữ liệu

```
df=read.csv("SeoulBikeData.csv")
head(df,10)
```

```
         Date Rented.Bike.Count Hour Temperature..C. Humidity... Wind.speed..m.s. Visibility..10m. Dew.point.temperature..C.
1  01/12/2017               254    0           -5.2          37              2.2            2000                      -17.6
2  01/12/2017               204    1           -5.5          38              0.8            2000                      -17.6
3  01/12/2017               173    2           -6.0          39              1.0            2000                      -17.7
4  01/12/2017               107    3           -6.2          40              0.9            2000                      -17.6
5  01/12/2017                78    4           -6.0          36              2.3            2000                      -18.6
6  01/12/2017               100    5           -6.4          37              1.5            2000                      -18.7
7  01/12/2017               181    6           -6.6          35              1.3            2000                      -19.5
8  01/12/2017               460    7           -7.4          38              0.9            2000                      -19.3
9  01/12/2017               930    8           -7.6          37              1.1            2000                      -19.8
10 01/12/2017               490    9           -6.5          27              0.5            1928                      -22.4
   Solar.Radiation..MJ.m2. Rainfall.mm. Snowfall..cm. Seasons    Holiday Functioning.Day
1                     0.00            0            0  Winter No Holiday             Yes
2                     0.00            0            0  Winter No Holiday             Yes
3                     0.00            0            0  Winter No Holiday             Yes
4                     0.00            0            0  Winter No Holiday             Yes
5                     0.00            0            0  Winter No Holiday             Yes
6                     0.00            0            0  Winter No Holiday             Yes
7                     0.00            0            0  Winter No Holiday             Yes
8                     0.00            0            0  Winter No Holiday             Yes
9                     0.01            0            0  Winter No Holiday             Yes
10                    0.23            0            0  Winter No Holiday             Yes
```

Hình 21: Kết quả khi xem 10 dòng đầu tiên của file SeoulBikeData

- Ta tiếp tục thực hiện các thống kê mô tả trên tập dữ liệu

```
summary(df)
```

```
    Date           Rented.Bike.Count      Hour       Temperature..C.    Humidity...    Wind.speed..m.s. Visibility..10m.
Length:8760        Min.   :   0.0    Min.   : 0.00   Min.   :-17.80   Min.   : 0.00   Min.   :0.000    Min.   :  27
Class :character   1st Qu.: 191.0    1st Qu.: 5.75   1st Qu.:  3.50   1st Qu.:42.00   1st Qu.:0.900    1st Qu.: 940
Mode  :character   Median : 504.5    Median :11.50   Median : 13.70   Median :57.00   Median :1.500    Median :1698
                   Mean   : 704.6    Mean   :11.50   Mean   : 12.88   Mean   :58.23   Mean   :1.725    Mean   :1437
                   3rd Qu.:1065.2    3rd Qu.:17.25   3rd Qu.: 22.50   3rd Qu.:74.00   3rd Qu.:2.300    3rd Qu.:2000
                   Max.   :3556.0    Max.   :23.00   Max.   : 39.40   Max.   :98.00   Max.   :7.400    Max.   :2000
Dew.point.temperature..C. Solar.Radiation..MJ.m2.  Rainfall.mm.      Snowfall..cm.       Seasons            Holiday
Min.   :-30.600           Min.   :0.0000          Min.   : 0.0000   Min.   :0.00000   Length:8760        Length:8760
1st Qu.: -4.700           1st Qu.:0.0000          1st Qu.: 0.0000   1st Qu.:0.00000   Class :character   Class :character
Median :  5.100           Median :0.0100          Median : 0.0000   Median :0.00000   Mode  :character   Mode  :character
Mean   :  4.074           Mean   :0.5691          Mean   : 0.1487   Mean   :0.07507
3rd Qu.: 14.800           3rd Qu.:0.9300          3rd Qu.: 0.0000   3rd Qu.:0.00000
Max.   : 27.200           Max.   :3.5200          Max.   :35.0000   Max.   :8.80000
Functioning.Day
Length:8760
Class :character
```

Hình 22: Kết quả khi thực hiện các thống kê mô tả trên tập dữ liệu

Qua quan sát, chúng ta có thể thấy được rằng tập dữ liệu bao gồm 8760 quan sát, 14 thuộc tính trong đó bao gồm 9 biến số và 4 biến phân loại (Hour, Seasons, Holiday và Functioning Day)

- Ta tiến hành kiểm tra các dữ liệu khuyết của tập dữ liệu

```
apply(is.na(df),2,sum)
```

```
                    Date           Rented.Bike.Count                        Hour             Temperature..C. 
                       0                           0                           0                           0 
           Humidity...          Wind.speed..m.s.      Visibility..10m. Dew.point.temperature..C. 
                       0                           0                           0                           0 
Solar.Radiation..MJ.m2.              Rainfall.mm.         Snowfall..cm.                     Seasons 
                       0                           0                           0                           0 
               Holiday          Functioning.Day 
                       0                           0
```

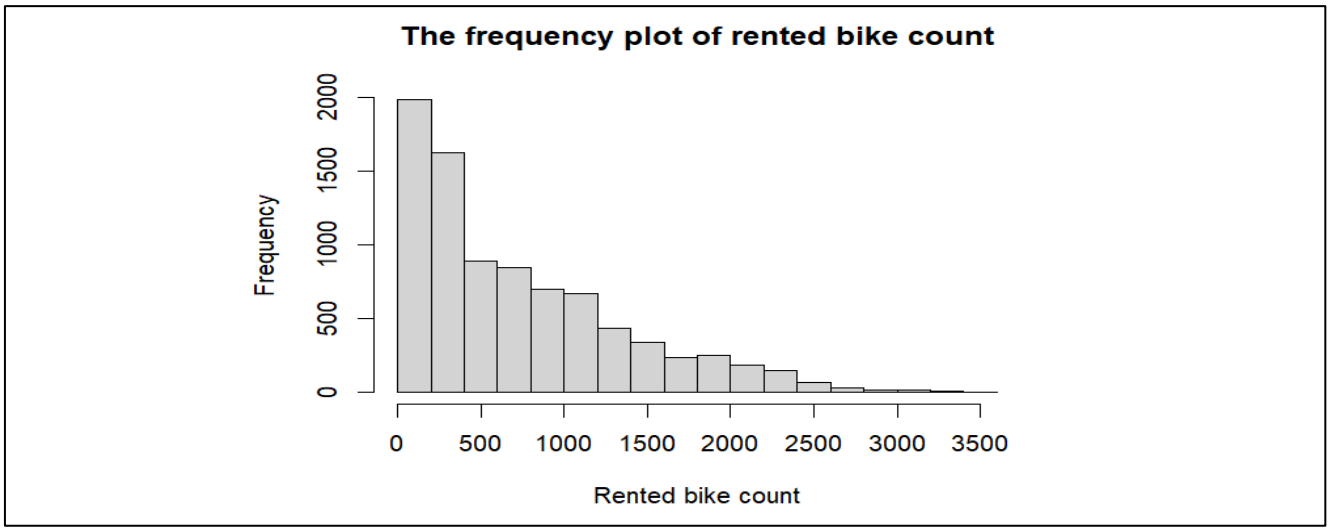
Hình 23: Kết quả khi kiểm tra các dữ liệu khuyết của tập dữ liệu

Ta nhận thấy không có dữ liệu khuyết trên tập dữ liệu, tuy nhiên biến Functioning Day mô tả ngày hoạt động tức có nghĩa khi nó có giá trị "No" thì giá trị của Rented Bike Count mặc định bằng 0 mà không cần quan tâm đến các biến khác, vậy nên để tăng độ chính xác trong quá trình trực quan hóa và xây dựng mô hình, ta sẽ loại bỏ những quan sát có giá trị Functioning Day là "No" .

```
df=df[df$Functioning.Day=="Yes",]
```

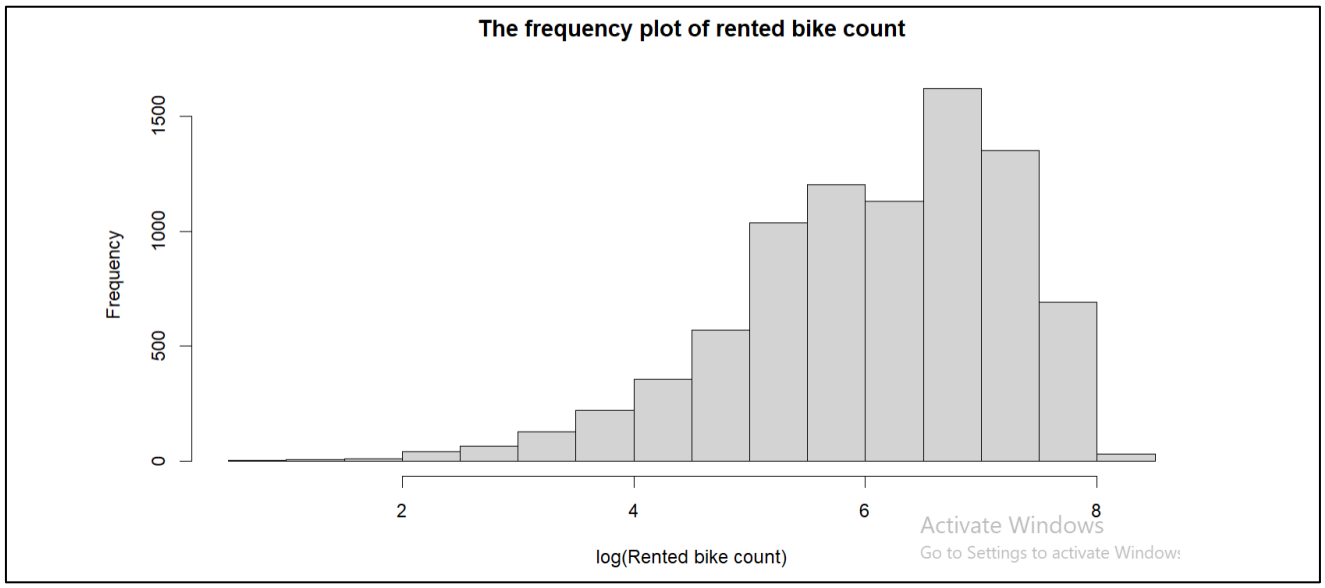
- Ta cùng xem sơ qua về lượng thuê xe đạp – đối tượng chính của tập dữ liệu thông qua biểu đồ tần số bên dưới

```
hist(df$Rented.Bike.Count,main="The frequency plot of rented bike count",xlab="Rented bike count")
```



Hình 24: Biểu đồ tần số cho biến Rented Bike Count

Ta có thể thấy, số lượng thuê xe đạp dao động cao chủ yếu ở mức 0-500 xe/giờ và giảm dần ở các mức sau, nó không tuân theo phân phối chuẩn, ta xem xét thêm biểu đồ tần số logarit của số lượng xe cho thuê:

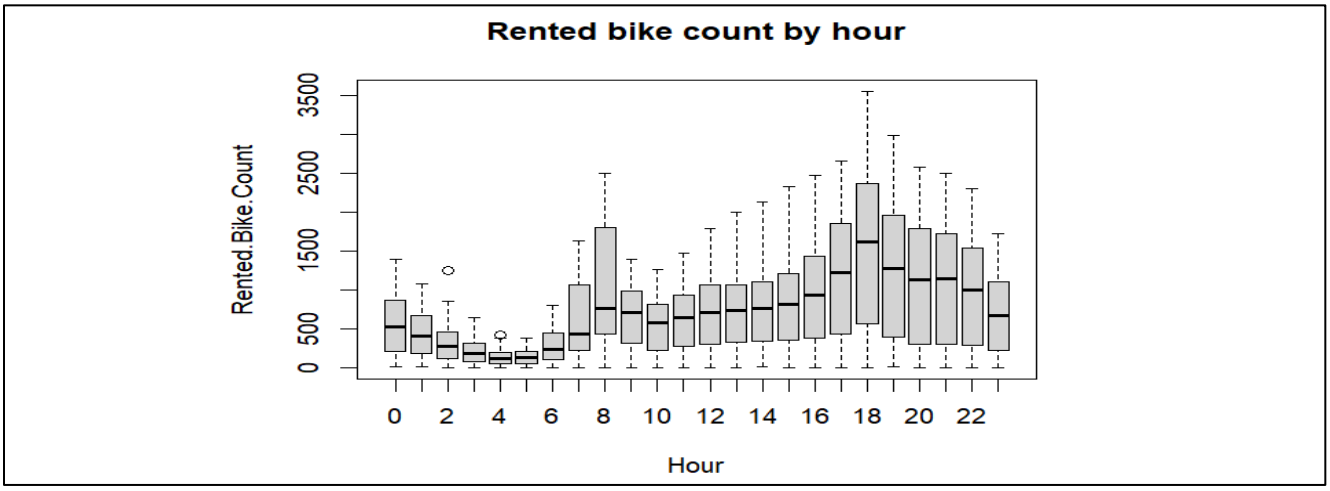


Hình 25: Biểu đồ tần số logarit của số lượng xe cho thuê

Biểu đồ thể hiện một phân phối lệch, dù vậy cũng khá tương đồng với phân phối chuẩn, phù hợp để huấn luyện mô hình.

- Ta cùng bắt đầu đi vào khảo sát biến Hour (giờ) thông qua biểu đồ hộp về lượng xe thuê theo giờ

```
boxplot(Rented.Bike.Count~Hour,data=df,main="Rented bike count by hour")
```

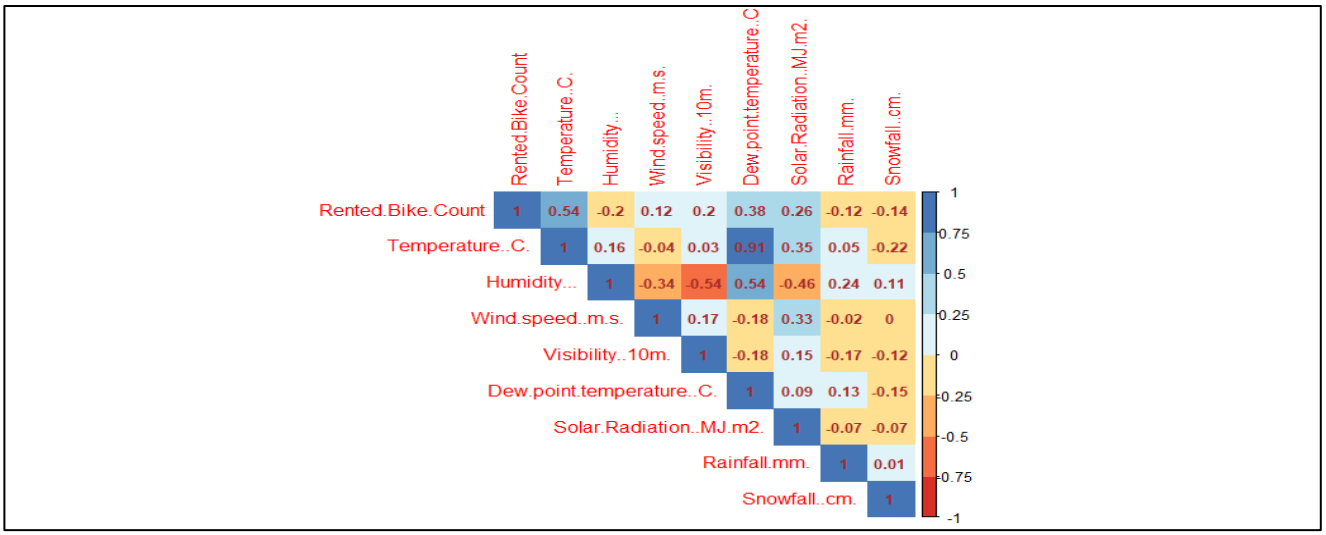


Hình 26: Biểu đồ hộp về lượng xe thuê theo giờ

Ta có thể quan sát thấy, lượng thuê xe thay đổi theo giờ và phân bố không đều giữa các khoảng thời gian, lượng thuê xe tăng nhanh từ thời điểm 4h đến 8h sáng sau đó giảm và tiếp tục tăng đến 6h chiều, đây cũng là thời điểm có lượng thuê xe cao nhất. Và lại tiếp tục giảm dần sau đó

- Ta tiếp tục khảo sát mối tương quan giữa các biến số thông qua biểu đồ tương quan nhằm loại bỏ một số biến không phù hợp

```
df1=df[,!names(df)%in%c("Seasons","Holiday","Functioning.Day","Date","Hour")]
corrplot(cor(df1),type="upper",tl.pos="td",method="color",addCoef.col =
"brown",number.cex=0.6,col=brewer.pal(n=8, name="RdYlBu"))
```

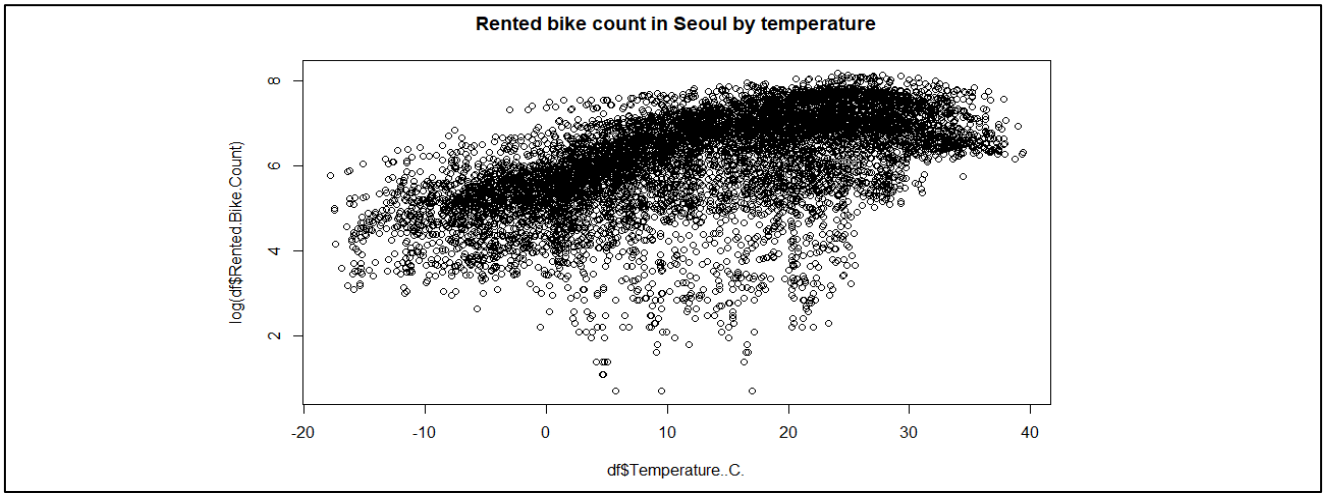


Hình 27: Biểu đồ tương quan

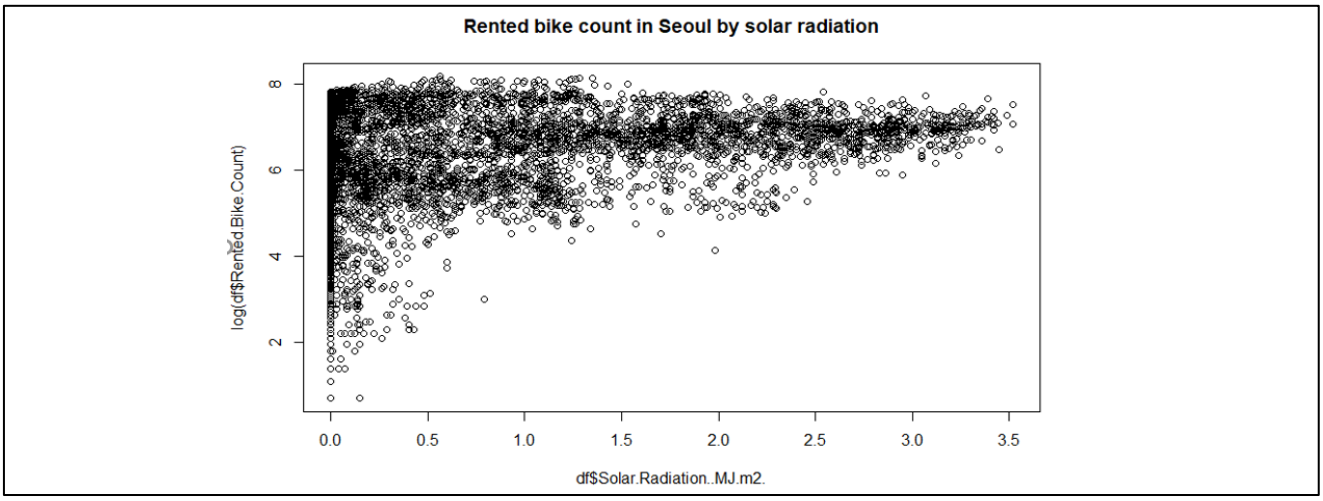
Quan sát biểu đồ, ta thấy rằng biến Temperature (nhiệt độ) và Dew point temperature (nhiệt độ điểm sương) có mối tương quan cao nhất là 0.91, vì vậy ta sẽ cân nhắc loại bỏ biến Dew point temperature khi xây dựng mô hình. Biến Humidity (độ ẩm) và Visibility (tầm nhìn) có mối tương quan nghịch cao nhất là -0.55 và cuối cùng là 2 biến không tương quan là Windspeed (tốc độ gió) và Snowfall (lượng tuyết).

Bên cạnh đó, biểu đồ trên cũng chỉ ra rằng 2 biến có mối tương quan thuận cao nhất với lượng thuê xe lần lượt là Temperature(nhiệt độ), và Solar radiation(bức xạ mặt trời). Ta cùng đi làm rõ mối tương quan giữa chung thông qua các biểu đồ phân tán bên dưới. Do sự chênh lệch về giá trị của lượng thuê xe so với các biến này nên để dễ quan sát, ta sẽ lấy logarit cơ số tự nhiên của biến Rented bike count.

```
plot(df$Temperature..C.,log(df$Rented.Bike.Count),main="Rented bike count in Seoul by
temperature")
plot(df$Solar.Radiation..MJ.m2.,log(df$Rented.Bike.Count),main="Rented bike count in Seoul by
solar radiation"
```



Hình 29: Biểu đồ phân tán của Temperature so với log (Rented bike count)

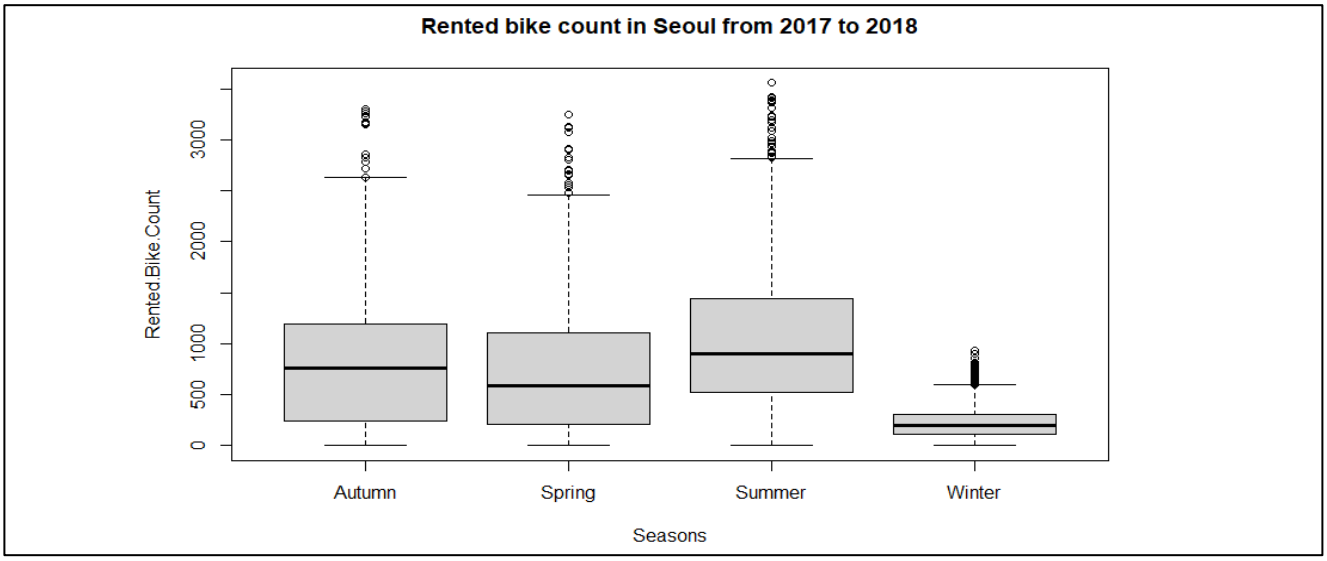


Hình 28: Biểu đồ phân tán của Solar radiation so với log (Rented bike count)

Biểu đồ phân tán của Temperature (nhiệt độ) so với log(Rented bike count) có mối quan hệ tuyến tính cao, biểu đồ phân tán của Solar radiation cho thấy sự tập trung ở gần giá trị 0 và có độ dốc tuyến tính nhỏ.

- Tiếp theo, ta cùng đi khảo sát biến chuỗi Season(mùa) thông qua biểu đồ hộp bên dưới

```
boxplot(Rented.Bike.Count~Seasons,data=df,main="Rented bike count in Seoul from 2017 to 2018")
```

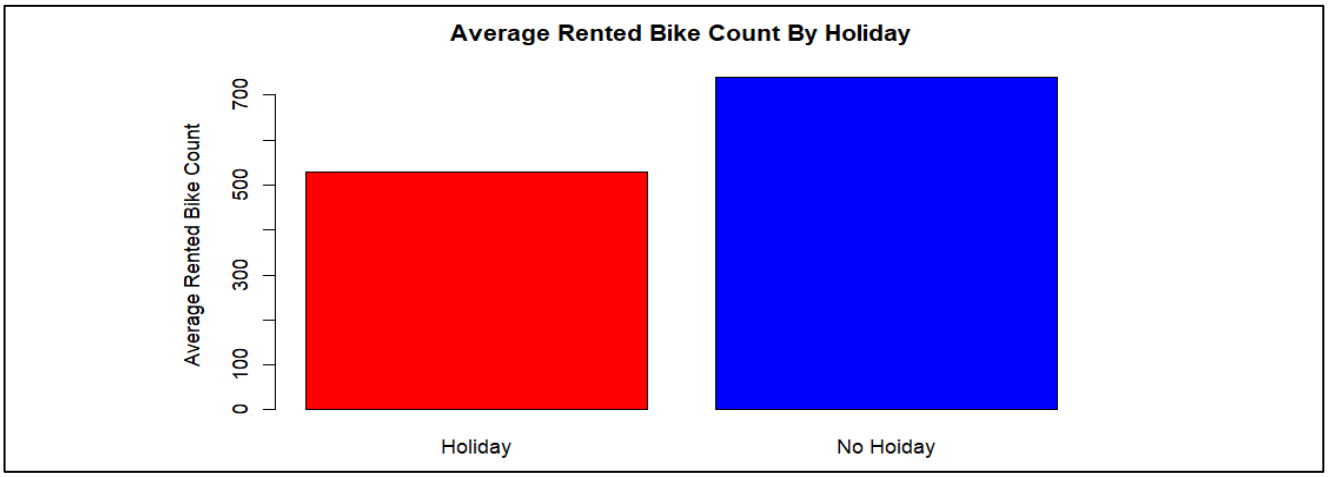


Hình 30: Biểu đồ hộp thể hiện phân phối của biến Rented bike count theo phân loại của biến Seasons

Qua quan sát ta có thể thấy rõ lượng chênh lệch lớn giữa số lượng thuê xe của mùa đông so với các mùa còn lại, đồng thời khoảng giá trị ở mùa đông cũng nhỏ hơn.

- Cuối cùng, ta cùng đi khảo sát biến chuỗi Holiday(ngày lễ) thông qua biểu đồ thanh về lượng thuê xe trung bình theo mùa giữa ngày lễ và ngày thường

```
Holiday=mean(df[df$Holiday=="Holiday",]$Rented.Bike.Count)
noHoliday=mean(df[df$Holiday=="No Holiday",]$Rented.Bike.Count)
barplot(c(Holiday,noHoliday),beside=T,col=c("red","blue"),main="Average Rented Bike Count By Holiday",names.arg=c("Holiday","No Hoiday"),ylab="Average Rented Bike Count")
```



Hình 31: Biểu đồ thanh về lượng thuê xe trung bình theo mùa giữa ngày lễ và ngày thường

Thông qua biểu đồ trên, ta có thể lượng thuê xe trung bình ở ngày lễ dao động ở mức trên 500 xe/giờ còn ngày thường là mức trên 700 xe/giờ, lượng thuê xe trung bình giữa ngày lễ là thấp hơn so với ngày thường .

3. Xây dựng mô hình

a) Chọn biến cho mô hình

Để dự đoán về số liệu của lượng thuê xe trong tương lai, chúng ta cùng đi xây dựng mô hình hồi quy tuyến tính bội với biến phụ thuộc Rented Bike Count và các biến độc lập sẽ được lựa chọn bên dưới.

Việc lựa chọn biến cho mô hình là một việc hết sức quan trọng vì nó giúp ta loại bỏ các biến dư thừa hoặc có mối tương quan yếu đối với biến phụ thuộc ta mà đang xét, những biến này không những không tăng thêm độ chính xác của mô hình mà còn có thể làm gia tăng sai số. Vì vậy, ta cần phải lựa chọn biến sao cho phù hợp với mô hình đang xét.

– Đầu tiên, ta cần loại bỏ các biến số có mối tương quan yếu đối với biến phụ thuộc Rented Bike Count. Cùng nhìn lại biểu đồ tương quan giữa các biến số được trình bày ở phần 21, ta tiến hành loại bỏ các biến có mối tương quan nằm trong khoảng (-0.2 ; 0.2) đối với biến phụ thuộc bao gồm : Windspeed (0.13), Rainfall (-0.13) và Snowfall (-0.15)

– Tiếp theo, ta loại bỏ các biến độc lập có độ tương quan cao đối với nhau như biến Dew point temperature (có hệ số tương quan 0.91 với biến temperature)

```
df=df[,!names(df)%in%c("Wind.speed..m.s.","Rainfall.mm.","Snowfall..cm.","Date","Functioning.
Day","Dew.point.temperature..C.")]
```

– Sau khi đã loại bỏ các biến không phù hợp, ta chia tập dữ liệu ra thành tập huấn luyện và tập kiểm thử với tỉ lệ 7:3 với thứ tự các quan sát ngẫu nhiên trước khi đi vào xây dựng mô hình. Việc phân chia tập dữ liệu như vậy cũng khá quan trọng trong việc xây dựng mô hình, tập huấn luyện chúng ta dùng để huấn luyện mô hình, trong khi đó tập kiểm thử dùng để kiểm tra lại độ chính xác của mô hình mà chúng ta vừa xây dựng dựa trên tập huấn luyện

```
set.seed(1)
split= sample(c(rep(0, 0.7 * nrow(df)), rep(1, 0.3 * nrow(df))))
train_data=df[split==0,]
test_data=df[split==1,]
```

b) Xây dựng mô hình

Lưu ý: khi xây dựng mô hình ta chọn biến phụ thuộc là logarit của Rented Bike Count do có phân phối xác suất tương đồng với phân phối chuẩn hơn

- Ta tiến hành xây dựng mô hình hồi quy tuyến tính thứ nhất :

$$log(Rented\ Bike\ Count) = \beta_0 + \beta_1 \times Hour + \beta_2 \times Temperature + \beta_3 \times Humidity + \beta_5 \times Visibility + \beta_6 \times Solar\ Radition + \beta_7 \times SeasonsSpring \times + \beta_8 \times SeasonsSummer + \beta_9 \times SeasonsWinter + \beta_{10} \times SeasonsAutumn + \beta_{11} \times HolidayHoliday + \beta_{12} \times HolidayNoHoliday$$

Với:

- Biến phụ thuộc : log(Rented Bike Count)
- Biến độc lập : Hour, Temperature, Humidity, Visibility, Solar Radiation, SeasonsSpring, SeasonsSummer, SeasonsWinter, SeasonsAutumn , HolidayHoliday, HolidayNoHoliday.

Lưu ý : Đối với các biến phân loại trong tập dữ liệu ban đầu như Seasons hay Holiday, khi xây dựng mô hình ta phải lượng hóa nó bằng cách chia nó thành các biến giả phân biệt tương ứng với các giá trị trong tập dữ liệu ban đầu của nó. Khi đó, ở một quan sát bất kì, giá trị ở quan sát đó tương ứng với biến giả nào thì biến giả đó sẽ mang giá trị 1 ngược lại sẽ mang giá trị 0.

Ví dụ: Khi Seasons nhận giá trị là Spring thì SeasonsSpring = 1, SeasonsWinter = SeasonsSummer = 0

Xây dựng mô hình:

```
model1=lm(log(Rented.Bike.Count)~.,data=train_data)
summary(model1)
```

```
Call:
lm(formula = log(Rented.Bike.Count) ~ ., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0843 -0.3528  0.0749  0.4646  2.4979

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.434e+00  9.778e-02  65.806  < 2e-16 ***
Hour                     3.896e-02  1.576e-03  24.724  < 2e-16 ***
Temperature..C.          4.292e-02  1.915e-03  22.412  < 2e-16 ***
Humidity...             -2.181e-02  7.955e-04 -27.422  < 2e-16 ***
Visibility..10m.         2.129e-05  2.172e-05   0.980    0.327
Solar.Radiation..MJ.m2. -6.431e-02  1.584e-02  -4.061 4.95e-05 ***
SeasonsSpring           -3.820e-01  3.046e-02 -12.542  < 2e-16 ***
SeasonsSummer           -2.893e-01  3.752e-02  -7.710 1.47e-14 ***
SeasonsWinter           -8.539e-01  4.329e-02 -19.726  < 2e-16 ***
HolidayNo Holiday        3.353e-01  4.868e-02   6.888 6.25e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7863 on 5916 degrees of freedom
Multiple R-squared:  0.5519,    Adjusted R-squared:  0.5512
F-statistic: 809.6 on 9 and 5916 DF,  p-value: < 2.2e-16
```

Hình 32: Kết quả mô hình hồi quy tuyến tính model1

Kiểm định các hệ số hồi quy :

- Giả thuyết H_0 : Hệ số hồi quy không có ý nghĩa thống kê ($\beta i = 0$)

- Giả thuyết H_1 : Hệ số hồi quy ý nghĩa thống kê ($\beta i \neq 0$)

Với :

- Estimate : giá trị ước tính các hệ số hồi quy βi
- Std. Error : độ lệch chuẩn của ước lượng hệ số βi
- t value : giá trị thống kê kiểm định hệ số hồi quy tương ứng
- Pr ($> |t|$) : giá trị p_{value}

Ta sẽ chọn mức ý nghĩa α = 5%

Khi đó, đối với các biến $Pr(> |t|)$ bé hơn mức ý nghĩa α = 5% ta bác bỏ giả thuyết H_0, chấp nhận giả thuyết H_1. Do đó hệ số ứng với các biến này có ý nghĩa với mô hình hồi quy ta xây dựng. Biến Visibility có $Pr(> |t|)$ lớn hơn mức ý nghĩa α = 5%, ta có thể cân nhắc loại bỏ khỏi mô hình.

Ta loại bỏ biến Visibility ra khỏi mô hình ban đầu để tiếp tục xây dựng mô hình thứ 2

```
train_data=train_data[,!names(train_data)%in%c("Visibility..10m.")]
test_data=test_data[,!names(test_data)%in%c("Visibility..10m.")]
```

- Ta xây dựng mô hình hồi quy thứ 2:

$$log(Rented\ Bike\ Count) = \beta_0 + \beta_1 \times Hour + \beta_2 \times Temperature + \beta_3 \times Humidity + \beta_5 \times Solar\ Radition + \beta_6 \times SeasonsSpring \times + \beta_7 \times SeasonsSummer + \beta_8 \times SeasonsWinter + \beta_9 \times HolidayNoHoliday$$

```
model2=lm(log(Rented.Bike.Count)~.,data=train_data_2)
summary(model2)
```

```
Call:
lm(formula = log(Rented.Bike.Count) ~ ., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0865 -0.3541  0.0764  0.4600  2.4912

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.4985798  0.0723790  89.785  < 2e-16 ***
Hour                     0.0388928  0.0015743  24.704  < 2e-16 ***
Temperature..C.          0.0428871  0.0019148  22.398  < 2e-16 ***
Humidity...             -0.0222745  0.0006415 -34.724  < 2e-16 ***
Solar.Radiation..MJ.m2. -0.0673311  0.0155330  -4.335 1.48e-05 ***
SeasonsSpring           -0.3883897  0.0297636 -13.049  < 2e-16 ***
SeasonsSummer           -0.2864277  0.0374073  -7.657 2.21e-14 ***
SeasonsWinter           -0.8617501  0.0425464 -20.254  < 2e-16 ***
HolidayNo Holiday        0.3338362  0.0486536   6.861 7.51e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7862 on 5917 degrees of freedom
Multiple R-squared:  0.5518,    Adjusted R-squared:  0.5512
F-statistic: 910.7 on 8 and 5917 DF,  p-value: < 2.2e-16
```

Hình 33: Kết quả mô hình hồi quy tuyến tính model2

- So sánh giữa mô hình 1 và mô hình 2.

  Giả thuyết H_0: Mô hình 2 hiệu quả hơn.

  Giả thuyết H_1: Mô hình 1 hiệu quả hơn.

```
anova(model1,model2)
```

```
Model 1: log(Rented.Bike.Count) ~ Hour + Temperature..C. + Humidity... +
    Visibility..10m. + Solar.Radiation..MJ.m2. + Seasons + Holiday
Model 2: log(Rented.Bike.Count) ~ Hour + Temperature..C. + Humidity... +
    Solar.Radiation..MJ.m2. + Seasons + Holiday
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   5916 3657.2
2   5917 3657.8 -1   -0.5938 0.9605 0.3271
```

Hình 34: Kết quả so sánh model1 và model2

Vì *p_value* = 0.3271 lớn hơn mức ý nghĩa α = 0.05 nên chưa bác bỏ được giả thuyết H_0 nên mô hình 2 hiệu quả hơn do đó ta chọn mô hình 2.

$$log(Rented\ Bike\ Count) = 6.4986 + 0.039 \times Hour + 0.043 \times Temperature - 0.0223 \times Humidity - 0.067 \times Solar\ Radition - 0.388 \times SeasonsSpring - 0.286 \times SeasonsSummer - 0.86 \times SeasonsWinter + 0.334 \times HolidayNoHoliday$$

- Ta tiến hành tính toán các sai số trên cả tập huấn luyện lẫn tập kiểm thử

```
train_error=regressionMetrics(trues=log(train_data$Rented.Bike.Count),preds=predict(model2,newdata=train_data))
test_error=regressionMetrics(trues=log(test_data$Rented.Bike.Count),preds=predict(model2,newdata=test_data))
print(train_error)
print(test_error)
```

```
>   print(train_error)
         mae          mse         rmse         mape         <NA>
3.125005e+02 2.283001e+05 4.778076e+02 1.009047e+00           NA
>   print(test_error)
         mae          mse         rmse         mape         <NA>
3.091276e+02 2.160269e+05 4.647869e+02 8.787359e-01           NA
```

Hình 35: Kết quả các sai số trên cả tập huấn luyện lẫn tập kiểm thử

Ở đây, ta nhận thấy các thông số gần giống nhau, do đó mô hình hoạt động giống nhau trên cả tập huấn luyện và tập kiểm thử.

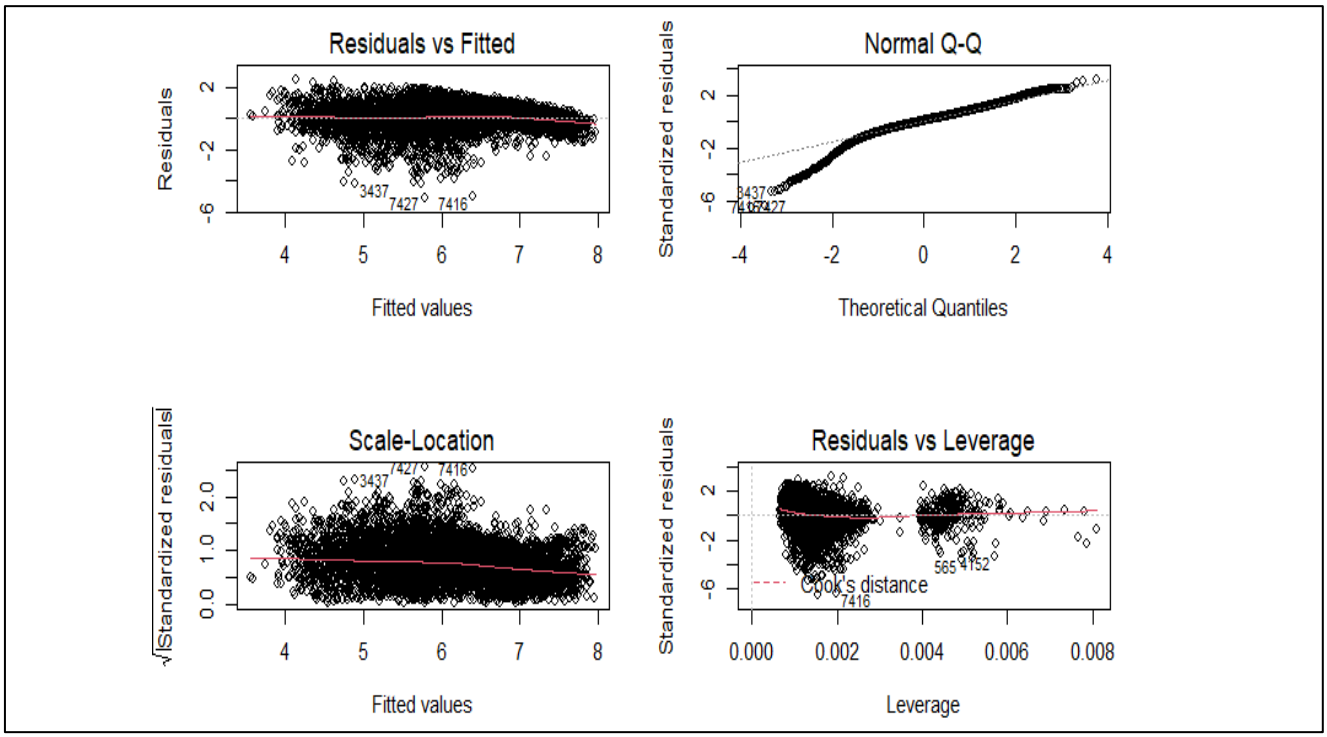
Kiểm tra giả định của mô hình

- Tính tuyến tính của dữ liệu: mối quan hệ giữa biến dự báo X và biến phụ thuộc Y được giả sử là tuyến tính

- Sai số có phân phối chuẩn
- Phương sai của các sai số là hằng số
- Các sai số ϵ có kỳ vọng = 0
- Các sai số ϵ1, ..., ϵn thì độc lập với nhau.

Ta thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình.

```
par(mfrow=c(2,2))
plot(model2)
```



Hình 36: Các đồ thị phân tích thặng dư

- Đồ thị thứ 1 (Residuals vs Fitted) kiểm tra tính tuyến tính của dữ liệu và tính đồng nhất của các phương sai sai số.
  - Quan sát đồ thị bên trên ta thấy rằng đường thẳng màu đỏ trên đồ thị phân bố khá thẳng nên ta kết luận tính tuyến tính của dữ liệu thỏa mãn.
  - Các điểm thặng dư trên đồ thị phân tán không đều đường thẳng y=0 nên giả định phương sai đồng nhất không thỏa.
- Đồ thị thứ 2 (Normal Q-Q) kiểm tra giả định về phân phối chuẩn của các sai số, có một số điểm lệch một ít về về phía dưới bên trái nên giả định sai số có phân phối chuẩn chưa thực sự thõa mãn.

- Đồ thị thứ 3 (Scale - Location) kiểm tra giả định tính đồng nhất phương sai của các sai số, các điểm thặng dư phân tán đều xung quanh đường thẳng màu đỏ nên giả định phương sai của các sai số là hằng số thõa mãn.
- Đồ thị thứ 4 (Residuals vs Leverage) chỉ ra các quan trắc 7416, 585,4152 có thể là các điểm có ảnh hưởng cao đến bộ dữ liệu. Tuy nhiên ta thấy được các điểm này chưa vượt qua đường thẳng Cook (đường thẳng nét đứt Cook's distance). Do vậy, các điểm này chưa thực sự là các điểm có ảnh hưởng cao đến bộ dữ liệu. Do đó ta không cần phải loại bỏ chúng khi phân tích.

c) Dự báo

Ta tiến hành đưa tập kiểm thử vào mô hình vừa mới xây dựng bên trên để so sánh với dữ liệu thực tế

```
cmp=data.frame(truth_value=test_data$Rented.Bike.Count,
               predict_value=exp(predict(model2,newdata=test_data)))
```

	truth_value	predict_value
2	204	137.99918
7	181	170.94797
10	490	227.01958
12	360	345.84350
15	447	402.09308
22	405	141.01518
24	323	139.54483
32	146	70.10369
35	334	151.24874
37	479	280.10008
41	618	324.87395
43	489	306.75967
47	354	321.68954
48	366	326.80843
50	227	103.53152
51	228	83.95587
53	76	104.61651
56	22	77.39200
60	273	105.89986

Hình 37: So sánh kết quả dự báo với dữ liệu thực tế

Nhận xét : giá trị dự đoán vẫn còn chênh lệch nhiều so với giá trị thực tế, mô hình vẫn chưa thật sự tốt.

PHỤ LỤC A – CODE HOẠT ĐỘNG 1

```
diem_so <- read.csv("C:/Users/Admin/Desktop/BTL_XSTK/diem_so.csv") #Doc du lieu
head(diem_so,10) #xem 10 dong dau tien
new_DF <- diem_so[,c("G1","G2","G3","studytime","failures","absences","paid","sex")] # trich cac bien
head(new_DF,10) #xem 10 dong dau tien

apply(is.na(new_DF),2,which) #kiem tra va xuat vị tri du lieu khuyet (NA)
apply(is.na(new_DF),2,sum) #kiem tra va dem gia tri NA
apply(is.na(new_DF),2,mean) #ti le NA
new_DF <- na.omit(new_DF)
apply(is.na(new_DF),2,sum) #kiem tra lai NA
head(new_DF,10) #xem 10 dong dau tien

mean <- apply(new_DF[,c("G1","G2","G3")],2,mean) #tinh trung binh mau
sd <- apply(new_DF[,c("G1","G2","G3")],2,sd) #tinh do lech chuan hieu chinh
Q1 <- apply(new_DF[,c("G1","G2","G3")],2,quantile,probs=0.25) #tinh diem phan vi 1
median <- apply(new_DF[,c("G1","G2","G3")],2,median) #tinh trung vi
Q3 <- apply(new_DF[,c("G1","G2","G3")],2,quantile,probs=0.75) #tinh diem phan vi 3
min <- apply(new_DF[,c("G1","G2","G3")],2,min) #tinh gia tri nho nhat
max <- apply(new_DF[,c("G1","G2","G3")],2,max) #tinh gia tri lon nhat
t(data.frame(mean, sd, Q1, median, Q3, min, max)) #tao bang voi cac bien

table(new_DF$studytime) #lap bang thong ke so luong cho cac phan loai
table(new_DF$failures) #lap bang thong ke so luong cho cac phan loai
table(new_DF$paid) #lap bang thong ke so luong cho cac phan loai
table(new_DF$sex) #lap bang thong ke so luong cho cac phan loai

hist(new_DF$G3,main ="Đồ thị phân bố tần số điểm thi cuối khoá G3",xlab ="Điểm thi G3",ylab ="Tần số (Số học sinh)",label =T,ylim=c(0,90),col ="royalblue")
```

```
    boxplot(G3~studytime,data=new_DF,main="Biểu đồ hộp thể hiện phân phối điểm thi theo thời gian tự học",xlab="Thời gian tự học", ylab="Điểm thi G3", col=c(2,3,4,7))
    boxplot(G3~failures,data=new_DF,main="Biểu đồ hộp thể hiện phân phối điểm thi theo số lần không qua môn",xlab="Số lần không qua môn", ylab="Điểm thi G3",col=c(2,3,4,7))
    boxplot(G3~paid,data=new_DF,main="Biểu đồ hộp thể hiện phân phối điểm thi theo phân loại tham gia các lớp học thêm môn Toán ngoài trường",xlab="Tham gia các lớp học thêm môn Toán ngoài trường", ylab="Điểm thi G3",col=c(3,7))
    boxplot(G3~sex,data=new_DF,main="Biểu đồ hộp thể hiện phân phối điểm thi theo giới tính",xlab="Giới tính", ylab="Điểm thi G3",col=c(2,4))

    par(mfrow=c(1,3)) #Xep 3 bieu do vao 1 hang
    plot(G3~G1,data=new_DF,main ="G1 and G3",col =4,pch=16) #Ve do thi phan tan cua G3 theo G1
    plot(G3~G2,data=new_DF,main ="G2 and G3",col=2,pch=16) #Ve do thi phan tan cua G3 theo G2
    plot(G3~absences,data=new_DF,main ="absences and G3",col=1,pch=16) #Ve do thi phan tan cua G3 theo absences

    model_1 <- lm(G3~G1 + G2 + studytime + failures + absences +paid + sex,new_DF) #Xay dung mo hinh 1 va luu voi ten model_1
    summary(model_1) #Ket qua mo hinh
    model_2 <- lm(G3~G1 + G2 + absences,new_DF) #Xay dung mo hinh 2 da loai bo bien studytime, failures, paid, sex tu mo hinh 1 va luu voi ten model_2
    summary(model_2) #Ket qua mo hinh 2

    anova(model_1,model_2) #So sanh mo hinh 1 va mo hinh 2

    par(mfrow = c(2, 2)) #xep 4 bieu do thanh 2 hang 2 cot
    plot(model_2) #ve do thi phan tich thang du

    X = data.frame("G1"= 15,"G2"= 12,"absences"= 2)
    predict(model_2,X,interval="confidence")  #thuc hien du bao cho bien X bang model_2
```

PHỤ LỤC B – CODE HOẠT ĐỘNG 2

```
library(dplyr)
library(mltools)
library(data.table)
library(corrplot)
library(RColorBrewer)
library(olsrr)
library(performanceEstimation)
library(caret)

rented_bike_count<-function(df){
  hist(df$Rented.Bike.Count,main="The frequency plot of rented bike count",xlab="Rented bike count") #biểu đồ tần số của lượng thuê xe
  hist(log(df$Rented.Bike.Count),main="The frequency plot of rented bike count")#biểu đồ tần số của logarit lượng thuê xe
}

hour<-function(df){
  boxplot(Rented.Bike.Count~Hour,data=df,main="Rented bike count by hour")#biểu đồ hộp của lượng thuê xe theo giờ
}

correlation<-function(df){
  df1=df[,!names(df)%in%c("Seasons","Holiday","Functioning.Day","Date","Hour")]#loại biến
  corrplot(cor(df1),type="upper",tl.pos="td",method="color",addCoef.col = "brown",number.cex=0.8,col=brewer.pal(n=8, name="RdYlBu"))#biểu đồ correlation
}

weather<-function(df){
  plot(df$Temperature..C.,log(df$Rented.Bike.Count),main="Rented bike count in Seoul by temperature")#biểu đồ phân tán của log lượng thuê xe với nhiệt độ
```

```
  plot(df$Solar.Radiation..MJ.m2.,log(df$Rented.Bike.Count),main="Rented bike count in Seoul by solar radiation")#biểu đồ phân tán của log lượng thuê xe với bức xạ mặt trời
}

season<-function(df){
  avg_winter=mean(df[df$Seasons=="Winter",]$Rented.Bike.Count) #trung bình lượng thuê xe theo mùa đông
  avg_spring=mean(df[df$Seasons=="Spring",]$Rented.Bike.Count) #...
  avg_summer=mean(df[df$Seasons=="Summer",]$Rented.Bike.Count) #...
  avg_autumn=mean(df[df$Seasons=="Autumn",]$Rented.Bike.Count) #...
  ss=c(avg_winter,avg_spring,avg_summer,avg_autumn)
  barplot(ss,names.arg=c("Winter","Spring","Summer","Autumn"),main="Average Rented Bike Count By Season",xlab="Season",ylab="Average Rented Bike Count")#biểu đố thanh trung bình lượng thuê xe theo mùa
  boxplot(Rented.Bike.Count~Seasons,data=df,main="Rented bike count in Seoul from 2017 to 2018")#biểu đồ hộp lượng thuê xe theo mùa
}

holiday<-function(df){
  Holiday=mean(df[df$Holiday=="Holiday",]$Rented.Bike.Count) #trung bình ngày lễ
  noHoliday=mean(df[df$Holiday=="No Holiday",]$Rented.Bike.Count) #trung bình ngày thường
  barplot(c(Holiday,noHoliday),beside=T,col=c("red","blue"),main="Average Rented Bike Count By Holiday",names.arg=c("Holiday","No Hoiday"),ylab="Average Rented Bike Count")#biểu đồ thanh về trung bình lượng xe theo ngày lễ
}

model<-function(df){

df=df[,!names(df)%in%c("Wind.speed..m.s.","Rainfall.mm.","Snowfall..cm.","Date","Functioning.Day","Dew.point.temperature..C.")]#loại biến
```

```
set.seed(1)
split= sample(c(rep(0, 0.7 * nrow(df)), rep(1, 0.3 * nrow(df)))) #chia tập dữ liệu
train_data=df[split==0,]
test_data=df[split==1,]

model1=lm(log(Rented.Bike.Count)~.,data=train_data)#mô hình 1
summary(model1)



train_data=train_data[,!names(train_data)%in%c("Visibility..10m.")]  #loại biến
test_data=test_data[,!names(test_data)%in%c("Visibility..10m.")]

model2=lm(log(Rented.Bike.Count)~.,data=train_data)
summary(model2)#mô hình 2

anova(model1,model2)#so sánh 2 mô hình


train_error=regressionMetrics(trues=train_data$Rented.Bike.Count,preds=exp(predict(model2,newdata=train_data))) #sai số của mô hình theo tập train

test_error=regressionMetrics(trues=test_data$Rented.Bike.Count,preds=exp(predict(model2,newdata=test_data)))#sai số của mô hình theo tập test
print(train_error)
print(test_error)

par(mfrow=c(2,2))
plot(model2) #các biểu đồ kiểm tra giả định mô hình

cmp=data.frame(truth_value=test_data$Rented.Bike.Count,
```

```
            predict_value=exp(predict(model2,newdata=test_data))) #so sánh giá trị thực tế so với
giá trị dự đoán
}


main<-function(){
  df=read.csv("SeoulBikeData.csv")
  head(df,10)
  apply(is.na(df),2,sum) #kiểm tra dữ liệu khuyết
  summary(df) #thống kê mô tả
  df=df[df$Functioning.Day=="Yes",] #loại hàng dữ liệu của những ngày không hoạt động
  rented_bike_count(df)
  hour(df)
  correlation(df)
  weather(df)
  season(df)
  holiday(df)
  model(df)
}

main()
```

TÀI LIỆU THAM KHẢO

[1] Nguyễn Thị Mộng Ngọc, *Bài giảng Xác Suất Thống Kê*.

[2] Nguyễn Tiến Dũng (chủ biên) & Nguyễn Đình Huy, (2019), *Xác suất – Thống kê & Phân tích số liệu*, Nxb. Đại học Quốc gia, Thành phố Hồ Chí Minh. từ https://www.kaggle.com/datasets/afumetto/3dprinter

[3] Pham Thi Hong Anh, 2020, *Xử lý missing data trong Data analysis*, truy cập từ https://viblo.asia/p/xu-ly-missing-data-trong-data-analysismaGK7qaAlj2?fbclid=IwAR1n1rDU2PWjpjIgWtc_UQKoNMHOfqqFYEFGKcvwFjPneJYz7XfQE9DL484

[4] Đạt Vũ, 22/07/2019, *Biểu đồ Boxplots (box and whiskers)*, truy cập từ https://www.diendat.net/bieu-do-box-and-whiskers/

[5] Minh Đức, 19/05/2021, *Những điều cần biết về biểu đồ Histogram*, truy cập từ https://vietquality.vn/nhung-dieu-can-biet-ve-histogram-diagram-bieu-do-phan-bo-tan-suat/

[6] *Hướng dẫn sử dụng phần mềm Rstudio*, otworzumysl.com, Truy cập từ https://otworzumysl.com/huong-dan-su-dung-phan-mem-r-studio/

[7] David Dalpiaz, *Applied Statistics with R*

[8] Peter Dalgaard, *Introductory Statistics with R*

[9] *Prediciting Number of Rented Bikes Using Machine Learning Algorithms* - https://rpubs.com/asmi2990/869714