**Báo cáo cuối kì**

**Thành viên:**

**Huỳnh Tâm An - 20110125**

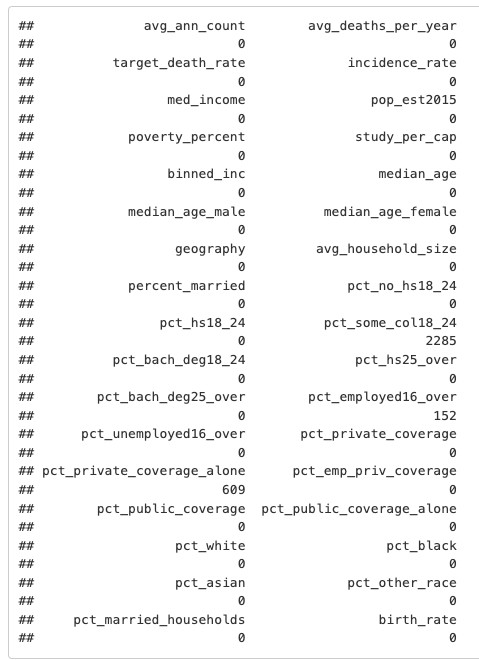
**Đinh Đình Thiên Phúc - 20110082**

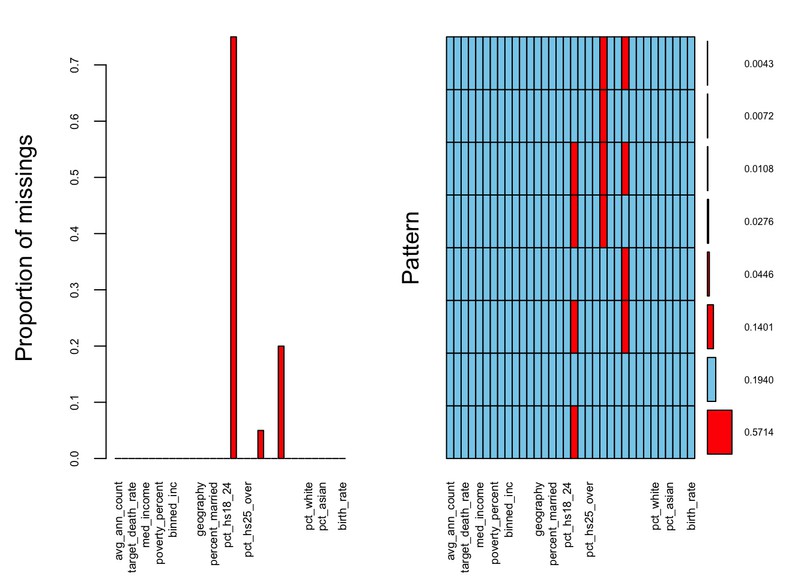
# Hoạt động 1:

# Phần 1: Dự đoán tỷ lệ tử vong do ung thư

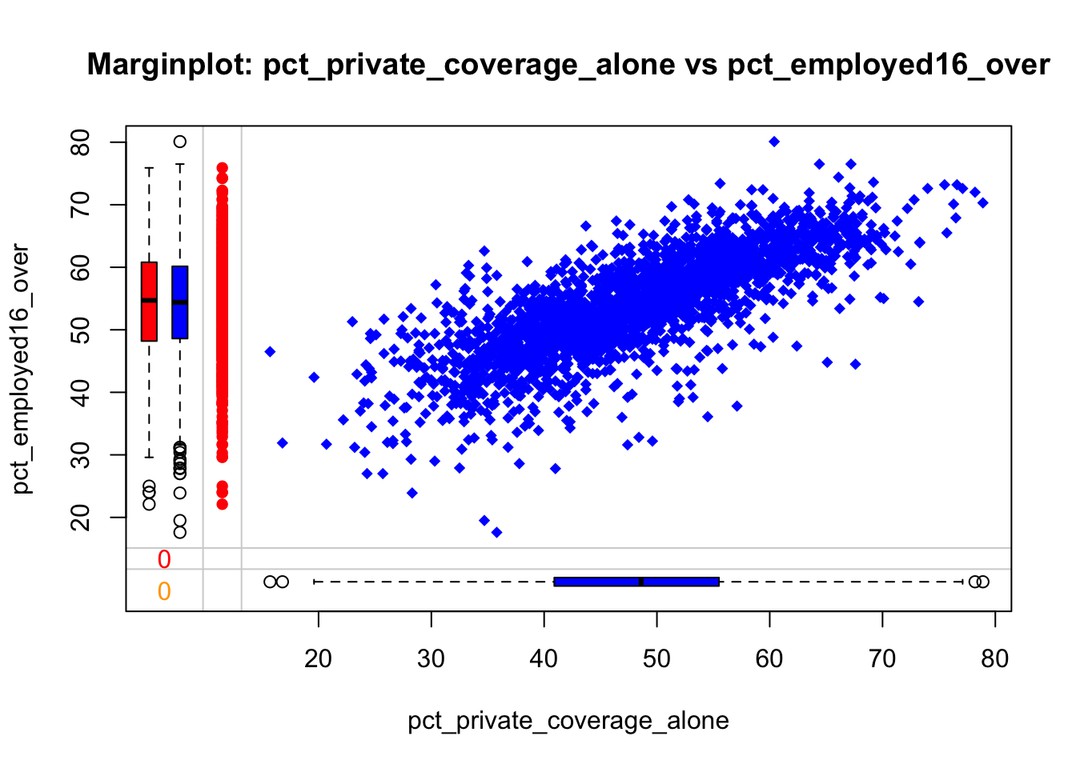
**Dataset: PredictCancerMR.csv**

### I. Tiền xử lý dữ liệu

Số lượng data bị thiếu ở từng cột Xu hướng dữ liệu khuyết

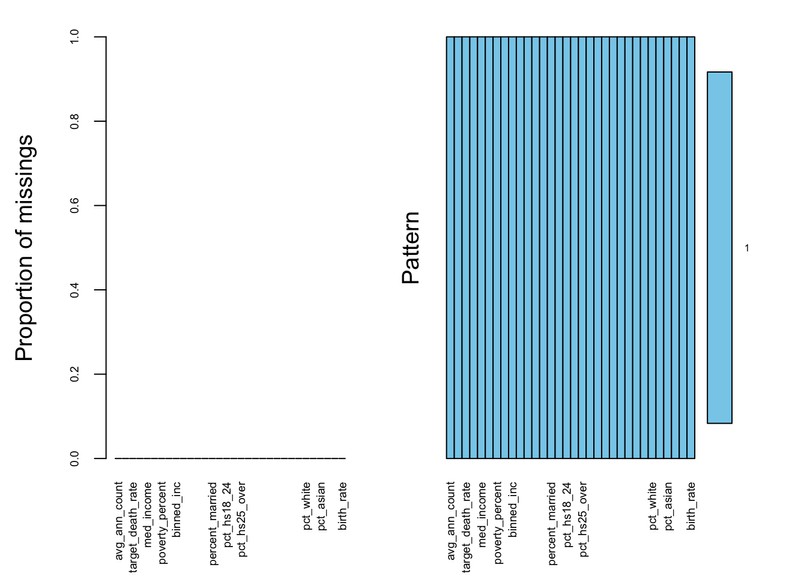


**Ta có thể dễ thấy pct\_employed16\_over có dữ liệu khuyết dưới 10% và pct\_some\_col18\_24 trên 70% nên xem xét loại bỏ luôn.**

Dữ liệu khuyết của pct\_private\_coverage\_alone với pct\_employed16\_over

=> Các giá trị khuyết pct\_private\_coverage\_alone không tương đồng. Điều này chỉ ra rằng dữ liệu có khả năng thuộc loại MAR

Xử lý dữ liệu khuyết: Xử lý MAR bằng hàm mice() sử dụng phương pháp 'pmm' cho các biến cần thiết.



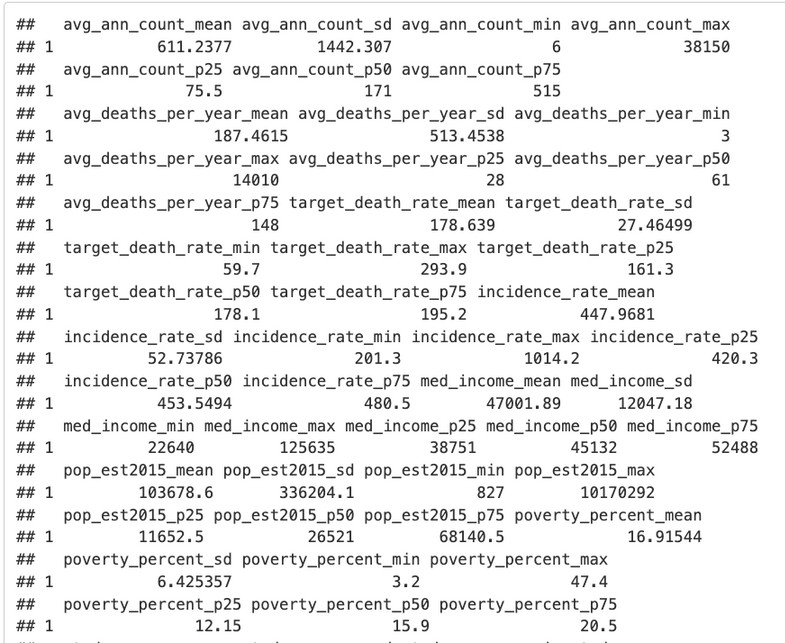
Ta có thể thấy bộ dữ liệu cơ bản đã sạch, không còn dữ liệu khuyết.

### Trực quan hoá dữ liệu

#### Chuyển đổi biến

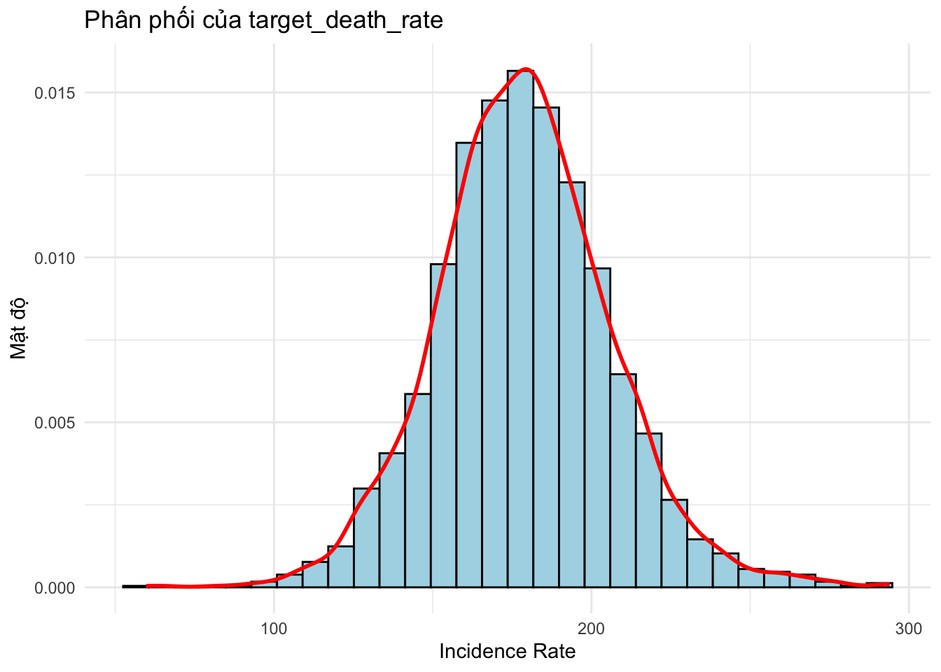
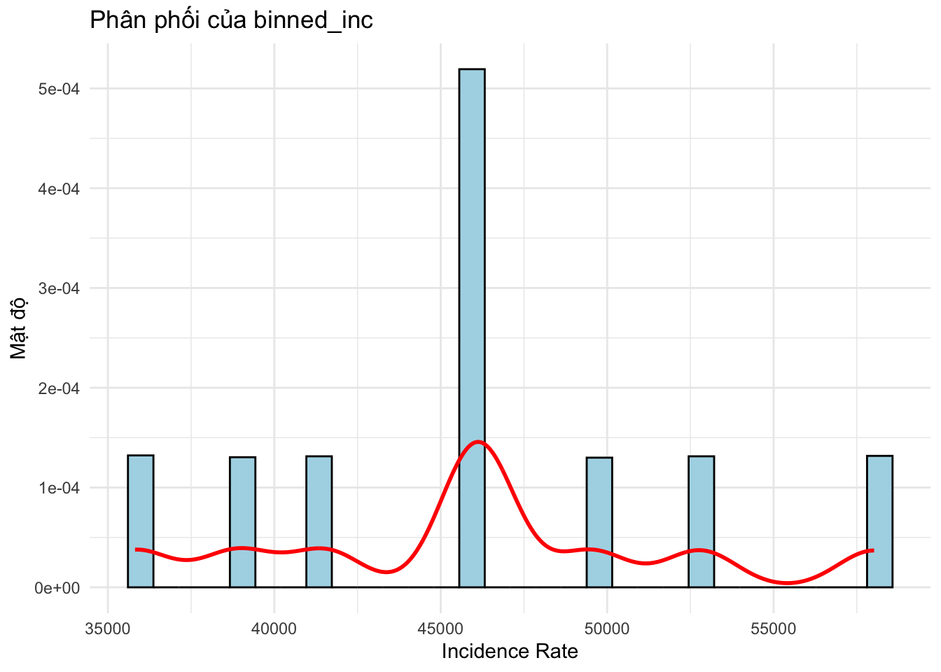
* 1. Thống kê mô tả

Thống kê mô tả của 7 cột numeric đầu tiên (bao gồm mean, sd, min, max, q25, q50, q75)



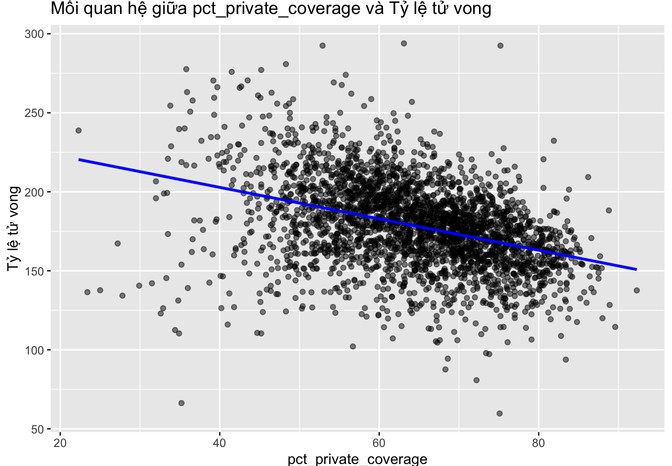
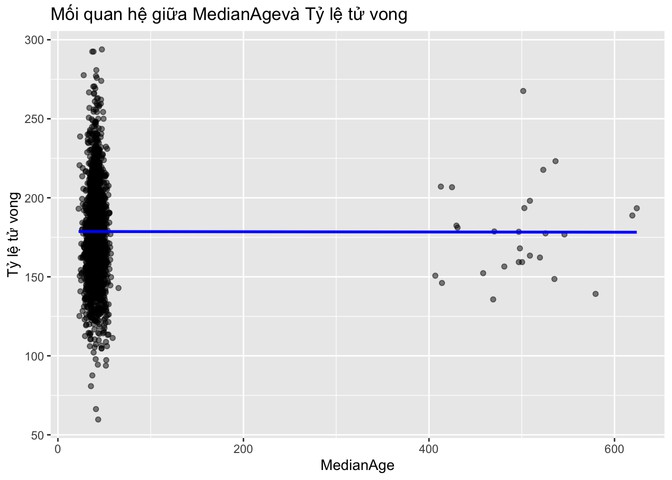
#### Trực quan hoá bằng đồ thị

Biểu đồ Bar: binned\_inc Biến mục tiêu

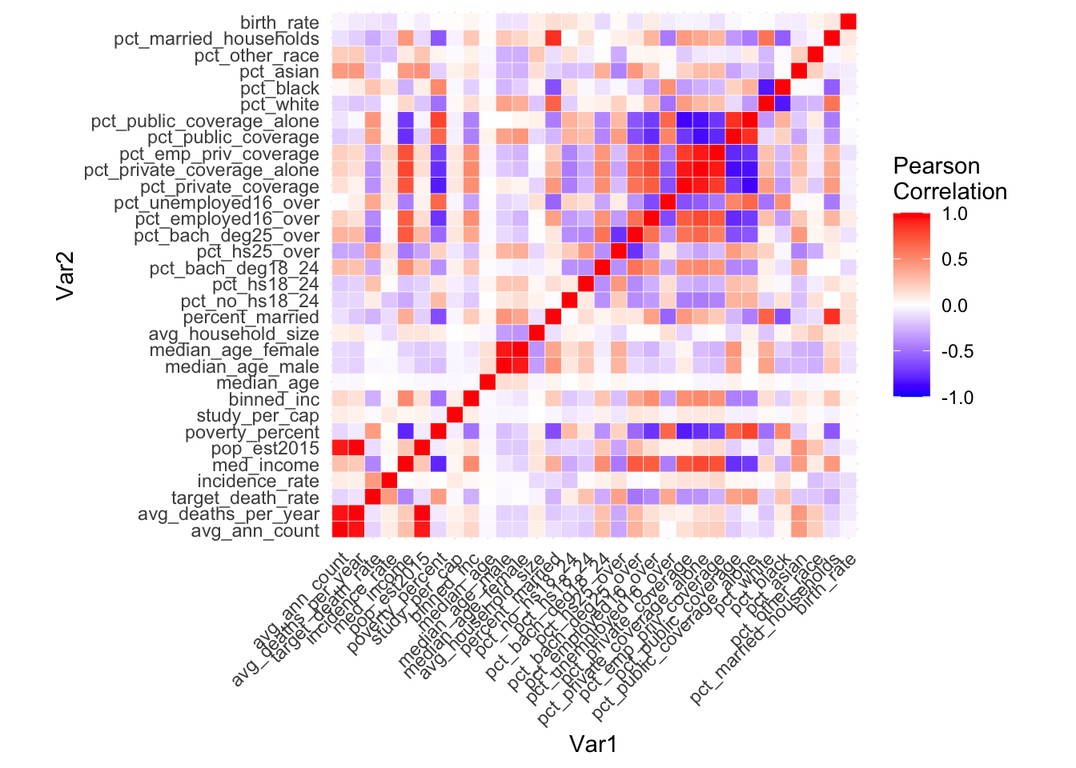


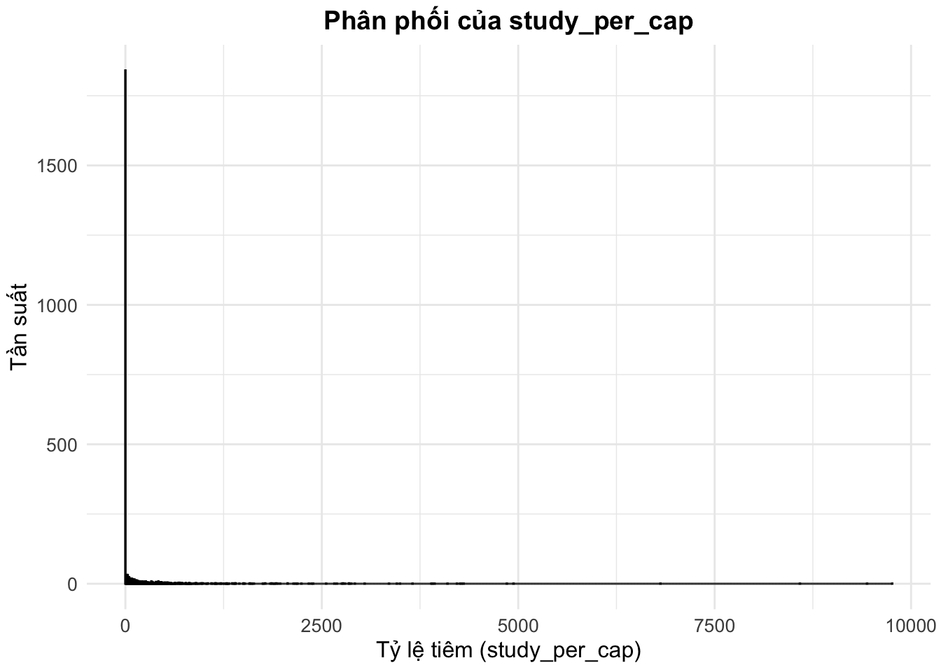
Mối quan hệ giữa một số biến liên tục và biến mục tiêu

#### 



Heatmap: Ma trận tương quan giữa các biến liên tục

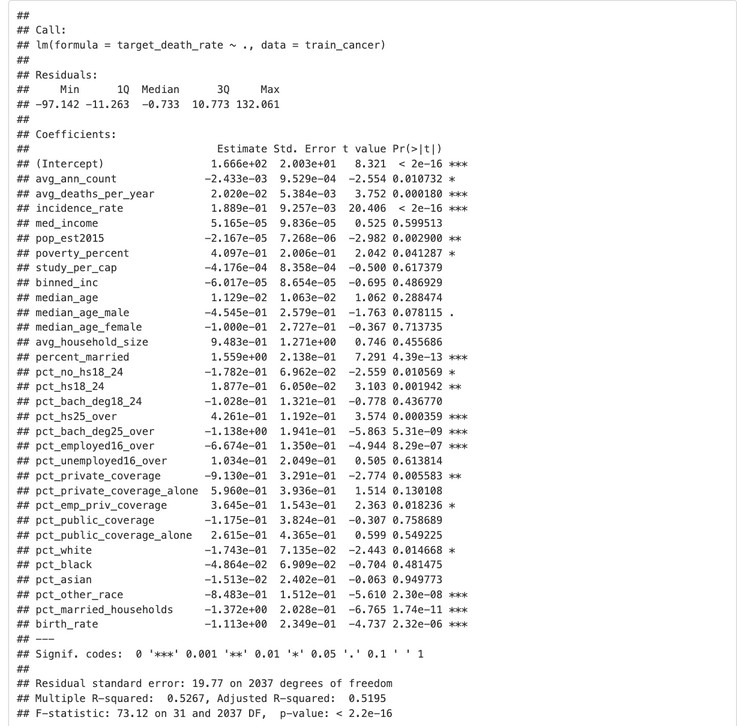




Ta thấy tỉ lệ tiêm ngừa khá thấp (đa số là không tiêm ngừa) - có vẽ sẽ không ảnh hưởng đến model dự đoán

### III. Modeling

#### Chia train - test: 70 - 30

1. Model linear cơ bản

Biến cần dự đoán: target\_death\_rate

Các biến còn lại là biến giải thích

Output trên R

#### 2. Model linear cơ bản: Kết luận

 **Multiple R-squared: Khoảng 0.5267 cho thấy mô hình giải thích được 52.67% sự biến thiên của biến mục tiêu target\_death\_rate.**

**Tuy nhiên, khoảng 47.33% còn lại vẫn do các yếu tố khác chưa được đưa vào mô hình.**

 **F-statistic (73.12, p-value < 2.2e-16): Mô hình tổng thể có ý nghĩa thống kê, tức là ít nhất một biến độc lập có ảnh hưởng đáng kể đến biến mục tiêu.**

 **Ý nghĩa cao với p-value < 0.001:**

 **avg\_deaths\_per\_year (Hệ số 0.3558): Mối quan hệ cùng chiều với tỷ lệ tử vong, tức là số ca tử vong trung bình mỗi năm càng cao, tỷ lệ tử vong càng cao.**

 **incidence\_rate: Hệ số 0.3578, cho thấy tỷ lệ mắc bệnh có ảnh hưởng mạnh mẽ đến tỷ lệ tử vong. percent\_married: Hệ số 0.3799, cho thấy tỷ lệ người kết hôn có tương quan dương đáng kể với tỷ lệ tử vong.**

 **pct\_bach\_deg25\_over: Hệ số -0.2179, chỉ ra rằng tỷ lệ người có bằng cử nhân trở lên từ 25 tuổi trở lên có mối quan hệ nghịch biến với tỷ lệ tử vong.**

 **pct\_employed16\_over: Hệ số -0.1994, cho thấy tỷ lệ người có việc làm từ 16 tuổi trở lên cũng có mối tương quan nghịch biến. ** **Tương tự với các biến còn lại**

 **Ý nghĩa với 0.001 < p-value < 0.01:**

 **pop\_est2015: Hệ số -0.2627, tức là mật độ dân số có mối quan hệ nghịch biến với tỷ lệ tử vong.**

 **pct\_hs18\_24 (Hệ số 0.0673): Mối quan hệ cùng chiều, cho thấy tỷ lệ người tốt nghiệp trung học từ 18-24 tuổi càng cao, tỷ lệ tử vong càng cao.**

 **pct\_private\_coverage**

**Ý nghĩa với 0.01 < p-value < 0.05: avg\_ann\_count, poverty\_percent, pct\_no\_hs18\_24, pct\_emp\_priv\_coverage, pct\_white**

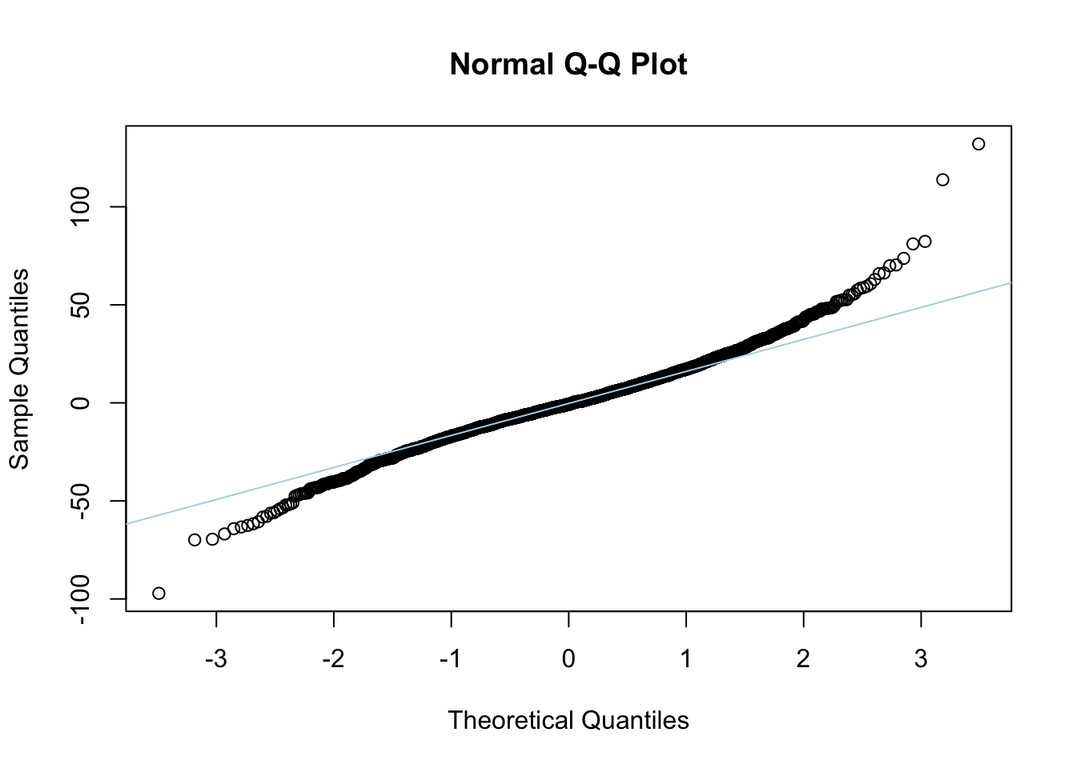


**** **Không có ý nghĩa thống kê: Không có**

#### Kiểm tra sự tuyến tính

Hình vẽ không cho thấy xu hướng đường cong nào đáng kể. Chứng tỏ giả định về tính tuyến tính của mô hình là phù hợp.

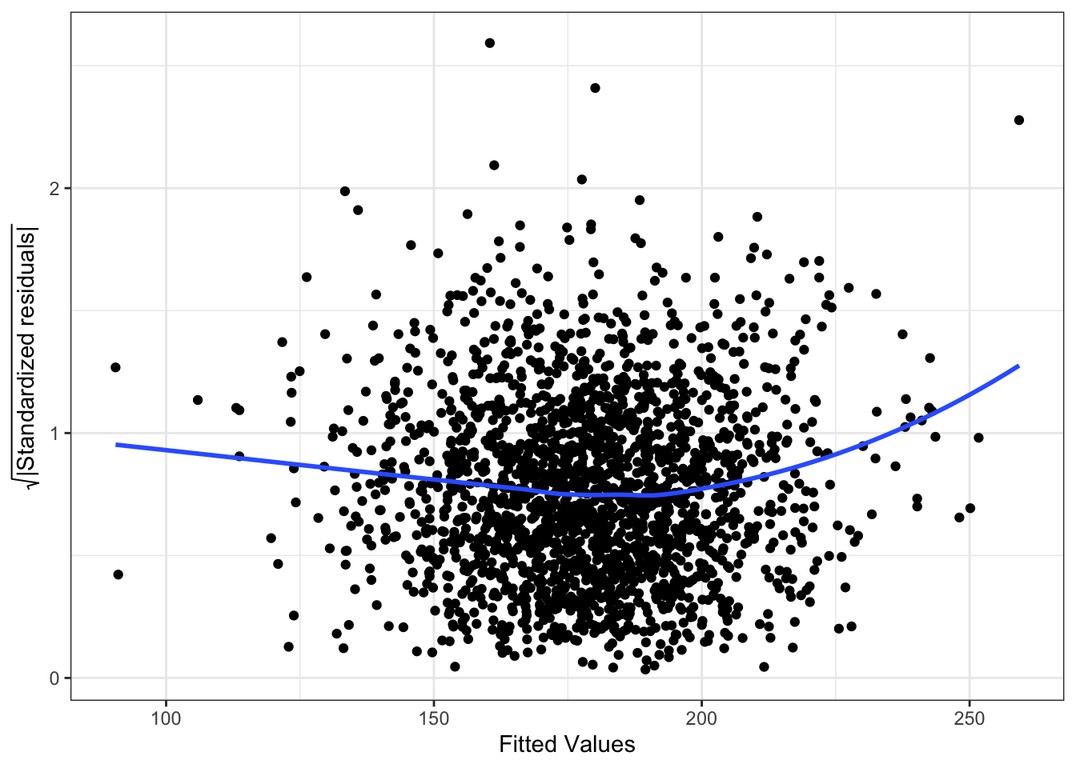
#### Kiểm tra phần dư

 Đa số các điểm nằm gần đường thẳng -> Điều này cho thấy rằng phần dư gần như tuân theo phân phối chuẩn, đặc biệt là ở vùng trung tâm.

 Phần đuôi và đầu cong không quá fit với đường cong lý thuyết -> Điều này gợi ý rằng có thể có hiện tượng skewness hoặc outlier ở phần dư, làm vi phạm nhẹ giả định về phân phối chuẩn.

=> Phần dư gần như tuân theo phân phối chuẩn, nhưng có một số ngoại lệ ở các điểm đầu và cuối.

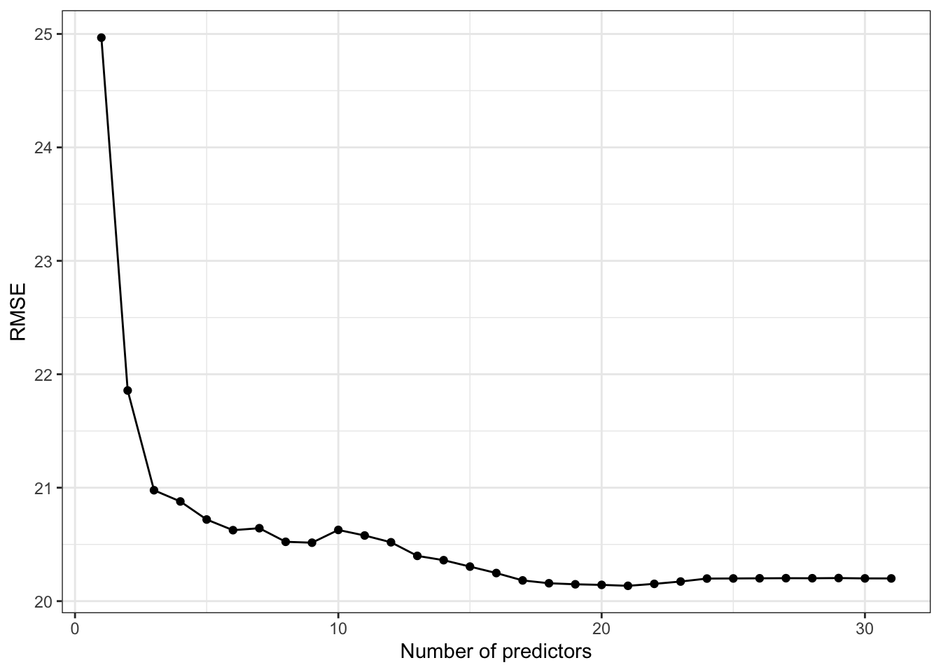
#### Kiểm tra sự đồng nhất phương sai

 Hình là đường cong xu hướng hơi lõm xuống ở giữa => vi phạm giả định đồng nhất phương sai

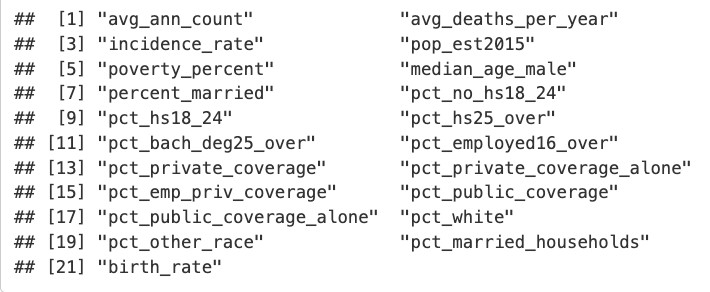
 Mô hình có vẽ chưa được tốt nên chúng ta cần cải tiến.

#### Model linear cải tiến:

Cải tiến bằng cách lựa chọn model tốt nhất sử dụng chỉ số RMSE nhỏ nhất cho từng số lượng biến sử dụng model

Với phương pháp exhaustive thì model được chọn với 21 biến tốt nhất

Các biến được chọn:



Kết luận sau khi sử dụng model được cải tiến:

 Residual Standard Error (RSE): 19.75. Điều này cho thấy độ lệch trung bình của dự đoán so với giá trị thực tế là 19.75 đơn vị. Mức độ sai số này có thể chấp nhận được.

 Multiple R-squared: 0.5251. Khoảng 52.51% phương sai của biến mục tiêu target\_death\_rate được giải thích bởi các biến độc lập trong mô hình. Tuy nhiên, vẫn còn 47.49% phương sai chưa được giải thích, điều này gợi ý rằng mô hình có thể cần được cải thiện.

 Adjusted R-squared: 0.5205. Giá trị này đã điều chỉnh theo số lượng biến độc lập trong mô hình. Vì nó gần với Multiple R-squared, điều này cho thấy các biến trong mô hình không bị dư thừa nhiều.

## Phần 2: Thông số và hiệu năng của máy tính phát triển

**thế nào theo thời gian**

**Dataset: ALL\_GPUs.csv**

### Tiền xử lý dữ liệu

Clean dữ liệu. Các thao tác bao gồm:

* Thay thế string trống thành NA

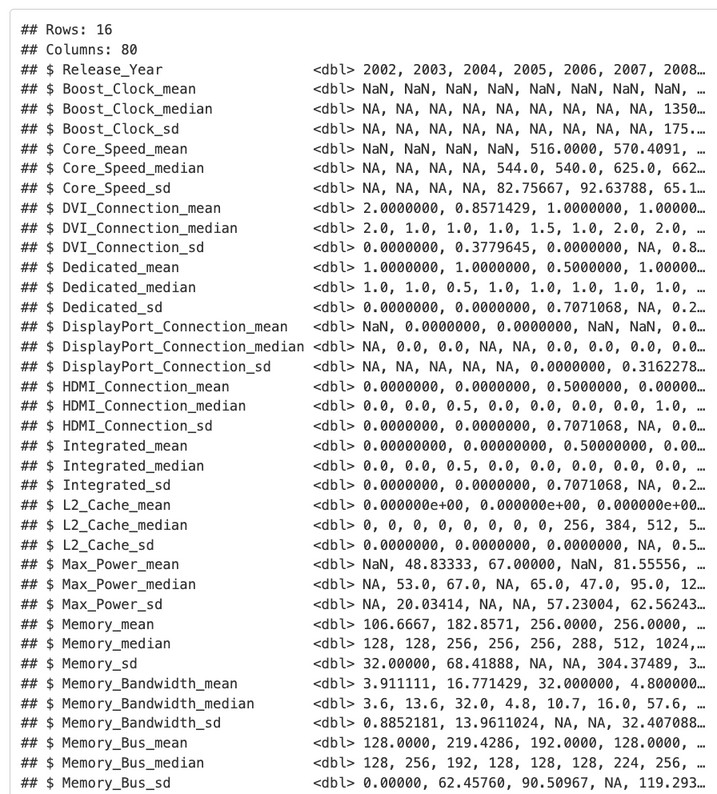
 Bỏ đi các hậu tố đơn vị ở các cột numeric như: Boost\_Clock (MHz), Pixel\_Rate (MHz),...

 Tách cột có hai đơn vị thành hai cột khác nhau: PSU thành PSU\_Watt và PSU\_Amps  Chuyển các cột Yes/No thành 1/0

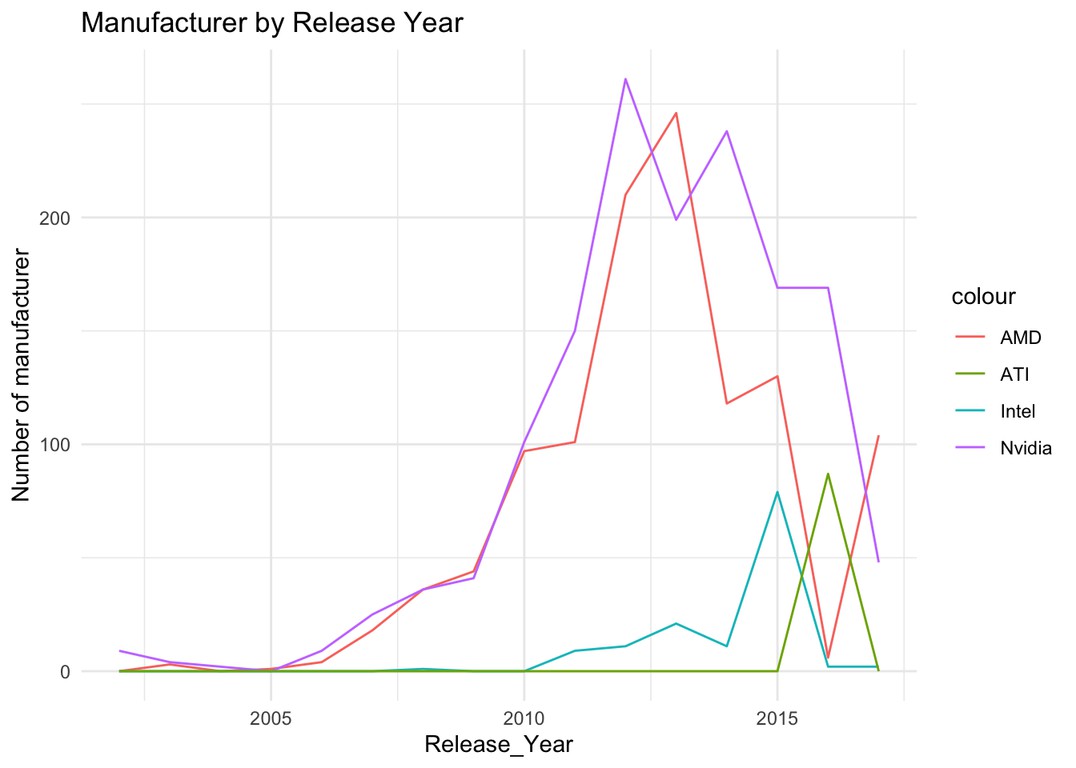
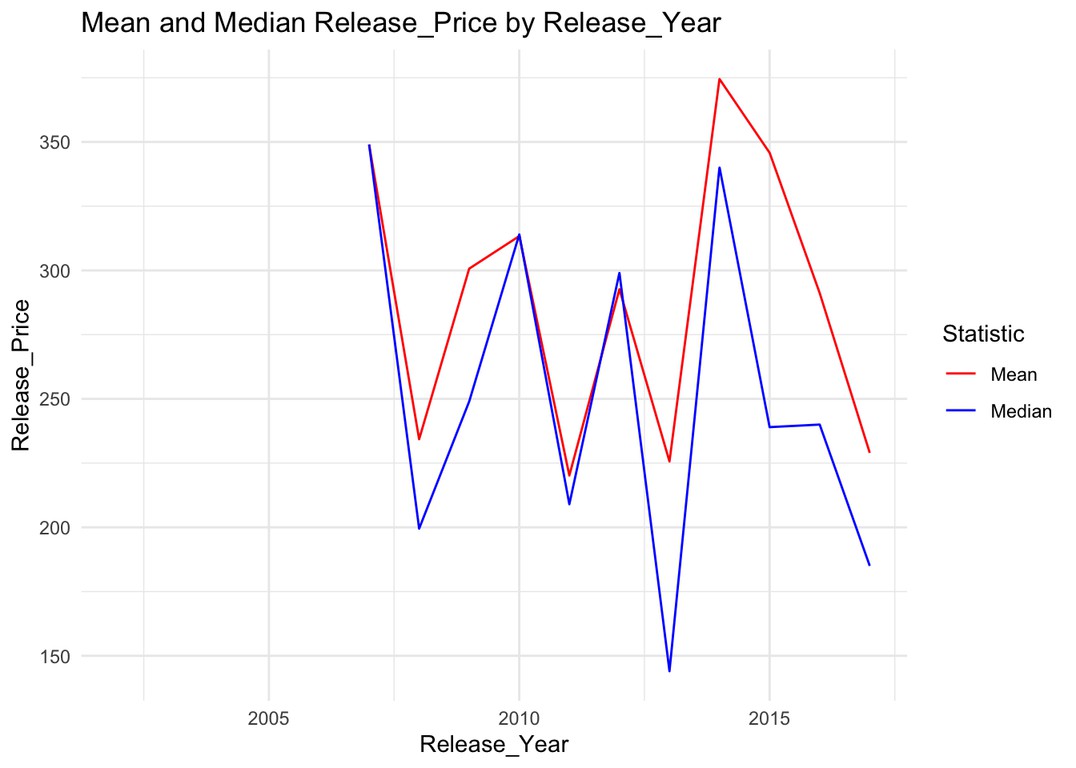
 Chuyển định dạng Release\_Date thành year và đổi tên cột biến thành Release\_Year  Loại bỏ NA của các cột: Release\_Year, Pixel\_Rate.

 Loại bỏ outlier cột Pixel\_Rate

### Thống kê mô tả

Dataset được phân nhóm theo cột Release\_Year

Ta có thể thấy trong data summary từ năm 2010 trở về trước có rất nhiều giá trị bị thiếu, nên ta có thể bỏ dữ liệu từ 2010 trở về trước.

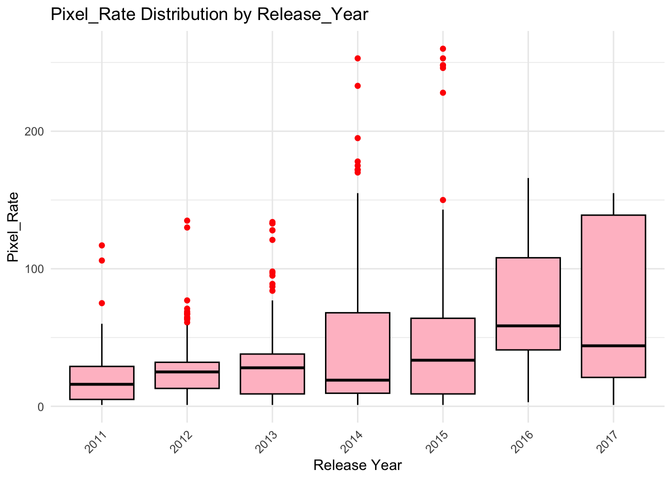
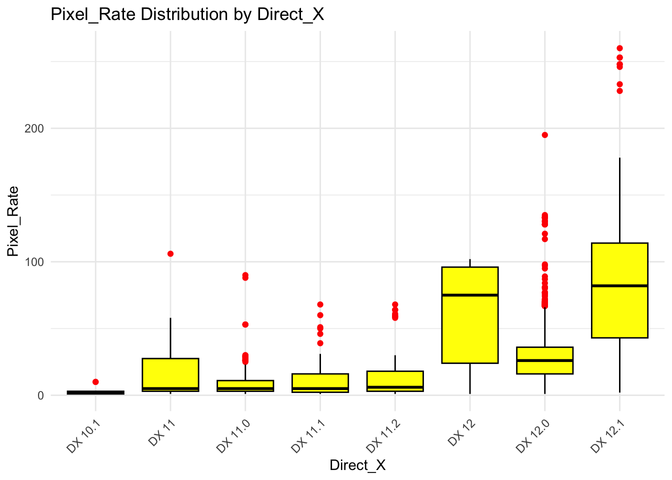
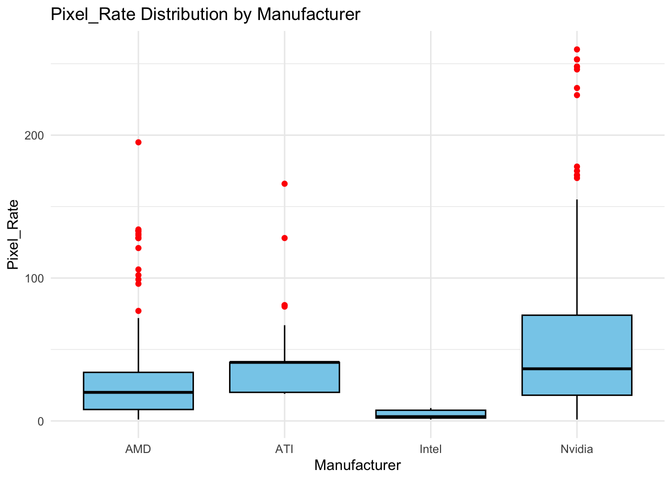
Một số quan hệ giữa các cột

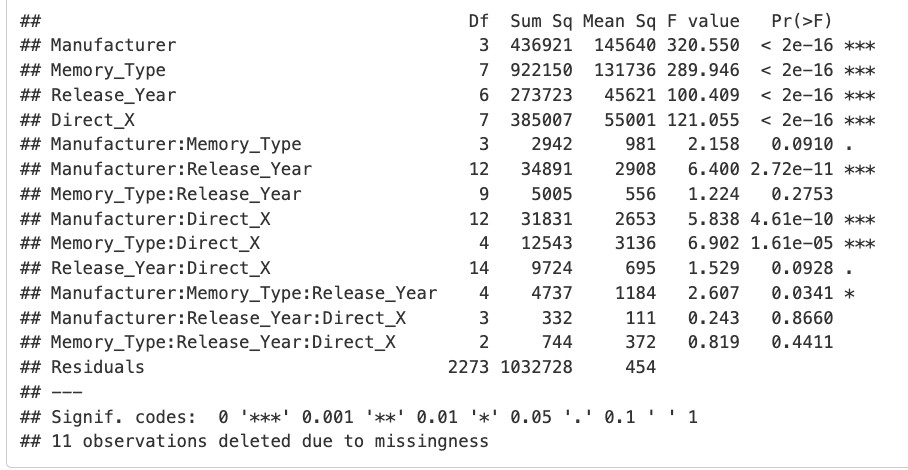
### III. Modeling

#### Chia train-test: 70-30

**Dataset khi chia train-test chỉ sử dụng các biến định lượng và các biến Manufacturer, Direct\_X, Memory\_Type được Encode để xây dựng model**

#### Xây dựng model:

a) Pixel\_Rate ANOVA: Ta có thể nhận thấy rằng Pixel\_Rate có thể được giải thích bởi các biến định tính như Manufacturer, Direct\_X, Release\_Year (định tính hoá chỉ để sử dụng ANOVA), Memory\_Type. Cho nên ta sử dụng 4-way ANOVA

 **Sau khi áp dụng 4-way ANOVA, ta có các kết luận sau:**

 **Các cột Manufacturer, Memory\_Type,**

**Release\_Year, Direct\_X đều có ảnh hưởng lớn tới Pixel\_Rate**

 **Cac cột có tương tác lẫn nhau gồm: ** **Manufacturer:Release\_Year**

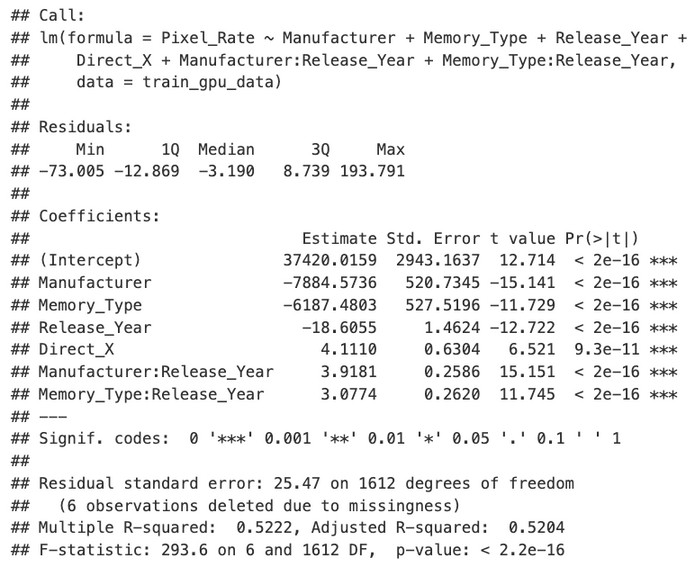
 **Memory\_Type:Release\_Year ** **Manufacturer:Direct\_X**

 **Release\_Year:Direct\_X**

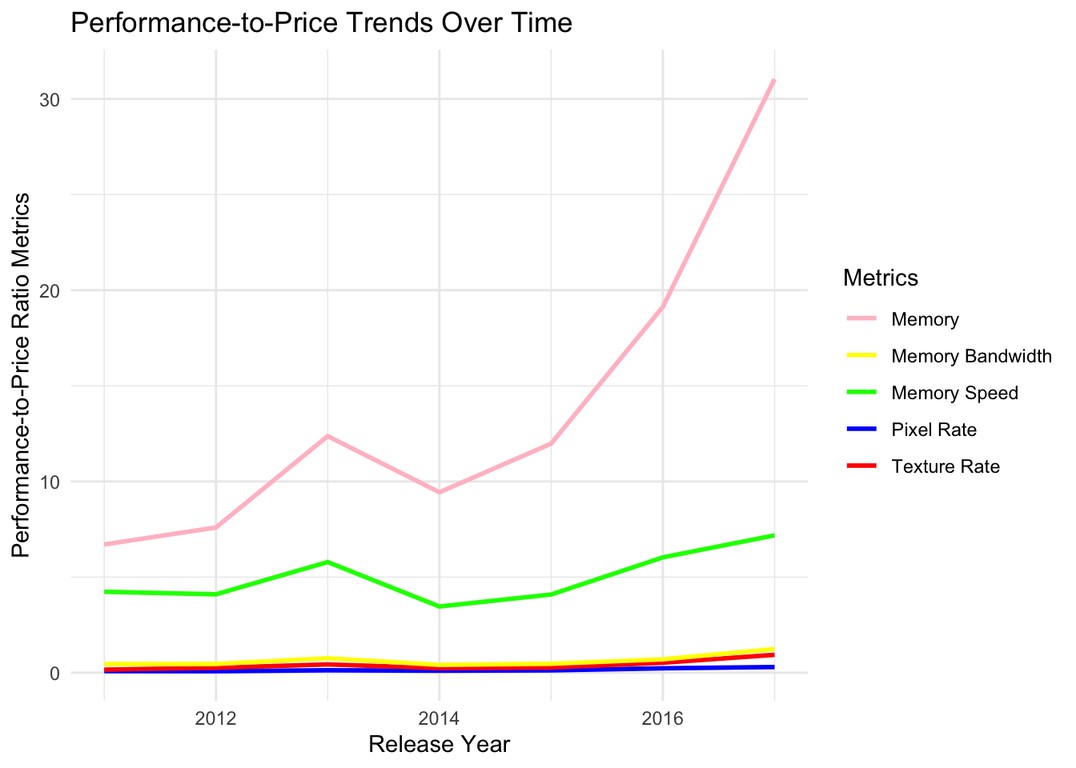
 **Manufacturer:Memory\_Type:Release\_Year**

b) Dự đoán GPU của tương lai (cột Pixel\_Rate)

Sử dụng mô hình hồi quy tuyến tính để dự đoán cột Pixel\_Rate

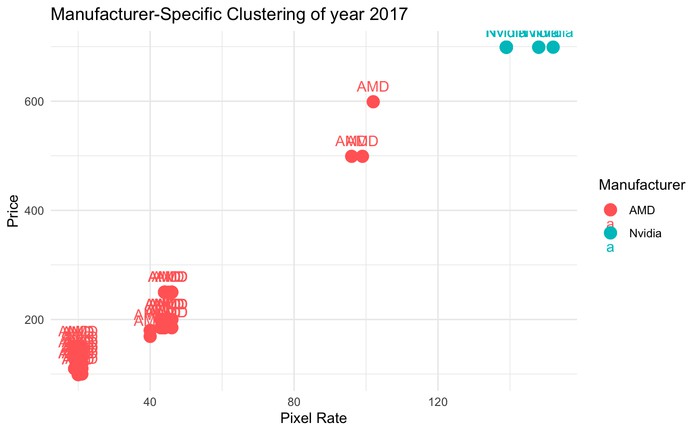
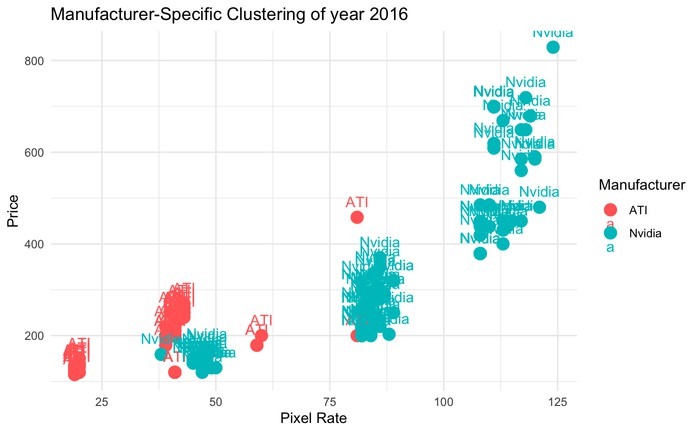
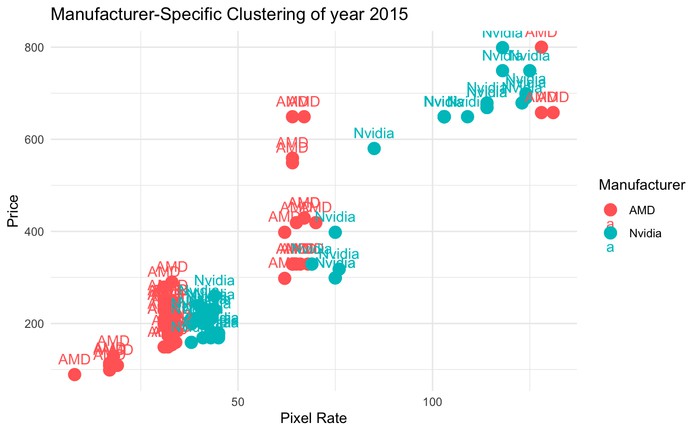
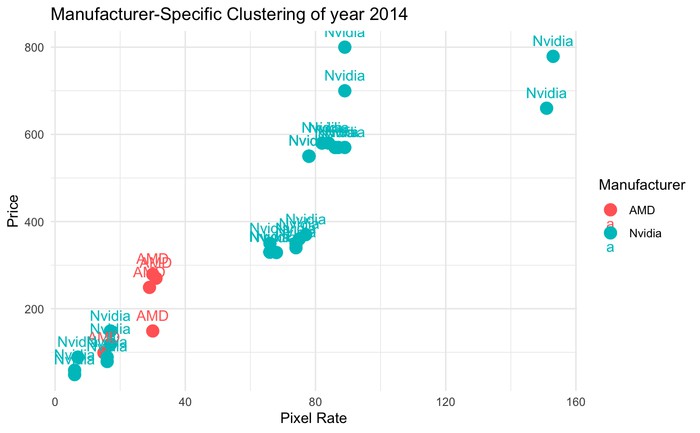
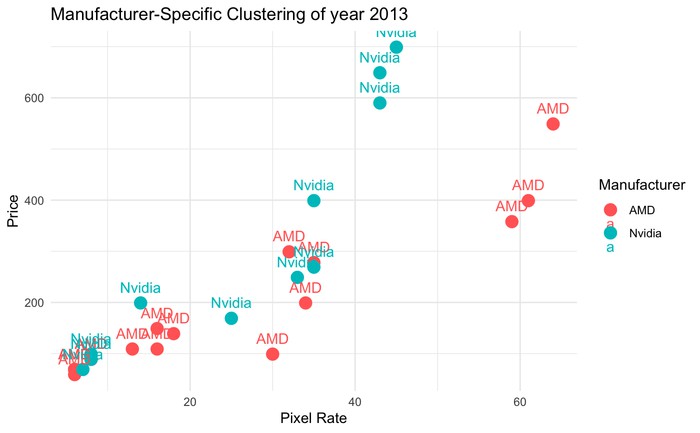
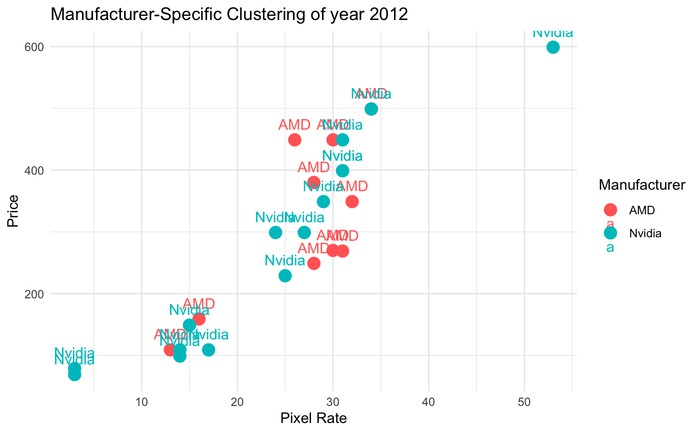
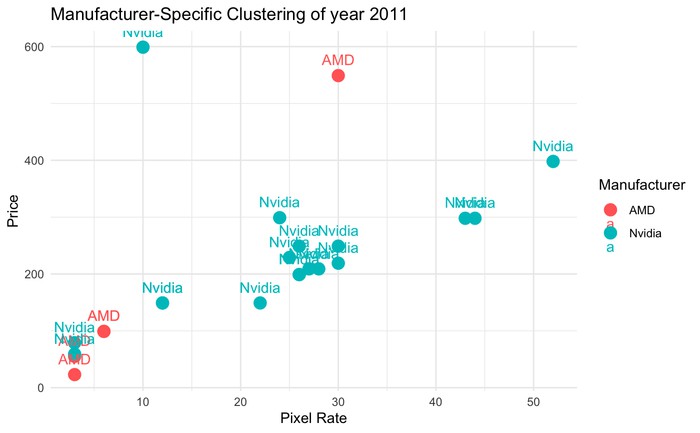
Đánh giá model khi dự đoán test data:



1. Tỷ lệ hiệu suất trên giá đã phát triển như thế nào theo thời gian?

Đồ thị cho ta thấy xu hướng đi lên từ các đường. Như vậy, theo thời gian, ta được hiệu suất tốt hơn với giá thành ít hơn.

4. Có nhà sản xuất nào được biết đến với một số phạm vi hiệu suất và giá cụ thể không?



Kết luận: Với mỗi cụm tượng trưng cho một phạm vi hiệu suất - giá thành:

 Năm 2011: Không có phạm vi hiệu suất và giá cụ thể đến từ các nhà sản xuất.

 Năm 2012: Có một số cụm nhưng đều có sự xuất hiện của 2 nhà sản xuất khác nhau nên không có phạm vi cụ thể.

 Năm 2013: Không có phạm vi hiệu suất và giá cụ thể đến từ các nhà sản xuất.

 Năm 2014: AMD có phạm vi hiệu suất cụ thể là Pixel\_Rate nằm ở khoảng 30, với giá thành dao động từ 150-300, còn lại là Nvidia với các cụm khác nhau.

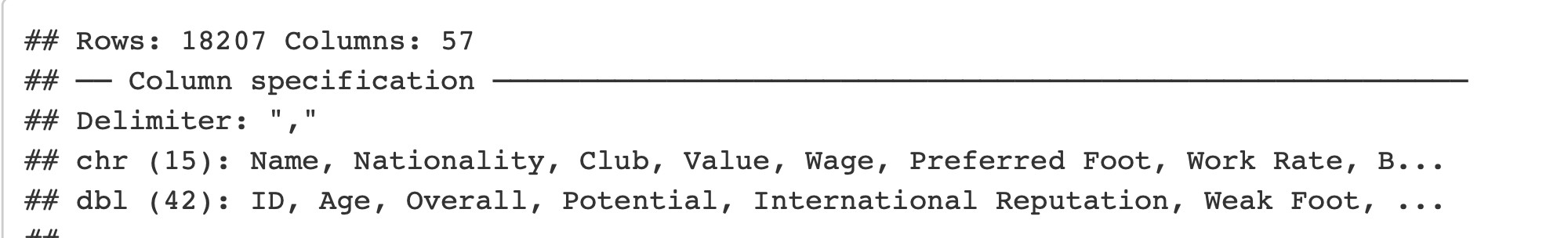
 Năm 2015: Cũng là AMD với ba cụm khác nhau.

 Năm 2016: ATI có hai cụm, còn Nvidia có bốn cụm.

 Năm 2017: AMD có ba cụm, Nvidia có một cụm dù số lượng ít.

# Hoạt động 2

Dữ liệu gồm thông tin của 18207 cầu thủ, được tổng hợp trong 01 file dữ liệu fifa\_eda\_stats.csv, bao gồm 57 biến.



Từ bộ dữ liệu thì ta có thể đặt ra những vấn đề rất hữu ích để câu lạc bộ có thể đưa ra những quyết định mua cầu thủ. Ở đây nhóm xin lựa chọn 1 bài toán demo “Dự đoán số tiền câu lạc bộ nên chi trả để mua cầu thủ dựa trên các chỉ số đo lường kỹ thuật…” Tuy nhiên việc tiền xử lý số liệu cũng như EDA gần như cho toàn bộ dữ liệu. Để có thể phát triển thêm những bài toán khác.

### Tiền xử lý dữ liệu

Clean dữ liệu. Các thao tác bao gồm:

 Đọc dataset và làm sạch tên cột

Hàm kiểm tra các giá trị NA (thiếu dữ liệu) sau đó xử lý

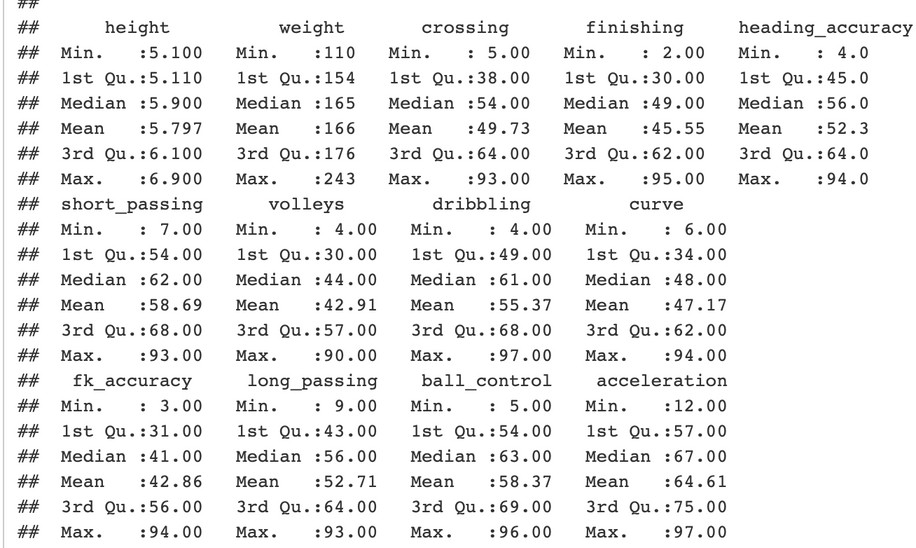
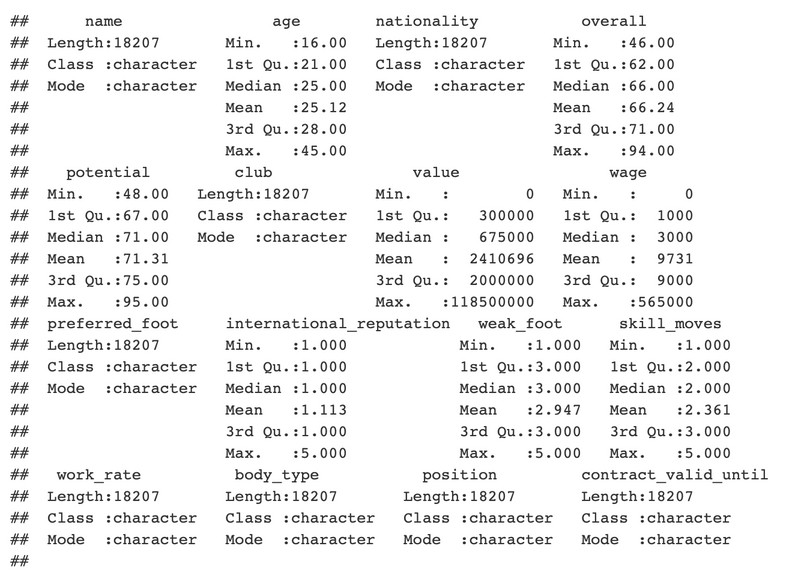


 Chuyển đổi dạng tiền tệ sang số

Đổi dạng các cột value, wage và release\_clause

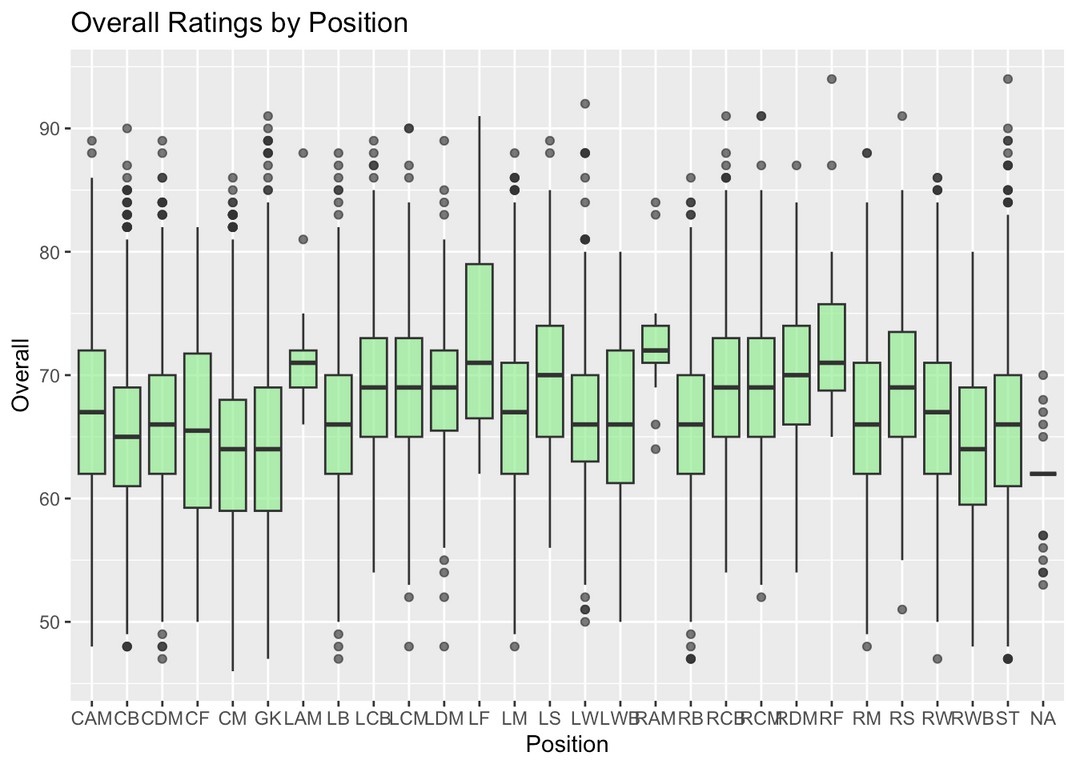


 Xử lý chiều cao (height) và cân nặng (weight)

 Loại bỏ các cột không cần thiết ("loaned\_from", "id", "jersey\_number", "joined")

### EDA

##### Boxplot Overall Ratings theo vị trí

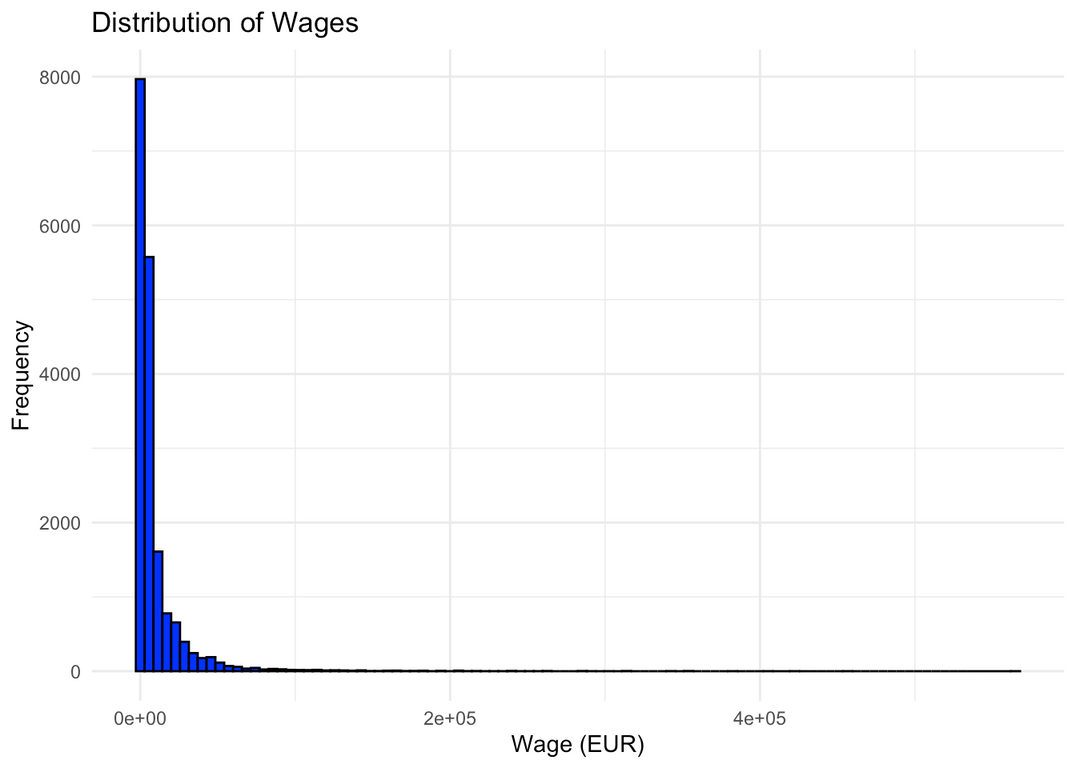
Nếu bài toán phân loại liên quan đến việc dự đoán nhóm cầu thủ (ví dụ: Cao, Trung bình, Thấp), dữ liệu có thể gặp vấn đề mất cân bằng (imbalance).

 Tạo nhóm hợp lý: Có thể chia thành 3 nhóm: Thấp: <60 Trung

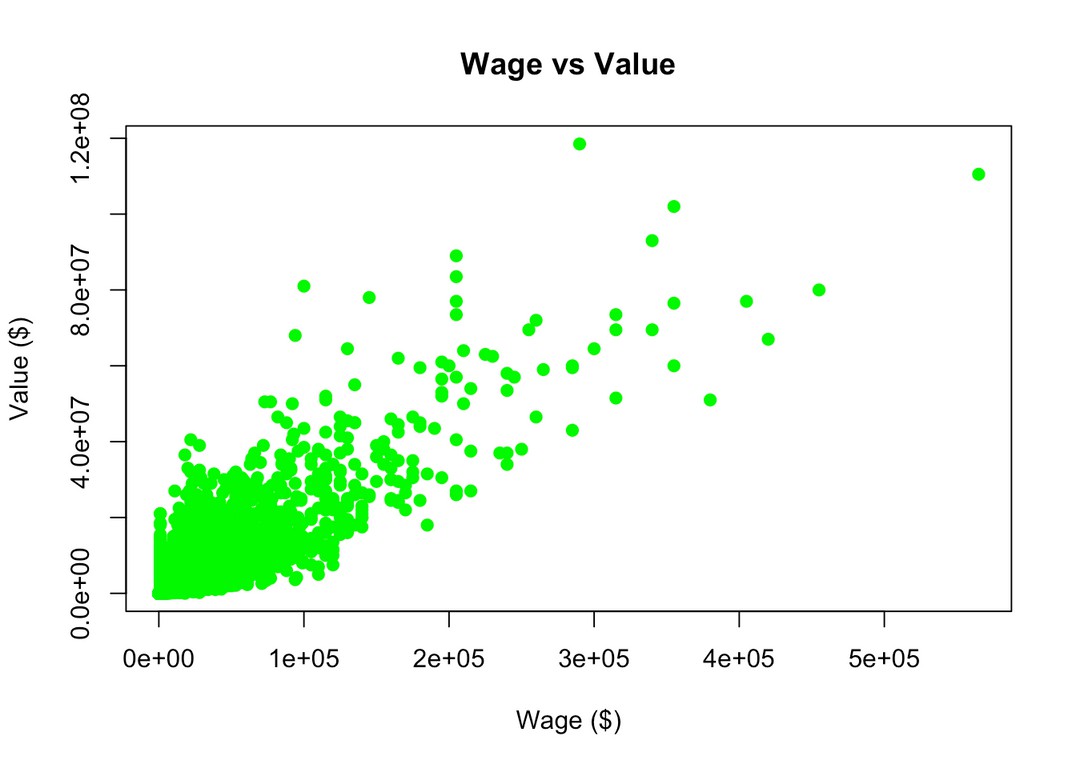
bình: 60-75 Cao: >75

 Điều này sẽ giúp bài toán phân loại trở nên rõ ràng và dễ huấn luyện hơn.

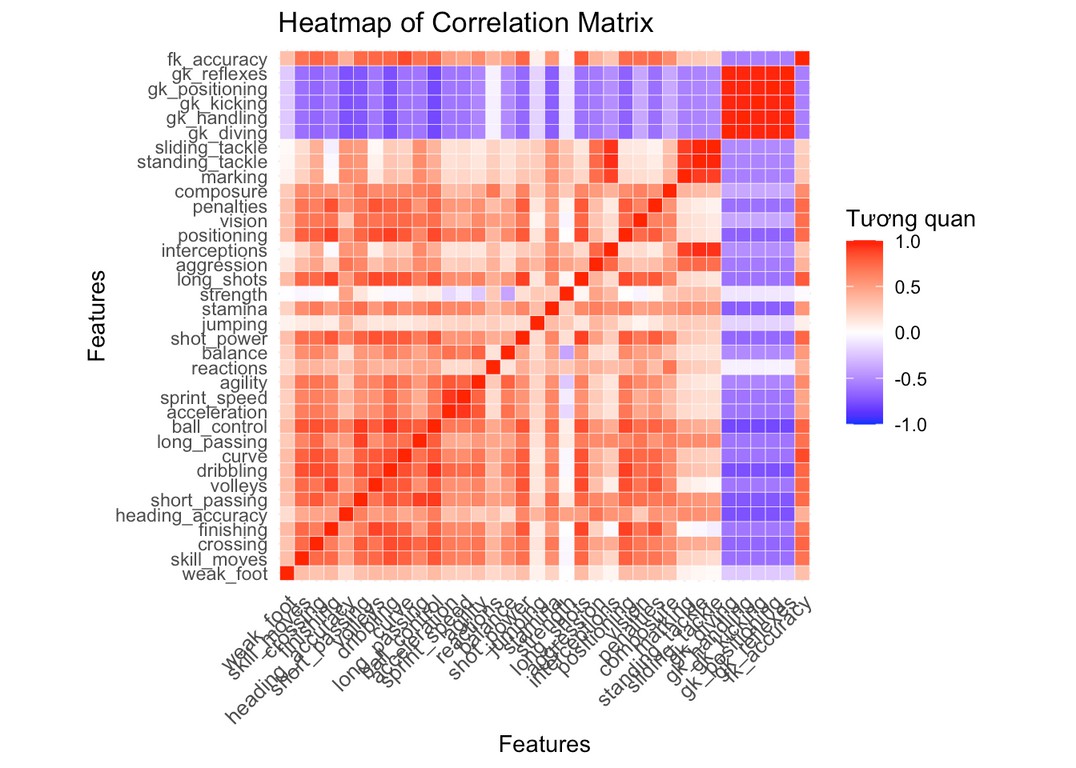
##### Biểu đồ phân phối Wages

Biểu đồ bị lệch phải và tồn tại nhiều giá trị ngoại lai (có thể sử dụng giá trị cook), có thể áp dụng bài toán phân loại bằng cách chia khoảng như ở trên

##### Biểu đồ Wages và Value

Biểu đồ bị lệch phải và tồn tại nhiều giá trị ngoại lai (có thể sử dụng giá trị cook), có thể áp dụng bài toán phân loại bằng cách chia khoảng như ở trên

##### Biểu diễn ma trận tương quan thông số kỹ thuật

Ta có thể dễ dàng thấy rất nhiều biến kỹ thuật có hệ số tương quan cao, điều đó dẫn đến nhiều hiện tượng (đa cộng tuyến,...) làm cho việc xây dựng model không được tốt, vì vậy cần xem xét lựa chọn model với các biến phù hợp

### Modeling

#### Chia train-test: 70-30

Dataset khi chia train-test chỉ sử dụng các biến định lượng.

#### Xây dựng model:

Xây dựng model với các biến giải thích: Tuổi, chiều cao, cân nặng, kỹ năng kỹ thuật (Finishing, Vision, Dribbling).

Biến mục tiêu total\_cost (wages + value)

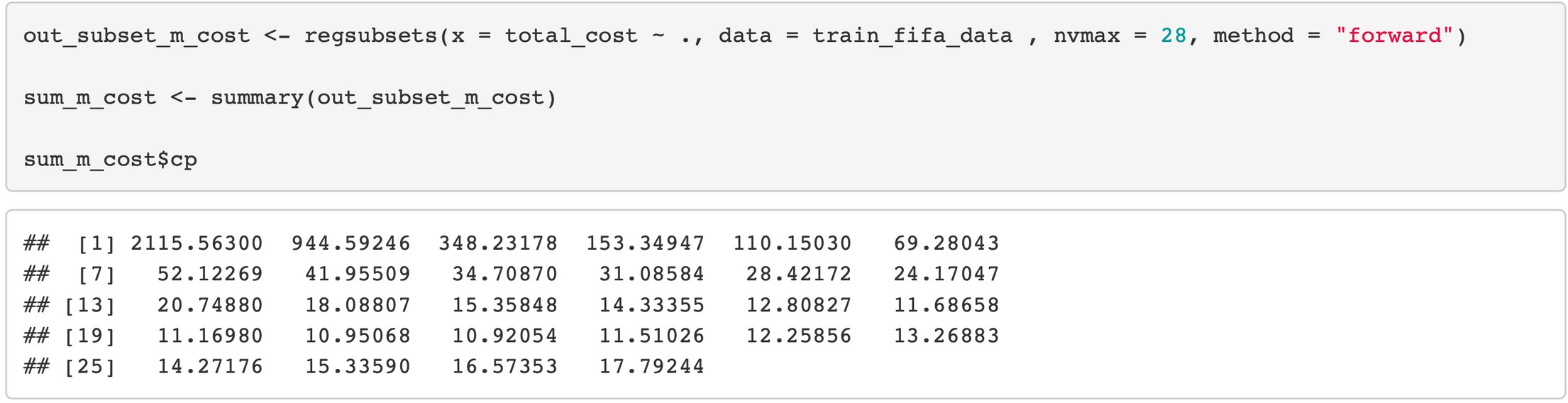
#### Kiểm tra tuyến tính

Có xu hướng là đường cong đáng kể nên giả định tuyến tính là không phù hợp

#### Kiểm đồng nhất phương sai

Biểu đồ cho thấy rằng phần dư chuẩn hóa có sự thay đổi (đường cong màu xanh lam chỉ ra một xu hướng phi tuyến tính). Điều này chỉ ra vi phạm tính đồng nhất phương sai, tức là mô hình có thể có phương sai không đồng nhất (heteroscedasticity)

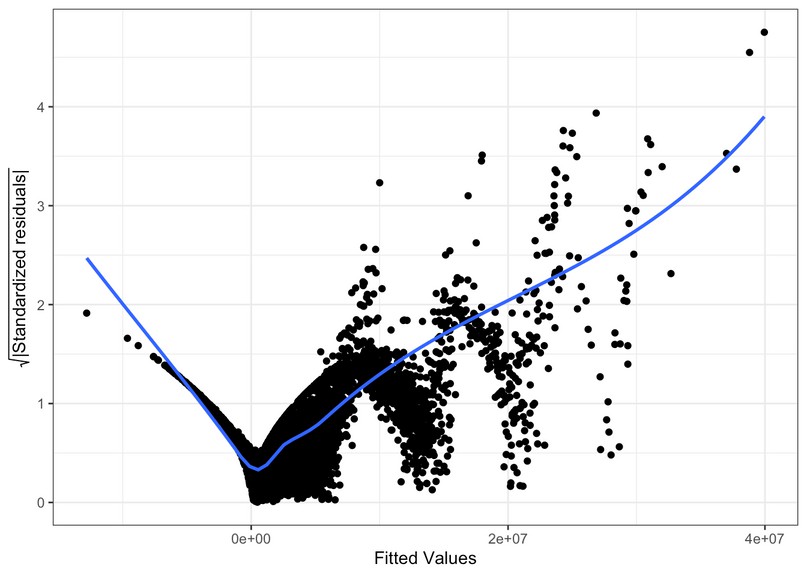
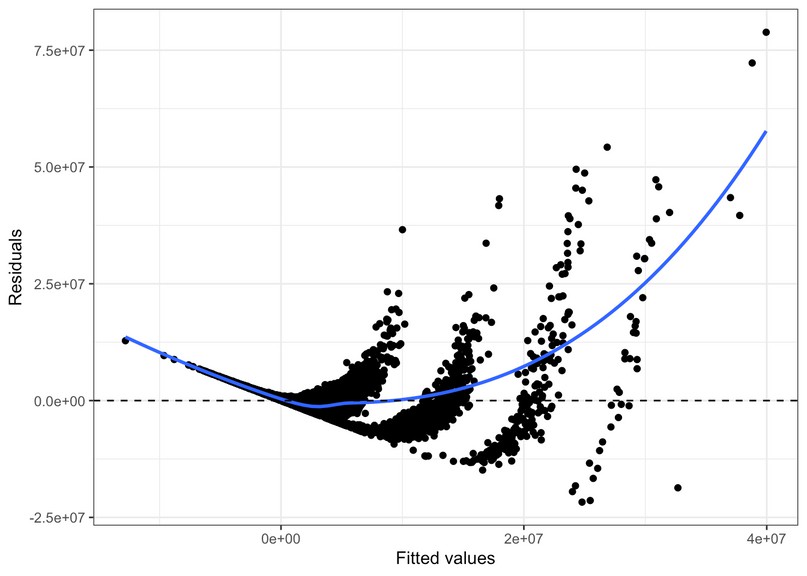
**Lựa chọn mô hình**



Dựa vào kết quả sử dụng Mallow’s cp ta dễ thấy sử dụng 21 biến là phù hợp

**Áp dụng cải tiến mô hình**

**Kiểm tra giả định tuyến tính Kiểm tra đồng nhất phương sai**



**Có xu hướng đường cong đáng kể nên giả định tuyến tính là chưa phù hợp**

**Biểu đồ cho thấy rằng phần dư chuẩn hóa có sự thay đổi (đường cong màu xanh lam chỉ ra một xu hướng phi tuyến tính). Điều này chỉ ra vi phạm tính đồng nhất phương sai, tức là mô hình có thể có phương sai không đồng nhất (heteroscedasticity).**

**Kết luận**

 Sau khi áp dụng cải tiến mô hình:

 Model có chỉ số RMSE thấp hơn model trước nhưng vẫn cao (3464457 so với 5037073 của mô hình trước) có thể cho thấy sự overfit của model khi sử dụng 21 biến giải thích.

 R-square được cải tiến đáng kể (0.6074266 so với 0.1701358) cho thấy model hoạt động khá tốt.

 Không giải quyết được tình trạng giả định tuyến tính và đồng nhất phương sai.

=> Mô hình dù đã vận hành tốt hơn nhưng vẫn chưa được cải tiến một cách tối ưu nhất có thể. Hoặc có thể data vẫn chưa được xử lý tốt dẫn đến model hoạt động không như mong muốn