Đồ Án Nhóm 1

Thanh Thảo - Bích Trâm

2024-07-23

Contents

PHÂN CÔNG VIỆC **3**

Bộ dữ liệu: CHOCOLATE ========== **4**
MÔ TẢ DỮ LIỆU . . . 4
YÊU CẦU . . . 4
ĐỌC DỮ LIỆU . . . 5
TIỀN XỬ LÝ DỮ LIỆU . . . 5
Loại bỏ dữ liệu trùng . . . 5
Loại bỏ biến không có giá trị phân tích . . . 5
Chuyển đổi kiểu dữ liệu . . . 6
Loại bỏ dữ liệu khuyết . . . 8
Kiểm tra phần dư . . . 9
Loại bỏ outlier . . . 10
Kiểm tra tương tác dữ liệu . . . 11
XÂY DỰNG MÔ HÌNH . . . 13

Bộ dữ liệu: FISH MARKET ========== **18**
MÔ TẢ DỮ LIỆU . . . 18
YÊU CẦU . . . 18
ĐỌC DỮ LIỆU . . . 19
TIỀN XỬ LÝ DỮ LIỆU . . . 19
Loại bỏ dữ liệu trùng . . . 19
Loại bỏ biến không có giá trị phân tích . . . 19
Thay thế dữ liệu bất thường . . . 19
Chuyển đổi kiểu dữ liệu . . . 19
Loại bỏ dữ liệu khuyết . . . 20
Quy tâm dữ liệu . . . 20
Loại bỏ outlier . . . 21
CHIA DỮ LIỆU . . . 22
XÂY DỰNG MÔ HÌNH . . . 23
Kiểm tra tương quan . . . 23
Kiểm tra đa cộng tuyến . . . 24
Xây dựng mô hình bằng phương pháp hồi quy bội . . . 25
So sánh mô hình xây dựng với mô hình tạo bằng phương pháp Stepswise . . . 25
Kiểm tra giả thiết tuyến tính của dữ liệu và giả thiết phần dư có trung bình bằng 0 . . . 27
Kiểm tra phần dư có phân phối chuẩn . . . 28
Kiểm định giả thiết phương sai của phần dư không đổi . . . 29
Kiểm tra sự ảnh hưởng của dữ liệu . . . 30
DỰ BÁO . . . 31

Bộ dữ liệu: CSM ========== **32**
MÔ TẢ DỮ LIỆU . . . 32

YÊU CẦU . . . 32
ĐỌC DỮ LIỆU . . . 33
TIỀN XỬ LÝ DỮ LIỆU . . . 33
Loại bỏ dữ liệu trùng . . . 33
Loại bỏ biến không có giá trị phân tích . . . 33
Chuyển đổi kiểu dữ liệu . . . 34
Loại bỏ dữ liệu khuyết . . . 34
Thay thế dữ liệu khuyết . . . 38
Quy tâm dữ liệu . . . 39
Loại bỏ outlier . . . 41
CHIA DỮ LIỆU . . . 42
XÂY DỰNG MÔ HÌNH . . . 42
Kiểm tra tương quan . . . 42
Kiểm tra đa cộng tuyến . . . 43
Xây dựng mô hình bằng phương pháp hồi quy bội . . . 44
So sánh mô hình xây dựng với mô hình tạo bằng phương pháp Stepswise . . . 45
Kiểm tra giả thiết tuyến tính của dữ liệu và giả thiết phần dư có trung bình bằng 0 . . . 47
Kiểm tra phần dư có phân phối chuẩn . . . 48
Kiểm định giả thiết phương sai của phần dư không đổi . . . 49
Kiểm tra sự ảnh hưởng của dữ liệu . . . 50
DỰ BÁO . . . 51

DỮ LIỆU **52**

THAM KHẢO **53**

LINK LẤY DỮ LIỆU **53**

PHÂN CÔNG VIỆC

1. ĐỖ THỊ THANH THẢO - 23C23009

- Làm bộ dữ liệu: Csm, Fish Market, Chocolate Bar Rating
- Tổng hợp báo cáo
- Hoàn thành: 100%

2. NGUYỄN BÍCH TRÂM - 23c23010

- Làm bộ dữ liệu: IslanderOrg, Heart Disease, Chocolate Bar Rating
- Tìm dữ liệu

Bộ dữ liệu: CHOCOLATE ==========

MÔ TẢ DỮ LIỆU

Sô cô la là một trong những loại kẹo phổ biến nhất trên thế giới. Mỗi năm, cư dân Hoa Kỳ ăn tổng cộng hơn 2,8 tỷ pound. Tuy nhiên, không phải tất cả các thanh sô cô la đều được tạo ra như nhau. Bộ dữ liệu này chứa xếp hạng của chuyên gia đối với hơn 1.700 thanh sô cô la riêng lẻ, cùng với thông tin về nguồn gốc khu vực, tỷ lệ ca cao, loại hạt sô cô la được sử dụng và nơi trồng hạt.

- "CompanyMaker": tên công ty sản xuất
- "SpecificBeanOrigin": Xuất xứ
- "REF": giá trị liên kết với thời điểm đánh giá, cao hơn là gần đây hơn
- "ReviewDate": ngày công bố đánh giá
- "CocoaPercent": tỉ lệ ca cao
- "CompanyLocation": quốc gia công ty sản xuất
- "Rating": đánh giá của chuyên gia
- "BeanType": loại hạt
- "BroadBeanOrigin": nguồn gốc hạt

YÊU CẦU

- Kiểm tra tỉ lệ ca cao (CocoaPercent) và quốc gia sản xuất (CompanyLocaion) có ảnh hướng đến đánh giá của chuyên gia không (Rating)

ĐỌC DỮ LIỆU

```
chocolateOrgData = docDuLieu("data","Chocolate.csv")

chocolateOrgData = rename(chocolateOrgData,c('CompanyMaker'='Company...Maker.if.known.'))
chocolateOrgData = rename(chocolateOrgData,
                          c('SpecificBeanOrigin'='Specific.Bean.Origin.or.Bar.Name'))
chocolateOrgData = rename(chocolateOrgData,c('ReviewDate'='Review.Date'))
chocolateOrgData = rename(chocolateOrgData,c('CocoaPercent'='Cocoa.Percent'))
chocolateOrgData = rename(chocolateOrgData,c('CompanyLocation'='Company.Location'))
chocolateOrgData = rename(chocolateOrgData,c('BeanType'='Bean.Type'))
chocolateOrgData = rename(chocolateOrgData,c('BroadBeanOrigin'='Broad.Bean.Origin'))
```

TIỀN XỬ LÝ DỮ LIỆU

Loại bỏ dữ liệu trùng

```
isTRUE(duplicated(chocolateOrgData))
```

```
## [1] FALSE
```

Nhận xét

- Không có dữ liệu trùng

Loại bỏ biến không có giá trị phân tích

```
summary(chocolateOrgData)
```

```
##  CompanyMaker       SpecificBeanOrigin      REF         ReviewDate
##  Length:1795        Length:1795        Min.   :   5   Min.   :2006
##  Class :character   Class :character   1st Qu.: 576   1st Qu.:2010
##  Mode  :character   Mode  :character   Median :1069   Median :2013
##                                        Mean   :1036   Mean   :2012
##                                        3rd Qu.:1502   3rd Qu.:2015
##                                        Max.   :1952   Max.   :2017
##  CocoaPercent       CompanyLocation        Rating        BeanType
##  Length:1795        Length:1795        Min.   :1.000   Length:1795
##  Class :character   Class :character   1st Qu.:2.875   Class :character
##  Mode  :character   Mode  :character   Median :3.250   Mode  :character
##                                        Mean   :3.186
##                                        3rd Qu.:3.500
##                                        Max.   :5.000
##  BroadBeanOrigin
##  Length:1795
##  Class :character
##  Mode  :character
##
##
##
```

```
chocolateData = chocolateOrgData[,c(5, 6, 7)]
```

Nhận xét

- Do bài toán cần 3 biến "CocoaPercent", "CompanyLocaion", "Rating" để đánh giá nên ta giữ lại 3 biến này. Trong đó, "Rating" là biến phụ thuộc, "CocoaPercent", "CompanyLocaion" là biến độc lập

Chuyển đổi kiểu dữ liệu

```
str(chocolateData)
```

```
## 'data.frame':    1795 obs. of  3 variables:
##  $ CocoaPercent   : chr  "63%" "70%" "70%" "70%" ...
##  $ CompanyLocation: chr  "France" "France" "France" "France" ...
##  $ Rating         : num  3.75 2.75 3 3.5 3.5 2.75 3.5 3.5 3.75 4 ...
```

```
chocolateData$CocoaPercent = sub("%", "", chocolateData$CocoaPercent)
arrange(tabyl(chocolateData$CocoaPercent), desc(percent))
```

```
##  chocolateData$CocoaPercent   n      percent
##                          70 672 0.3743732591
##                          75 222 0.1236768802
##                          72 189 0.1052924791
##                          65  78 0.0434540390
##                          80  72 0.0401114206
##                          74  50 0.0278551532
##                          68  47 0.0261838440
##                          60  43 0.0239554318
##                          73  40 0.0222841226
##                          85  36 0.0200557103
##                          64  34 0.0189415042
##                          77  33 0.0183844011
##                          71  31 0.0172701950
##                          67  27 0.0150417827
##                          66  23 0.0128133705
##                          76  23 0.0128133705
##                         100  20 0.0111420613
##                          78  17 0.0094707521
##                          82  17 0.0094707521
##                          55  16 0.0089136490
##                          62  14 0.0077994429
##                          63  12 0.0066852368
##                          69  10 0.0055710306
##                          58   8 0.0044568245
##                          61   8 0.0044568245
##                          88   8 0.0044568245
##                          90   8 0.0044568245
##                          81   5 0.0027855153
##                        72.5   4 0.0022284123
##                          83   4 0.0022284123
##                          84   4 0.0022284123
##                          91   3 0.0016713092
##                          56   2 0.0011142061
##                        73.5   2 0.0011142061
##                          89   2 0.0011142061
##                          99   2 0.0011142061
##                          42   1 0.0005571031
##                          46   1 0.0005571031
##                          50   1 0.0005571031
##                          53   1 0.0005571031
##                          57   1 0.0005571031
##                        60.5   1 0.0005571031
##                          79   1 0.0005571031
##                          86   1 0.0005571031
```

```
##                           87   1 0.0005571031
```

```
chocolateData$CocoaPercent = case_when(
                                      chocolateData$CocoaPercent == 70 ~ 70,
                                      chocolateData$CocoaPercent == 75 ~ 75,
                                      chocolateData$CocoaPercent == 72 ~ 72,
                                      TRUE ~ 65)
chocolateData$CocoaPercent = as.factor(chocolateData$CocoaPercent)

arrange(tabyl(chocolateData$CompanyLocation), desc(percent))
```

```
##  chocolateData$CompanyLocation   n      percent
##                        U.S.A. 764 0.4256267409
##                         France 156 0.0869080780
##                         Canada 125 0.0696378830
##                           U.K.  96 0.0534818942
##                          Italy  63 0.0350974930
##                        Ecuador  54 0.0300835655
##                      Australia  49 0.0272980501
##                        Belgium  40 0.0222841226
##                    Switzerland  38 0.0211699164
##                        Germany  35 0.0194986072
##                        Austria  26 0.0144846797
##                          Spain  25 0.0139275766
##                       Colombia  23 0.0128133705
##                        Hungary  22 0.0122562674
##                      Venezuela  20 0.0111420613
##                         Brazil  17 0.0094707521
##                          Japan  17 0.0094707521
##                     Madagascar  17 0.0094707521
##                    New Zealand  17 0.0094707521
##                           Peru  17 0.0094707521
##                        Denmark  15 0.0083565460
##                        Vietnam  11 0.0061281337
##                      Guatemala  10 0.0055710306
##                       Scotland  10 0.0055710306
##                      Argentina   9 0.0050139276
##                     Costa Rica   9 0.0050139276
##                         Israel   9 0.0050139276
##                         Poland   8 0.0044568245
##                       Honduras   6 0.0033426184
##                      Lithuania   6 0.0033426184
##                Domincan Republic   5 0.0027855153
##                        Nicaragua   5 0.0027855153
##                    South Korea   5 0.0027855153
##                         Sweden   5 0.0027855153
##                      Amsterdam   4 0.0022284123
##                           Fiji   4 0.0022284123
##                        Ireland   4 0.0022284123
##                         Mexico   4 0.0022284123
##                    Netherlands   4 0.0022284123
##                    Puerto Rico   4 0.0022284123
##                       Sao Tome   4 0.0022284123
##                        Grenada   3 0.0016713092
##                        Iceland   3 0.0016713092
##                       Portugal   3 0.0016713092
##                      Singapore   3 0.0016713092
```

```
##                    South Africa   3 0.0016713092
##                         Bolivia   2 0.0011142061
##                           Chile   2 0.0011142061
##                         Finland   2 0.0011142061
##                        St. Lucia   2 0.0011142061
##                  Czech Republic   1 0.0005571031
##                         Eucador   1 0.0005571031
##                           Ghana   1 0.0005571031
##                           India   1 0.0005571031
##                      Martinique   1 0.0005571031
##                       Niacragua   1 0.0005571031
##                     Philippines   1 0.0005571031
##                          Russia   1 0.0005571031
##                        Suriname   1 0.0005571031
##                           Wales   1 0.0005571031
```

```
chocolateData$CompanyLocation = case_when(
                                        chocolateData$CompanyLocation == 'U.S.A.' ~ 1,
                                        chocolateData$CompanyLocation == 'France' ~ 2,
                                        chocolateData$CompanyLocation == 'Canada' ~ 3,
                                        TRUE ~ 4)
chocolateData$CompanyLocation = as.factor(chocolateData$CompanyLocation)
```

Nhận xét

- Biến "CocoaPercent" đang lưu trữ dạng ký tự và do số lượng giá trị phân biệt khá lớn nên ta chuyển về dạng factor. Với những dữ liệu có phần trăm ca cao bằng 70, 72, 75, do phần trăm dữ liệu của 3 loại này khá cao so với các loại khác nên ta giữ nguyên giá trị, các dữ liệu còn lại ta thay bằng giá trị 65 để phân loại

- Biến "CompanyLocation" đang lưu trữ dạng ký tự và do số lượng giá trị phân biệt khá lớn nên ta chuyển về dạng factor. Với những dữ liệu là "U.S.A.", "France", "Canada", do phần trăm dữ liệu của 3 loại này cao nhất so với các loại khác nên ta thay lần lượt bởi các giá trị 1, 2, 3 và những dữ liệu còn lại ta thay bằng 4 để phân loại

Loại bỏ dữ liệu khuyết

```
anyNA(chocolateData)
```

```
## [1] FALSE
```

Nhận xét

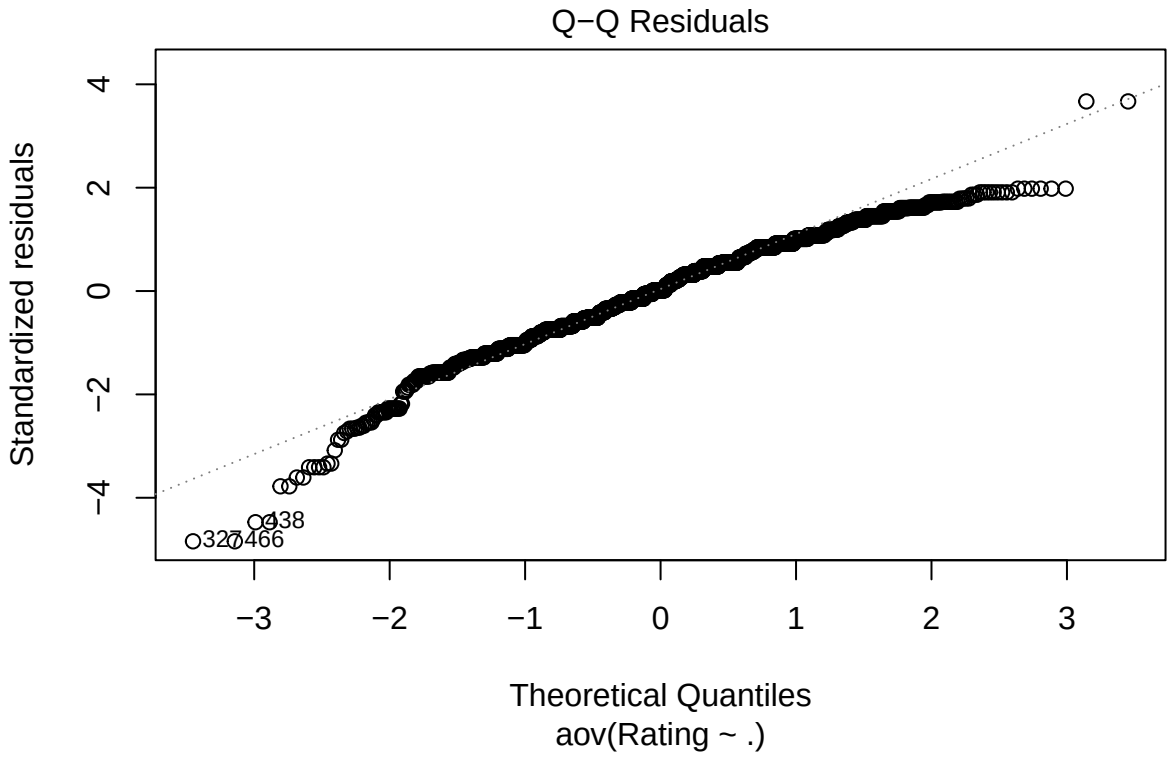
- Không có dữ liệu khuyết

Kiểm tra phần dư

```
model = aov(Rating ~ ., data = chocolateData)
aovResidual = rstandard(model)
shapiro.test(aovResidual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aovResidual
## W = 0.97584, p-value < 2.2e-16
```

```
plot(model, 2)
```

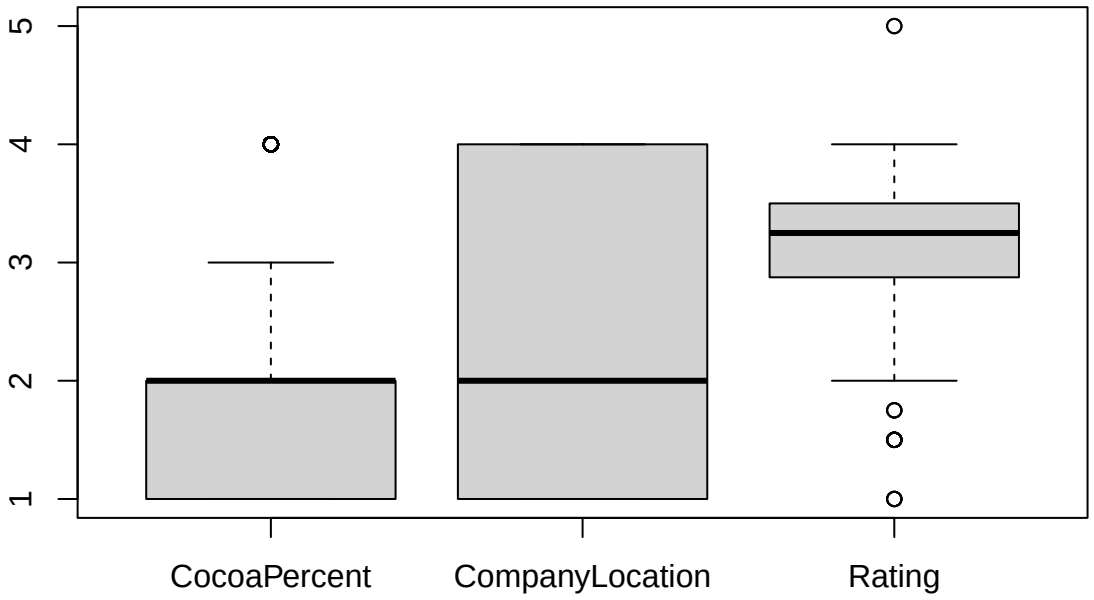


Nhận xét

- Từ biểu đồ và kiểm định Shapiro-Wilk (p_value < 2.2e-16), ta thấy phần dư không có phân phối chuẩn

Loại bỏ outlier

```
boxplot(chocolateData)
```



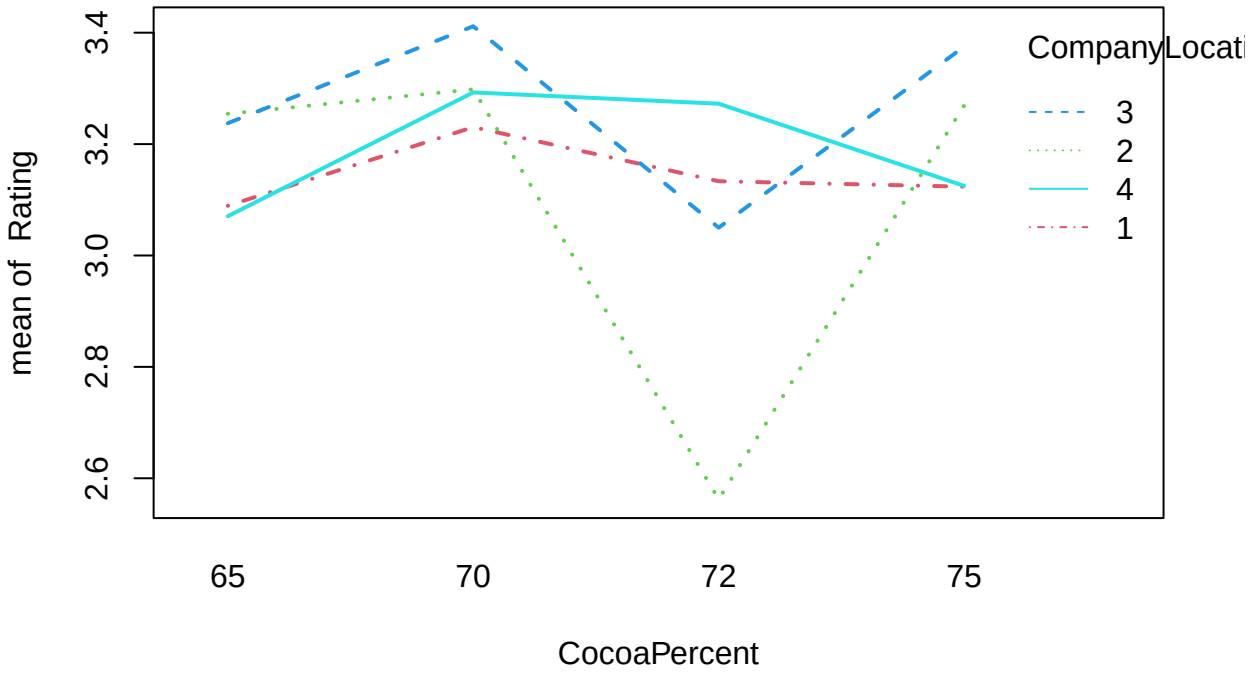

```
ratingOutlierFrame = subset(chocolateData, Rating < 2 | Rating > 4)
ratingOutlierDataIndex = as.numeric(rownames(ratingOutlierFrame))
duplicatedOutlierAmount = length(ratingOutlierDataIndex)
duplicatedOutlierPercentage = round(duplicatedOutlierAmount/dim(chocolateData)[1]*100, 2)
```

Nhận xét

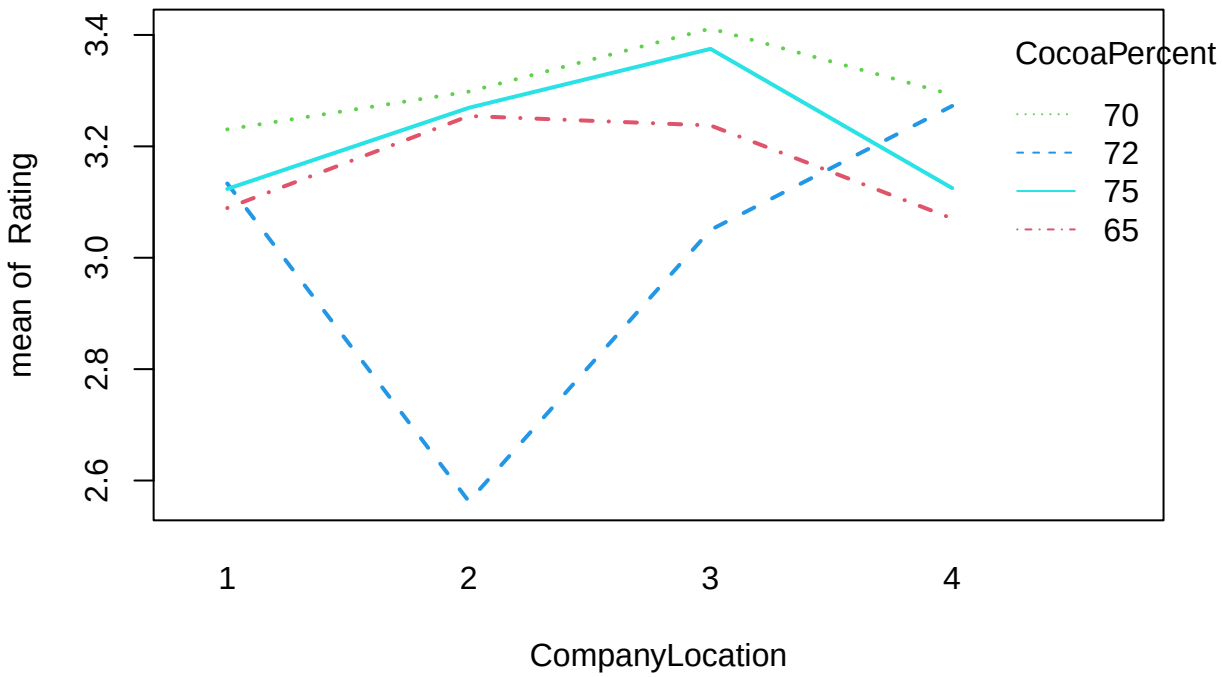
- Qua biểu đồ hộp, ta thấy biến "CocoaPercent", "Rating" có outlier
- Biến "Rating" có 19 dòng có outlier, chiếm 1.06% dữ liệu
- Các dòng cần loại khỏi bộ dữ liệu là: 79, 87, 126, 133, 246, 250, 325, 327, 438, 450, 466, 555, 829, 989, 1130, 1176, 1412, 1493, 1695

Kiểm tra tương tác dữ liệu

```
with(chocolateData, interaction.plot(CocoaPercent, CompanyLocation, Rating, lwd = 2, col = 2:9))
```




```
with(chocolateData, interaction.plot(CompanyLocation, CocoaPercent, Rating, lwd = 2, col = 2:9))
```



Nhận xét

- Từ hai biểu đồ ta thấy, nhân tố "CompanyLocation" và nhân tố "CocoaPercent" đều có tương tác lẫn nhau

XÂY DỰNG MÔ HÌNH

```
str(chocolateData)
```

```
## 'data.frame':    1795 obs. of  3 variables:
##  $ CocoaPercent   : Factor w/ 4 levels "65","70","72",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ CompanyLocation: Factor w/ 4 levels "1","2","3","4": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Rating         : num  3.75 2.75 3 3.5 3.5 2.75 3.5 3.5 3.75 4 ...
```

```
model = aov(Rating ~ CocoaPercent * CompanyLocation, data = chocolateData)
summary(model)
```

```
##                                Df Sum Sq Mean Sq F value   Pr(>F)
## CocoaPercent                    3   10.5    3.49  15.871 3.48e-10 ***
## CompanyLocation                 3    3.8    1.26   5.731 0.000669 ***
## CocoaPercent:CompanyLocation    9    4.6    0.51   2.319 0.013605 *
## Residuals                    1779  391.2    0.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Rating ~ CocoaPercent * CompanyLocation, data = chocolateData)
##
## $CocoaPercent
##               diff         lwr          upr     p adj
## 70-65  0.17386470  0.10901365  0.238715749 0.0000000
## 72-65  0.08829922 -0.01036622  0.186964671 0.0980869
## 75-65  0.07575096 -0.01693854  0.168440463 0.1529570
## 72-70 -0.08556548 -0.18484499  0.013714035 0.1191854
## 75-70 -0.09811374 -0.19145663 -0.004770852 0.0349721
## 75-72 -0.01254826 -0.13188856  0.106792034 0.9930892
##
## $CompanyLocation
##            diff         lwr          upr     p adj
## 2-1  0.11606290  0.01012325  0.222002544 0.0252340
## 3-1  0.15330369  0.03696537  0.269642012 0.0040001
## 4-1  0.03410924 -0.02787186  0.096090328 0.4899823
## 3-2  0.03724080 -0.10750616  0.181987755 0.9114638
## 4-2 -0.08195366 -0.18806084  0.024153518 0.1935592
## 4-3 -0.11919446 -0.23568535 -0.002703562 0.0426097
##
## $`CocoaPercent:CompanyLocation`
##                     diff          lwr           upr     p adj
## 70:1-65:1   0.141230437  0.010286200  0.2721746732 0.0202209
## 72:1-65:1   0.044107744 -0.163803484  0.2520189723 0.9999969
## 75:1-65:1   0.034303823 -0.163602343  0.2322099880 0.9999998
## 65:2-65:1   0.165238696 -0.069146644  0.3996240366 0.5334350
## 70:2-65:1   0.209161508 -0.094502403  0.5128254182 0.5758637
## 72:2-65:1  -0.526725589 -1.336549057  0.2830978782 0.6745070
## 75:2-65:1   0.180005180 -0.040305075  0.4003154347 0.2678621
## 65:3-65:1   0.148274411 -0.122696119  0.4192449406 0.8861145
## 70:3-65:1   0.322312872  0.102002618  0.5426231270 0.0000658
## 72:3-65:1  -0.039225589 -0.556482490  0.4780313116 1.0000000
## 75:3-65:1   0.285774411 -0.231482490  0.8030313116 0.8780963
```

```
## 65:4-65:1 -0.018692674 -0.148420195  0.1110348475 1.0000000
## 70:4-65:1  0.203525340  0.068109590  0.3389410907 0.0000325
## 72:4-65:1  0.183274411 -0.002733904  0.3692827260 0.0585492
## 75:4-65:1  0.035774411 -0.188866311  0.2604151322 0.9999999
## 72:1-70:1 -0.097122693 -0.304350100  0.1101047141 0.9670772
## 75:1-70:1 -0.106926614 -0.304114265  0.0902610366 0.8933039
## 65:2-70:1  0.024008260 -0.209770711  0.2577872301 1.0000000
## 70:2-70:1  0.067931071 -0.235265054  0.3711271952 0.9999935
## 72:2-70:1 -0.667956026 -1.477604201  0.1416921492 0.2524797
## 75:2-70:1  0.038774743 -0.180890292  0.2584397783 0.9999998
## 65:3-70:1  0.007043974 -0.263402226  0.2774901744 1.0000000
## 70:3-70:1  0.181082435 -0.038582600  0.4007474706 0.2536950
## 72:3-70:1 -0.180456026 -0.697438444  0.3365263922 0.9983498
## 75:3-70:1  0.144543974 -0.372438444  0.6615263922 0.9998839
## 65:4-70:1 -0.159923111 -0.288551837 -0.0312943849 0.0022225
## 70:4-70:1  0.062294903 -0.072068577  0.1966583836 0.9701005
## 72:4-70:1  0.042043974 -0.143199684  0.2272876314 0.9999922
## 75:4-70:1 -0.105456026 -0.329464001  0.1185519494 0.9657583
## 75:1-72:1 -0.009803922 -0.264683610  0.2450757666 1.0000000
## 65:2-72:1  0.121130952 -0.163005094  0.4052669986 0.9863622
## 70:2-72:1  0.165053763 -0.178469999  0.5085775264 0.9591142
## 72:2-72:1 -0.570833333 -1.396430604  0.2547639374 0.5689878
## 75:2-72:1  0.135897436 -0.136744112  0.4085389834 0.9443527
## 65:3-72:1  0.104166667 -0.210828306  0.4191616398 0.9991126
## 70:3-72:1  0.278205128  0.005563581  0.5508466757 0.0399134
## 72:3-72:1 -0.083333333 -0.624952671  0.4582860040 1.0000000
## 75:3-72:1  0.241666667 -0.299952671  0.7832860040 0.9789373
## 65:4-72:1 -0.062800418 -0.269261152  0.1436603165 0.9996716
## 70:4-72:1  0.159417596 -0.050663890  0.3694990816 0.3965360
## 72:4-72:1  0.139166667 -0.106589103  0.3849224365 0.8555592
## 75:4-72:1 -0.008333333 -0.284485934  0.2678192673 1.0000000
## 65:2-75:1  0.130934874 -0.145964141  0.4078338886 0.9643996
## 70:2-75:1  0.174857685 -0.162704675  0.5124200455 0.9243361
## 72:2-75:1 -0.561029412 -1.384164045  0.2621052218 0.5947195
## 75:2-75:1  0.145701357 -0.119389542  0.4107922569 0.8824681
## 65:3-75:1  0.113970588 -0.194512158  0.4224533343 0.9968987
## 70:3-75:1  0.288009050  0.022918150  0.5530999492 0.0183592
## 72:3-75:1 -0.073529412 -0.611387458  0.4643286347 1.0000000
## 75:3-75:1  0.251470588 -0.286387458  0.7893286347 0.9677771
## 65:4-75:1 -0.052996496 -0.249378284  0.1433852910 0.9999258
## 70:4-75:1  0.169221518 -0.030963402  0.3694064374 0.2157130
## 72:4-75:1  0.148970588 -0.088380776  0.3863219521 0.7297333
## 75:4-75:1  0.001470588 -0.267230045  0.2701712216 1.0000000
## 70:2-65:2  0.043922811 -0.316241019  0.4040866412 1.0000000
## 72:2-65:2 -0.691964286 -1.524622837  0.1406942652 0.2411722
## 75:2-65:2  0.014766484 -0.278564085  0.3080970518 1.0000000
## 65:3-65:2 -0.016964286 -0.350027706  0.3160991346 1.0000000
## 70:3-65:2  0.157074176 -0.136256392  0.4504047441 0.9026803
## 72:3-65:2 -0.204464286 -0.756787484  0.3478589126 0.9968309
## 75:3-65:2  0.120535714 -0.431787484  0.6728589126 0.9999954
## 65:4-65:2 -0.183931370 -0.417031014  0.0491682731 0.3264232
## 70:4-65:2  0.038286644 -0.198025945  0.2745992321 0.9999999
## 72:4-65:2  0.018035714 -0.250488600  0.2865600284 1.0000000
## 75:4-65:2 -0.129464286 -0.426061096  0.1671325243 0.9828769
## 72:2-70:2 -0.735887097 -1.590635917  0.1188617232 0.1900552
## 75:2-70:2 -0.029156328 -0.380323104  0.3220104494 1.0000000
```

```
## 65:3-70:2 -0.060887097 -0.445863102  0.3240889088 0.9999999
## 70:3-70:2  0.113151365 -0.238015412  0.4643181417 0.9993456
## 72:3-70:2 -0.248387097 -0.833482130  0.3367079368 0.9869108
## 75:3-70:2  0.076612903 -0.508482130  0.6617079368 1.0000000
## 65:4-70:2 -0.227854181 -0.530526821  0.0748184581 0.4112403
## 70:4-70:2 -0.005636167 -0.310790100  0.2995177649 1.0000000
## 72:4-70:2 -0.025887097 -0.356614462  0.3048402686 1.0000000
## 75:4-70:2 -0.173387097 -0.527286730  0.1805125368 0.9517587
## 75:2-72:2  0.706730769 -0.122075832  1.5355373709 0.2033011
## 65:3-72:2  0.675000000 -0.168687622  1.5186876215 0.3026369
## 70:3-72:2  0.849038462  0.020231860  1.6778450632 0.0381554
## 72:3-72:2  0.487500000 -0.464308004  1.4393080038 0.9307442
## 75:3-72:2  0.812500000 -0.139308004  1.7643080038 0.2017738
## 65:4-72:2  0.508032915 -0.301419371  1.3174852021 0.7297611
## 70:4-72:2  0.730250929 -0.080132437  1.5406342956 0.1341590
## 72:4-72:2  0.710000000 -0.110355330  1.5303553303 0.1830161
## 75:4-72:2  0.562500000 -0.267468209  1.3924682093 0.6047184
## 65:3-75:2 -0.031730769 -0.355043872  0.2915823339 1.0000000
## 70:3-75:2  0.142307692 -0.139903121  0.4245185053 0.9388496
## 72:3-75:2 -0.219230769 -0.765729659  0.3272681202 0.9925819
## 75:3-75:2  0.105769231 -0.440729659  0.6522681202 0.9999991
## 65:4-75:2 -0.198697854 -0.417639774  0.0202440663 0.1265273
## 70:4-75:2  0.023520160 -0.198839369  0.2458796891 1.0000000
## 72:4-75:2  0.003269231 -0.253061470  0.2595999311 1.0000000
## 75:4-75:2 -0.144230769 -0.429835021  0.1413734821 0.9381204
## 70:3-65:3  0.174038462 -0.149274642  0.4973515646 0.8988553
## 72:3-65:3 -0.187500000 -0.756314078  0.3813140784 0.9991452
## 75:3-65:3  0.137500000 -0.431314078  0.7063140784 0.9999821
## 65:4-65:3 -0.166967085 -0.436826278  0.1028921084 0.7503113
## 70:4-65:3  0.055250929 -0.217388356  0.3278902145 0.9999983
## 72:4-65:3  0.035000000 -0.265988119  0.3359881187 1.0000000
## 75:4-65:3 -0.112500000 -0.438779340  0.2137793405 0.9985644
## 72:3-70:3 -0.361538462 -0.908037351  0.1849604278 0.6464173
## 75:3-70:3 -0.036538462 -0.583037351  0.5099604278 1.0000000
## 65:4-70:3 -0.341005546 -0.559947466 -0.1220636260 0.0000121
## 70:4-70:3 -0.118787532 -0.341147061  0.1035719968 0.9043784
## 72:4-70:3 -0.139038462 -0.395369162  0.1172922388 0.8930755
## 75:4-70:3 -0.286538462 -0.572142713 -0.0009342103 0.0482495
## 75:3-72:3  0.325000000 -0.394499221  1.0444992211 0.9764004
## 65:4-72:3  0.020532915 -0.496142667  0.5372084979 1.0000000
## 70:4-72:3  0.242750929 -0.275382117  0.7608839753 0.9671803
## 72:4-72:3  0.222500000 -0.311094904  0.7560949035 0.9890317
## 75:4-72:3  0.075000000 -0.473258955  0.6232589548 1.0000000
## 65:4-75:3 -0.304467085 -0.821142667  0.2122084979 0.8130185
## 70:4-75:3 -0.082249071 -0.600382117  0.4358839753 0.9999999
## 72:4-75:3 -0.102500000 -0.636094904  0.4310949035 0.9999992
## 75:4-75:3 -0.250000000 -0.798258955  0.2982589548 0.9742680
## 70:4-65:4  0.222218014  0.089040007  0.3553960211 0.0000014
## 72:4-65:4  0.201967085  0.017581485  0.3863526844 0.0164473
## 75:4-65:4  0.054467085 -0.168831840  0.2777660094 0.9999799
## 72:4-70:4 -0.020250929 -0.208681938  0.1681800795 1.0000000
## 75:4-70:4 -0.167750929 -0.394401774  0.0588999156 0.4427972
## 75:4-72:4 -0.147500000 -0.407562057  0.1125620567 0.8540240
```

Nhận xét

Từ bảng anova, ta thấy:

- Ở mức ý nghĩa 5%, nếu xét từng nhân tố thì cả 2 nhân tố "CocoaPercent" và "CompanyLocation" đều có ý nghĩa thống kê do p_value tương ứng lần lượt là 3.48e-10, 0.000669 đều nhỏ hơn 0.05. Nghĩa là có sự khác biệt về đánh giá của chuyên gia ("Rating") đối với chocolate có tỉ lệ ca cao khác nhau ("CocoaPercent"), quốc giá công ty sản xuất ("CompanyLocation")

- Ở mức ý nghĩa 5%, nếu xét cả 2 nhân tố thì ta thấy có sự tương tác lẫn nhau giữa nhân tố "CocoaPercent" và "CompanyLocation" (p_value = 0.013605 < 0.05). Và có 11 cặp tương tác có ý nghĩa như bên dưới: 70:1-65:1; 70:3-65:1; 70:4-65:1; 65:4-70:1; 70:3-72:1; 70:3-75:1; 70:3-72:2; 65:4-70:3; 75:4-70:3; 70:4-65:4; 72:4-65:4

Diễn giải

- 70:4-72:1 có nghĩa là điểm đánh giá của chuyên gia (Rating) cho chocolate có phần trăm ca cao bằng 70 (CocoaPercent) sản xuất ở các nước không phải là "USA", "France", "Canada" khác với chocolate có phần trăm ca cao bằng 72 sản xuất ở "USA"
  => Ta cần phân tích hồi quy để tìm ra sự ảnh hưởng của phần trăm ca cao có trong chocolate (CocoaPercent), nước sản xuất (CompanyLocation) đến điểm đánh giá của chuyên gia (Rating)

```
summary(lm(Rating ~ CocoaPercent + CompanyLocation, data = chocolateData))
```

```
##
## Call:
## lm(formula = Rating ~ CocoaPercent + CompanyLocation, data = chocolateData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.27564 -0.31904  0.00773  0.35486  1.72436
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.06904    0.02199 139.588  < 2e-16 ***
## CocoaPercent70     0.17323    0.02542   6.816 1.28e-11 ***
## CocoaPercent72     0.09311    0.03859   2.413 0.015935 *
## CocoaPercent75     0.05696    0.03726   1.529 0.126564
## CompanyLocation2   0.12202    0.04280   2.851 0.004406 **
## CompanyLocation3   0.15288    0.04548   3.362 0.000791 ***
## CompanyLocation4   0.03338    0.02425   1.376 0.168963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4705 on 1788 degrees of freedom
## Multiple R-squared:  0.03476,    Adjusted R-squared:  0.03152
## F-statistic: 10.73 on 6 and 1788 DF,  p-value: 9.506e-12
```

Nhận xét

- Do biến "CocoaPercent75", "CompanyLocation4" có p_value > 0.05, không có ý nghĩa thống kê nên ta không đưa vào mô hình.

- Từ thống kê mô tả, ta có mô hình hồi quy tuyến tính sau:
  Rating = 3.06904 + 0.17323*(CocoaPercent70) + 0.09311*(CocoaPercent72) + 0.12202*(CompanyLocation2) + 0.15288*(CompanyLocation3)

Diễn giải

- Điểm đánh giá của chuyên gia phụ thuộc vào phần trăm ca cao có trong chocolate là 70% hay 72% và nước sản xuất là "France" hay "Canada"

- Khi chocolate có "70%" ca cao thì điểm đánh giá tăng 0.17323 đơn vị

- Khi chocolate có "72%" ca cao thì điểm đánh giá tăng 0.09311 đơn vị

- Khi nước sản xuất là "France" thì điểm đánh giá tăng 0.12202 đơn vị

- Khi nước sản xuất là "Canada" thì điểm đánh giá tăng 0.15288 đơn vị

Bộ dữ liệu: FISH MARKET ==========

MÔ TẢ DỮ LIỆU

Bộ dữ liệu thị trường cá là tập hợp dữ liệu liên quan đến nhiều loài cá và đặc điểm của chúng. Sau đây là mô tả về từng cột trong bộ dữ liệu:

- Loài: Cột này biểu thị loài cá. Đây là biến phân loại phân loại từng loài cá thành một trong bảy loài. Các loài có thể bao gồm tên như “Perch,” “Bream,” “Roach,” “Pike,” “Smelt,” “Parkki,” and “Whitefish”
- Trọng lượng: Cột này biểu thị trọng lượng của cá. Đây là biến số thường được đo bằng gam.
- Chiều dài1: Cột này biểu thị phép đo đầu tiên về chiều dài của cá. Đây là biến số, thường được đo bằng cm.
- Chiều dài2: Cột này biểu thị phép đo thứ hai về chiều dài của cá. Đây là biến số, thường được đo bằng cm.
- Chiều dài 3: Cột này biểu thị phép đo thứ ba về chiều dài của cá. Đây là biến số, thường được đo bằng cm.
- Chiều cao: Cột này biểu thị chiều cao của cá. Đây là biến số, thường được đo bằng cm.
- Chiều rộng: Cột này biểu thị chiều rộng của cá. Đây là biến số, thường được đo bằng cm.

YÊU CẦU

- Dự đoán trọng lượng của một con cá dựa trên loài của nó và các phép đo vật lý được cung cấp.

ĐỌC DỮ LIỆU

```
fishMarketOrg = docDuLieu("data","FishMarket.csv")
```

TIỀN XỬ LÝ DỮ LIỆU

Loại bỏ dữ liệu trùng

```
isTRUE(duplicated(fishMarketOrg))
```

```
## [1] FALSE
```

Nhận xét

- Không có dữ liệu trùng

Loại bỏ biến không có giá trị phân tích

```
summary(fishMarketOrg)
```

```
##    Species              Weight          Length1         Length2
##  Length:159         Min.   :   0.0   Min.   : 7.50   Min.   : 8.40
##  Class :character   1st Qu.: 120.0   1st Qu.:19.05   1st Qu.:21.00
##  Mode  :character   Median : 273.0   Median :25.20   Median :27.30
##                     Mean   : 398.3   Mean   :26.25   Mean   :28.42
##                     3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.50
##                     Max.   :1650.0   Max.   :59.00   Max.   :63.40
##     Length3          Height           Width
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048
##  1st Qu.:23.15   1st Qu.: 5.945   1st Qu.:3.386
##  Median :29.40   Median : 7.786   Median :4.248
##  Mean   :31.23   Mean   : 8.971   Mean   :4.417
##  3rd Qu.:39.65   3rd Qu.:12.366   3rd Qu.:5.585
##  Max.   :68.00   Max.   :18.957   Max.   :8.142
```

Nhận xét

- Không có biến nào cần loại ra khỏi bộ dữ liệu

Thay thế dữ liệu bất thường

```
fishMarketOrg$Weight = replace(fishMarketOrg$Weight, fishMarketOrg$Weight == 0,
                               median(fishMarketOrg$Weight))
```

Nhận xét

- Biến "Weight" có dữ liệu bất thường (Min = 0.0) => để không ảnh hưởng đến phân bố dữ liệu, ta thay đổi giá trị 0.0 bằng giá trị của Median

Chuyển đổi kiểu dữ liệu

```
str(fishMarketOrg)
```

```
## 'data.frame':    159 obs. of  7 variables:
##  $ Species: chr  "Bream" "Bream" "Bream" "Bream" ...
##  $ Weight : num  242 290 340 363 430 450 500 390 450 500 ...
##  $ Length1: num  23.2 24 23.9 26.3 26.5 26.8 26.8 27.6 27.6 28.5 ...
##  $ Length2: num  25.4 26.3 26.5 29 29 29.7 29.7 30 30 30.7 ...
##  $ Length3: num  30 31.2 31.1 33.5 34 34.7 34.5 35 35.1 36.2 ...
```

```
##  $ Height : num  11.5 12.5 12.4 12.7 12.4 ...
##  $ Width  : num  4.02 4.31 4.7 4.46 5.13 ...
```

```
fishMarketOrg$Species = as.factor(fishMarketOrg$Species)
```

Nhận xét

- Biến "Species" (loài) là biến định tính => ta chuyển sang dạng factor

Loại bỏ dữ liệu khuyết

```
anyNA(fishMarketOrg)
```

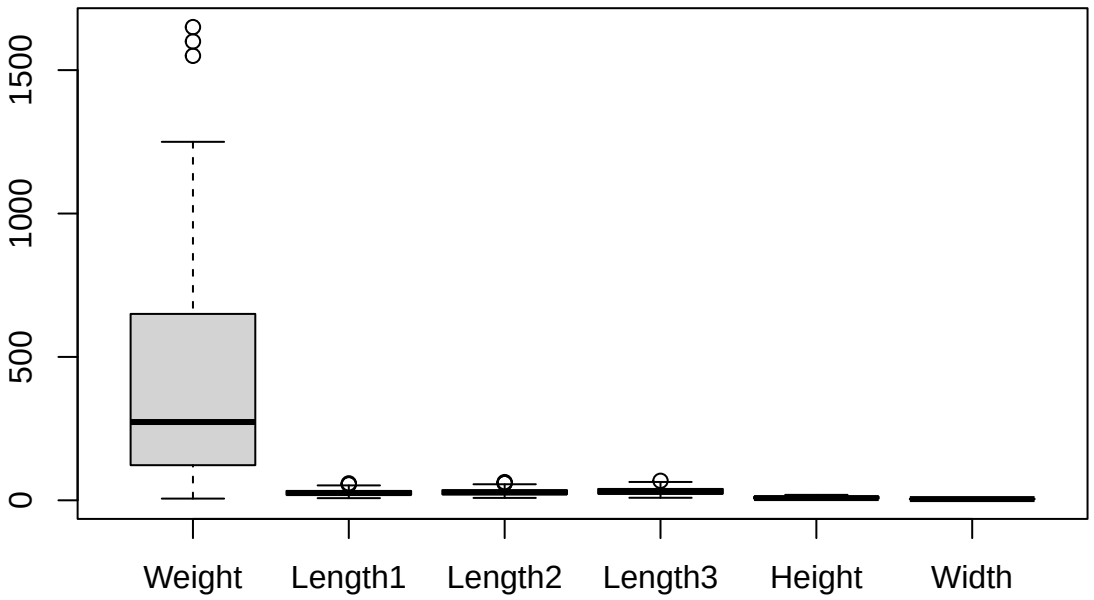
```
## [1] FALSE
```

Nhận xét

- Không có dữ liệu khuyết

Quy tâm dữ liệu

```
fishMarketQuantitativeData = fishMarketOrg[,-1]
boxplot(fishMarketQuantitativeData)
```




```
apply(fishMarketQuantitativeData, 2, mean)
```

```
##      Weight    Length1    Length2    Length3     Height      Width
## 400.043396  26.247170  28.415723  31.227044   8.970994   4.417486
```

```
apply(fishMarketQuantitativeData, 2, sd)
```

```
##      Weight    Length1    Length2    Length3     Height      Width
```

```
## 356.708166   9.996441  10.716328  11.610246   4.286208   1.685804
```

```
apply(fishMarketQuantitativeData, 2, range)
```

```
##      Weight Length1 Length2 Length3  Height  Width
## [1,]    5.9     7.5     8.4     8.8  1.7284 1.0476
## [2,] 1650.0    59.0    63.4    68.0 18.9570 8.1420
```

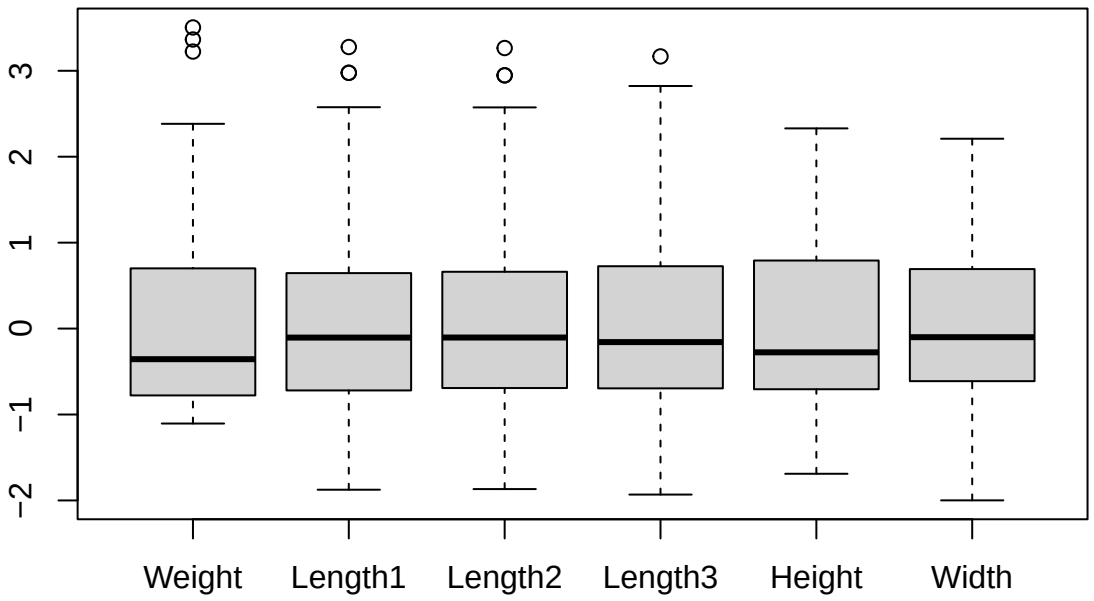
```
fishMarketStdData = as.data.frame(scale(fishMarketQuantitativeData, scale = TRUE))
```

Nhận xét

- Từ BoxPlot, ta thấy range của biến "Weight" khá lớn so với các biến khác
- Trung bình, độ lệch chuẩn của các biến không tương đồng. Do đó, ta cần quy tâm dữ liệu. Đưa dữ liệu có trung bình của biến về 0, phương sai về 1

Loại bỏ outlier

```
boxplot(fishMarketStdData)
```




```
outlierIndexList = list()
for(i in 1:length(fishMarketStdData)){
  quantitleValue = quantile(fishMarketStdData[[i]])
  upperValue = quantitleValue[4] + (quantitleValue[4]-quantitleValue[2])*1.5

  outlierIndexList = append(outlierIndexList, which(fishMarketStdData[i] > upperValue))
}

duplicatedOutlierIndex = unique(outlierIndexList)
duplicatedOutlierAmount = length(duplicatedOutlierIndex)
```

```
duplicatedOutlierPercentage = round(duplicatedOutlierAmount/dim(fishMarketStdData)[1]*100, 2)

fishMarketFinalData = data.frame(fishMarketStdData,
                                 fishMarketOrg[,1])[c(-unlist(duplicatedOutlierIndex)), ]
fishMarketFinalData = rename(fishMarketFinalData,c('Species'='fishMarketOrg...1.'))
```

Nhận xét

- Sau khi quy tâm dữ liệu, ta thấy biến "Weight", "Length1", "Length2", "Length3" có outlier
- Có tổng cộng 3 dòng mà một trong các biến trên cùng có outlier, chiếm 1.89% dữ liệu
- Các dòng cần loại khỏi bộ dữ liệu là: 143, 144, 145

CHIA DỮ LIỆU

```
set.seed(1234)
trainingSamples = fishMarketFinalData$Weight  %>%  createDataPartition(p = 0.8, list = FALSE)
trainingData  = fishMarketFinalData[trainingSamples, ]
validationData = fishMarketFinalData[-trainingSamples, ]
```

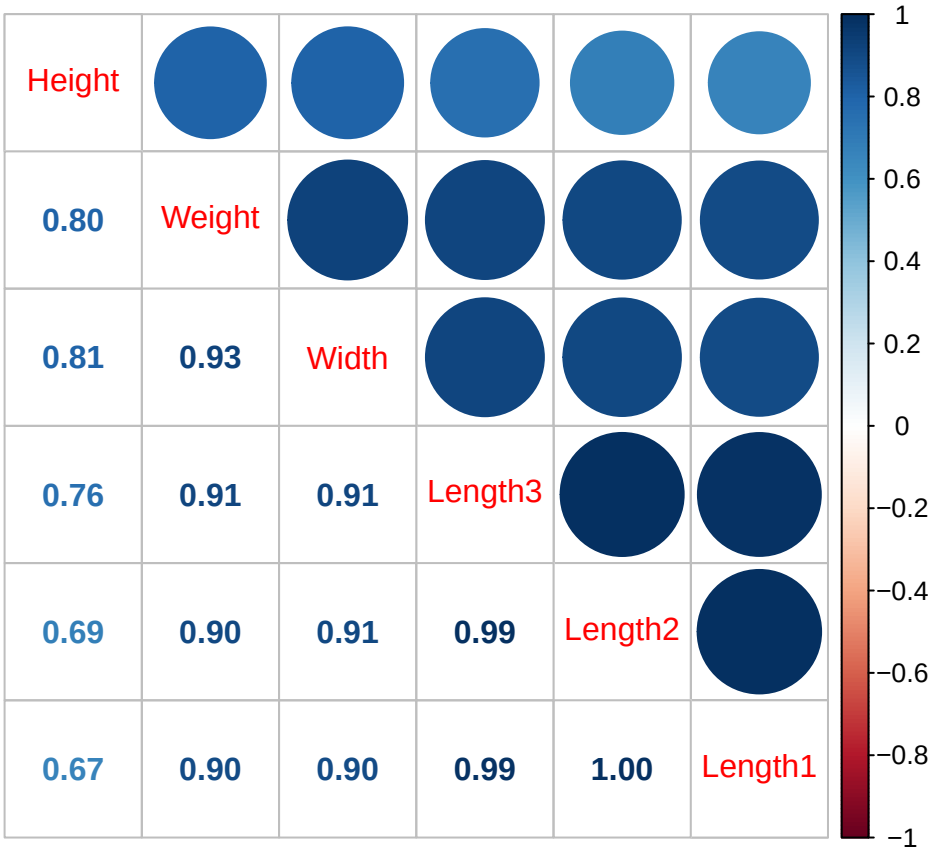
XÂY DỰNG MÔ HÌNH

Kiểm tra tương quan

```
cor(trainingData[,1:6])
```

```
##             Weight   Length1   Length2   Length3    Height     Width
## Weight  1.0000000 0.8969616 0.9013905 0.9122473 0.8035356 0.9268454
## Length1 0.8969616 1.0000000 0.9993676 0.9898115 0.6695944 0.8986681
## Length2 0.9013905 0.9993676 1.0000000 0.9925619 0.6865672 0.9054014
## Length3 0.9122473 0.9898115 0.9925619 1.0000000 0.7558713 0.9115069
## Height  0.8035356 0.6695944 0.6865672 0.7558713 1.0000000 0.8057907
## Width   0.9268454 0.8986681 0.9054014 0.9115069 0.8057907 1.0000000
```

```
corrplot.mixed(cor(trainingData[,1:6]), order = 'AOE')
```



Nhận xét

- Các biến có mối tương quan rất mạnh với nhau
  => cần kiểm tra các biến có xảy ra hiện tượng đa cộng tuyến không

Kiểm tra đa cộng tuyến

```
model = lm(Weight~., data = trainingData)
vif(model)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## Length1 2051.61594  1       45.294767
## Length2 3479.24923  1       58.985161
## Length3 1973.68879  1       44.426217
## Height    64.72285  1        8.045051
## Width     37.39627  1        6.115249
## Species 2379.85042  6        1.911521
```

```
model2 = lm(Weight~. -Length2, data = trainingData)
vif(model2)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## Length1  973.91728  1       31.207648
## Length3 1293.30483  1       35.962548
## Height    63.77365  1        7.985841
## Width     37.34333  1        6.110919
## Species 1100.17777  6        1.792484
```

```
model3 = lm(Weight~. -Length2 -Length3, data = trainingData)
vif(model3)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## Length1  41.56395  1        6.447011
## Height   55.71838  1        7.464475
## Width    34.85806  1        5.904072
## Species 262.77831  6        1.590862
```

```
model4 = lm(Weight~. -Length2 -Length3 - Species, data = trainingData)
vif(model4)
```

```
##  Length1   Height    Width
## 5.437370 2.982954 8.552831
```

```
model5 = lm(Weight~. -Length2 -Length3 - Species - Width, data = trainingData)
vif(model5)
```

```
##  Length1   Height
## 1.812765 1.812765
```

Nhận xét

- Ở mô hình 1, ta thấy các giá trị GVIF đều rất lớn (>10) nên chắc chắn xảy ra hiện tượng đa cộng tuyến. Biến "Length2" có GVIF lớn nhất nên ta loại biến "Length2" ra khỏi mô hình
- Ở mô hình 2, ta thấy các giá trị GVIF đều rất lớn (>10) nên chắc chắn xảy ra hiện tượng đa cộng tuyến. Biến "Length3" có GVIF lớn nhất nên ta tiếp tục loại biến "Length3" ra khỏi mô hình
- Ở mô hình 3, ta thấy các giá trị GVIF đều rất lớn (>10) nên chắc chắn xảy ra hiện tượng đa cộng tuyến. Biến "Spices" có GVIF lớn nhất nên ta tiếp tục loại biến "Spices" ra khỏi mô hình
- Ở mô hình 4, ta thấy giá trị GVIF của biến "Width" lớn hơn 5 và nhỏ hơn 10, có thể xảy ra hiện tượng đa cộng tuyến. Ta loại tiếp biến "Width" ra khỏi mô hình
- Ở mô hình 5, ta thấy các giá trị GVIF đều nhỏ hơn 2, ta dừng kiểm tra hiện tượng đa cộng tuyến

Xây dựng mô hình bằng phương pháp hồi quy bội

```
summary(model5)
```

```
##
## Call:
## lm(formula = Weight ~ . - Length2 - Length3 - Species - Width,
##     data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71246 -0.19663 -0.06964  0.18722  0.87965
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04099    0.02812  -1.458    0.147
## Length1      0.65477    0.04229  15.482  < 2e-16 ***
## Height       0.32270    0.03686   8.754 1.26e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3155 on 124 degrees of freedom
## Multiple R-squared:  0.8792, Adjusted R-squared:  0.8772
## F-statistic: 451.2 on 2 and 124 DF,  p-value: < 2.2e-16
```

```
finalModel = lm(Weight ~ Length1 + Height, data = trainingData)
```

Nhận xét

- Các biến "Length1", "Height" có ý nghĩa thống kê (Pr<0.05) ở mức $\alpha = 5\%$ nên ta giữ lại để xây dựng mô hình.

So sánh mô hình xây dựng với mô hình tạo bằng phương pháp Stepswise

```
comparisonModel = stepAIC(model, direction = "both",trace = FALSE)
anova(model5, comparisonModel)
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ (Length1 + Length2 + Length3 + Height + Width + Species) -
##     Length2 - Length3 - Species - Width
## Model 2: Weight ~ Length2 + Length3 + Height + Width + Species
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    124 12.3426
## 2    116  4.3711  8    7.9716 26.444 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(comparisonModel)
```

```
##
## Call:
## lm(formula = Weight ~ Length2 + Length3 + Height + Width + Species,
##     data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48987 -0.12373 -0.01156  0.09498  0.65165
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.3386     0.1915  -1.768   0.0797 .
## Length2            1.7246     0.7829   2.203   0.0296 *
## Length3           -1.3140     0.8393  -1.566   0.1202
## Height             0.5844     0.1304   4.482 1.75e-05 ***
## Width              0.2492     0.1104   2.258   0.0258 *
## SpeciesParkki      0.1515     0.1718   0.882   0.3796
## SpeciesPerch       0.2650     0.2750   0.964   0.3372
## SpeciesPike        0.2397     0.3311   0.724   0.4705
## SpeciesRoach       0.3600     0.2032   1.772   0.0790 .
## SpeciesSmelt       1.2263     0.2663   4.606 1.06e-05 ***
## SpeciesWhitefish   0.4659     0.2289   2.035   0.0441 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1941 on 116 degrees of freedom
## Multiple R-squared:  0.9572, Adjusted R-squared:  0.9535
## F-statistic: 259.5 on 10 and 116 DF,  p-value: < 2.2e-16
```

Nhận xét

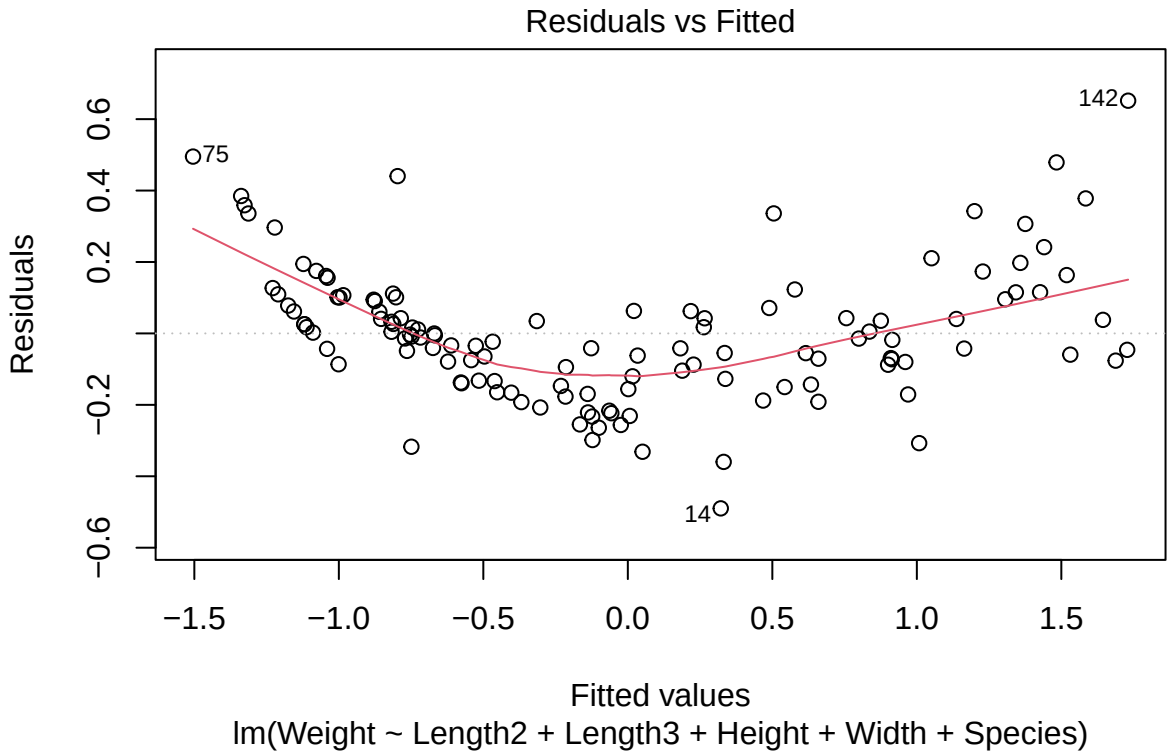
- Mô hình 2 hiệu quả hơn mô hình 1 do Pr < 2.2e-16 nên ta chọn mô hình 2
- Adjusted R-squared = 95.35% => giải thích được 95.35% sự phụ thuộc của biến "Weight" vào các biến "Length2", "Height", "Width", "SpeciesSmelt", "SpeciesWhitefish"
- Sự phụ thuộc của các biến tỉ lệ như sau:
  Weight = -0.3386 + 1.7246*(Length2) + 0.5844*(Height) + 0.2492*(Width) + 1.2263*(SpeciesSmelt) + 0.4659*(SpeciesWhitefish)

Diễn giải

- Trọng lượng cá phụ thuộc vào chiều dài 2, chiều cao, chiều rộng, loài cá là "Smelt" hay "Whitefish"
- Khi chiều dài cá tăng 1 đơn vị thì trọng lượng cá tăng 1.7246*(Length2) đơn vị
- Khi chiều cao cá tăng 1 đơn vị thì trọng lượng cá tăng 0.5844*(Height) đơn vị
- Khi chiều rộng cá tăng 1 đơn vị thì trọng lượng cá tăng 0.2492*(Width) đơn vị
- Khi loài cá là "Smelt" thì trọng lượng cá tăng 1.2263*(SpeciesSmelt) đơn vị
- Khi loài cá là "SpeciesWhitefish" thì trọng lượng cá tăng 0.4659*(SpeciesWhitefish)

Kiểm tra giả thiết tuyến tính của dữ liệu và giả thiết phần dư có trung bình bằng 0

```
plot(comparisonModel, 1)
```



Nhận xét

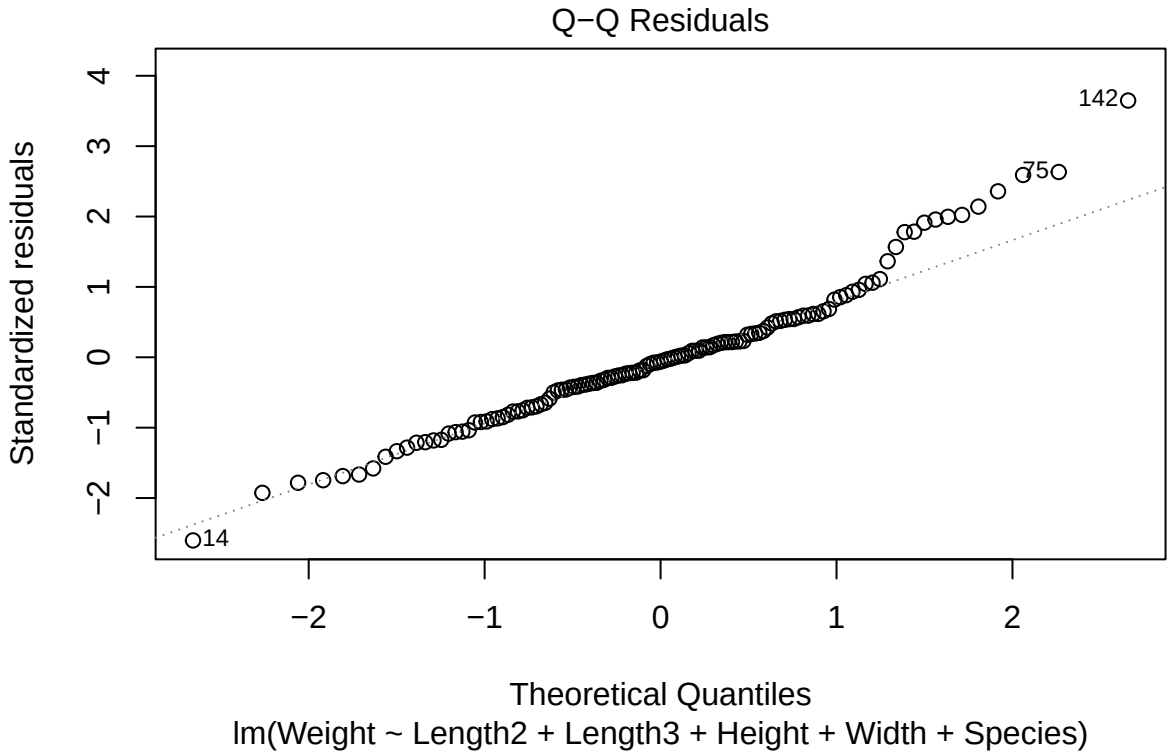
- Đồ thị cho thấy giả thiết về tính tuyến tính của dữ liệu hơi bị vi phạm. Tuy nhiên giả thiết trung bình của phần dư có thể coi là thỏa mãn

Kiểm tra phần dư có phân phối chuẩn

```
residus = residuals(comparisonModel)
shapiro.test(residus)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residus
## W = 0.96906, p-value = 0.005243
```

```
plot(comparisonModel, 2)
```

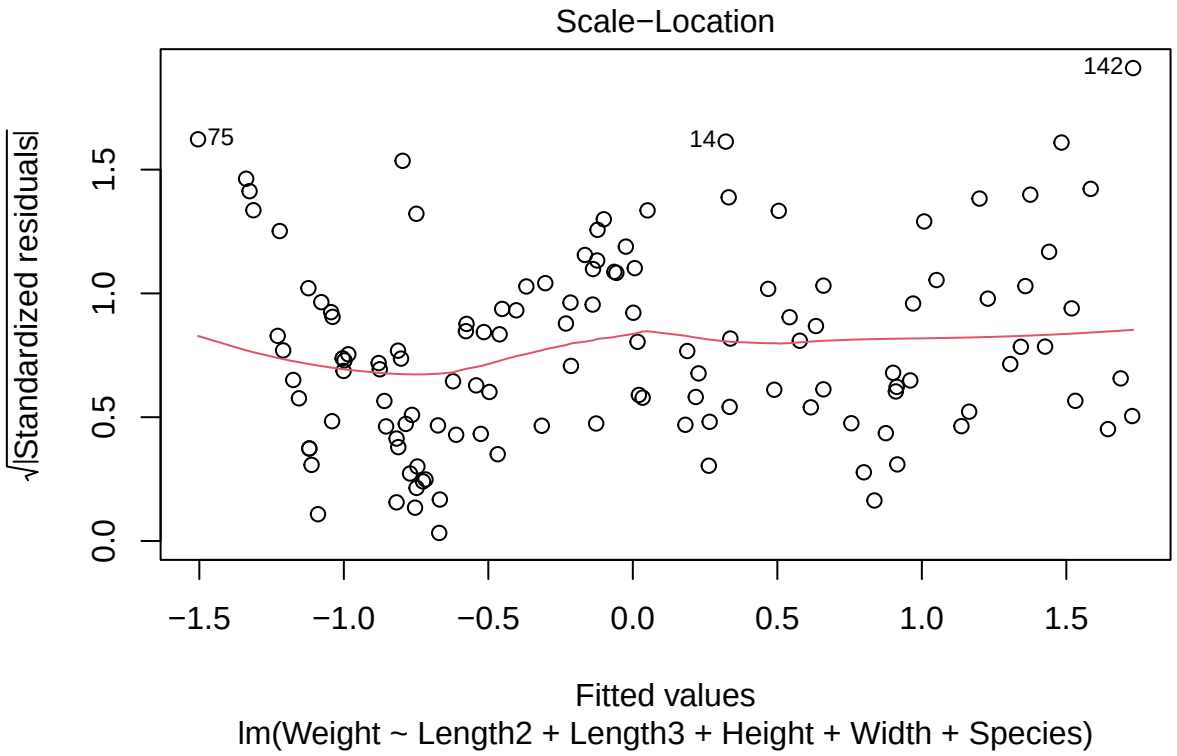


Nhận xét

- Từ đồ thị và kiểm định Shapiro Wilk (p-value = 0.005243 < 0.05) cho thấy phần dư không tuân theo theo phân phối chuẩn

Kiểm định giả thiết phương sai của phần dư không đổi

```
plot(comparisonModel, 3)
```

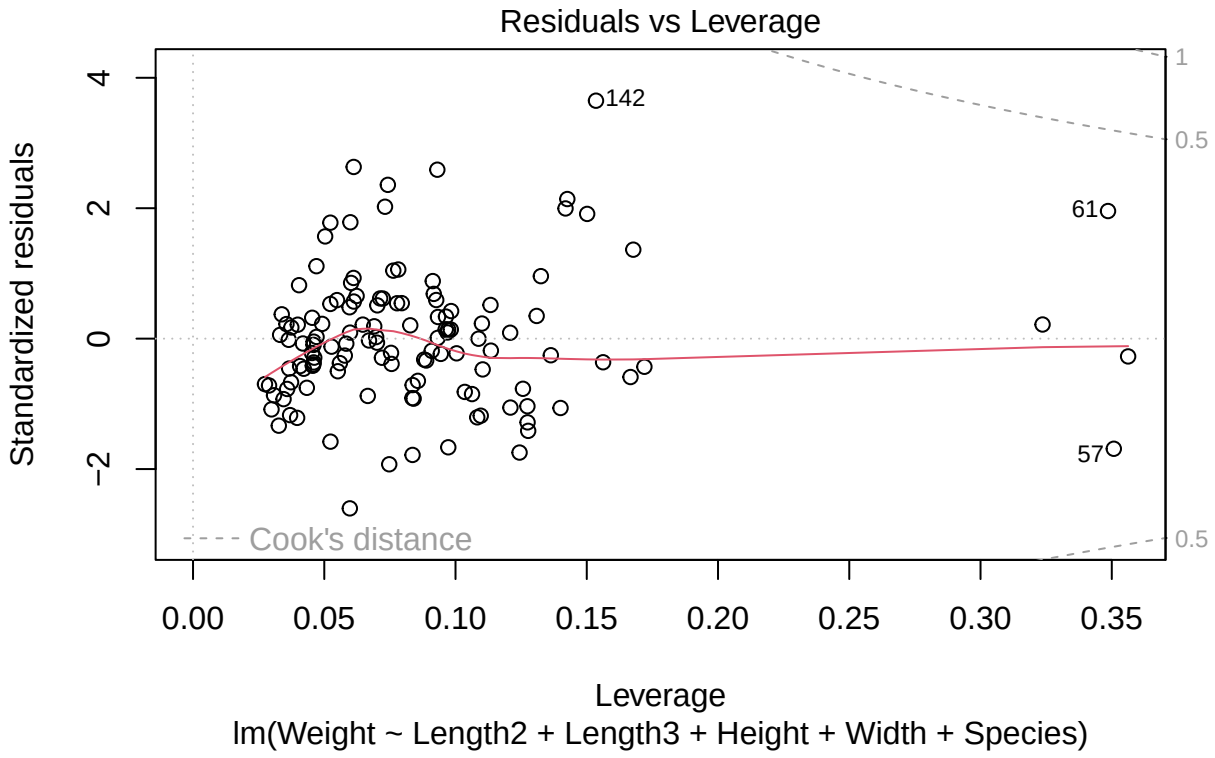


Nhận xét

- Từ đồ thị ta thấy phương sai của phần dư có thay đổi

Kiểm tra sự ảnh hưởng của dữ liệu

```
plot(comparisonModel, 5)
```

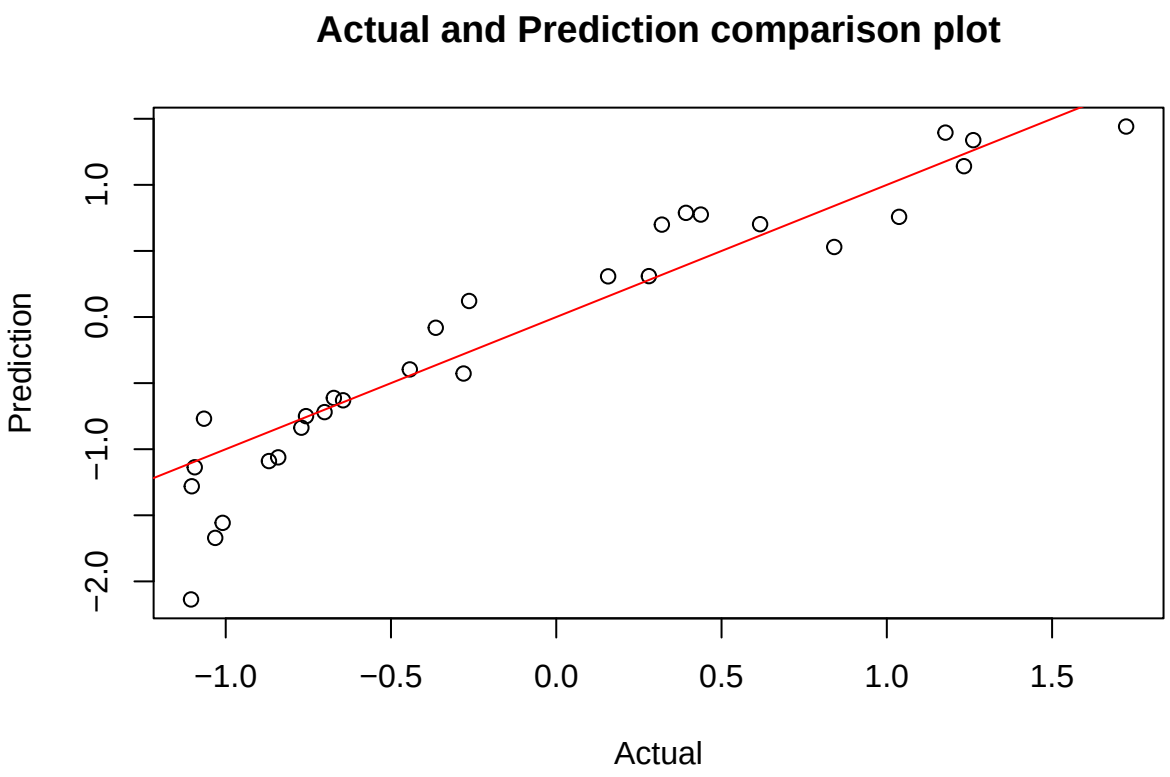


Nhận xét

- Từ đồ thị ta thấy quan sát 142, 61, 57 có thể là các quan sát có ảnh hưởng cao trong bộ dữ liệu

DỰ BÁO

```
predictions =  predict(comparisonModel, validationData)
rmse = RMSE(predictions, validationData$Weight)
r2 = R2(predictions, validationData$Weight)

plot(validationData$Weight, predictions, xlab = "Actual", ylab = "Prediction",
     main = "Actual and Prediction comparison plot")
abline(0, 1, col = "red")
```



Nhận xét

- Độ lệch trung bình giữa các giá trị dự đoán và các giá trị thực tế là RMSE = 0.3224472
- R2 = 0.9017254 cho biết 90.17% biến thiên của biến phụ thuộc có thể được giải thích bởi các biến độc lập được sử dụng trong mô hình. Từ đây cho thấy mô hình phù hợp chặt chẽ với dữ liệu.
- Qua đồ thị ta thấy, mô hình dự báo không có quá nhiều sai lệch

Bộ dữ liệu: CSM ==========

MÔ TẢ DỮ LIỆU

Bộ dữ liệu CSM (Conventional and Social Media Movies) cung cấp một số thuộc tính của phim ảnh lấy từ nguồn UCI Machine Learning Repository. Bộ dữ liệu gồm 231 quan trắc trên 14 biến:

1. "Movie": tên phim
2. "Year": năm phát hành
3. "Ratings": điểm đánh giá
4. "Genre": thể loại phim
5. "Gross": tổng doanh thu
6. "Budget": tổng chi phí
7. "Screens": số rạp chiều
8. "Sequel": phần phim
9. "Sentiment": ý kiến khán giả
10. "Views": số lượt xem
11. "Likes": số lượt thích
12. "Dislikes": số lượt chê
13. "Comments": số bình luận
14. "Aggregate Followers": số người theo dõi

YÊU CẦU

- (a) Tiền xử lý dữ liệu nếu cần
- (b) Chia bộ dữ liệu làm 2 phần: mẫu huấn luyện (training dataset) và mẫu kiểm tra (validation dataset)
- (c) Chọn mô hình tốt nhất giải thích cho biến phụ thuộc là biến doanh thu "Gross"thông qua việc chọn lựa các biến độc lập phù hợp trong các biến còn lại từ mẫu huấn luyện. Cần trình bày từng bước phương pháp chọn, tiêu chuẩn chọn mô hình, lý do chọn phương pháp đó.
  Kiểm tra các giả định (giả thiết) của mô hình (nếu giả thiết không thoả, có thể biến đổi "transformation"biến (bằng phương pháp Box-Cox, ...), hoặc có thể dùng phương pháp phi tham số, để giải quyết vấn đề này). Nêu ý nghĩa của mô hình đã chọn.
- (d) Dự báo (Prediction): sử dụng mẫu kiểm tra (validation dataset) và dựa vào mô hình tốt nhất được chọn trên đưa số liệu dự báo cho biến phụ thuộc "Gross".
- (e) So sánh kết quả dự báo với giá trị thực tế của "Gross". Rút ra nhận xét?
- (f) Có thể đề xuất những phương pháp cải tiến /phân tích khác có thể cho kết quả tốt hơn nếu có thể.

ĐỌC DỮ LIỆU

```
csmOrg = docDuLieu("data","csm.xlsx")
```

TIỀN XỬ LÝ DỮ LIỆU

Loại bỏ dữ liệu trùng

```
isTRUE(duplicated(csmOrg))
```

```
## [1] FALSE
```

Nhận xét

- Không có dữ liệu trùng

Loại bỏ biến không có giá trị phân tích

```
summary(csmOrg)
```

```
##     Movie                Year         Ratings          Genre
##  Length:231         Min.   :2014   Min.   :3.100   Min.   : 1.000
##  Class :character   1st Qu.:2014   1st Qu.:5.800   1st Qu.: 1.000
##  Mode  :character   Median :2014   Median :6.500   Median : 3.000
##                     Mean   :2014   Mean   :6.442   Mean   : 5.359
##                     3rd Qu.:2015   3rd Qu.:7.100   3rd Qu.: 8.000
##                     Max.   :2015   Max.   :8.700   Max.   :15.000
##
##      Gross               Budget             Screens         Sequel
##  Min.   :     2470   Min.   :    70000   Min.   :   2   Min.   :1.000
##  1st Qu.: 10300000   1st Qu.:  9000000   1st Qu.: 449   1st Qu.:1.000
##  Median : 37400000   Median : 28000000   Median :2777   Median :1.000
##  Mean   : 68066033   Mean   : 47921730   Mean   :2209   Mean   :1.359
##  3rd Qu.: 89350000   3rd Qu.: 65000000   3rd Qu.:3372   3rd Qu.:1.000
##  Max.   :643000000   Max.   :250000000   Max.   :4324   Max.   :7.000
##                      NA's   :1           NA's   :10
##    Sentiment           Views               Likes            Dislikes
##  Min.   :-38.00   Min.   :     698   Min.   :     1   Min.   :    0.0
##  1st Qu.:  0.00   1st Qu.:  623302   1st Qu.:  1776   1st Qu.:  105.5
##  Median :  0.00   Median : 2409338   Median :  6096   Median :  341.0
##  Mean   :  2.81   Mean   : 3712851   Mean   : 12732   Mean   :  679.1
##  3rd Qu.:  5.50   3rd Qu.: 5217380   3rd Qu.: 15248   3rd Qu.:  697.5
##  Max.   : 29.00   Max.   :32626778   Max.   :370552   Max.   :13960.0
##
##     Comments       Aggregate Followers
##  Min.   :    0.0   Min.   :    1066
##  1st Qu.:  248.5   1st Qu.:  183025
##  Median :  837.0   Median : 1052600
##  Mean   : 1825.7   Mean   : 3038193
##  3rd Qu.: 2137.0   3rd Qu.: 3694500
##  Max.   :38363.0   Max.   :31030000
##                    NA's   :35
```

```
csmData = csmOrg[,-1]
```

Nhận xét

- Biến "Movie" (tên phim) không ảnh hưởng đến quá trình phân tích nên ta loại ra khỏi bộ dữ liệu

Chuyển đổi kiểu dữ liệu

```
str(csmOrg)
```

```
## tibble [231 x 14] (S3: tbl_df/tbl/data.frame)
##  $ Movie              : chr [1:231] "13 Sins" "22 Jump Street" "3 Days to Kill" "300: Rise of an Empire
##  $ Year               : num [1:231] 2014 2014 2014 2014 2014 ...
##  $ Ratings            : num [1:231] 6.3 7.1 6.2 6.3 4.7 4.6 6.1 7.1 6.5 6.1 ...
##  $ Genre              : num [1:231] 8 1 1 1 8 3 8 1 10 8 ...
##  $ Gross              : num [1:231] 9.13e+03 1.92e+08 3.07e+07 1.06e+08 1.73e+07 2.90e+04 4.26e+07 5.75
##  $ Budget             : num [1:231] 4.00e+06 5.00e+07 2.80e+07 1.10e+08 3.50e+06 5.00e+05 4.00e+07 2.00
##  $ Screens            : num [1:231] 45 3306 2872 3470 2310 ...
##  $ Sequel             : num [1:231] 1 2 1 2 2 1 1 1 1 1 ...
##  $ Sentiment          : num [1:231] 0 2 0 0 0 0 0 2 3 0 ...
##  $ Views              : num [1:231] 3280543 583289 304861 452917 3145573 ...
##  $ Likes              : num [1:231] 4632 3465 328 2429 12163 ...
##  $ Dislikes           : num [1:231] 425 61 34 132 610 7 419 197 419 532 ...
##  $ Comments           : num [1:231] 636 186 47 590 1082 ...
##  $ Aggregate Followers: num [1:231] 1120000 12350000 483000 568000 1923800 ...
```

```
csmData$Year = as.factor(csmData$Year)
csmData$Genre = as.factor(csmData$Genre)
csmData$Sequel = as.factor(csmData$Sequel)
```

Nhận xét

- Biến "Year" (năm phát hành) là biến định tính => ta chuyển sang dạng factor
- Biến "Genre" (thể loại phim) là biến định tính => ta chuyển sang dạng factor
- Biến "Sequel" (phần phim) là biến định tính => ta chuyển sang dạng factor

Loại bỏ dữ liệu khuyết

```
anyNA(csmData)
```

```
## [1] TRUE
```

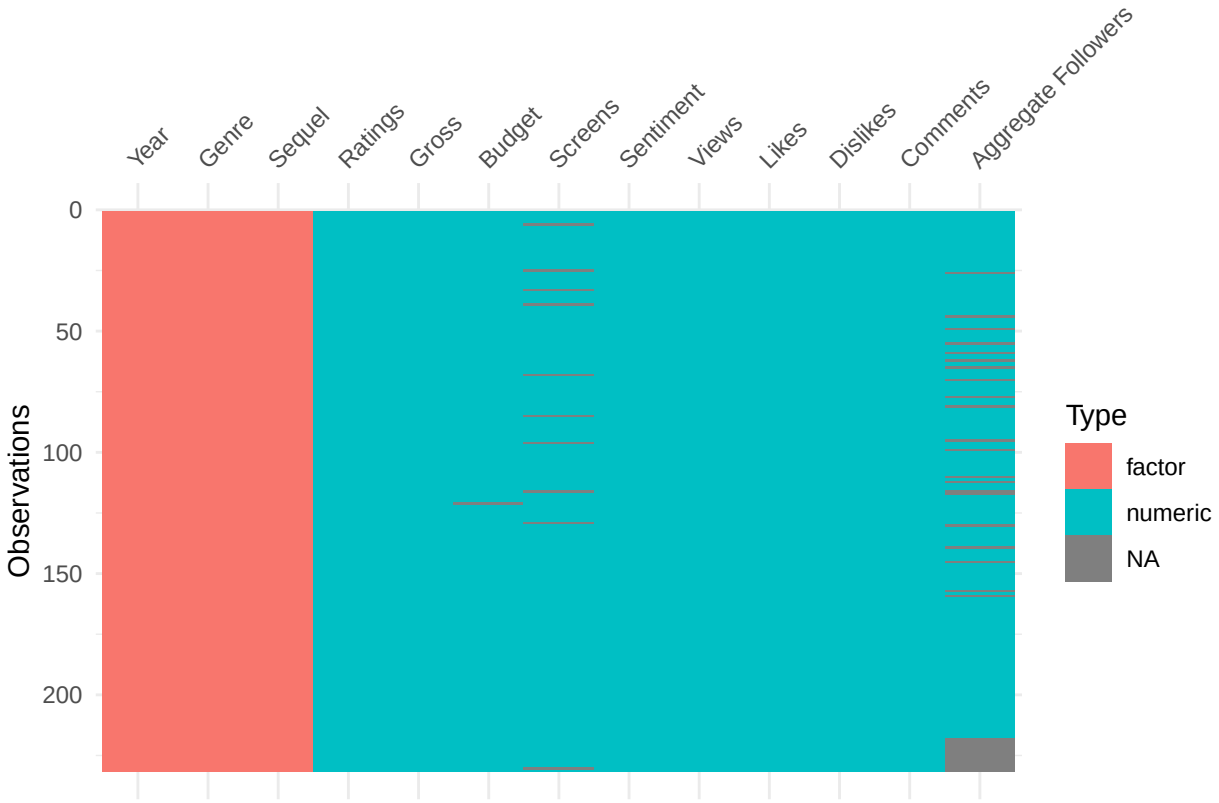
```
miss_var_table(csmData)
```

```
## # A tibble: 4 x 3
##   n_miss_in_var n_vars pct_vars
##           <int>  <int>    <dbl>
## 1             0     10    76.9
## 2             1      1     7.69
## 3            10      1     7.69
## 4            35      1     7.69
```

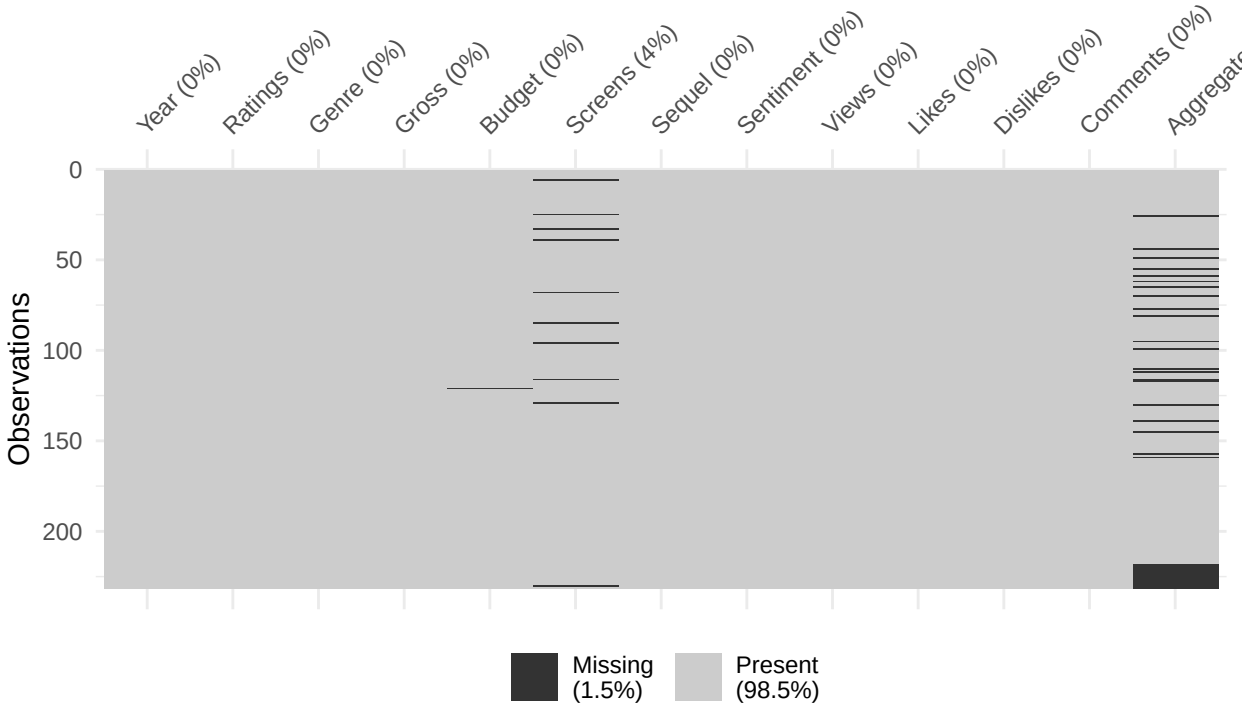
```
colSums(is.na(csmData))
```

```
##                 Year              Ratings                Genre                Gross
##                    0                    0                    0                    0
##               Budget              Screens               Sequel            Sentiment
##                    1                   10                    0                    0
##                Views                Likes             Dislikes             Comments
##                    0                    0                    0                    0
## Aggregate Followers
##                   35
```

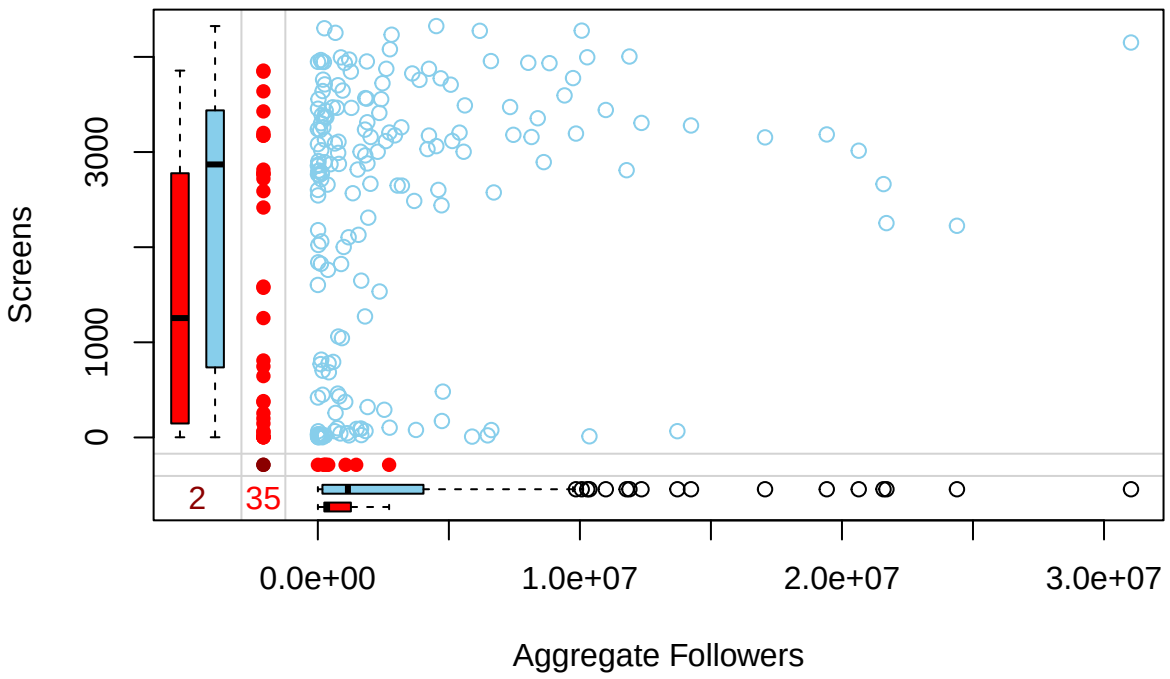
```
budgetMissing = n_miss(csmData$Budget)
screensMissing = n_miss(csmData$Screens)
aggregateFollowersMissing= n_miss(csmData$`Aggregate Followers`)
vis_dat(csmData)
```



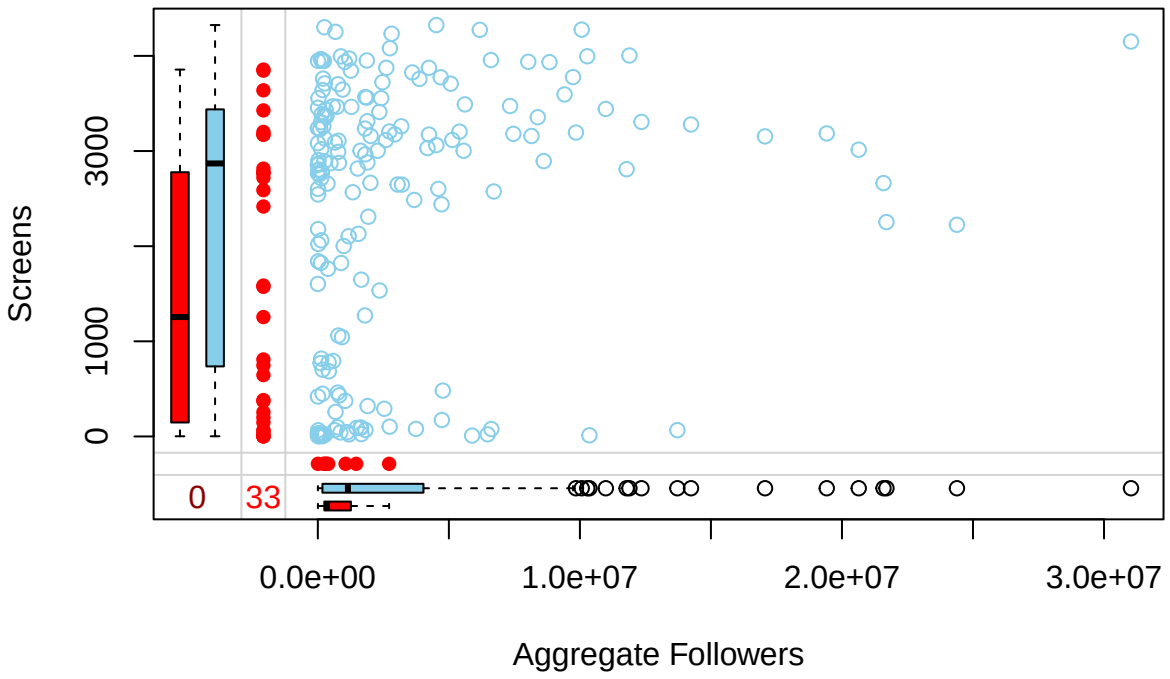

```
vis_miss(csmData)
```




```
missingObservations = sum(!complete.cases(csmData))

missCases = miss_case_table(csmData)
missingPercentage = round(pct_miss_case(csmData), 2)

csmData = subset(csmData, Screens != "NA" | `Aggregate Followers` != "NA")

marginplot(csmOrg[,c("Aggregate Followers", "Screens")])
```




```
marginplot(csmData[,c("Aggregate Followers", "Screens")])
```



Nhận xét

- Biến "Budget" có 1 dữ liệu khuyết (NA)
- Biến "Screens" có 10 dữ liệu khuyết (NA)
- Biến "Aggregate Followers" có 35‘ dữ liệu khuyết (NA)
- Bộ dữ liệu có 44 quan sát bị khuyết, chiếm 19.05% dữ liệu. Trong đó, có 2 dòng dữ liệu bị khuyết 2 biến, 42 dòng dữ liệu bị khuyết 1 biến (biến bị khuyết dữ liệu có cùng quan sát sẽ tính là 1 quan sát bị khuyết, ví dụ: dòng dữ liệu của phim "The Devil's Hand", có dữ liệu khuyết ở 2 biến "Screens", "Aggregate Followers" ta tính 1 quan sát bị khuyết) => loại bỏ 2 dòng cùng có 2 biến bị khuyết dữ liệu
- Từ 2 biểu đồ Margin trước và sau khi loại bỏ 2 dòng dữ liệu khuyết, ta thấy, sau khi loại bỏ 2 dòng dữ liệu khuyết không ảnh hưởng đến sự phân bố dữ liệu

Thay thế dữ liệu khuyết

```
summary(csmData$Budget)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##     70000   9375000  28000000  48298236  65000000 250000000         1
```

```
summary(csmData$Screens)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       2     449    2777    2209    3372    4324       8
```

```
summary(csmData$`Aggregate Followers`)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##     1066   183025  1052600  3038193  3694500 31030000       33
```

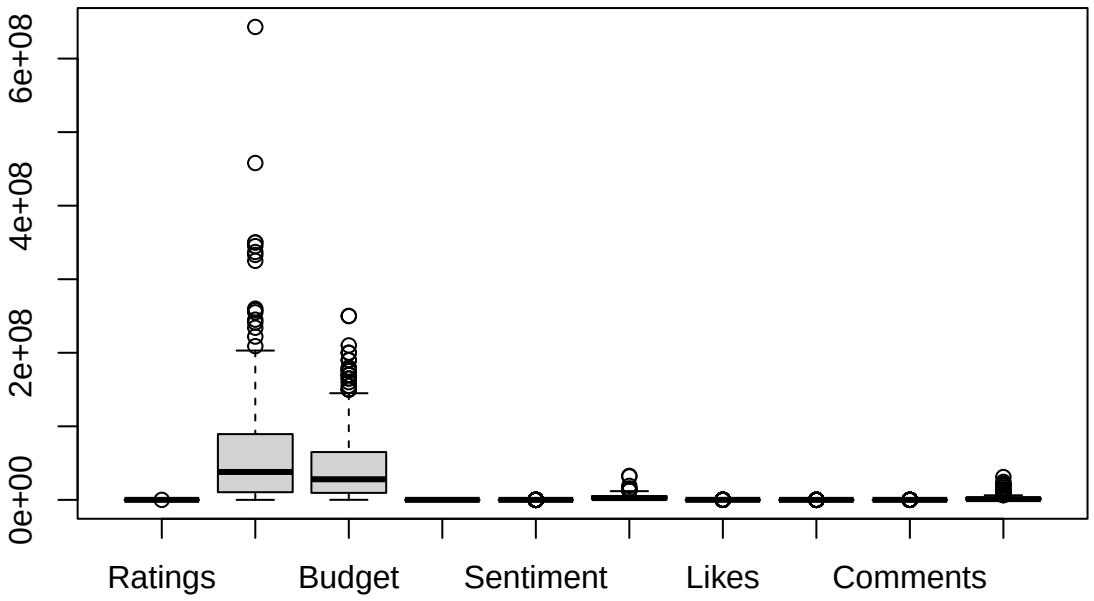
```
csmData = mutate(csmData, Budget = case_when(is.na(Budget) ~ median(Budget, na.rm = TRUE),
                                             TRUE ~ Budget))
csmData = mutate(csmData, Screens = case_when(is.na(Screens) ~ median(Screens, na.rm = TRUE),
                                              TRUE ~ Screens))
csmData = mutate(csmData, `Aggregate Followers` =
                   case_when(is.na(`Aggregate Followers`) ~
                               median(`Aggregate Followers`, na.rm = TRUE),
                             TRUE ~ `Aggregate Followers`))
```

Nhận xét

- Biến "Budget" có biên độ dao động từ 70.000 đến 250.000.000 => để không ảnh hưởng đến phân bố của dữ liệu, ta sử dụng giá trị của Median thay thế các giá trị khuyết
- Biến "Screens" có biên độ dao động từ 2 đến 4.324 => để không ảnh hưởng đến phân bố của dữ liệu, ta sử dụng giá trị của Median thay thế các giá trị khuyết
- Biến "Aggregate Followers" có biên độ dao động từ 1.066 đến 31.030.000 => để không ảnh hưởng đến phân bố của dữ liệu, ta sử dụng giá trị của Median thay thế các giá trị khuyết

Quy tâm dữ liệu

```
csmQuantitativeData = csmData[,c(-1, -3, -7)]
boxplot(csmQuantitativeData)
```




```
apply(csmQuantitativeData, 2, mean)
```

```
##             Ratings               Gross              Budget             Screens
##        6.453275e+00        6.860677e+07        4.820960e+07        2.229079e+03
##           Sentiment               Views               Likes            Dislikes
##        2.790393e+00        3.743543e+06        1.284022e+04        6.844629e+02
##            Comments Aggregate Followers
##        1.840367e+03        2.752060e+06
```

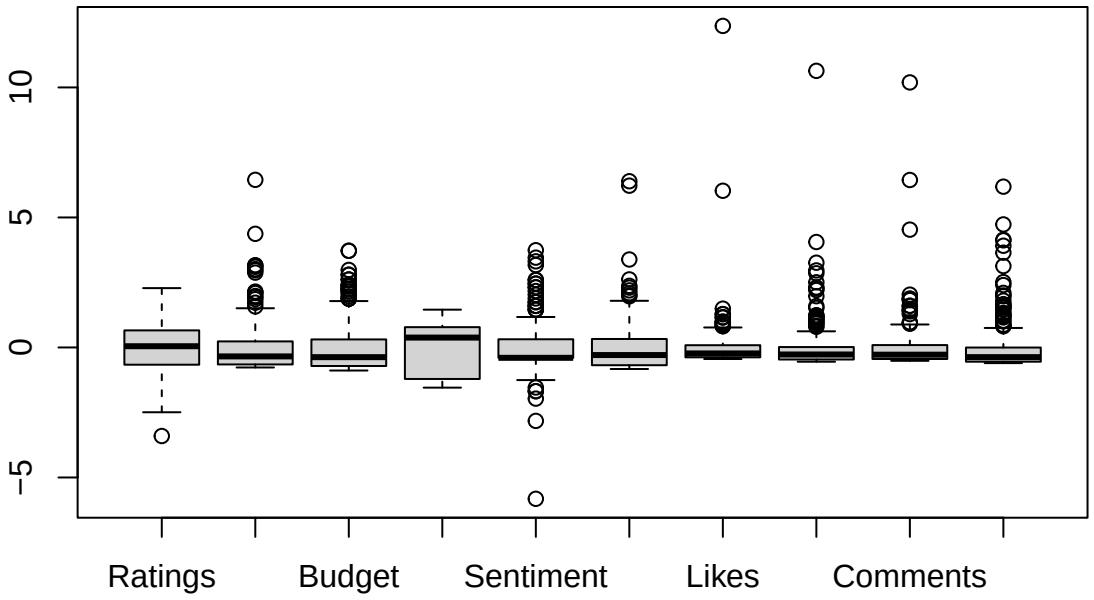
```
apply(csmQuantitativeData, 2, sd)
```

```
##             Ratings               Gross              Budget             Screens
##        9.846409e-01        8.909997e+07        5.427340e+07        1.441649e+03
##           Sentiment               Views               Likes            Dislikes
##        7.008747e+00        4.518754e+06        2.892840e+04        1.248013e+03
##            Comments Aggregate Followers
##        3.583185e+03        4.572573e+06
```

```
apply(csmQuantitativeData, 2, range)
```

```
##      Ratings    Gross  Budget Screens Sentiment    Views  Likes Dislikes
## [1,]     3.1 2.47e+03 7.0e+04       2       -38      698      1        0
## [2,]     8.7 6.43e+08 2.5e+08    4324        29 32626778 370552    13960
##      Comments Aggregate Followers
## [1,]        0                1066
## [2,]    38363            31030000
```

```r
csmStdData = as.data.frame(scale(csmQuantitativeData, scale = TRUE))
boxplot(csmStdData)
```

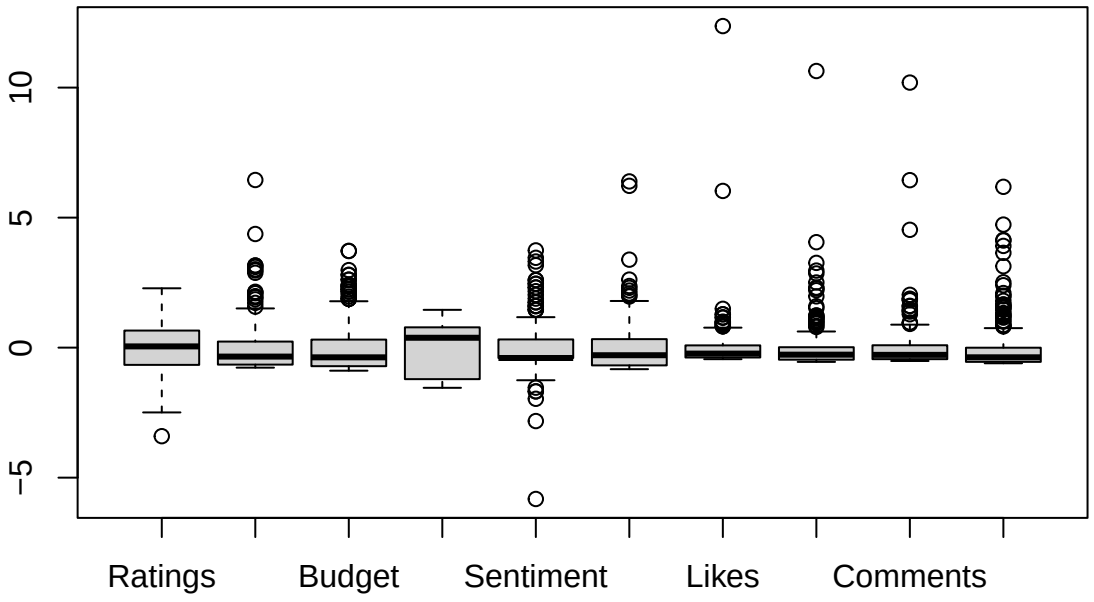


Nhận xét

- Từ BoxPlot, ta thấy range của biến "Gross", "Bugdet" khá lớn so với các biến khác
- Trung bình, độ lệch chuẩn của các biến không tương đồng. Do đó, ta cần quy tâm dữ liệu. Đưa dữ liệu có trung bình của biến về 0, phương sai về 1

Loại bỏ outlier

```
boxplot(csmStdData)
```




```
duplicatedOutlierIndex = list()

for(i in 1:length(csmStdData)){
  if(is.integer(which(csmStdData[i] < -3))){
    duplicatedOutlierIndex = append(duplicatedOutlierIndex, which(csmStdData[i] < -3))
  }

  if(is.integer(which(csmStdData[i] > 5))){
    duplicatedOutlierIndex = append(duplicatedOutlierIndex, which(csmStdData[i] > 5))
  }
}

duplicatedOutlierIndex = unique(duplicatedOutlierIndex)
duplicatedOutlierAmount = length(duplicatedOutlierIndex)
duplicatedOutlierPercentage = round(duplicatedOutlierAmount/dim(csmStdData)[1]*100, 2)

csmFinalData = data.frame(csmStdData, csmData[,c(1, 3, 7)])[c(-unlist(duplicatedOutlierIndex)), ]
```

Nhận xét

- Sau khi quy tâm dữ liệu, ta thấy biến "Ratings", "Gross", "Sentiment", "Views", "Likes", "Dislikes", "Comment" có outlier rõ ràng so với các biến còn lại
- Có tổng cộng 7 dòng mà một trong các biến trên cùng có outlier hoặc outlier rõ ràng, chiếm 3.06% dữ liệu
- Các dòng cần loại khỏi bộ dữ liệu là: 66, 163, 202, 119, 171, 85, 127

CHIA DỮ LIỆU

```
set.seed(123)
trainingSamples = csmFinalData$Gross  %>%  createDataPartition(p = 0.8, list = FALSE)
trainingData  = csmFinalData[trainingSamples, ]
validationData = csmFinalData[-trainingSamples, ]
```

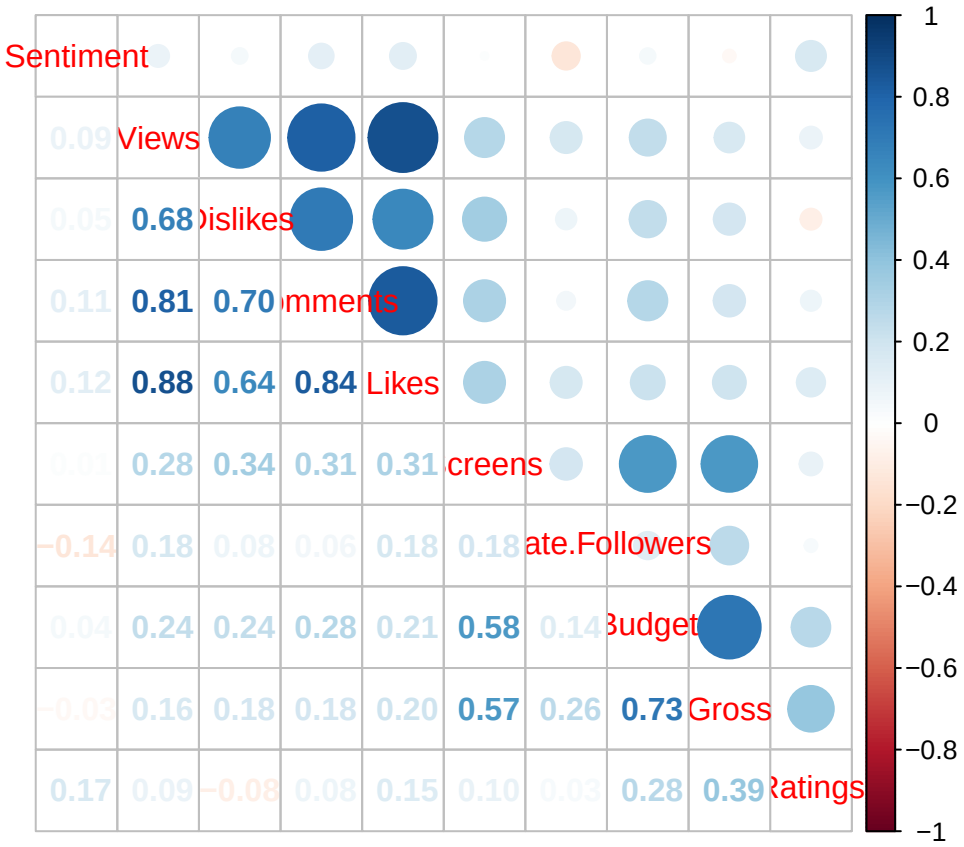
XÂY DỰNG MÔ HÌNH

Kiểm tra tương quan

```
cor(trainingData[,1:10])
```

```
##                          Ratings       Gross     Budget    Screens   Sentiment
## Ratings               1.00000000  0.38605790 0.27744682 0.09674048  0.16691755
## Gross                 0.38605790  1.00000000 0.72659539 0.57217513 -0.03009594
## Budget                0.27744682  0.72659539 1.00000000 0.57693268  0.04452301
## Screens               0.09674048  0.57217513 0.57693268 1.00000000  0.01069364
## Sentiment             0.16691755 -0.03009594 0.04452301 0.01069364  1.00000000
## Views                 0.08836633  0.16321195 0.24044872 0.28190950  0.08952311
## Likes                 0.14897781  0.20109574 0.21313514 0.31153845  0.12471526
## Dislikes             -0.08182375  0.18094831 0.24432774 0.34478704  0.04581954
## Comments              0.07547420  0.18354866 0.28401036 0.31243979  0.11231060
## Aggregate.Followers   0.03125705  0.26339038 0.13893902 0.18414924 -0.13753742
##                            Views     Likes    Dislikes   Comments
## Ratings               0.08836633 0.1489778 -0.08182375 0.07547420
## Gross                 0.16321195 0.2010957  0.18094831 0.18354866
## Budget                0.24044872 0.2131351  0.24432774 0.28401036
## Screens               0.28190950 0.3115384  0.34478704 0.31243979
## Sentiment             0.08952311 0.1247153  0.04581954 0.11231060
## Views                 1.00000000 0.8788005  0.67630766 0.81352722
## Likes                 0.87880048 1.0000000  0.64365701 0.83692497
## Dislikes              0.67630766 0.6436570  1.00000000 0.70082401
## Comments              0.81352722 0.8369250  0.70082401 1.00000000
## Aggregate.Followers   0.17790491 0.1783993  0.07755907 0.05995278
##                      Aggregate.Followers
## Ratings                       0.03125705
## Gross                         0.26339038
## Budget                        0.13893902
## Screens                       0.18414924
## Sentiment                    -0.13753742
## Views                         0.17790491
## Likes                         0.17839927
## Dislikes                      0.07755907
## Comments                      0.05995278
## Aggregate.Followers           1.00000000
```

```
corrplot.mixed(cor(trainingData[,1:10]), order = 'AOE')
```



Nhận xét

- Biến "Views" tương quan mạnh với biến "Dislikes", "Comments", "Likes"
- Biến "Dislike" tương quan mạnh với biến "Comments", "Likes"
- Biến "Comments" tương quan mạnh với biến "Likes"
- Biến "Budget" tương quan mạnh với biến "Gross"
  => cần kiểm tra các biến có xảy ra hiện tượng đa cộng tuyến không

Kiểm tra đa cộng tuyến

```
model = lm(Gross~., data = trainingData)
vif(model)
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## Ratings              1.549811  1        1.244914
## Budget               2.609415  1        1.615368
## Screens              2.093452  1        1.446877
## Sentiment            1.358853  1        1.165699
## Views                5.600779  1        2.366597
## Likes                6.870104  1        2.621088
## Dislikes             2.981482  1        1.726697
## Comments             5.796923  1        2.407680
## Aggregate.Followers  1.275588  1        1.129419
## Year                 1.757571  1        1.325734
## Genre                4.562937 10        1.078853
## Sequel               2.373171  5        1.090267
```

```r
model2 = lm(Gross~. -Likes, data = trainingData)
vif(model2)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## Ratings             1.492889  1        1.221838
## Budget              2.544447  1        1.595132
## Screens             2.076558  1        1.441027
## Sentiment           1.358520  1        1.165556
## Views               3.741644  1        1.934333
## Dislikes            2.951937  1        1.718120
## Comments            4.430130  1        2.104787
## Aggregate.Followers 1.252181  1        1.119009
## Year                1.629706  1        1.276599
## Genre               4.435682 10        1.077328
## Sequel              2.311124  5        1.087382
```

Nhận xét

- Ở mô hình 1, có hiện tượng đa cộng tuyến xảy ra, mạnh nhất ở biến "Likes" (GVIF = 6.870104) nên ta loại biến "Likes" ra khỏi mô hình
- Ở mô hình 2, các giá trị GVIF đều nhỏ hơn 5, có thể xảy ra hiện tượng đa cộng tuyến nhưng không nghiêm trọng lắm nên ta dừng kiểm tra hiện tượng đa cộng tuyến

Xây dựng mô hình bằng phương pháp hồi quy bội

```r
summary(model2)
```

```
##
## Call:
## lm(formula = Gross ~ . - Likes, data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20827 -0.28650 -0.05575  0.23231  2.75730
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.009773   0.096625  -0.101 0.919567
## Ratings              0.278511   0.053990   5.159 7.61e-07 ***
## Budget               0.422996   0.068643   6.162 6.07e-09 ***
## Screens              0.209224   0.059414   3.521 0.000566 ***
## Sentiment           -0.050302   0.048309  -1.041 0.299408
## Views               -0.088560   0.099982  -0.886 0.377137
## Dislikes             0.083479   0.113374   0.736 0.462668
## Comments            -0.051918   0.176866  -0.294 0.769504
## Aggregate.Followers  0.109523   0.048687   2.250 0.025907 *
## Year2015             0.033387   0.113452   0.294 0.768938
## Genre2              -0.104084   0.266550  -0.390 0.696719
## Genre3              -0.147581   0.140873  -1.048 0.296466
## Genre4              -0.413261   0.571864  -0.723 0.470995
## Genre6              -0.665914   0.399355  -1.667 0.097466 .
## Genre7               0.395474   0.570614   0.693 0.489319
## Genre8              -0.115700   0.139093  -0.832 0.406809
## Genre9              -0.072463   0.197251  -0.367 0.713853
## Genre10             -0.277118   0.199203  -1.391 0.166204
## Genre12              0.050064   0.218747   0.229 0.819277
```

```
## Genre15              0.220302   0.254649   0.865 0.388327
## Sequel2              0.282526   0.143007   1.976 0.049998 *
## Sequel3              0.347839   0.275153   1.264 0.208092
## Sequel4             -0.615361   0.576247  -1.068 0.287258
## Sequel5             -0.372742   0.300427  -1.241 0.216614
## Sequel7              1.090126   0.614901   1.773 0.078244 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5451 on 153 degrees of freedom
## Multiple R-squared:  0.6806, Adjusted R-squared:  0.6305
## F-statistic: 13.58 on 24 and 153 DF,  p-value: < 2.2e-16
```

```
model3 = lm(Gross~Ratings + Budget + Screens + Aggregate.Followers + Sequel,
            data = trainingData)
```

Nhận xét

- Các biến "Ratings", "Budget", "Screens", "Aggregate.Followers", "Sequel2" có ý nghĩa thống kê (Pr< 0.05) ở mức $\alpha = 5\%$ nên ta giữ lại để xây dựng mô hình.
- Các biến còn lại không có ý nghĩa thống kê nên ta loại khỏi mô hình.

So sánh mô hình xây dựng với mô hình tạo bằng phương pháp Stepswise

```
comparisonModel = stepAIC(model, direction = "both",trace = FALSE)
anova(model3, comparisonModel)
```

```
## Analysis of Variance Table
##
## Model 1: Gross ~ Ratings + Budget + Screens + Aggregate.Followers + Sequel
## Model 2: Gross ~ Ratings + Budget + Screens + Aggregate.Followers + Sequel
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    168 48.883
## 2    168 48.883  0         0
```

```
summary(comparisonModel)
```

```
##
## Call:
## lm(formula = Gross ~ Ratings + Budget + Screens + Aggregate.Followers +
##     Sequel, data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24625 -0.29737 -0.05147  0.18438  2.81284
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.07895    0.04583  -1.723   0.0868 .
## Ratings              0.22645    0.04687   4.832 3.03e-06 ***
## Budget               0.44619    0.05768   7.736 9.03e-13 ***
## Screens              0.21711    0.05348   4.060 7.52e-05 ***
## Aggregate.Followers  0.09588    0.04544   2.110   0.0363 *
## Sequel2              0.31512    0.13618   2.314   0.0219 *
## Sequel3              0.29428    0.26162   1.125   0.2623
## Sequel4             -0.41566    0.55304  -0.752   0.4534
## Sequel5             -0.37721    0.27915  -1.351   0.1784
```

```
## Sequel7              1.39480    0.56394   2.473   0.0144 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5394 on 168 degrees of freedom
## Multiple R-squared:  0.6565, Adjusted R-squared:  0.6381
## F-statistic: 35.68 on 9 and 168 DF,  p-value: < 2.2e-16
```

Nhận xét

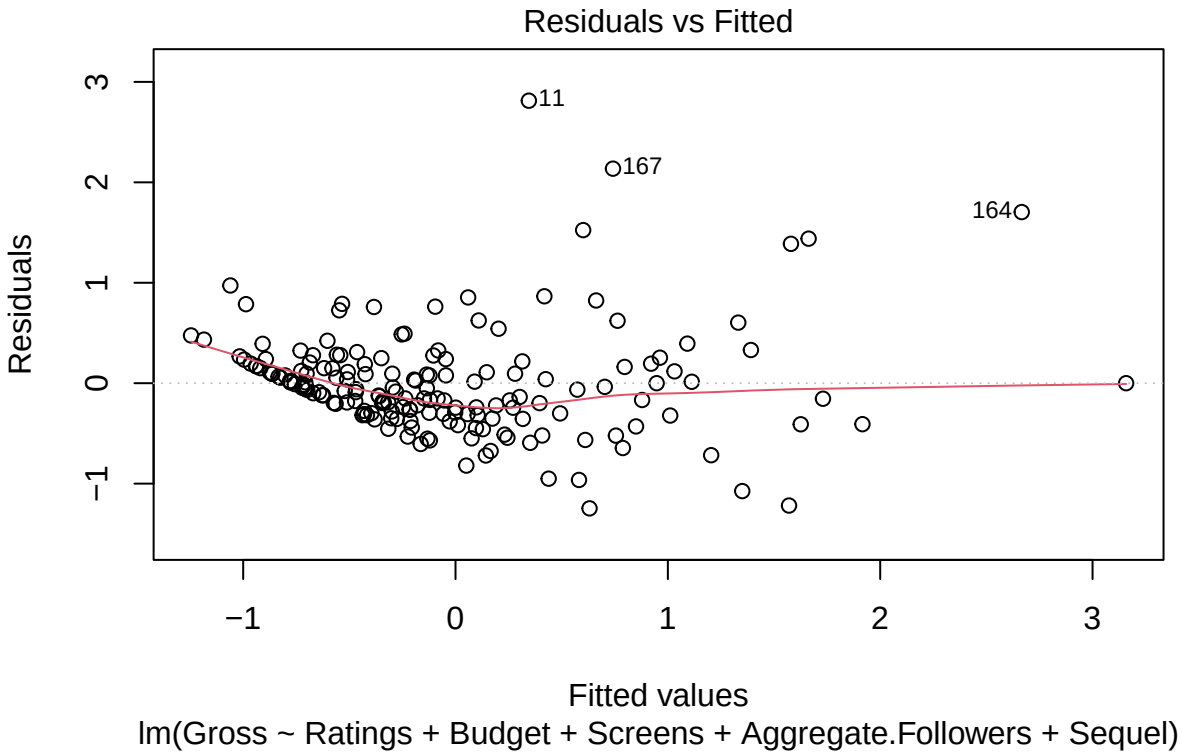
- Mô hình xây dựng từ hai cách giống nhau
- Adjusted R-squared = 63.81% => giải thích được 63.81% sự phụ thuộc của biến "Gross" vào các biến "Ratings", "Budget", "Screens", "Aggregate.Followers", "Sequel2"
- Sự phụ thuộc của các biến tỉ lệ như sau:
  Gross = -0.07895 + 0.22645*(Ratings) + 0.44619*(Budget) + 0.21711*(Screens) + 0.09588*(Aggregate.Followers) + 0.31512*(Sequel2) + 1.39480*(Sequel7)

Diễn giải

- Tổng doanh thu (Gross) phụ thuộc vào Điểm đánh giá, Tổng chi phí, Số lượng rạp chiếu, Số người theo dõi, Phần phim sẽ chiếu
- Khi Điểm đánh giá tăng 1 đơn vị thì tổng doanh thu tăng 0.22645*(Ratings) đơn vị
- Khi Tổng chi phí tăng 1 đơn vị thì tổng doanh thu tăng 0.44619*(Budget) đơn vị
- Khi số lượng rạp chiếu tăng 1 đơn vị thì tổng doanh thu tăng 0.21711*(Screens) đơn vị
- Khi số lượng người theo dõi tăng 1 người thì tổng doanh thu tăng 0.09588*(Aggregate.Followers) đơn vị
- Khi phần phim sẽ chiếu là 2 hoặc 7 thì tổng doanh thu tăng tương ứng là 0.31512*(Sequel2), 1.39480*(Sequel7) đơn vị

Kiểm tra giả thiết tuyến tính của dữ liệu và giả thiết phần dư có trung bình bằng 0

```
plot(model3, 1)
```



Nhận xét

- Đồ thị cho thấy giả thiết về tính tuyến tính của dữ liệu hơi bị vi phạm. Tuy nhiên giả thiết trung bình của phần dư có thể coi là thỏa mãn

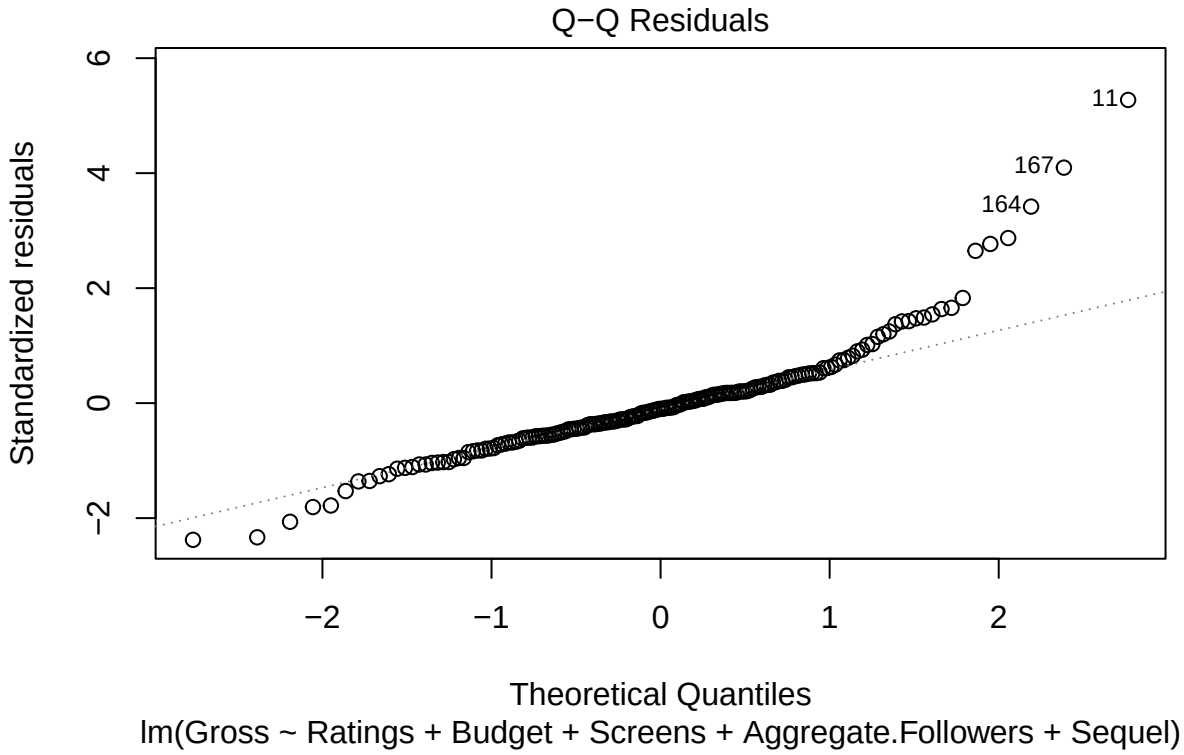
Kiểm tra phần dư có phân phối chuẩn

```
residus = residuals(model3)
shapiro.test(residus)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residus
## W = 0.88711, p-value = 2.371e-10
```

```
plot(model3, 2)
```

```
## Warning: not plotting observations with leverage one:
##   126, 133
```



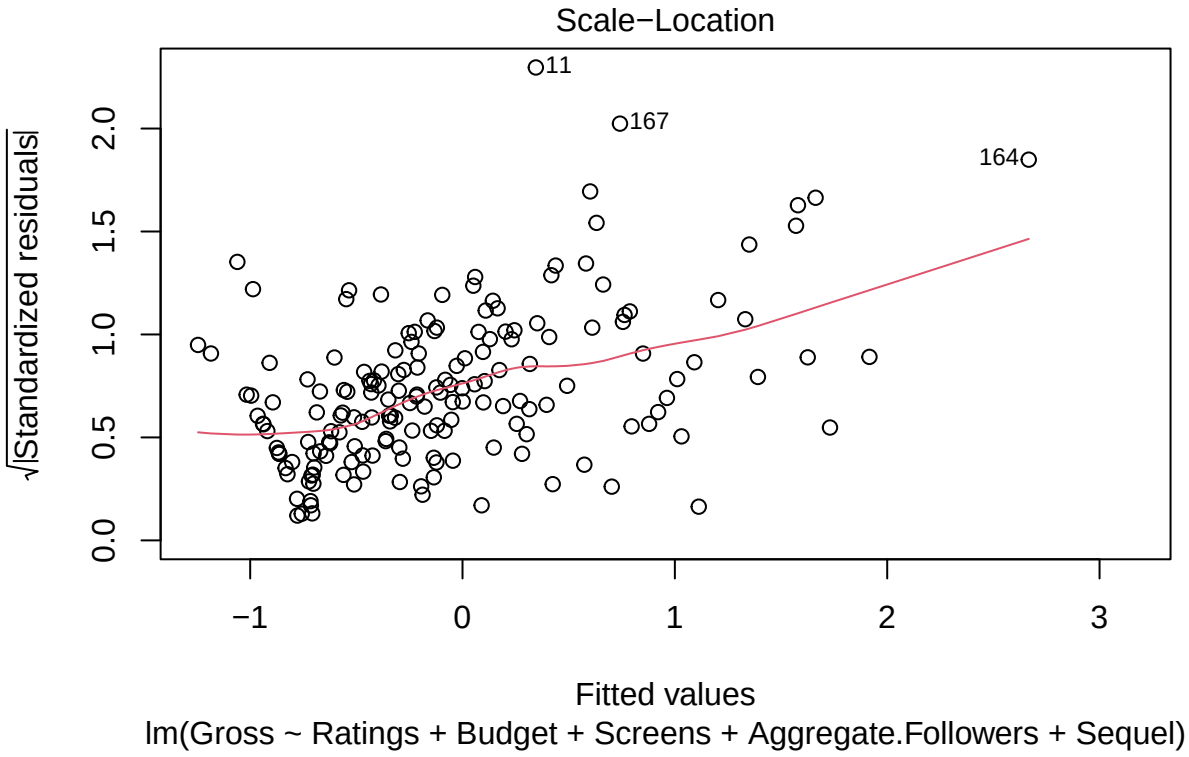
Nhận xét

- Từ đồ thị và kiểm định Shapiro Wilk (p-value = 2.371e-10 < 0.05) cho thấy phần dư không tuân theo theo phân phối chuẩn

Kiểm định giả thiết phương sai của phần dư không đổi

```
plot(model3, 3)
```

```
## Warning: not plotting observations with leverage one:
##   126, 133
```



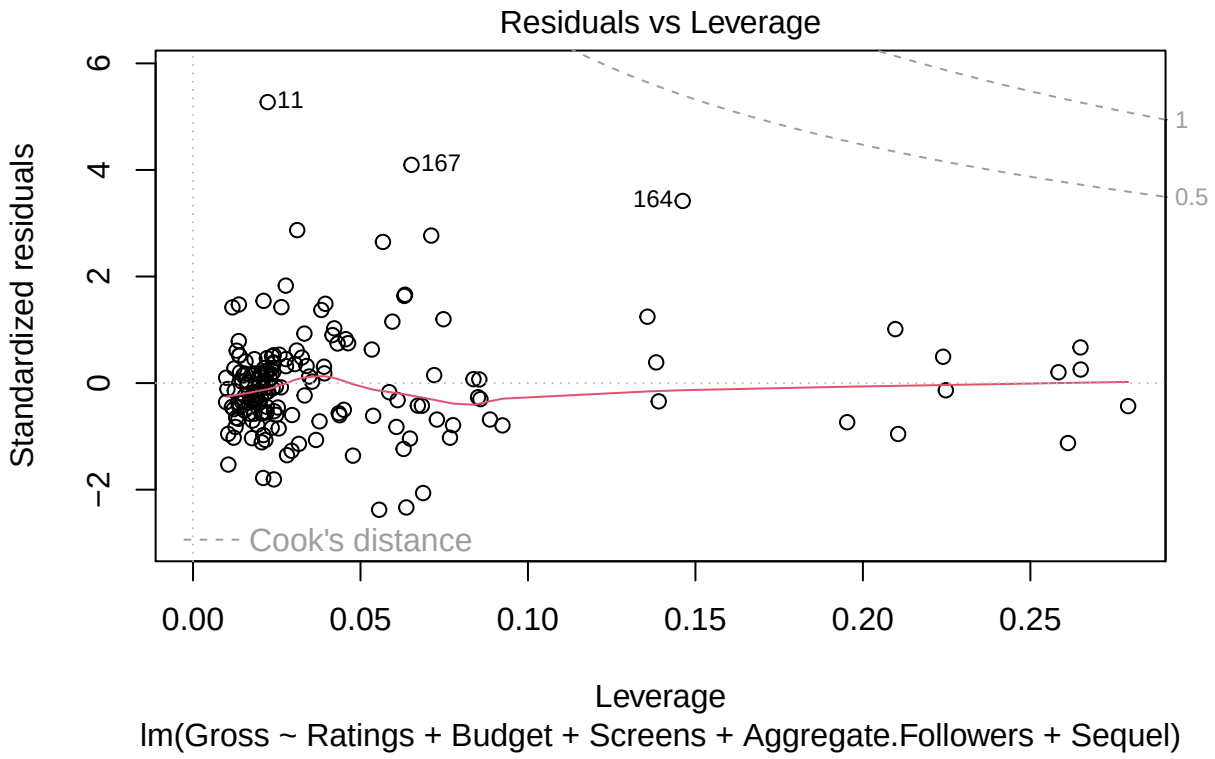
Nhận xét

- Từ đồ thị ta thấy phương sai của phần dư có thay đổi

Kiểm tra sự ảnh hưởng của dữ liệu

```
plot(model3, 5)
```

```
## Warning: not plotting observations with leverage one:
##   126, 133
```

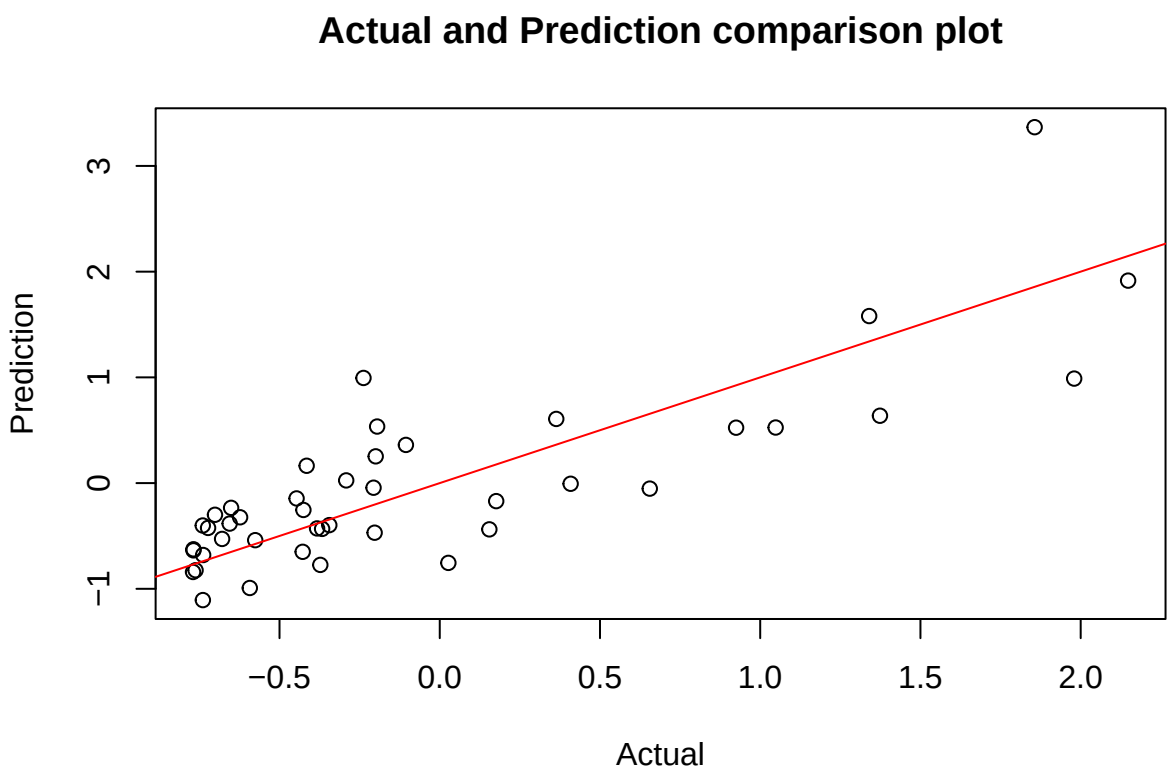


Nhận xét

- Từ đồ thị ta thấy quan sát 11, 167, 164 có thể là các quan sát có ảnh hưởng cao trong bộ dữ liệu

DỰ BÁO

```
validationData = validationData[-which(!(validationData$Sequel %in% unique(trainingData$Sequel))),]
trainingData = trainingData[-which(!(trainingData$Sequel %in% unique(validationData$Sequel))),]

predictions =  predict(model3, validationData)
rmse = RMSE(predictions, validationData$Gross)
r2 = R2(predictions, validationData$Gross)

plot(validationData$Gross, predictions, xlab = "Actual", ylab = "Prediction",
     main = "Actual and Prediction comparison plot")
abline(0, 1, col = "red")
```



Nhận xét

- Độ lệch trung bình giữa các giá trị dự đoán và các giá trị thực tế là RMSE = 0.4961573
- R2 = 0.667295 cho biết 66.73% biến thiên của biến phụ thuộc có thể được giải thích bởi các biến độc lập được sử dụng trong mô hình. Từ đây cho thấy mô hình phù hợp chặt chẽ với dữ liệu.
- Qua đồ thị ta thấy, mô hình dự báo không có quá nhiều sai lệch

DỮ LIỆU

- https://archive.ics.uci.edu/dataset/45/heart+disease
- https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings/data
- https://www.kaggle.com/datasets/vipullrathod/fish-market

THAM KHẢO

- Làm sạch dữ liệu: https://bigdatauni.com/tin-tuc/data-cleaning-lam-sach-du-lieu-xu-ly-missing-values-p2.html
- Xử lý dữ liệu khuyết: https://www.youtube.com/watch?v=2XcMzpeHNZU
- Tìm giá trị khuyết: https://rpubs.com/nlxbach/540930
- Hồi quy bội: https://rpubs.com/TKUD/810985
- Dummy variable: https://www.youtube.com/watch?v=2s8AwoKZ-UE
- Hồi quy đa biến: https://www.youtube.com/watch?v=-hYhY_IriuQ
- PCA: https://www.youtube.com/watch?v=tNw9yHrqFl8&t=5795s
- Tứ phân vị: https://phantichspss.com/cach-doc-bieu-do-hop-boxplot.html
- Box cox: https://www.r-bloggers.com/2022/10/box-cox-transformation-in-r/
- Thống kê mô tả: https://rpubs.com/R_stats/917311

LINK LẤY DỮ LIỆU

- https://archive.ics.uci.edu/dataset/45/heart+disease
- https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings/data
- https://www.kaggle.com/datasets/vipullrathod/fish-market