ĐẠI HỌC KINH TẾ QUỐC DÂN

-------***-------

BÀI TẬP GIỮA KỲ
MÔN PHÂN TÍCH THỐNG KÊ NHIỀU CHIỀU

Họ và tên : Bùi Văn Thái

MSV : 11233249

Lớp học phần : TOKT1143(125)_01

Giảng viên : TS. Nguyễn Mạnh Thế

Mục lục

1. Bài tập ANOVA

1.1 Mô tả dữ liệu và thống kê mô tả

Dựa trên dữ liệu thử nghiệm từ *Consumer Reports* đối với 16 mẫu xe đại diện cho các phân khúc thị trường chính (Small Car, Midsize Car, Small SUV, Midsize SUV), báo cáo này tập trung phân tích hiệu quả thực tế của công nghệ động cơ Hybrid so với động cơ xăng truyền thống (Conventional).

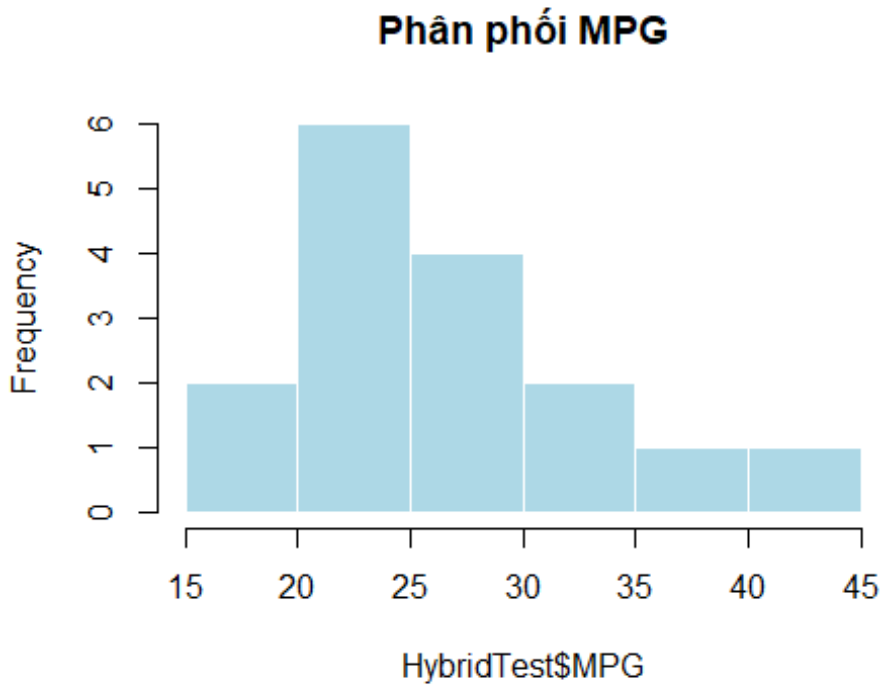
Mục tiêu cốt lõi là trả lời câu hỏi chiến lược: Công nghệ Hybrid có thực sự mang lại lợi thế vượt trội về tiết kiệm nhiên liệu cho mọi dòng xe hay không, hay chỉ hiệu quả ở một số phân khúc nhất định? Ở dưới là phần thống kê mô tả và kiểm định phân phối chuẩn Jarque–Bera. Kiểm định cho kết quả: p-value = 0,2085 > 0,05

Kết luận: Không bác bỏ giả thuyết phân phối chuẩn, do đó giả định phân phối chuẩn của ANOVA được thỏa mãn.

```
summary(HybridTest)
```

```
##   Make/Model                Class              Type         MPG        
##  Length:16          Midsize Car:4   Conventional:8   Min.   :18.00  
##  Class :character   Midsize SUV:4   Hybrid      :8   1st Qu.:22.75  
##  Mode  :character   Small Car  :4                    Median :26.00  
##                     Small SUV  :4                    Mean   :26.88  
##                                                      3rd Qu.:29.00  
##                                                      Max.   :44.00
```

```
# Kiểm tra phân phối chuẩn
hist(HybridTest$MPG, main="Phân phối MPG", col="lightblue", border="whit
e")
```




```
jarque.bera.test(HybridTest$MPG)
```

```
##
##  Jarque Bera Test
##
## data:  HybridTest$MPG
## X-squared = 3.1361, df = 2, p-value = 0.2085
```

So sánh MPG theo loại xe , cỡ xe:

- Xu hướng chủ đạo: Nhìn chung, các mẫu xe Hybrid luôn nằm ở mức MPG cao hơn so với đối thủ chạy xăng cùng phân khúc.
- Xe hybrid có mức tiêu hao nhiên liệu hiệu quả hơn rõ rệt so với xe xăng truyền thống. Xe cỡ nhỏ, đặc biệt là Small Car, có mức MPG cao nhất; trong khi SUV cỡ trung tiêu hao nhiên liệu nhiều nhất.
- Độ biến thiên: Phân khúc xe nhỏ (Small Car) có dải biến động MPG lớn nhất, cho thấy sự chênh lệch công nghệ tác động mạnh mẽ nhất ở nhóm này. Trong khi đó, ở nhóm xe gầm cao (SUV), khoảng cách giữa hai loại động cơ dường như thu hẹp lại.

```
# Tính trung bình theo từng nhóm (Code của bạn)
print("--- Trung bình MPG theo Loại động cơ ---")
```

```
## [1] "--- Trung bình MPG theo Loại động cơ ---"
```

```r
print(by(HybridTest$MPG, HybridTest$Type, mean))
```

```
## HybridTest$Type: Conventional
## [1] 23.5
## HybridTest$Type: Hybrid
## [1] 30.25
```

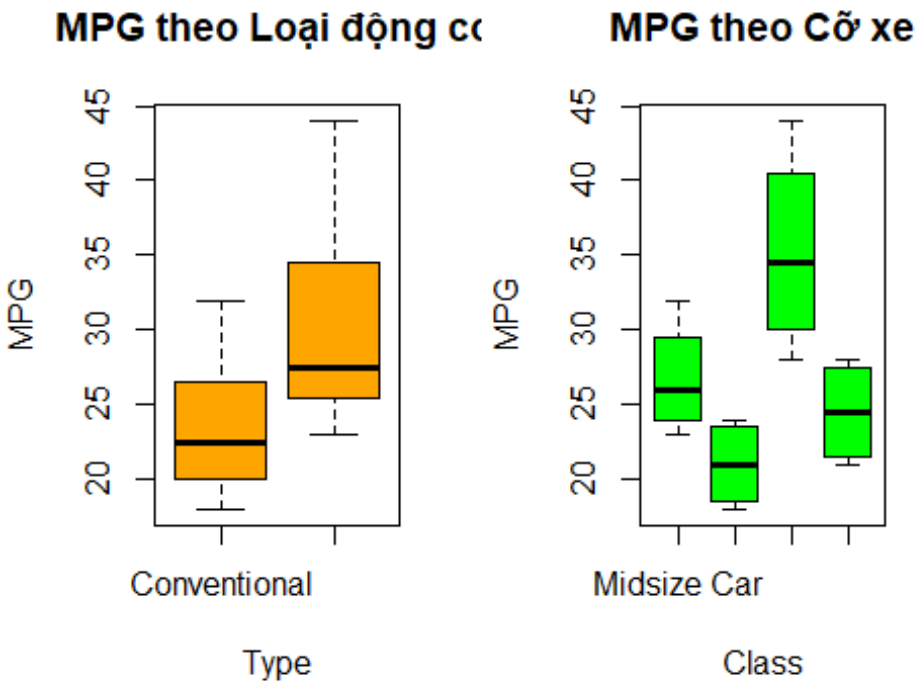
```r
print(" Trung bình MPG theo Cỡ xe")
```

```
## [1] " Trung bình MPG theo Cỡ xe"
```

```r
print(by(HybridTest$MPG, HybridTest$Class, mean))
```

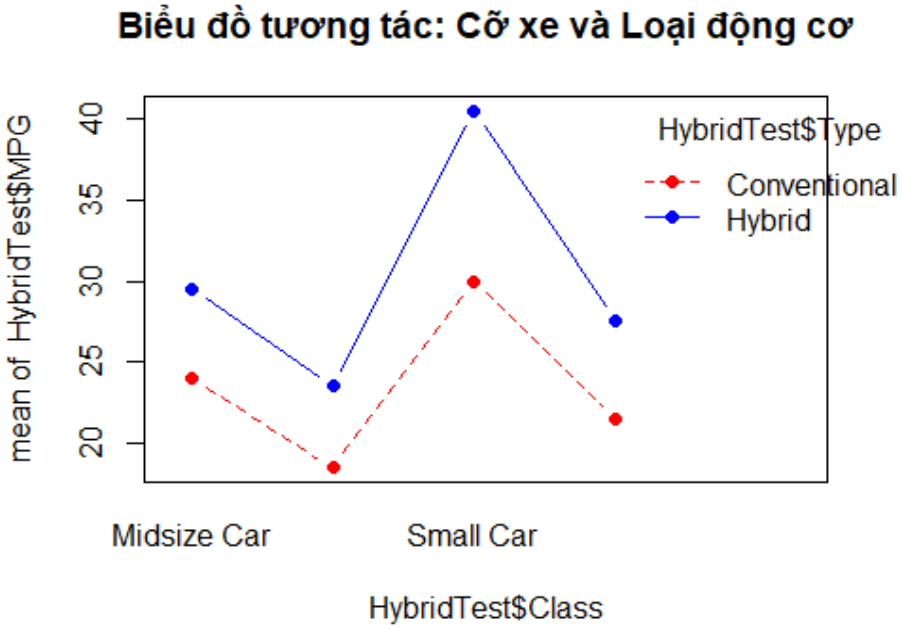
```
## HybridTest$Class: Midsize Car
## [1] 26.75
## HybridTest$Class: Midsize SUV
## [1] 21
## HybridTest$Class: Small Car
## [1] 35.25
## HybridTest$Class: Small SUV
## [1] 24.5
```

```r
# Vẽ biểu đồ hộp (Boxplot) để so sánh trực quan
par(mfrow=c(1,2))
boxplot(MPG ~ Type, data = HybridTest, main="MPG theo Loại động cơ", col
="orange")
boxplot(MPG ~ Class, data = HybridTest, main="MPG theo Cỡ xe", col="gree
n")
```




```r
par(mfrow=c(1,1))

# Biểu đồ tương tác (Interaction Plot)
```

```r
interaction.plot(x.factor = HybridTest$Class,
                 trace.factor = HybridTest$Type,
                 response = HybridTest$MPG,
                 fun = mean,
                 type = "b", col = c("red", "blue"), pch = 19,
                 fixed = TRUE,
                 legend = TRUE,
                 main = "Biểu đồ tương tác: Cỡ xe và Loại động cơ")
```



1.2. Kiểm định tác động bằng ANOVA hai nhân tố (α = 0,05)

1.2.1 Mô hình ANOVA

Mô hình được sử dụng:

$$MPG = \mu + \alpha_{Class} + \beta_{Type} + (\alpha\beta)_{Class \times Type} + \varepsilon$$

Trong đó:

- Class: cỡ xe
- Type: loại xe
- *Class×Type: tác động tương tác*

1.2.2 Diễn giải kết quả

- Tác động của cỡ xe (Class): p-value < 0,05 cho thấy có sự khác biệt có ý nghĩa thống kê về MPG giữa các cỡ xe.
- Tác động của loại xe (Type): p-value < 0,05 cho thấy có sự khác biệt có ý nghĩa thống kê về MPG giữa xe hybrid và xe xăng.
- *Tác động tương tác giữa cỡ xe và loại xe: p-value > 0,05 cho thấy không có bằng chứng thống kê cho thấy tồn tại tương tác giữa cỡ xe và loại xe .*

```r
# KIỂM ĐỊNH ANOVA 2 YẾU TỐ
anova.2way <- aov(MPG ~ Class * Type, data = HybridTest)

print("--- Kết quả phân tích phương sai (ANOVA) ---")
```

```
## [1] "--- Kết quả phân tích phương sai (ANOVA) ---"
```

```r
summary(anova.2way)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Class        3  441.3  147.08  24.014 0.000236 ***
## Type         1  182.2  182.25  29.755 0.000605 ***
## Class:Type   3   19.3    6.42   1.048 0.422860
## Residuals    8   49.0    6.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.3. Kiểm định so sánh cặp (Multiple Comparisons – Tukey HSD)

1.3.1 So sánh giữa các cỡ xe

Kết quả Tukey HSD cho thấy Small Car có MPG cao hơn có ý nghĩa thống kê so với:

- Midsize Car
- Midsize SUV
- Small SUV

Midsize SUV có MPG thấp hơn đáng kể so với Midsize Car và Small Car. Điều này khẳng định xe cỡ nhỏ tiết kiệm nhiên liệu hơn xe cỡ lớn, đặc biệt là SUV.

1.3.2 So sánh giữa các loại xe

- Chênh lệch trung bình MPG giữa Hybrid và Conventional: +6,75 MPG
- p-value = 0,0006052 < 0,05

Xe hybrid tiết kiệm nhiên liệu hơn xe xăng truyền thống một cách có ý nghĩa thống kê.

1.3.3 So sánh cặp theo tương tác

Do tác động tương tác không có ý nghĩa thống kê, nên các so sánh cặp theo tổ hợp *Class* × *Type* chỉ mang tính tham khảo và không phải là kết luận chính của nghiên cứu.

```r
# SO SÁNH CẶP

print("--- Kết quả so sánh cặp Tukey HSD ---")
```

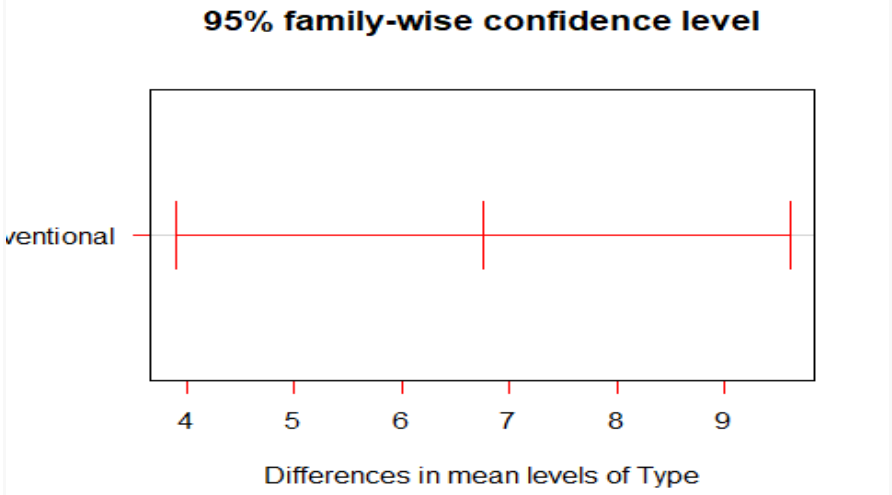
```
## [1] "--- Kết quả so sánh cặp Tukey HSD ---"
```

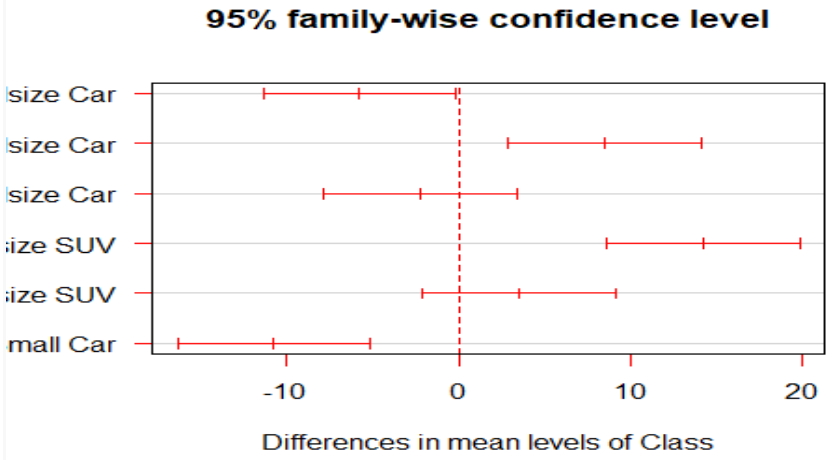
```r
TukeyHSD(anova.2way)
```

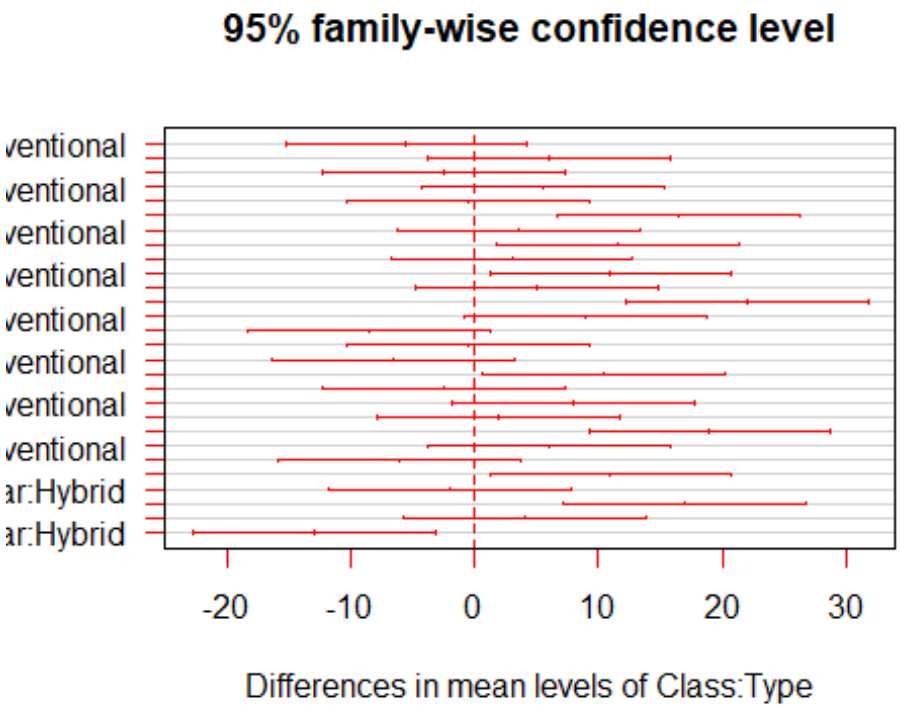
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = MPG ~ Class * Type, data = HybridTest)
##
## $Class
##                            diff        lwr       upr     p adj
## Midsize SUV-Midsize Car  -5.75 -11.354116 -0.145884 0.0444736
## Small Car-Midsize Car     8.50   2.895884 14.104116 0.0055113
## Small SUV-Midsize Car    -2.25  -7.854116  3.354116 0.5956797
## Small Car-Midsize SUV    14.25   8.645884 19.854116 0.0001773
## Small SUV-Midsize SUV     3.50  -2.104116  9.104116 0.2640524
## Small SUV-Small Car     -10.75 -16.354116 -5.145884 0.0012450
##
## $Type
##                      diff      lwr      upr     p adj
## Hybrid-Conventional 6.75 3.896465 9.603535 0.0006052
##
## $`Class:Type`
##                                                    diff         lwr
upr
## Midsize SUV:Conventional-Midsize Car:Conventional  -5.5 -15.2933155
4.293316
## Small Car:Conventional-Midsize Car:Conventional     6.0  -3.7933155 1
5.793316
## Small SUV:Conventional-Midsize Car:Conventional    -2.5 -12.2933155
7.293316
## Midsize Car:Hybrid-Midsize Car:Conventional         5.5  -4.2933155 1
5.293316
## Midsize SUV:Hybrid-Midsize Car:Conventional        -0.5 -10.2933155
9.293316
## Small Car:Hybrid-Midsize Car:Conventional          16.5   6.7066845 2
6.293316
## Small SUV:Hybrid-Midsize Car:Conventional           3.5  -6.2933155 1
3.293316
## Small Car:Conventional-Midsize SUV:Conventional    11.5   1.7066845 2
1.293316
## Small SUV:Conventional-Midsize SUV:Conventional     3.0  -6.7933155 1
2.793316
## Midsize Car:Hybrid-Midsize SUV:Conventional        11.0   1.2066845 2
0.793316
## Midsize SUV:Hybrid-Midsize SUV:Conventional         5.0  -4.7933155 1
4.793316
## Small Car:Hybrid-Midsize SUV:Conventional          22.0  12.2066845 3
1.793316
## Small SUV:Hybrid-Midsize SUV:Conventional           9.0  -0.7933155 1
8.793316
## Small SUV:Conventional-Small Car:Conventional      -8.5 -18.2933155
1.293316
## Midsize Car:Hybrid-Small Car:Conventional          -0.5 -10.2933155
9.293316
## Midsize SUV:Hybrid-Small Car:Conventional          -6.5 -16.2933155
3.293316
## Small Car:Hybrid-Small Car:Conventional            10.5   0.7066845 2
```

```
0.293316
## Small SUV:Hybrid-Small Car:Conventional            -2.5 -12.2933155
7.293316
## Midsize Car:Hybrid-Small SUV:Conventional           8.0  -1.7933155 1
7.793316
## Midsize SUV:Hybrid-Small SUV:Conventional           2.0  -7.7933155 1
1.793316
## Small Car:Hybrid-Small SUV:Conventional            19.0   9.2066845 2
8.793316
## Small SUV:Hybrid-Small SUV:Conventional             6.0  -3.7933155 1
5.793316
## Midsize SUV:Hybrid-Midsize Car:Hybrid              -6.0 -15.7933155
3.793316
## Small Car:Hybrid-Midsize Car:Hybrid                11.0   1.2066845 2
0.793316
## Small SUV:Hybrid-Midsize Car:Hybrid                -2.0 -11.7933155
7.793316
## Small Car:Hybrid-Midsize SUV:Hybrid                17.0   7.2066845 2
6.793316
## Small SUV:Hybrid-Midsize SUV:Hybrid                 4.0  -5.7933155 1
3.793316
## Small SUV:Hybrid-Small Car:Hybrid                 -13.0 -22.7933155 -
3.206684
##                                                        p adj
## Midsize SUV:Conventional-Midsize Car:Conventional 0.4250359
## Small Car:Conventional-Midsize Car:Conventional   0.3400389
## Small SUV:Conventional-Midsize Car:Conventional   0.9597903
## Midsize Car:Hybrid-Midsize Car:Conventional       0.4250359
## Midsize SUV:Hybrid-Midsize Car:Conventional       0.9999984
## Small Car:Hybrid-Midsize Car:Conventional         0.0022665
## Small SUV:Hybrid-Midsize Car:Conventional         0.8294285
## Small Car:Conventional-Midsize SUV:Conventional   0.0212356
## Small SUV:Conventional-Midsize SUV:Conventional   0.9072882
## Midsize Car:Hybrid-Midsize SUV:Conventional       0.0271801
## Midsize SUV:Hybrid-Midsize SUV:Conventional       0.5220242
## Small Car:Hybrid-Midsize SUV:Conventional         0.0003054
## Small SUV:Hybrid-Midsize SUV:Conventional         0.0752288
## Small SUV:Conventional-Small Car:Conventional     0.0974534
## Midsize Car:Hybrid-Small Car:Conventional         0.9999984
## Midsize SUV:Hybrid-Small Car:Conventional         0.2684550
## Small Car:Hybrid-Small Car:Conventional           0.0349135
## Small SUV:Hybrid-Small Car:Conventional           0.9597903
## Midsize Car:Hybrid-Small SUV:Conventional         0.1261941
## Midsize SUV:Hybrid-Small SUV:Conventional         0.9874561
## Small Car:Hybrid-Small SUV:Conventional           0.0008643
## Small SUV:Hybrid-Small SUV:Conventional           0.3400389
## Midsize SUV:Hybrid-Midsize Car:Hybrid             0.3400389
## Small Car:Hybrid-Midsize Car:Hybrid               0.0271801
## Small SUV:Hybrid-Midsize Car:Hybrid               0.9874561
## Small Car:Hybrid-Midsize SUV:Hybrid               0.0018547
## Small SUV:Hybrid-Midsize SUV:Hybrid               0.7326560
## Small SUV:Hybrid-Small Car:Hybrid                 0.0103746
```

```r
plot(TukeyHSD(anova.2way), las=1, col="red")
```


95% family-wise confidence level
ventional
4
5
6
7
8
9
Differences in mean levels of Type


95% family-wise confidence level
size Car
size Car
size Car
ize SUV
ize SUV
mall Car
-10
0
10
20
Differences in mean levels of Class



1.4. Kết luận chung cho Bài 1

- Mức tiêu hao nhiên liệu MPG phụ thuộc đáng kể vào cỡ xe và loại xe.
- Xe hybrid có hiệu quả sử dụng nhiên liệu vượt trội so với xe xăng.
- Xe cỡ nhỏ, đặc biệt là Small Car, đạt mức MPG cao nhất.
- Không có bằng chứng cho thấy hiệu quả của xe hybrid thay đổi khác nhau theo từng cỡ xe.

Khuyến nghị: Người tiêu dùng quan tâm đến tiết kiệm nhiên liệu nên ưu tiên xe hybrid và xe cỡ nhỏ, trong khi SUV cỡ trung là phân khúc có mức tiêu hao nhiên liệu cao nhất.

2. Bài tập PCA và phân tích nhân tố

2.1. Bối cảnh và mục tiêu nghiên cứu

Tệp dữ liệu *car_sale* bao gồm thông tin về doanh số bán ô tô (sales), giá niêm yết và nhiều thông số kỹ thuật của các mẫu xe khác nhau như dung tích động cơ, công suất, kích thước, mức tiêu hao nhiên liệu,…

Nhà phân tích mong muốn dự đoán doanh số bán xe từ tập hợp các yếu tố kỹ thuật và giá cả. Tuy nhiên, các biến giải thích này có mối tương quan chặt chẽ với nhau, có thể dẫn đến hiện tượng đa cộng tuyến, làm sai lệch kết quả hồi quy truyền thống.

Mục tiêu của bài toán là:

- Sử dụng Phân tích thành phần chính (PCA) để giảm chiều dữ liệu,
- Xác định các nhân tố tổng hợp đại diện cho nhóm đặc tính của xe,
- Từ đó đánh giá ảnh hưởng của các nhân tố này đến doanh số bán xe.

2.2 Chuẩn bị dữ liệu và phương pháp phân tích

2.2.1 Xử lý dữ liệu và kiểm định điều kiện để chạy PCA

- Các quan sát thiếu dữ liệu được loại bỏ. Phân tích PCA được thực hiện trên 11 biến định lượng (từ cột 15 đến 25), phản ánh:

- Giá bán
- Động cơ
- Công suất
- Kích thước
- Trọng lượng
- Mức tiêu hao nhiên liệu,…

Do các biến có đơn vị đo khác nhau, PCA được thực hiện với chuẩn hóa dữ liệu (scale = TRUE) nhằm đảm bảo các biến có vai trò tương đương trong phân tích.

-Kiểm định Bartlett về tính cầu (Bartlett's Test of Sphericity):

- Giả thuyết H_0: Ma trận tương quan là ma trận đơn vị (các biến không có tương quan với nhau).
- Kết quả: giá trị p-value < 0.05
- Kết luận: Bác bỏ H_0. Các biến quan sát có tương quan với nhau, thỏa mãn điều kiện để phân tích nhân tố.

- Hệ số KMO (Kaiser-Meyer-Olkin):

- Hệ số KMO tổng thể của mô hình là 0.84.
- Giá trị này nằm trong khoảng $0.8 \leq KMO < 0.9$, được đánh giá là Rất tốt (Meritorious).
- Kiểm tra từng biến riêng lẻ (MSA) cho thấy tất cả các biến đều có giá trị > 0.5, biến thấp nhất là ztype (0.64) và cao nhất là zwidth (0.95).

```r
#Đọc và xử lí dữ liệu ban đầu
car_sales <- read_excel("D:/Phân tích thống kê nhiều chiều/BT PTTKNC KTK64/car_sales.xlsx")
car_cleaned <- car_sales %>%
  mutate(across(everything(), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))
data_car = car_cleaned[,15:25]
```

```r
# Kiểm định điều kiện để chạy PCA
# Nếu không làm bước này, bài làm thiếu căn cứ khoa học
print("Kiểm định KMO (Yêu cầu > 0.5)")
```

```
## [1] "Kiểm định KMO (Yêu cầu > 0.5)"
```

```r
KMO(data_car)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_car)
## Overall MSA =  0.84
## MSA for each item =
##  zresale    ztype   zprice zengine_ zhorsepo zwheelba   zwidth  zleng
th
##     0.78     0.64     0.72     0.89     0.82     0.79     0.95     0.
73
## zcurb_wg zfuel_ca     zmpg
##     0.86     0.93     0.93
```

```r
print("--- Kiểm định Bartlett (Yêu cầu p.value < 0.05)")
```

```
## [1] "--- Kiểm định Bartlett (Yêu cầu p.value < 0.05)"
```

```r
cortest.bartlett(data_car)
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 1812.788
##
## $p.value
## [1] 0
##
## $df
## [1] 55
```

2.2.2 Phương pháp

Áp dụng Phân tích thành phần chính (PCA) để trích xuất các nhân tố không tương quan. Lựa chọn số lượng nhân tố dựa trên:

- Tỷ lệ phương sai giải thích
- Scree plot
- Sử dụng các thành phần chính thu được để hồi quy doanh số bán xe.

2.3. Kết quả phân tích PCA

2.3.1 Phương sai giải thích của các thành phần chính

Thành phần	Tỷ lệ phương sai	Phương sai tích lũy
PC1	56,6%	56,6%
PC2	19,17%	75,77%
PC3	10,77%	86,54%

Ba thành phần chính đầu tiên giải thích khoảng 86,54% tổng phương sai, cho thấy chúng nắm giữ hầu hết thông tin của dữ liệu ban đầu.

Biểu đồ scree plot thể hiện điểm "gãy" rõ rệt sau PC3, do đó việc lựa chọn 3 thành phần chính là hợp lý.

2.3.2 Ý nghĩa kinh tế của các thành phần chính

Dựa trên đặc điểm dữ liệu và các biến gốc, có thể diễn giải khái quát:

- PC1: Nhân tố phản ánh quy mô và sức mạnh kỹ thuật của xe (động cơ lớn, công suất cao, kích thước lớn, trọng lượng nặng)
- PC2: Nhân tố liên quan đến giá bán và định vị phân khúc thị trường
- PC3: Phản ánh các đặc điểm kỹ thuật bổ sung như hiệu suất nhiên liệu, thiết kế, hoặc các yếu tố phụ trợ khác

Các thành phần chính không tương quan với nhau, giúp loại bỏ vấn đề đa cộng tuyến.

2.4. Hồi quy doanh số bán xe theo các thành phần chính

2.4.1 Mô hình hồi quy

Doanh số bán xe được hồi quy theo bốn thành phần chính đầu tiên:

$$Sales = \beta_0 + \beta_1 PC1 + \beta_2 PC2 + \beta_3 PC3 + \beta_4 PC4 + \varepsilon$$

2.4.2 Kết quả hồi quy

Biến	Hệ số	p-value	Ý nghĩa
PC1	1,514	0,447	Không có ý nghĩa
PC2	-21,381	**< 0,01**	**Có ý nghĩa**
PC3	5,309	0,245	Không có ý nghĩa

- R^2 hiệu chỉnh = 19,69%
- Kiểm định F có p-value < 0,01 ⇒ mô hình có ý nghĩa tổng thể

2.4.3 Diễn giải kết quả

- PC2 là nhân tố có ảnh hưởng có ý nghĩa thống kê đến doanh số bán xe.
- Hệ số âm cho thấy: Khi nhân tố phản ánh giá bán/định vị cao cấp tăng, doanh số có xu hướng giảm, phù hợp với quy luật cầu trong kinh tế học.
- Các nhân tố còn lại (PC1, PC3) không có ảnh hưởng đáng kể đến doanh số trong mô hình.

2.5. Kết luận và hàm ý phân tích

- PCA đã giúp giảm chiều dữ liệu từ 11 biến xuống còn 3 nhân tố chính, đồng thời loại bỏ hiện tượng đa cộng tuyến.
- Doanh số bán xe không phụ thuộc trực tiếp vào quy mô hay sức mạnh kỹ thuật, mà chịu ảnh hưởng mạnh bởi nhân tố giá và phân khúc thị trường.
- Kết quả cho thấy giá bán là yếu tố then chốt quyết định doanh số, trong khi các đặc tính kỹ thuật cao cấp không đảm bảo bán chạy.

Hàm ý thực tiễn:

- Các hãng xe nên cân nhắc chiến lược giá phù hợp với thị trường mục tiêu.
- Nâng cấp kỹ thuật cần đi kèm với định vị giá hợp lý để tránh làm giảm doanh số.

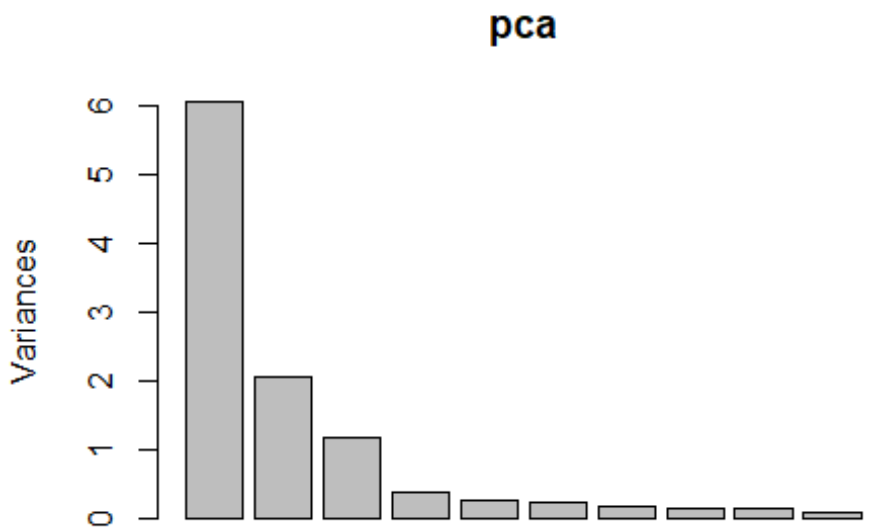
```
#PCA
pca = prcomp(data_car)
pca

## Standard deviations (1, .., p=11):
##  [1] 2.4592662 1.4312384 1.0729190 0.6186425 0.5024863 0.4665175 0.41
94226
##  [8] 0.3806824 0.3536970 0.2873493 0.2386491
##
## Rotation (n x k) = (11 x 11):
##                PC1         PC2         PC3         PC4         PC5
PC6
## zresale   0.1582503  0.46979045 -0.08343085 -0.41063248 -0.08907878
0.39534925
## ztype     0.1772056 -0.34431836 -0.65195848 -0.04400952  0.17531336
0.49616954
## zprice    0.2526043  0.49467933 -0.03695500 -0.29090635 -0.16567726 -
0.06021775
## zengine_  0.3565980  0.14803030  0.04978350  0.46673092  0.41693096
0.09877509
## zhorsepo  0.3126656  0.38534863  0.11338746  0.17161540  0.28989636
0.08584561
## zwheelba  0.2858918 -0.36399221  0.27476047 -0.42747978  0.17306745
```

```
0.24551021
## zwidth    0.3275826 -0.13653498  0.28067621  0.42598783 -0.65553562
0.39371345
## zlength   0.2832825 -0.25718667  0.49373439 -0.22850789  0.27810735 -
0.10381271
## zcurb_wg  0.3737221 -0.08407988 -0.11536317 -0.03795245 -0.17252386 -
0.37648978
## zfuel_ca  0.3543292 -0.14433284 -0.21521219 -0.21433709 -0.30462296 -
0.31078345
## zmpg     -0.3468622  0.02119240  0.30596995 -0.18364857 -0.12816899
0.33521495
##                  PC7         PC8         PC9        PC10        PC11
## zresale  -0.42344137 -0.29101770  0.22102773  0.23310933  0.21737260
## ztype     0.16161847  0.03195720  0.20232754 -0.24515221 -0.13401316
## zprice    0.34222562  0.36342086 -0.01764759 -0.03252869 -0.56908126
## zengine_  0.19248456 -0.36008135 -0.07508561  0.47305344 -0.22559818
## zhorsepo  0.11315268  0.10387554 -0.19419409 -0.58972205  0.45929468
## zwheelba  0.08097413  0.32571911 -0.43083801  0.32294474  0.18641677
## zwidth   -0.10387039  0.10053461  0.03693490 -0.06055516 -0.05602221
## zlength  -0.18427612 -0.18415124  0.45057100 -0.32307711 -0.30697451
## zcurb_wg  0.36401275  0.08745039  0.51155436  0.24140717  0.45484742
## zfuel_ca  0.01677015 -0.56816258 -0.45822409 -0.19065606 -0.05071166
## zmpg      0.66596891 -0.40149208  0.06994787 -0.07735924  0.08665511
```

```
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
PC7
## Standard deviation     2.459 1.4312 1.0729 0.61864 0.50249 0.46652 0.
41942
## Proportion of Variance 0.566 0.1917 0.1077 0.03582 0.02363 0.02037 0.
01646
## Cumulative Proportion  0.566 0.7577 0.8654 0.90122 0.92485 0.94521 0.
96167
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.38068 0.35370 0.28735 0.23865
## Proportion of Variance 0.01356 0.01171 0.00773 0.00533
## Cumulative Proportion  0.97524 0.98694 0.99467 1.00000
```

```
screeplot(pca)
```




```
#Regression
print(pca$rotation[, 1:3], cutoff = 0.3)
```

```
##                 PC1         PC2         PC3
## zresale   0.1582503  0.46979045 -0.08343085
## ztype     0.1772056 -0.34431836 -0.65195848
## zprice    0.2526043  0.49467933 -0.03695500
## zengine_  0.3565980  0.14803030  0.04978350
## zhorsepo  0.3126656  0.38534863  0.11338746
## zwheelba  0.2858918 -0.36399221  0.27476047
## zwidth    0.3275826 -0.13653498  0.28067621
## zlength   0.2832825 -0.25718667  0.49373439
## zcurb_wg  0.3737221 -0.08407988 -0.11536317
## zfuel_ca  0.3543292 -0.14433284 -0.21521219
## zmpg     -0.3468622  0.02119240  0.30596995
```

```
reg = lm(car_cleaned$sales~pca$x[,1:3])
summary(reg)
```

```
##
## Call:
## lm(formula = car_cleaned$sales ~ pca$x[, 1:3])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -98.60 -33.02 -14.99  15.02 405.06
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       52.998      4.866  10.893  < 2e-16 ***
## pca$x[, 1:3]PC1    1.514      1.985   0.763    0.447
```

```
## pca$x[, 1:3]PC2  -21.381      3.410  -6.270 3.52e-09 ***
## pca$x[, 1:3]PC3    5.309      4.549   1.167    0.245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.96 on 153 degrees of freedom
## Multiple R-squared:  0.2124, Adjusted R-squared:  0.1969
## F-statistic: 13.75 on 3 and 153 DF,  p-value: 5.486e-08
```

3. Bài tập phân tích cụm

3.1. Bối cảnh và mục tiêu nghiên cứu

Trong bối cảnh thị trường ô tô cạnh tranh gay gắt, các nhà sản xuất cần hiểu rõ cấu trúc thị trường nhằm xác định vị thế cạnh tranh của từng dòng xe. Nếu các mẫu xe có thể được nhóm tự động dựa trên giá bán và đặc tính kỹ thuật, doanh nghiệp có thể:

- Nhận diện các phân khúc sản phẩm tương đồng,
- So sánh xe của mình với đối thủ cạnh tranh trực tiếp,
- Hỗ trợ xây dựng chiến lược giá và phát triển sản phẩm.

Mục tiêu của bài tập là sử dụng phân tích cụm phân cấp (hierarchical clustering) để nhóm các mẫu ô tô bán chạy dựa trên giá cả và các đặc tính kỹ thuật, sử dụng dữ liệu *car_sale*.

3.2. Chuẩn bị dữ liệu và phương pháp phân tích

3.2.1 Dữ liệu sử dụng

Phân tích được thực hiện trên **các biến định lượng đã chuẩn hóa** (z-score), bao gồm:

- Giá bán
- Dung tích động cơ
- Công suất
- Kích thước (chiều dài, chiều rộng, chiều dài cơ sở)
- Trọng lượng xe
- Dung tích bình nhiên liệu
- Mức tiêu hao nhiên liệu (MPG)

Việc chuẩn hóa là cần thiết do các biến có đơn vị đo khác nhau, tránh biến có thang đo lớn chi phối kết quả phân cụm.

3.2.2 Phương pháp phân tích cụm

Quy trình phân tích cụm phân cấp gồm các bước:

B1: Lựa chọn thước đo khoảng cách: Khoảng cách Euclidean.

B2: So sánh các phương pháp liên kết (linkage):

- Single
- Average
- Complete
- Ward

B3: Tiêu chí lựa chọn phương pháp tối ưu: Hệ số kết dính (Agglomerative Coefficient – AC) càng cao càng tốt.

3.3. Lựa chọn phương pháp phân cụm

3.3.1 So sánh hệ số kết dính

Phương pháp	**Agglomerative Coefficient**
Single	0,779
Average	0,872
Complete	0,924
Ward	**0,967**

Phương pháp Ward cho hệ số kết dính cao nhất, cho thấy các quan sát trong cùng cụm có mức độ tương đồng cao, do đó được lựa chọn cho phân tích tiếp theo.

3.4. Xác định số lượng cụm tối ưu

3.4.1 Dendrogram

Biểu đồ cây phân cấp (dendrogram) cho thấy dữ liệu có thể được chia thành 2 nhóm lớn với khoảng cách phân tách rõ rệt.

3.4.2 Chỉ số Gap Statistic

Kết quả Gap Statistic xác nhận rằng 2 cụm là lựa chọn tối ưu, vì tại mức này khoảng cách giữa các cụm được tối đa hóa so với dữ liệu ngẫu nhiên.

Kết luận: Chọn 2 cụm để phân tích thị trường ô tô.

3.5. Kết quả phân cụm

Sau khi cắt cây phân cấp tại k = 2, số lượng quan sát trong mỗi cụm là:

- Cụm 1: 49 mẫu xe
- Cụm 2: 68 mẫu xe

Các nhãn cụm được gắn vào dữ liệu gốc để phục vụ phân tích đặc điểm.

3.6. Đặc điểm của các cụm xe

3.6.1 So sánh giá trị trung bình các biến theo cụm

Cụm 1 – Nhóm xe phổ thông, tiết kiệm

- Giá bán trung bình: ≈ 19,55
- Dung tích động cơ: ≈ 2,2
- Công suất: ≈ 143 mã lực
- Trọng lượng: ≈ 2,79
- MPG trung bình: ≈ 26,96

Đặc trưng:

- Giá thấp
- Động cơ nhỏ
- Tiết kiệm nhiên liệu
- Phù hợp với phân khúc đại chúng

Cụm 2 – Nhóm xe cao cấp, hiệu suất cao

- Giá bán trung bình: ≈ 33,07
- Dung tích động cơ: ≈ 3,69
- Công suất: ≈ 216 mã lực
- Trọng lượng: ≈ 3,71
- MPG trung bình: ≈ 21,59

Đặc trưng:

- Giá cao
- Động cơ lớn
- Công suất mạnh
- Tiêu hao nhiên liệu nhiều hơn
- Thuộc phân khúc cao cấp / SUV / xe hiệu suất cao

3.7. Diễn giải và hàm ý quản trị

Phân tích cụm phân cấp đã phân chia thị trường ô tô thành hai phân khúc rõ rệt:

- Phân khúc xe phổ thông – tiết kiệm nhiên liệu
- Phân khúc xe cao cấp – hiệu suất cao

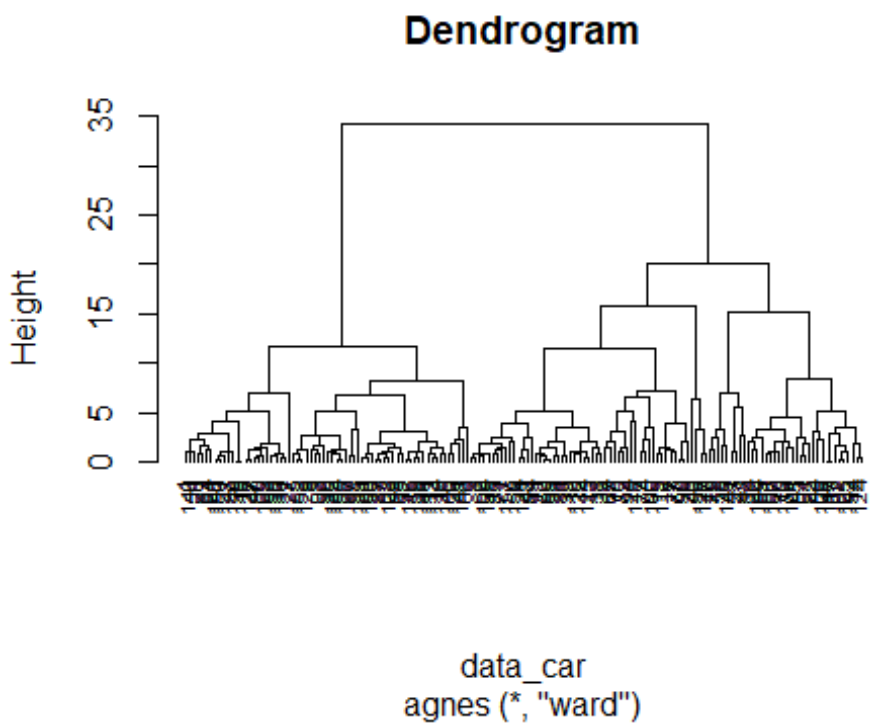
Các nhà sản xuất có thể:

- So sánh sản phẩm của mình với các xe trong cùng cụm,
- Xác định đối thủ cạnh tranh trực tiếp,
- Điều chỉnh chiến lược giá và cấu hình kỹ thuật phù hợp với phân khúc mục tiêu.

3.8. Kết luận

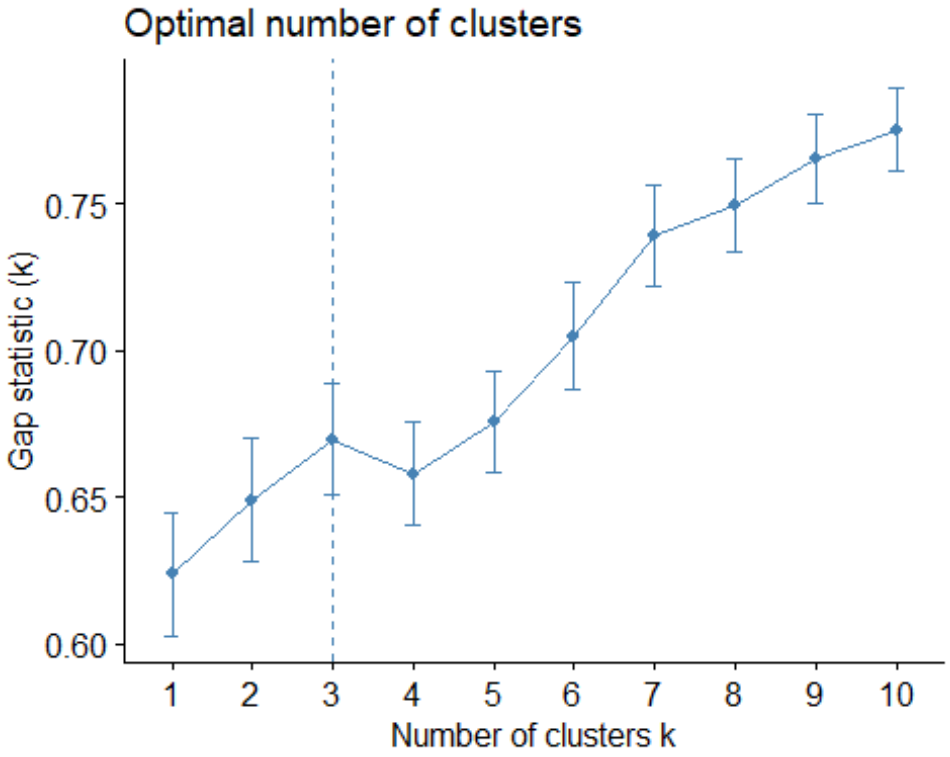
Phân tích cụm phân cấp là công cụ hữu hiệu trong phân đoạn thị trường ô tô. Kết quả nghiên cứu cho thấy thị trường có thể được chia thành hai nhóm xe có đặc điểm giá cả và kỹ thuật khác biệt rõ ràng, từ đó hỗ trợ doanh nghiệp trong việc định vị sản phẩm và nâng cao năng lực cạnh tranh.

```
#Bài tập 3: Phân tích cụm
#Xác định các phương pháp liên kết
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")

#Hàm để tính hệ số kết tụ
ac <- function(x) {
  agnes(data_car, method = x)$ac
}

#Tính toán hệ số kết tụ cho mỗi phương pháp liên kết phân cụm
sapply(m, ac)

##   average    single  complete      ward
## 0.8719614 0.7795969 0.9242269 0.9676587

#Wald cho hệ số lớn nhất->dùng Wald cho hierachical cluster

#Thực hiện phân cụm phân cấp bằng cách sử dụng phương sai tối thiểu của
Ward
clust <- agnes(data_car, method = "ward")

#Produce dendrogram
pltree(clust, cex = 0.6, hang = -1, main = "Dendrogram")
```




```
#Tính toán thống kê khoảng cách cho mỗi số lượng cụm (tối đa 10 cụm)
gap_stat <- clusGap(data_car, FUN = hcut, nstart = 25, K.max = 10, B = 5
0)

#Produce plot of clusters và gap statistic
fviz_gap_stat(gap_stat)
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## ℹ Please use tidy evaluation idioms with `aes()`.
## ℹ See also `vignette("ggplot2-in-packages")` for more information.
## ℹ The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextr
a/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning
was
## generated.
```




```
#compute distance matrix
d <- dist(data_car, method = "euclidean")

#Thực hiện phân cụm phân cấp bằng phương pháp Ward
final_clust <- hclust(d, method = "ward.D2" )

#Cut the dendrogram into 2 clusters
groups <- cutree(final_clust, k=2)

#Find number of observations in each cluster
table(groups)

## groups
##  1  2
## 66 91

#Thêm nhãn cụm vào dữ liệu gốc
final_data <- cbind(car_cleaned, cluster = groups)

#Display first six rows of final data
head(final_data)

##     model  sales resale type    price engine_s horsepow wheelbas widt
h length
## 1 Integra 16.919 16.360    0 21.50000      1.8      140    101.2  67.
3  172.4
## 2      TL 39.384 19.875    0 28.40000      3.2      225    108.1  70.
3  192.9
## 3      CL 14.114 18.225    0 27.39075      3.2      225    106.9  70.
6  192.0
## 4      RL  8.588 29.725    0 42.00000      3.5      210    114.6  71.
```

```
4  196.6
## 5      A4 20.397 22.255    0 23.99000      1.8      150    102.6  68.
2  178.0
## 6      A6 18.780 23.555    0 33.95000      2.8      200    108.7  76.
1  192.0
##   curb_wgt fuel_cap mpg  lnsales     zresale      ztype        zprice
## 1    2.639     13.2  28 2.828437 -0.14956062 -0.5926188 -4.104583e-01
## 2    3.517     17.2  25 3.673360  0.15733558 -0.5926188  7.032257e-02
## 3    3.470     17.2  26 2.647167  0.01327335 -0.5926188 -3.622718e-16
## 4    3.850     18.0  22 2.150366  1.01734341 -0.5926188  1.017949e+00
## 5    2.998     16.4  27 3.015388  0.36513442 -0.5926188 -2.369591e-01
## 6    3.561     18.5  22 2.932792  0.47863799 -0.5926188  4.570376e-01
##     zengine_   zhorsepo    zwheelba      zwidth    zlength   zcurb_wg
## 1 -1.2070012 -0.8103784 -0.82278892 -1.11533688 -1.1125568 -1.1721235
## 2  0.1331567  0.6887312  0.08019843 -0.24624321  0.4136772  0.2204185
## 3  0.1331567  0.6887312 -0.07684285 -0.15933384  0.3466718  0.1458746
## 4  0.4203334  0.4241825  0.93083869  0.07242447  0.6891438  0.7485693
## 5 -1.2070012 -0.6340126 -0.63957410 -0.85460878 -0.6956344 -0.6027356
## 6 -0.2497456  0.2478166  0.15871907  1.43400456  0.3466718  0.2902042
##      zfuel_ca       zmpg cluster
## 1 -1.22222719  0.9705266       1
## 2 -0.19339977  0.2700372       2
## 3 -0.19339977  0.5035336       2
## 4  0.01236571 -0.4304523       2
## 5 -0.39916525  0.7370301       1
## 6  0.14096914 -0.4304523       2

#Find mean values for each cluster
aggregate(final_data, by=list(cluster=final_data$cluster), mean)

## Warning in mean.default(X[[i]], ...): argument is not numeric or logi
cal:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logi
cal:
## returning NA

##   cluster model    sales   resale       type    price engine_s horsep
ow
## 1       1    NA 50.24198 14.05895 0.07575758 19.55362 2.192424 143.37
88
## 2       2    NA 54.99700 20.98425 0.39560440 33.07483 3.690779 216.82
36
##   wheelbas    width   length curb_wgt fuel_cap      mpg  lnsales    z
resale
## 1 102.5667 68.45455 177.9242 2.849061 15.04091 26.95606 3.347247 -0.3
504665
## 2 111.0559 73.10495 194.1752 3.761671 20.06321 21.58605 3.258622  0.2
541845
##        ztype     zprice   zengine_   zhorsepo   zwheelba     zwidth
zlength
## 1 -0.4207024 -0.5460788 -0.8313509 -0.7507882 -0.6439364 -0.7808675 -
0.7012745
## 2  0.3051248  0.3960572  0.6029578  0.5445277  0.4670308  0.5663435
```

```
0.5086167
##     zcurb_wg   zfuel_ca       zmpg cluster
## 1 -0.8389593 -0.7487328  0.7267704       1
## 2  0.6084760  0.5430369 -0.5271082       2
```

4. Bài tập phân tích khác biệt

4.1. Bối cảnh và mục tiêu nghiên cứu

Hãng hàng không tiến hành thu thập dữ liệu về nhân viên thuộc ba nhóm công việc:

- Nhân viên dịch vụ khách hàng
- Công nhân kỹ thuật
- Nhân viên điều phối

Mục tiêu là xác định liệu ba nhóm công việc này có thể được phân biệt dựa trên các đặc điểm tính cách của nhân viên hay không, từ đó đánh giá tính hợp lý của việc phân loại công việc hiện tại.

Ba đặc điểm tính cách được đo lường bao gồm:

- OUTDOOR: Tính hướng ngoại
- SOCIAL: Tính hòa đồng
- CONSERVATIVE: Tính bảo thủ

Mục tiêu nghiên cứu: Sử dụng phân tích khác biệt tuyến tính (Linear Discriminant Analysis – LDA) để phân loại nhân viên thành 3 nhóm công việc dựa trên các đặc điểm tính cách.

4.2. Chuẩn bị dữ liệu và phương pháp phân tích

4.2.1 Xử lý dữ liệu

- Ba biến định lượng OUTDOOR, SOCIAL và CONSERVATIVE được chuẩn hóa (standardization) nhằm loại bỏ ảnh hưởng của đơn vị đo.
- Dữ liệu được chia thành:
  - Tập huấn luyện (training set): 70%
  - Tập kiểm tra (test set): 30%
- Phân tích khác biệt được thực hiện trên tập huấn luyện để xây dựng mô hình.

4.2.2 Phương pháp

- Áp dụng phân tích khác biệt tuyến tính (LDA) để tìm các hàm phân biệt tối ưu.

- Các hàm phân biệt là tổ hợp tuyến tính của các biến tính cách, nhằm tối đa hóa sự khác biệt giữa các nhóm công việc.

4.3. Kết quả phân tích khác biệt

4.3.1 Xác suất tiên nghiệm (Prior probabilities)

Nhóm công việc	Xác suất
Nhóm 1	0,395
Nhóm 2	0,355
Nhóm 3	0,250

Các xác suất phản ánh tỷ trọng quan sát của từng nhóm trong tập huấn luyện.

4.3.2 Giá trị trung bình theo nhóm

Kết quả cho thấy sự khác biệt rõ rệt về đặc điểm tính cách giữa các nhóm:

- **Nhóm 1**:
    - Ít hướng ngoại
    - Hòa đồng cao
    - Ít bảo thủ
- **Nhóm 2**:
    - Hướng ngoại cao
    - Hòa đồng ở mức trung bình
    - Ít bảo thủ
- **Nhóm 3**:
    - Ít hướng ngoại
    - Ít hòa đồng
    - Mức độ bảo thủ cao

Điều này cho thấy mỗi loại công việc có đặc điểm tính cách điển hình riêng.

4.3.3 Các hàm phân biệt tuyến tính

Mô hình LDA trích xuất **2 hàm phân biệt**:

Hàm	Tỷ lệ giải thích
LD1	77,52%

Hàm	Tỷ lệ giải thích
LD2	22,48%

Hàm phân biệt thứ nhất (LD1) giải thích phần lớn sự khác biệt giữa các nhóm công việc.

4.3.4 Hệ số của các hàm phân biệt

Biến	LD1	LD2
OUTDOOR	-0,26	1,13
SOCIAL	1,06	0,081
CONSERVATIVE	-0,63	-0,193

Diễn giải

- LD1 phân biệt chủ yếu dựa trên:
    - Giảm tính hòa đồng
    - Tăng tính bảo thủ và hướng ngoại
- LD2 phản ánh sự đối lập giữa:
    - Tính hướng ngoại
    - Các yếu tố xã hội và bảo thủ

4.4. Đánh giá khả năng phân loại của mô hình

Mô hình LDA được sử dụng để dự đoán nhóm công việc cho tập kiểm tra.

- Tỷ lệ phân loại đúng: 77,78%

Tỷ lệ này tương đối cao so với phân loại ngẫu nhiên (≈33%), cho thấy mô hình có khả năng phân biệt tốt giữa ba nhóm công việc.

4.5. Trực quan hóa kết quả

Biểu đồ phân tán theo hai hàm phân biệt LD1 và LD2 cho thấy:

- Các nhóm công việc tách biệt tương đối rõ ràng,
- Một số điểm chồng lấn tồn tại, phản ánh sự giao thoa về đặc điểm tính cách giữa các nhóm nhân viên.

4.6. Kết luận và hàm ý thực tiễn

- Phân tích khác biệt cho thấy ba nhóm công việc trong hãng hàng không có thể được phân biệt tương đối tốt dựa trên đặc điểm tính cách.

- Các kết quả ủng hộ quan điểm rằng việc phân loại công việc hiện tại là hợp lý.
- Mô hình có thể được sử dụng để:
    - Hỗ trợ tuyển dụng và bố trí nhân sự,
    - Định hướng đào tạo phù hợp với từng nhóm tính cách.

Hàm ý thực tiễn:

- Nhân viên dịch vụ khách hàng phù hợp với tính cách hòa đồng, giao tiếp tốt.
- Công nhân kỹ thuật phù hợp với nhóm ít hòa đồng, bảo thủ hơn.
- Nhân viên điều phối nằm ở vị trí trung gian, yêu cầu cân bằng giữa giao tiếp và tính hệ thống.

```
## Bài tập 4: Discriminant analysis
discrim <- read_excel("D:/Phân tích thống kê nhiều chiều/BT PTTKNC KTK64
/discriminant_example.xlsx")
discrim[,1:3] = scale(discrim[,1:3])
attach(discrim)
set.seed(1)
#Chia data thành train/test
sample = sample(c(TRUE, FALSE), nrow(discrim), replace = TRUE, prob = c(
0.7, 0.3))
train = discrim[sample,]
test = discrim[!sample,]
dim(train)

## [1] 172   5

dim(test)

## [1] 72  5

boxM(
  train[, c("OUTDOOR", "SOCIAL", "CONSERVATIVE")],
  grouping = train$JOB
)

##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  train[, c("OUTDOOR", "SOCIAL", "CONSERVATIVE")]
## Chi-Sq (approx.) = 19.916, df = 12, p-value = 0.06868

# LDA
model = lda(JOB~OUTDOOR+SOCIAL+CONSERVATIVE, data = train)
model
```

```
## Call:
## lda(JOB ~ OUTDOOR + SOCIAL + CONSERVATIVE, data = train)
##
## Prior probabilities of groups:
##         1         2         3
## 0.3372093 0.3953488 0.2674419
##
## Group means:
##       OUTDOOR     SOCIAL CONSERVATIVE
## 1 -0.48476725  0.6412224   -0.4729481
## 2  0.59105280  0.1557112   -0.1741412
## 3 -0.08268996 -0.9605647    0.7107906
##
## Coefficients of linear discriminants:
##                     LD1         LD2
## OUTDOOR      -0.2594236  1.13881884
## SOCIAL        1.0599198  0.08134661
## CONSERVATIVE -0.6309429 -0.19312843
##
## Proportion of trace:
##    LD1    LD2
## 0.7752 0.2248
```

```
#Dự đoán trên test
predicted = predict(model, test)
names(predicted)
```

```
## [1] "class"     "posterior" "x"
```

```
head(predicted$class)
```

```
## [1] 1 2 1 1 1 1
## Levels: 1 2 3
```

```
head(predicted$posterior)
```

```
##           1           2           3
## 1 0.7657277 0.224722096 0.009550166
## 2 0.1927068 0.762466430 0.044826765
## 3 0.9166348 0.059946836 0.023418385
## 4 0.9358532 0.060618809 0.003527994
## 5 0.9933849 0.004310651 0.002304459
## 6 0.7859103 0.153692941 0.060396720
```

```
head(predicted$x)
```

```
##         LD1        LD2
## 1 1.4255621 -0.3990634
## 2 0.3301899  0.9533301
## 3 1.0808993 -2.0301704
## 4 1.8575495 -1.3623782
## 5 1.9583865 -3.6710842
## 6 0.6727544 -1.4144804
```

```
mean(predicted$class == test$JOB)
```

```
## [1] 0.7777778

# Plot
lda_plot <- cbind(train, predict(model)$x)

lda_plot$JOB <- factor(lda_plot$JOB, levels = c(1, 2, 3))

ggplot(lda_plot, aes(x = LD1, y = LD2)) +
  geom_point(aes(color = JOB)) +
  scale_color_manual(values = c("red", "green", "blue")) +
  theme() +
  labs(title = "LDA Plot", color = "JOB")
```