ĐẠI HỌC QUỐC GIA TP. HCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN – TIN HỌC

BÁO CÁO ĐỒ ÁN CUỐI KỲ

THỐNG KÊ NHIỀU CHIỀU

NHÓM: 7

Thành viên:

Đoàn Thị Kỳ Duyên – 21110280

Võ Thị Hồng Gấm – 21110281

Lớp: 21TTH_KDL

Giảng viên: Nguyễn Thị Mộng Ngọc

TP. Hồ Chí Minh – Năm 2024

ĐẠI HỌC QUỐC GIA TP. HCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN – TIN HỌC

BÁO CÁO ĐỒ ÁN CUỐI KỲ

THỐNG KÊ NHIỀU CHIỀU

NHÓM: 7

Thành viên:

Đoàn Thị Kỳ Duyên – 21110280

Võ Thị Hồng Gấm – 21110281

Lớp: 21TTH_KDL

Giảng viên: Nguyễn Thị Mộng Ngọc

TP. Hồ Chí Minh – Năm 2024

MỤC LỤC

MỤC LỤC 3

ĐỀ TÀI VÀ NGUỒN DỮ LIỆU 4

1. Mức độ căng thẳng của sinh viên. 4

2. Dự đoán hóa đơn giả 4

NỘI DUNG 6

A. PHẦN 1 6

1. PHÂN TÍCH DỮ LIỆU NGHIÊN CỨU VỀ MỨC ĐỘ CĂNG THẲNG (STRESS) CỦA SINH VIÊN. 6

1.1. Phân tích thành phần chính 6

1.2. Phân tích nhân tố 25

2. BÀI TẬP 38

2.1. Bài 4.21 trang 205 của sách Johnson, 2013. 38

2.2. Bài 6.33 trang 355-356 của sách Johnson, 2013. 39

B. PHẦN 2 50

1. PHÂN TÍCH DỮ LIỆU DỰ ĐOÁN HÓA ĐƠN GIẢ BẰNG PHƯƠNG PHÁP PHÂN TÍCH THÀNH PHẦN CHÍNH 51

2. PHÂN TÍCH DỮ LIỆU DỰ ĐOÁN HÓA ĐƠN GIẢ BẰNG PHƯƠNG PHÁP PHÂN TÍCH NHÂN TỐ 63

TÀI LIỆU THAM KHẢO 76

BẢNG PHÂN CÔNG CÔNG VIỆC 77

ĐỀ TÀI VÀ NGUỒN DỮ LIỆU

1. Mức độ căng thẳng của sinh viên.

Căng thẳng thường được mô tả là một tình trạng tiêu cực hay tích cực có ảnh hưởng đến sức khỏe tinh thần và thể chất của người đó.

Theo tâm lý học giải thích thì đây là một cảm giác căng thẳng và dồn ép. Áp lực với cường độ thấp có thể là một điều tốt và thậm chí có lợi ích trong công việc và sức khỏe.

Căng thẳng tích cực giúp tăng hiệu suất vận động thể thao. Nó cũng có vai trò trong động lực, thích nghi và phản ứng với môi trường xung quanh. Tuy nhiên với một lượng áp lực quá nhiều có thể dẫn đến nhiều vấn đề đối với cơ thể và điều đó có thể cực kì có hại.

Căng thẳng có thể từ bên ngoài và liên quan đến môi trường sống, nhưng cũng có thể được tạo ra từ sự nhìn nhận sinh bản thân dẫn đến lo âu hay các cảm xúc tiêu cực khác như dồn ép, không thoải mái quanh một tình huống mà sau đó họ sẽ cho là sự kiện áp lực.

Theo sinh lý học và sinh học, căng thẳng là một phản ứng của cơ thể sống đối với stressor (nghĩa là "căng thẳng nguyên") như là điều kiện môi trường hay một kích thích tố (stimulus). [1]

Ở sinh viên, hầu hết ai cũng sẽ bị căng thẳng tùy thuộc vào mức độ nặng hoặc nhẹ. Sự căng thẳng của sinh viên thường do một số nguyên nhân như áp lực thi cử, áp lực học tập, mâu thuẫn trong gia đình hoặc với bạn bè, vấn đề sức khỏe, …

Từ dữ liệu nghiên cứu về mức độ căng thẳng của sinh viên, nhóm em sẽ tiến hành phân tích thành phần chính (PCA) và phân tích nhân tố.

- **Nguồn dữ liệu:** *Student Stress Factors: A Comprehensive analysis*. (2023, October 14). Kaggle. https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis

2. Dự đoán hóa đơn giả

Theo khoản 8 Điều 3 Nghị định 123/2020/NĐ-CP, hóa đơn giả là hóa đơn được in hoặc khởi tạo theo mẫu hóa đơn đã được thông báo phát hành của tổ chức, cá nhân khác hoặc in, khởi tạo trùng số của cùng một ký hiệu hóa đơn hoặc làm giả hóa đơn điện tử.

Mặt khác, việc sử dụng hóa đơn giả là hành vi sử dụng hóa đơn không hợp pháp theo quy định tại điểm a Điều 4 Nghị định 125/2020/NĐ-CP. [2]

Vì vậy, việc phân biệt hóa đơn giả cũng rất quan trọng. Do đó, nhóm em đã chọn dữ liệu về dự đoán hóa đơn giả để phân tích.

- **Nguồn dữ liệu:** *Fake bills*. (2023, February 18). Kaggle. https://www.kaggle.com/datasets/alexandrepetit881234/fake-bills

NỘI DUNG

A. PHẦN 1

1. PHÂN TÍCH DỮ LIỆU NGHIÊN CỨU VỀ MỨC ĐỘ CĂNG THẲNG (STRESS) CỦA SINH VIÊN.

1.1. Phân tích thành phần chính

❖ Thư viện

```
library(janitor)
library(tidyverse)
library(FactoMineR)
library(ggplot2)
library(ggbiplot)
library(ggfortify)
library(dplyr)
library(psych)
library(GPArotation)
library(factoextra)    #fviz
library(base)
```

❖ Dữ liệu

```
stress = read.csv("D:/TKNC/Datasets/Datasets/StressLevelDataset.csv")
head(stress)
```

```
##   anxiety_level self_esteem mental_health_history depression headache
## 1            14          20                      0         11        2
## 2            15           8                      1         15        5
## 3            12          18                      1         14        2
## 4            16          12                      1         15        4
## 5            16          28                      0          7        2
## 6            20          13                      1         21        3
##   blood_pressure sleep_quality breathing_problem noise_level living_con
ditions
## 1              1             2                 4           2
3
## 2              3             1                 4           3
1
## 3              1             2                 2           2
2
## 4              3             1                 3           4
```

```
2
## 5              3             5                 1           3
2
## 6              3             1                 4           3
2
##   safety basic_needs academic_performance study_load
## 1      3           2                    3          2
## 2      2           2                    1          4
## 3      3           2                    2          3
## 4      2           2                    2          4
## 5      4           3                    4          3
## 6      2           1                    2          5
##   teacher_student_relationship future_career_concerns social_support
## 1                            3                       3              2
## 2                            1                       5              1
## 3                            3                       2              2
## 4                            1                       4              1
## 5                            1                       2              1
## 6                            2                       5              1
##   peer_pressure extracurricular_activities bullying stress_level
## 1             3                          3        2            1
## 2             4                          5        5            2
## 3             3                          2        2            1
## 4             4                          4        5            2
## 5             5                          0        5            1
## 6             4                          4        5            2
```

```
names(stress)
```

```
##  [1] "anxiety_level"                "self_esteem"
##  [3] "mental_health_history"        "depression"
##  [5] "headache"                     "blood_pressure"
##  [7] "sleep_quality"                "breathing_problem"
##  [9] "noise_level"                  "living_conditions"
## [11] "safety"                       "basic_needs"
## [13] "academic_performance"         "study_load"
## [15] "teacher_student_relationship" "future_career_concerns"
## [17] "social_support"               "peer_pressure"
## [19] "extracurricular_activities"   "bullying"
## [21] "stress_level"
```

```
str(stress)
```

```
## 'data.frame':    1100 obs. of  21 variables:
##  $ anxiety_level               : int  14 15 12 16 16 20 4 17 13 6 ...
##  $ self_esteem                 : int  20 8 18 12 28 13 26 3 22 8 ...
##  $ mental_health_history       : int  0 1 1 1 0 1 0 1 1 0 ...
##  $ depression                  : int  11 15 14 15 7 21 6 22 12 27 ...
##  $ headache                    : int  2 5 2 4 2 3 1 4 3 4 ...
##  $ blood_pressure              : int  1 3 1 3 3 3 2 3 1 3 ...
##  $ sleep_quality               : int  2 1 2 1 5 1 4 1 2 1 ...
##  $ breathing_problem           : int  4 4 2 3 1 4 1 5 4 2 ...
```

```
##  $ noise_level                 : int  2 3 2 4 3 3 1 3 3 0 ...
##  $ living_conditions           : int  3 1 2 2 2 2 4 1 3 5 ...
##  $ safety                      : int  3 2 3 2 4 2 4 1 3 2 ...
##  $ basic_needs                 : int  2 2 2 2 3 1 4 1 3 2 ...
##  $ academic_performance        : int  3 1 2 2 4 2 5 1 3 2 ...
##  $ study_load                  : int  2 4 3 4 3 5 1 3 3 2 ...
##  $ teacher_student_relationship: int  3 1 3 1 1 2 4 2 2 1 ...
##  $ future_career_concerns      : int  3 5 2 4 2 5 1 4 3 5 ...
##  $ social_support              : int  2 1 2 1 1 1 3 1 3 1 ...
##  $ peer_pressure               : int  3 4 3 4 5 4 2 4 3 5 ...
##  $ extracurricular_activities  : int  3 5 2 4 0 4 2 4 2 3 ...
##  $ bullying                    : int  2 5 2 5 5 5 1 5 2 4 ...
##  $ stress_level                : int  1 2 1 2 1 2 0 2 1 1 ...
```

```
dim(stress)
```

```
## [1] 1100   21
```

❖ Mô tả dữ liệu

Dữ liệu về nghiên cứu mức độ căng thẳng (stress) của sinh viên gồm 1100 quan trắc và 21 biến:

1. `anxiety_level`: mức độ lo lắng
2. `self_esteem`: lòng tự trọng
3. `mental_health_history`: tiền sử sức khỏe tâm thần
4. `depression`: trầm cảm
5. `headache`: đau đầu
6. `blood_pressure`: huyết áp
7. `sleep_quality`: chất lượng giấc ngủ
8. `breathing_problem`: vấn đề hô hấp
9. `noise_level`: mức độ ồn
10. `living_conditions`: điều kiện sống
11. `safety`: sự an toàn
12. `basic_needs`: nhu cầu cơ bản
13. `academic_performance`: thành tích học tập
14. `study_load`: tải học tập

15. `teacher_student_relationship`: quan hệ giữa giáo viên và học sinh
16. `future_career_concerns`: quan ngại về nghề nghiệp tương lai
17. `social_support`: hỗ trợ xã hội
18. `peer_pressure`: áp lực ngang hàng
19. `extracurricular_activities`: các hoạt động ngoại khóa
20. `bullying`: bắt nạt
21. `stress_level`: mức độ căng thẳng

Trong đó, các biến có thể chia thành năm nhóm:

- Nhóm "yếu tố tâm lý" gồm: `anxiety-level`, `self-esteem`, `mental-health-history` và `depression`.
- Nhóm "yếu tố sinh lý" gồm: `headache`, `blood-pressure`, `sleep-quality` và `breathing-problem`.
- Nhóm "yếu tố môi trường" gồm: `noise-level`, `living-conditions`, `safety` và `basic-needs`.
- Nhóm "yếu tố học tập" gồm: `academic-performance`, `study-load`, `teacher-student-relationship` và `future-career-concerns`.
- Nhóm "yếu tố xã hội" gồm: `social-support`, `peer-pressure`, `extracurricular-activities` và `bullying`.

```
apply(stress,2,mean)
```

```
##                anxiety_level                   self_esteem
##                   11.0636364                    17.7772727
##        mental_health_history                    depression
##                    0.4927273                    12.5554545
##                     headache                blood_pressure
##                    2.5081818                     2.1818182
##                sleep_quality            breathing_problem
##                    2.6600000                     2.7536364
##                  noise_level            living_conditions
##                    2.6490909                     2.5181818
##                       safety                   basic_needs
##                    2.7372727                     2.7727273
##         academic_performance                    study_load
##                    2.7727273                     2.6218182
## teacher_student_relationship       future_career_concerns
```

```
##                    2.6481818                    2.6490909
##               social_support               peer_pressure
##                    1.8818182                    2.7345455
##   extracurricular_activities                    bullying
##                    2.7672727                    2.6172727
##                  stress_level
##                    0.9963636
```

- Tập dữ liệu gồm 1100 sinh viên. Trong đó, mức độ lo lắng trung bình là 11.0636364.

❖ Kiểm tra giá trị khuyết

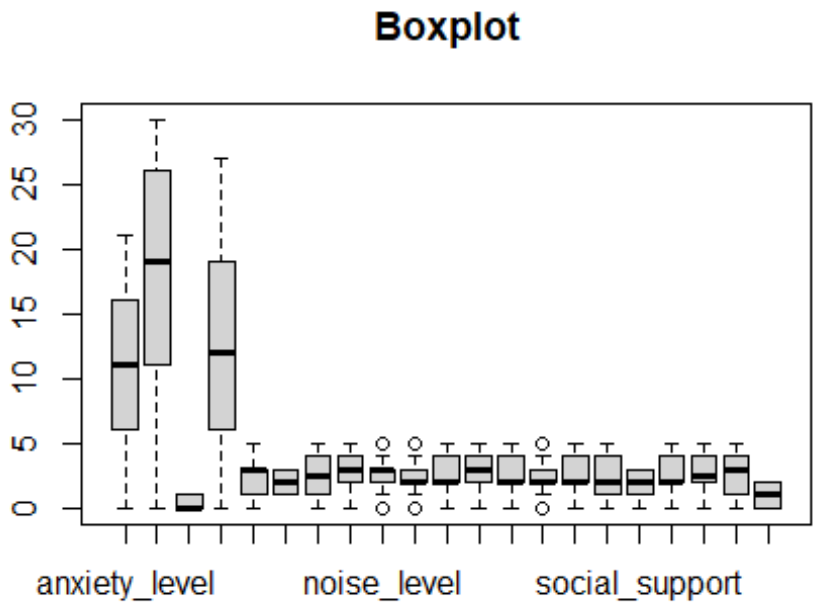
```
any(is.na(stress))
```

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

Nhận xét: Dữ liệu không có giá trị khuyết.

❖ Kiểm tra giá trị ngoại lai

```
#Boxplot
boxplot(stress, main = "Boxplot")
```



Nhận xét: Nhìn vào boxplot, ta thấy giá trị ngoại lai của ba biến `noise_level`, `living_conditions` và `study_load` đều là các giá trị 0 và 5, 0 và 5 đều có chênh lệch không đáng kể so với trung bình của ba biến. Do đó, ta có thể giữ lại các giá trị ngoại lai này.

Tiến hành phân tích dữ liệu bằng phương pháp phân tích thành phần chính.

- Biến `stress_level` được dùng để phân loại.

❖ Phương sai

```
dat_stress = stress[, -21]
apply(dat_stress, 2, var)
```

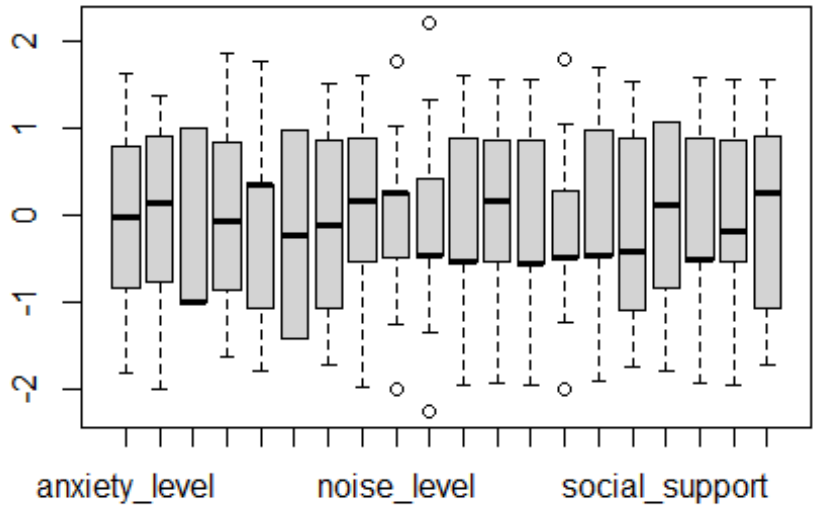
```
##                anxiety_level                  self_esteem
##                   37.4245182                   80.0058524
##        mental_health_history                   depression
##                    0.2501745                   59.7066581
##                     headache               blood_pressure
##                    1.9862842                    0.6948466
##                sleep_quality            breathing_problem
##                    2.3974886                    1.9619977
##                  noise_level            living_conditions
##                    1.7639209                    1.2526264
##                       safety                  basic_needs
##                    1.9773174                    2.0556704
##         academic_performance                   study_load
##                    2.0010754                    1.7312797
## teacher_student_relationship       future_career_concerns
##                    1.9170577                    2.3389892
##               social_support                peer_pressure
##                    1.0979403                    2.0313806
##   extracurricular_activities                     bullying
##                    2.0094830                    2.3438324
```

Nhận xét: Phương sai của ba biến `anxiety_level`, `self_esteem` và `depression` lần lượt là 37.4245182, 80.0058524 và 59.7066581, lớn hơn rất nhiều so với các phương sai của các biến còn lại. Do đó, cần phải chuẩn hóa dữ liệu trước khi phân tích thành phần chính.

❖ Chuẩn hóa dữ liệu

```
sc_stress = as.data.frame(scale(dat_stress, scale = T))
```

```r
#Vẽ lại boxplot
boxplot(sc_stress)
```



❖ Chọn các thành phần chính

```r
pca_stress = princomp(sc_stress, cor = T)
summary(pca_stress)
```

```
## Importance of components:
##                           Comp.1     Comp.2     Comp.3    Comp.4     Co
mp.5
## Standard deviation     3.4512862 1.09368465 0.83263645 0.7706050 0.7466
0020
## Proportion of Variance 0.5955688 0.05980731 0.03466417 0.0296916 0.0278
7059
## Cumulative Proportion  0.5955688 0.65537614 0.69004031 0.7197319 0.7476
0250
##                            Comp.6     Comp.7     Comp.8     Comp.9    C
omp.10
## Standard deviation     0.72515168 0.68857366 0.67428673 0.63683808 0.61
946764
## Proportion of Variance 0.02629225 0.02370668 0.02273313 0.02027814 0.01
918701
## Cumulative Proportion  0.77389475 0.79760144 0.82033457 0.84061270 0.85
979971
##                           Comp.11    Comp.12    Comp.13    Comp.14    C
omp.15
## Standard deviation     0.60089533 0.58668016 0.56966154 0.55958168 0.55
```

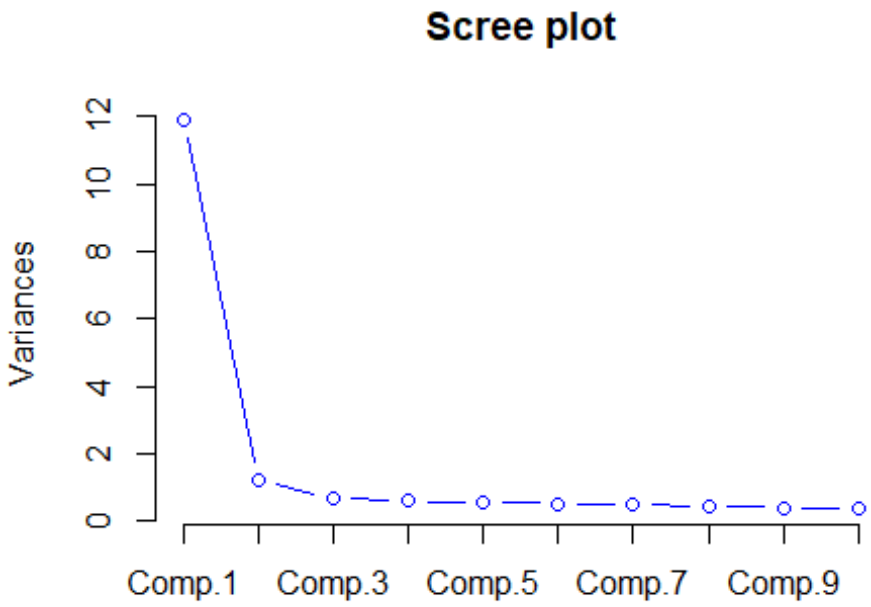
```
650319
## Proportion of Variance 0.01805376 0.01720968 0.01622571 0.01565658 0.01
548479
## Cumulative Proportion  0.87785347 0.89506315 0.91128886 0.92694545 0.94
243024
##                           Comp.16    Comp.17   Comp.18    Comp.19     C
omp.20
## Standard deviation     0.52865967 0.52270278 0.5158275 0.47937314 0.320
654590
## Proportion of Variance 0.01397405 0.01366091 0.0133039 0.01148993 0.005
140968
## Cumulative Proportion  0.95640429 0.97006520 0.9833691 0.99485903 1.000
000000
```

```
pca_stress$sdev^2
```

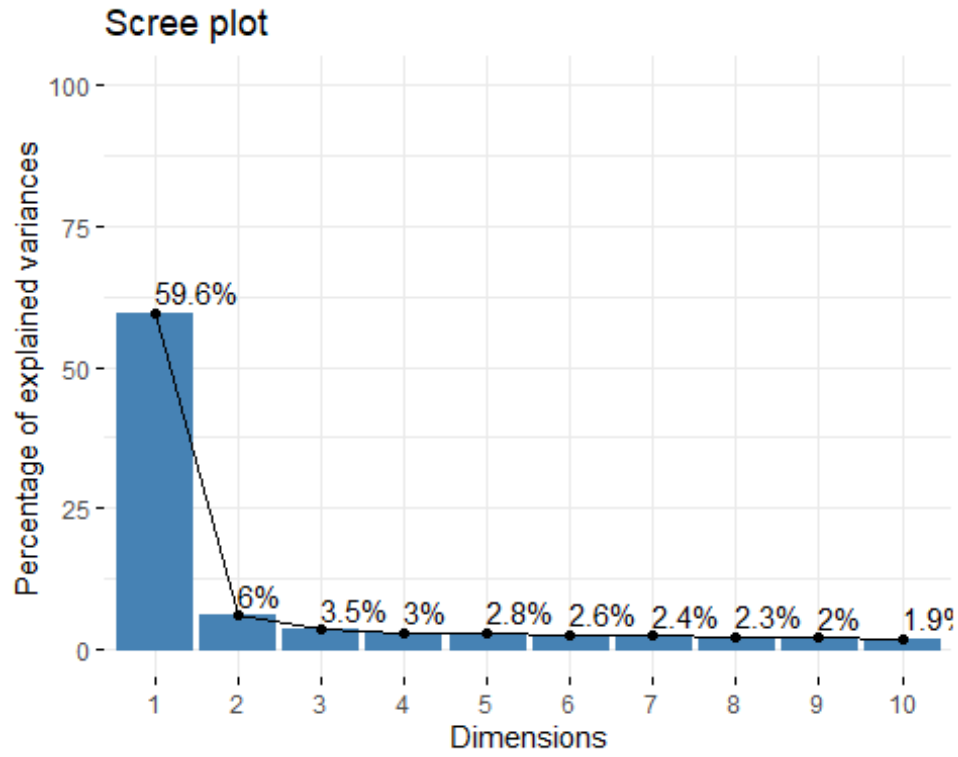
```
##     Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     C
omp.7
## 11.9113767  1.1961461  0.6932835  0.5938320  0.5574119  0.5258450  0.47
41337
##     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12    Comp.13    Co
mp.14
##  0.4546626  0.4055627  0.3837402  0.3610752  0.3441936  0.3245143  0.31
31317
##    Comp.15    Comp.16    Comp.17    Comp.18    Comp.19    Comp.20
##  0.3096958  0.2794810  0.2732182  0.2660780  0.2297986  0.1028194
```

Nhận xét: Theo tiêu chuẩn Kaiser, ta sẽ giữ lại hai thành phần chính đầu tiên do chúng có phương sai lớn hơn 1.

```
#screeplot
screeplot(pca_stress,type="lines",col="blue", main="Scree plot")
```




```
fviz_eig(pca_stress, addlabels = T, ylim = c(0,100))
```



Nhận xét: Qua đồ thị Screeplot, ta thấy sự thay đổi rõ ràng nhất về độ dốc ở thành phần chính thứ hai (PC2), thành phần chính thứ nhất (PC1) đóng góp 59.6% vào phương sai

suy rộng. Bên cạnh đó, PC2 đóng góp 6% vào phương sai suy rộng, ta có thể phân tích thêm thành phần chính này. Cả hai thành phần chính PC1 và PC2 giải thích được 65.6% phương sai.

❖ Loadings

```
loading = pca_stress$loadings
loading
```

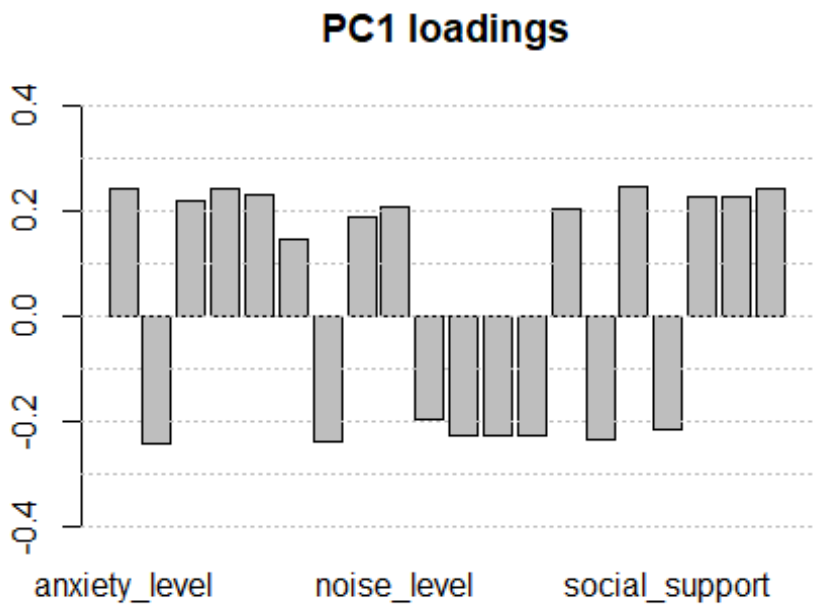
```
##
## Loadings:
##                               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Comp.7
## anxiety_level                  0.244  0.111
## self_esteem                   -0.242  0.150
## mental_health_history          0.219  0.118        -0.328 -0.209
## depression                     0.243                             -0.183
0.102
## headache                       0.229               -0.268 -0.158
0.161
## blood_pressure                 0.145 -0.744  0.131  0.147
## sleep_quality                 -0.240 -0.135
-0.234
## breathing_problem              0.190  0.300  0.115  0.627  0.345 -0.212
-0.205
## noise_level                    0.209         0.274         0.270  0.710
-0.275
## living_conditions             -0.199         0.197 -0.362  0.782 -0.224
0.230
## safety                        -0.229         0.346 -0.129
## basic_needs                   -0.227 -0.106  0.246  0.245 -0.218 -0.208
0.295
## academic_performance          -0.230 -0.136  0.263  0.150 -0.134  0.187
-0.139
## study_load                     0.204         0.407 -0.297 -0.137 -0.450
-0.600
## teacher_student_relationship  -0.237         0.342
## future_career_concerns         0.247
0.172
## social_support                -0.218  0.481  0.313        -0.127
0.116
## peer_pressure                  0.226         0.330                0.232
0.306
## extracurricular_activities     0.226         0.300  0.253
0.348
## bullying                       0.243
##                               Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp
.13
## anxiety_level                         0.141          0.276           0.1
```

```
93
## self_esteem                         -0.123  0.206          -0.286  -0.5
33
## mental_health_history        -0.515  0.635                          -0.2
61
## depression                                -0.194   0.434  -0.120   0.1
37
## headache                     -0.199 -0.515 -0.523                  -0.2
31
## blood_pressure               -0.165
## sleep_quality                -0.182        -0.130  -0.340   0.179  -0.1
47
## breathing_problem            -0.427 -0.107 -0.131          -0.124
## noise_level                   0.260  0.166 -0.238          -0.151
## living_conditions                                          0.168  -0.1
28
## safety                       -0.320                        -0.606   0.4
20
## basic_needs                   0.280  0.135 -0.131   0.233  -0.201  -0.2
16
## academic_performance         -0.172                 0.536   0.264  -0.2
94
## study_load                    0.242 -0.133                         -0.1
22
## teacher_student_relationship                       -0.185   0.442   0.3
07
## future_career_concerns                                      0.328   0.1
31
## social_support                                                      0.1
41
## peer_pressure                -0.122 -0.389  0.465
## extracurricular_activities    0.274  0.204 -0.258  -0.380          -0.1
67
## bullying                                  0.455  -0.243
##                            Comp.14 Comp.15 Comp.16 Comp.17 Comp.18 Co
mp.19
## anxiety_level                 0.479   0.124   0.214   0.236   0.643
## self_esteem                   0.290          -0.177  -0.195   0.229  -0
.510
## mental_health_history        -0.105  -0.138
## depression                   -0.126   0.193  -0.336  -0.664   0.129
## headache                      0.277  -0.163           0.125           0
.235
## blood_pressure                       -0.144                         -0
.119
## sleep_quality                         0.384   0.400  -0.422   0.250   0
.297
## breathing_problem                    -0.126
## noise_level                                   0.100  -0.124  -0.130
## living_conditions
## safety                        0.105   0.249           0.148  -0.103
```

```
## basic_needs                          -0.383   0.400  -0.119  -0.101   0
.154
## academic_performance                  0.308  -0.234   0.256  -0.156   0
.168
## study_load
## teacher_student_relationship  0.208  -0.426  -0.328  -0.143   0.194  -0
.140
## future_career_concerns        0.326   0.261   0.368  -0.122  -0.430  -0
.489
## social_support                                                -0.119
## peer_pressure                -0.417  -0.106   0.218           0.202
## extracurricular_activities   -0.170   0.371  -0.242   0.240   0.110
## bullying                      0.436          -0.235  -0.204  -0.316   0
.487
##                              Comp.20
## anxiety_level
## self_esteem
## mental_health_history
## depression
## headache
## blood_pressure                0.545
## sleep_quality
## breathing_problem
## noise_level
## living_conditions
## safety                       -0.189
## basic_needs                  -0.136
## academic_performance
## study_load
## teacher_student_relationship -0.247
## future_career_concerns
## social_support                0.726
## peer_pressure                -0.152
## extracurricular_activities
## bullying
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
Comp.9
## SS loadings      1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
1.00
## Proportion Var   0.05   0.05   0.05   0.05   0.05   0.05   0.05   0.05
0.05
## Cumulative Var   0.05   0.10   0.15   0.20   0.25   0.30   0.35   0.40
0.45
##                Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16
Comp.17
## SS loadings       1.00    1.00    1.00    1.00    1.00    1.00    1.00
1.00
## Proportion Var    0.05    0.05    0.05    0.05    0.05    0.05    0.05
0.05
## Cumulative Var    0.50    0.55    0.60    0.65    0.70    0.75    0.80
```

```
0.85
##                Comp.18 Comp.19 Comp.20
## SS loadings       1.00    1.00    1.00
## Proportion Var    0.05    0.05    0.05
## Cumulative Var    0.90    0.95    1.00

barplot(loading[,1],main="PC1 loadings", ylim = c(-0.4,0.4))
abline(h = seq(-0.4,0.4, by = 0.1), col = "gray",lty = "dotted")
```



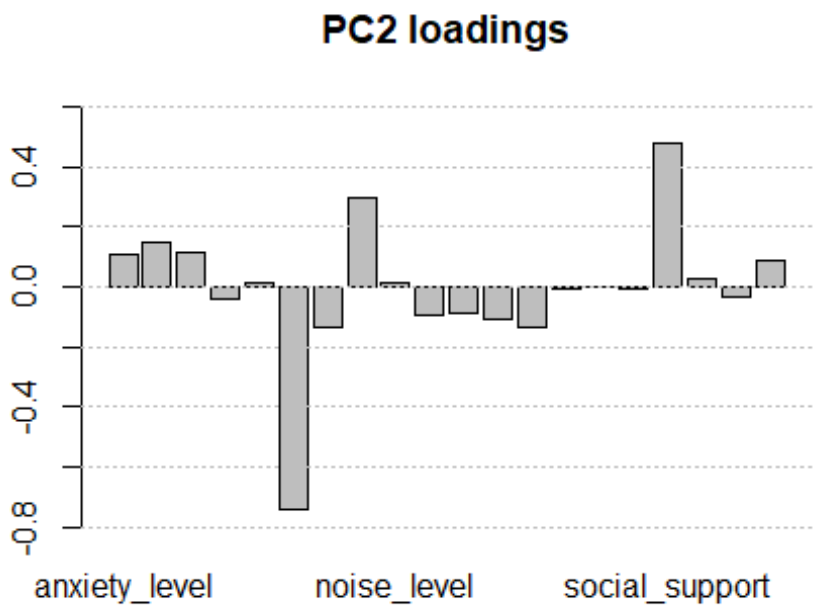
Nhận xét: Ở thành phần chính thứ nhất, các biến có độ lớn trọng số tương đối đồng đều nhau, chúng đều có vai trò trong PC1.

- Ở nhóm "yếu tố tâm lý" có 3 biến có trọng số dương là `anxiety_level`, `mental_health_history`, `depression` và 1 biến có trọng số âm là `self_esteem`, nghĩa là chúng đối nghịch nhau. Khi một sinh viên lo lắng quá nhiều, có tiền sử sức khỏe tâm thần và trầm cảm thì họ sẽ thiếu tự tin, ý thức kém về giá trị bản thân mình, lòng tự trọng sẽ thấp dần, và ngược lại.
- Ở nhóm "yếu tố sinh lý" có 3 biến có trọng số dương là `headache`, `blood_pressure`, `breathing_problem` và 1 biến có trọng số âm là `sleep_quality`. Khi sinh viên gặp phải các vấn đề về đau đầu, huyết áp, hô hấp thì sẽ bị ảnh hưởng đến giấc ngủ, chất lượng giấc ngủ sẽ kém, và ngược lại.

- Ở nhóm "yếu tố môi trường" có 1 biến có trọng số dương là `noise_level` và 3 biến có trọng số âm là `living_conditions`, `safety`, `basic_needs`. Khi mức độ ồn quá lớn sẽ kéo theo điều kiện sống, sự an toàn và các nhu cầu cơ bản thấp dần, và ngược lại.
- Ở nhóm "yếu tố học tập" có 2 biến có trọng số dương là `study_load`, `future_career_concerns` và 2 biến có trọng số âm là `academic_performance`, `teacher_student_relationship`. Khi mối quan hệ giữa giáo viên và sinh viên kém cũng như thành tích học tập không tốt sẽ dẫn đến quá tải học tập và ảnh hưởng đến quan ngại về nghề nghiệp trong tương lai, và ngược lại.
- Ở nhóm "yếu tố xã hội" có 3 biến có trọng số dương là `peer_pressure`, `extracurricular_activities`, `bullying` và 1 biến có trọng số âm là `social_support`. Khi không nhận được hỗ trợ xã hội, sinh viên sẽ cảm thấy áp lực và có thể dẫn đến bị bắt nạt.

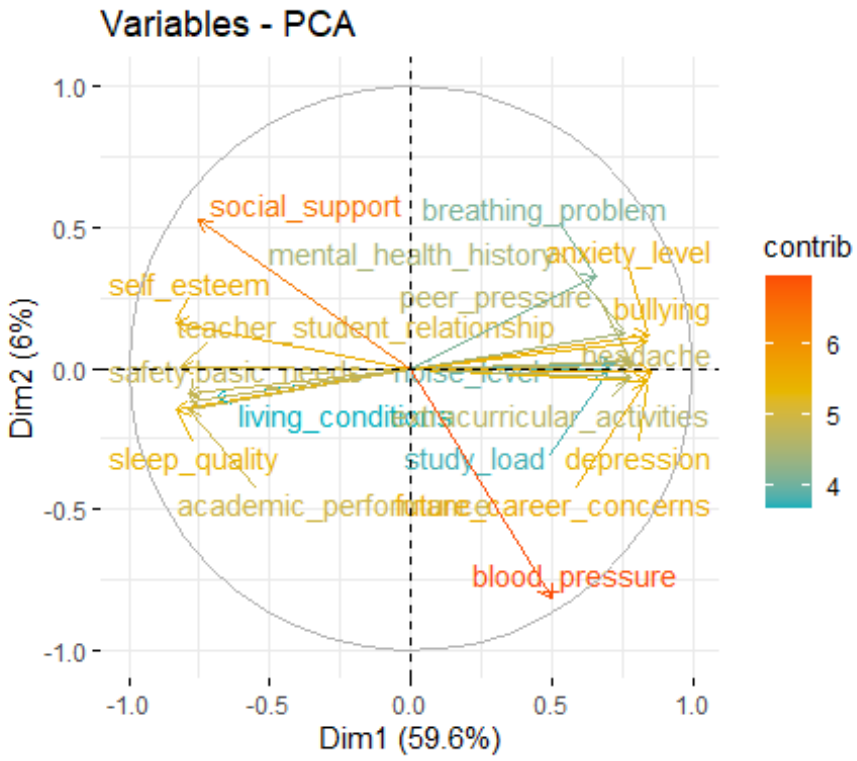
Tóm lại, PC1 có các biến có trọng số dương khá nhiều, nếu giá trị của chúng càng lớn thì mức độ căng thẳng càng cao, ảnh hưởng không tốt cho sinh viên.

```
barplot(loading[,2],main="PC2 loadings", ylim = c(-0.8,0.6))
abline(h = seq(-0.8,0.6, by = 0.2), col = "gray",lty = "dotted")
```



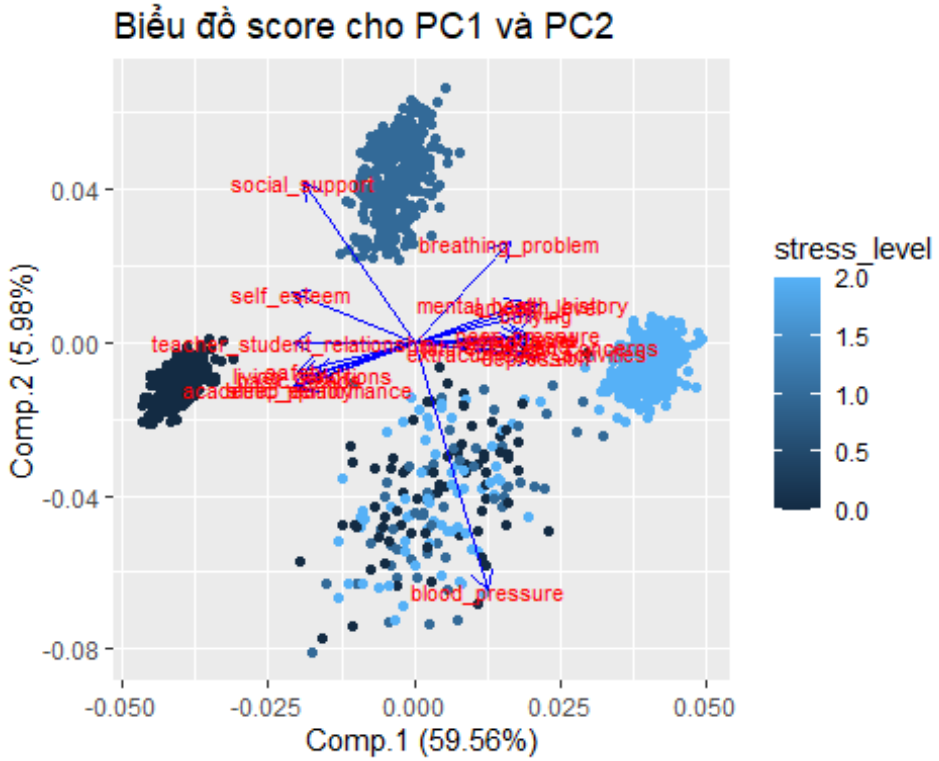
Nhận xét: Ở thành phần chính thứ hai, có 3 biến có độ lớn trọng số khá lớn là `blood_pressure` (-0.744), `breathing_problem` (0.300) và `social_support` (0.481), chúng có đóng góp vai trò trong PC2. PC2 giải thích về sức khỏe và xã hội.

```
fviz_pca_var(pca_stress,
             axes = c(1,2),
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
```



❖ Scores

```
autoplot(pca_stress, loadings = TRUE, loadings.colour = 'blue', loadings.l
abel = TRUE, loadings.label.size = 3,
         colour = "stress_level", data = stress, main="Biểu đồ score cho P
C1 và PC2")
```

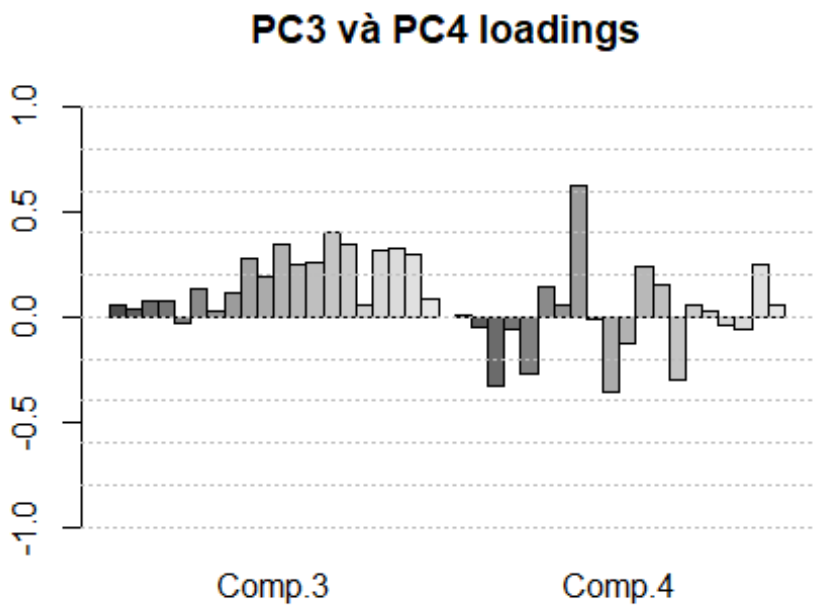


Nhận xét: Nhìn vào biểu đồ, ta thấy các điểm bên phải có màu xanh nhạt là mức độ căng thẳng ở mức 2, có thể coi là mức độ nặng (giá trị PC1 lớn). Các điểm bên trái có màu xanh đậm là mức độ căng thẳng ở mức 0, có thể coi là không bị căng thẳng (giá trị PC1 nhỏ). Các điểm phía bên trên có màu xanh trung bình là mức độ căng thẳng ở mức 1, có thể coi là nhẹ hoặc trung bình (giá trị PC1 xấp xỉ 0). Với PC2, ta có thể thấy yếu tố xã hội có thể gây ra căng thẳng ở sinh viên ở mức độ vừa phải để tạo nên động lực có sinh viên cố gắng học tập và làm việc.

❖ Phân tích thêm về một số thành phần chính khác

Hai thành phần chính đầu tiên chỉ đóng góp 65.6% vào phương sai suy rộng, tám thành phần chính đầu tiên mới đóng góp được 82%. Do đó, ta phân tích thêm về sáu thành phần chính tiếp theo.

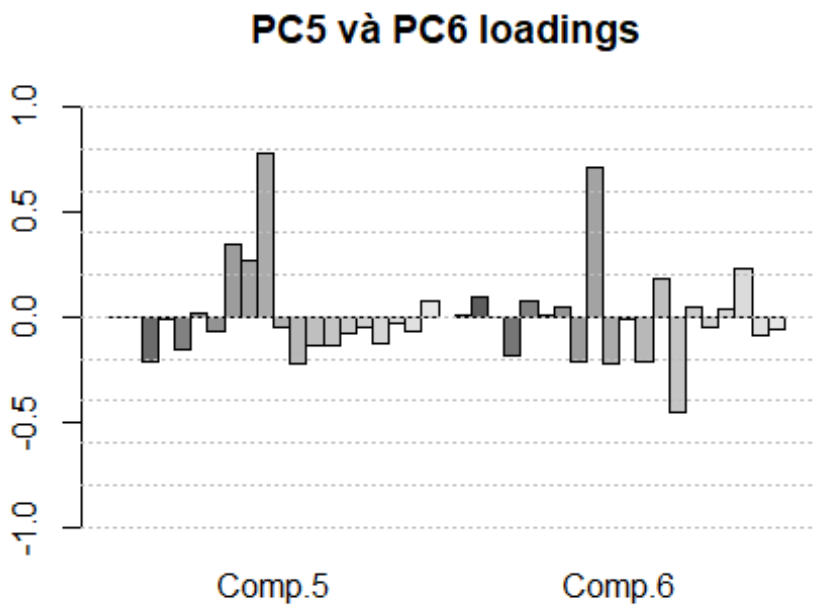
```
barplot(loading[,3:4],beside = T, ylim = c(-1,1), main = "PC3 và PC4 loadi
ngs")
abline(h = seq(-1,1, by = 0.2), col = "gray",lty = "dotted")
```



Nhận xét:

- Thành phần chính thứ ba (PC3) có 9 biến có trọng số lớn là `noise_level` (0.274), `safety` (0.346), `basic_needs` (0.246), `academic_performance` (0.263), `study_load` (0.407), `teacher_student_relationship` (0.342), `social_support` (0.313), `peer_pressure` (0.330), `extracurricular_activities` (0.300). PC3 giải thích chủ yếu về các yếu tố môi trường, học tập và xã hội.
- Thành phần chính thứ tư (PC4) có biến `breathing_problem` có trọng số cao nhất (0.627), giải thích chủ yếu về vấn đề sức khỏe hô hấp.

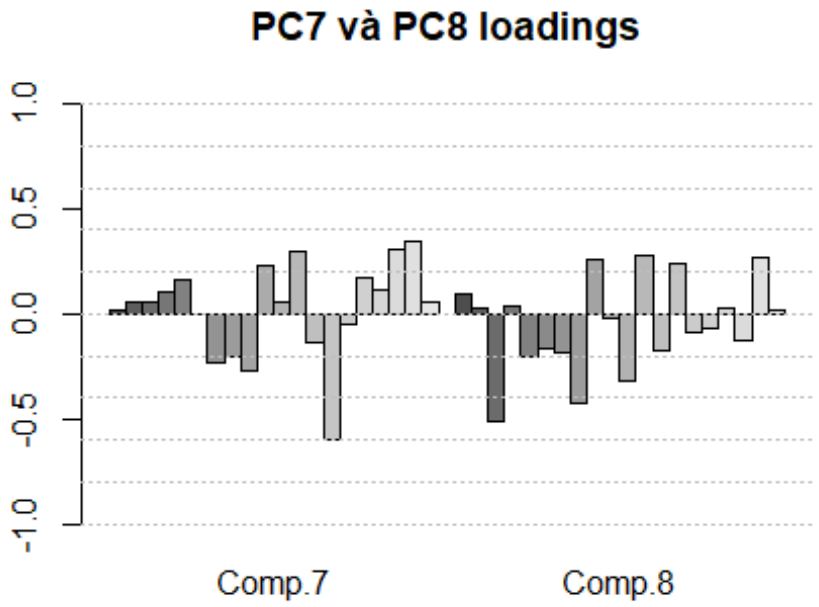
```
barplot(loading[,5:6],beside = T, ylim = c(-1,1), main = "PC5 và PC6 loadi
ngs")
abline(h = seq(-1,1, by = 0.2), col = "gray",lty = "dotted")
```



Nhận xét:

- Thành phần chính thứ năm (PC5) có biến `living_conditions` có trọng số cao nhất (0.782), giải thích chủ yếu về vấn đề điều kiện sống.
- Thành phần chính thứ sáu (PC6) có biến `noise_level` có trọng số cao nhất (0.710), giải thích chủ yếu về mức độ ồn.

```
barplot(loading[,7:8],beside = T, ylim = c(-1,1), main = "PC7 và PC8 loadi
ngs")
abline(h = seq(-1,1, by = 0.2), col = "gray",lty = "dotted")
```



Nhận xét:

- Thành phần chính thứ bảy (PC7) có biến `study_load` có trọng số cao nhất (0.600), giải thích chủ yếu về vấn đề tải học tập.
- Thành phần chính thứ tám (PC8) có biến `mental_health_history` có trọng số cao nhất (-0.515), giải thích chủ yếu về vấn đề sức khỏe tâm thần.

1.2. Phân tích nhân tố

❖ Ma trận hệ số tương quan

```
(R = cor(stress))
```

```
##                              anxiety_level self_esteem mental_health_hi
story
## anxiety_level                    1.0000000  -0.6727453             0.63
44496
## self_esteem                     -0.6727453   1.0000000            -0.60
35019
## mental_health_history            0.6344496  -0.6035019             1.00
00000
## depression                       0.6943404  -0.6996018             0.61
58823
```

```
## headache                          0.6327376  -0.6260585            0.60
48258
## blood_pressure                    0.3308669  -0.5146920            0.29
56173
## sleep_quality                    -0.7102923   0.6626928           -0.61
41460
## breathing_problem                 0.5616538  -0.5105143            0.46
43473
## noise_level                       0.6086243  -0.5711687            0.51
52899
## living_conditions                -0.5684344   0.5505350           -0.50
85252
## safety                           -0.6512198   0.6439807           -0.54
67313
## basic_needs                      -0.6370785   0.6312116           -0.60
11955
## academic_performance             -0.6496011   0.6390452           -0.58
61931
## study_load                        0.5860642  -0.5751117            0.53
22665
## teacher_student_relationship     -0.6631765   0.6529343           -0.58
77277
## future_career_concerns            0.7170156  -0.7125195            0.62
59094
## social_support                   -0.5697479   0.6792110           -0.48
25596
## peer_pressure                     0.6429101  -0.6071179            0.58
06019
## extracurricular_activities        0.6410219  -0.6412018            0.55
45762
## bullying                          0.7099815  -0.6407371            0.62
43658
## stress_level                      0.7367954  -0.7561951            0.64
86440
##                             depression   headache blood_pressure sleep
_quality
## anxiety_level                0.6943404  0.6327376      0.3308669    -0
.7102923
## self_esteem                 -0.6996018 -0.6260585     -0.5146920     0
.6626928
## mental_health_history        0.6158823  0.6048258      0.2956173    -0
.6141460
## depression                   1.0000000  0.6576999      0.4360842    -0
.6931609
## headache                     0.6576999  1.0000000      0.3619863    -0
.6387709
## blood_pressure               0.4360842  0.3619863      1.0000000    -0
.3003234
## sleep_quality               -0.6931609 -0.6387709     -0.3003234     1
.0000000
## breathing_problem            0.5225402  0.4617188      0.1623084    -0
```

```
.5416866
## noise_level                    0.5662495  0.5435573      0.3527440    -0
.5766455
## living_conditions             -0.5303507 -0.5328251     -0.2746855     0
.5354615
## safety                        -0.6258572 -0.5891365     -0.2883536     0
.6576863
## basic_needs                   -0.6087761 -0.6231989     -0.2805900     0
.6209547
## academic_performance          -0.6331743 -0.6220585     -0.2627854     0
.6713263
## study_load                     0.6024984  0.5428897      0.3489643    -0
.5517751
## teacher_student_relationship -0.6738530 -0.6259282     -0.3521230     0
.6775691
## future_career_concerns         0.7065606  0.6793072      0.4340874    -0
.6821298
## social_support                -0.6179720 -0.5729879     -0.7525310     0
.5545531
## peer_pressure                  0.6355437  0.6225807      0.4013921    -0
.6490981
## extracurricular_activities     0.6485506  0.5825619      0.4262545    -0
.6230923
## bullying                       0.6657899  0.6097755      0.3704402    -0
.6994272
## stress_level                   0.7343786  0.7134840      0.3941999    -0
.7490679
##                              breathing_problem noise_level living_condi
tions
## anxiety_level                        0.5616538   0.6086243        -0.56
84344
## self_esteem                         -0.5105143  -0.5711687         0.55
05350
## mental_health_history                0.4643473   0.5152899        -0.50
85252
## depression                           0.5225402   0.5662495        -0.53
03507
## headache                             0.4617188   0.5435573        -0.53
28251
## blood_pressure                       0.1623084   0.3527440        -0.27
46855
## sleep_quality                       -0.5416866  -0.5766455         0.53
54615
## breathing_problem                    1.0000000   0.4592345        -0.44
89970
## noise_level                          0.4592345   1.0000000        -0.45
23616
## living_conditions                   -0.4489970  -0.4523616         1.00
00000
## safety                              -0.5193476  -0.5366296         0.56
35710
```

```
## basic_needs                          -0.5081721  -0.5723266         0.50
32750
## academic_performance                 -0.5072509  -0.5137298         0.50
72206
## study_load                            0.4287910   0.4936254        -0.43
77325
## teacher_student_relationship         -0.4988945  -0.5387583         0.54
93316
## future_career_concerns                0.5453451   0.5754391        -0.56
50713
## social_support                       -0.3651734  -0.4920940         0.46
65942
## peer_pressure                         0.4927288   0.5838173        -0.50
17946
## extracurricular_activities            0.5168844   0.5636136        -0.51
57936
## bullying                              0.5763408   0.5854583        -0.55
11387
## stress_level                          0.5739837   0.6633713        -0.58
17231
##                                   safety basic_needs academic_performanc
e
## anxiety_level                -0.6512198  -0.6370785           -0.649601
1
## self_esteem                   0.6439807   0.6312116            0.639045
2
## mental_health_history        -0.5467313  -0.6011955           -0.586193
1
## depression                   -0.6258572  -0.6087761           -0.633174
3
## headache                     -0.5891365  -0.6231989           -0.622058
5
## blood_pressure               -0.2883536  -0.2805900           -0.262785
4
## sleep_quality                 0.6576863   0.6209547            0.671326
3
## breathing_problem            -0.5193476  -0.5081721           -0.507250
9
## noise_level                  -0.5366296  -0.5723266           -0.513729
8
## living_conditions             0.5635710   0.5032750            0.507220
6
## safety                        1.0000000   0.6247745            0.642846
2
## basic_needs                   0.6247745   1.0000000            0.639387
5
## academic_performance          0.6428462   0.6393875            1.000000
0
## study_load                   -0.4939029  -0.5134595           -0.520416
6
## teacher_student_relationship  0.6633279   0.6495188            0.669469
```

```
3
## future_career_concerns       -0.6581057  -0.6393479           -0.643804
9
## social_support                0.6149881   0.5841414            0.567501
2
## peer_pressure                -0.5569454  -0.5870365           -0.562947
7
## extracurricular_activities   -0.5803042  -0.5064258           -0.588612
2
## bullying                     -0.6456733  -0.6448862           -0.666228
7
## stress_level                 -0.7096016  -0.7089676           -0.720922
4
##                             study_load teacher_student_relationship
## anxiety_level                0.5860642                  -0.6631765
## self_esteem                 -0.5751117                   0.6529343
## mental_health_history        0.5322665                  -0.5877277
## depression                   0.6024984                  -0.6738530
## headache                     0.5428897                  -0.6259282
## blood_pressure               0.3489643                  -0.3521230
## sleep_quality               -0.5517751                   0.6775691
## breathing_problem            0.4287910                  -0.4988945
## noise_level                  0.4936254                  -0.5387583
## living_conditions           -0.4377325                   0.5493316
## safety                      -0.4939029                   0.6633279
## basic_needs                 -0.5134595                   0.6495188
## academic_performance        -0.5204166                   0.6694693
## study_load                   1.0000000                  -0.5141228
## teacher_student_relationship -0.5141228                  1.0000000
## future_career_concerns       0.5760782                  -0.6702550
## social_support              -0.4733121                   0.6812878
## peer_pressure                0.5441890                  -0.5877702
## extracurricular_activities   0.5435431                  -0.5823108
## bullying                     0.5866686                  -0.6559604
## stress_level                 0.6341555                  -0.6801627
##                             future_career_concerns social_support
## anxiety_level                            0.7170156     -0.5697479
## self_esteem                             -0.7125195      0.6792110
## mental_health_history                    0.6259094     -0.4825596
## depression                               0.7065606     -0.6179720
## headache                                 0.6793072     -0.5729879
## blood_pressure                           0.4340874     -0.7525310
## sleep_quality                           -0.6821298      0.5545531
## breathing_problem                        0.5453451     -0.3651734
## noise_level                              0.5754391     -0.4920940
## living_conditions                       -0.5650713      0.4665942
## safety                                  -0.6581057      0.6149881
## basic_needs                             -0.6393479      0.5841414
## academic_performance                    -0.6438049      0.5675012
## study_load                               0.5760782     -0.4733121
## teacher_student_relationship            -0.6702550      0.6812878
```

```
## future_career_concerns                  1.0000000     -0.6027916
## social_support                         -0.6027916      1.0000000
## peer_pressure                           0.6668725     -0.4901717
## extracurricular_activities              0.6665646     -0.5300474
## bullying                                0.7112781     -0.5670783
## stress_level                            0.7426186     -0.6324970
##                            peer_pressure extracurricular_activities
## anxiety_level                 0.6429101                  0.6410219
## self_esteem                  -0.6071179                 -0.6412018
## mental_health_history         0.5806019                  0.5545762
## depression                    0.6355437                  0.6485506
## headache                      0.6225807                  0.5825619
## blood_pressure                0.4013921                  0.4262545
## sleep_quality                -0.6490981                 -0.6230923
## breathing_problem             0.4927288                  0.5168844
## noise_level                   0.5838173                  0.5636136
## living_conditions            -0.5017946                 -0.5157936
## safety                       -0.5569454                 -0.5803042
## basic_needs                  -0.5870365                 -0.5064258
## academic_performance         -0.5629477                 -0.5886122
## study_load                    0.5441890                  0.5435431
## teacher_student_relationship -0.5877702                 -0.5823108
## future_career_concerns        0.6668725                  0.6665646
## social_support               -0.4901717                 -0.5300474
## peer_pressure                 1.0000000                  0.6183706
## extracurricular_activities    0.6183706                  1.0000000
## bullying                      0.6610577                  0.6519786
## stress_level                  0.6906840                  0.6929769
##                              bullying stress_level
## anxiety_level               0.7099815    0.7367954
## self_esteem                -0.6407371   -0.7561951
## mental_health_history       0.6243658    0.6486440
## depression                  0.6657899    0.7343786
## headache                    0.6097755    0.7134840
## blood_pressure              0.3704402    0.3941999
## sleep_quality              -0.6994272   -0.7490679
## breathing_problem           0.5763408    0.5739837
## noise_level                 0.5854583    0.6633713
## living_conditions          -0.5511387   -0.5817231
## safety                     -0.6456733   -0.7096016
## basic_needs                -0.6448862   -0.7089676
## academic_performance       -0.6662287   -0.7209224
## study_load                  0.5866686    0.6341555
## teacher_student_relationship -0.6559604   -0.6801627
## future_career_concerns      0.7112781    0.7426186
## social_support             -0.5670783   -0.6324970
## peer_pressure               0.6610577    0.6906840
## extracurricular_activities  0.6519786    0.6929769
## bullying                    1.0000000    0.7511623
## stress_level                0.7511623    1.0000000
```

Nhận xét: Ta thấy hầu như các biến đều có hệ số tương quan cao, nên ta có thể nói biến này có liên hệ với nhau.

❖ Kiểm tra dữ liệu có đủ để phân tích nhân tố:

- Sử dụng phương pháp kiểm định KMO

```
KMO(stress)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = stress)
## Overall MSA =  0.97
## MSA for each item =
##                anxiety_level                   self_esteem
##                         0.99                          0.98
##        mental_health_history                    depression
##                         0.99                          0.99
##                     headache                blood_pressure
##                         0.99                          0.76
##                sleep_quality             breathing_problem
##                         0.99                          0.99
##                  noise_level             living_conditions
##                         0.99                          0.99
##                       safety                   basic_needs
##                         0.97                          0.97
##         academic_performance                    study_load
##                         0.98                          0.99
## teacher_student_relationship       future_career_concerns
##                         0.97                          0.98
##               social_support                 peer_pressure
##                         0.88                          0.97
##   extracurricular_activities                     bullying
##                         0.98                          0.98
##                 stress_level
##                         0.98
```

Nhận xét: Với MSA = 0.97 từ bảng kết quả trên được xem là kết quả khá tốt, cùng với các KMO của các biến riêng lẻ đều lớn hơn 0.6 cũng là kết quả tốt. Từ đó ta kết luận được rằng dữ liệu là đủ để phân tích nhân tố.

❖ Xác định số nhân tố:

Sử dụng kiểm định Kaiser để xác định số nhân tố:

```
ev = eigen(R)
print(ev$values)
```

```
##  [1] 12.7029447  1.1986181  0.6939477  0.5952937  0.5592153  0.5262102
##  [7]  0.4742388  0.4580151  0.4063027  0.3859583  0.3643367  0.3483829
## [13]  0.3287309  0.3132260  0.3122791  0.2824861  0.2732262  0.2664635
## [19]  0.2329641  0.1748916  0.1022682
```

Nhận xét: Với kết quả có 2 giá trị riêng đầu tiên lớn hơn 1, nên ta giữ lại hai nhân tố.

- ❖ Phân tích nhân tố với m = 2:

Ta phân tích 2 nhân tố với phép quay Varimax.

```
fa_stress = fa(R,nfactors = 2, rotate = "varimax", residuals = TRUE, fm ="
ml")
fa_stress
```

```
## Factor Analysis using method =  ml
## Call: fa(r = R, nfactors = 2, rotate = "varimax", residuals = TRUE,
##     fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                                ML2   ML1   h2    u2 com
## anxiety_level                 0.80  0.24 0.70 0.297 1.2
## self_esteem                  -0.71 -0.44 0.70 0.299 1.7
## mental_health_history         0.71  0.22 0.55 0.450 1.2
## depression                    0.75  0.36 0.69 0.315 1.4
## headache                      0.73  0.28 0.61 0.392 1.3
## blood_pressure                0.11  0.99 1.00 0.005 1.0
## sleep_quality                -0.81 -0.21 0.70 0.305 1.1
## breathing_problem             0.65  0.09 0.44 0.563 1.0
## noise_level                   0.65  0.28 0.50 0.503 1.4
## living_conditions            -0.63 -0.21 0.44 0.557 1.2
## safety                       -0.76 -0.21 0.62 0.376 1.1
## basic_needs                  -0.76 -0.20 0.61 0.391 1.1
## academic_performance         -0.78 -0.18 0.64 0.362 1.1
## study_load                    0.63  0.28 0.47 0.528 1.4
## teacher_student_relationship -0.76 -0.27 0.64 0.355 1.3
## future_career_concerns        0.77  0.35 0.71 0.286 1.4
## social_support               -0.51 -0.70 0.75 0.251 1.8
## peer_pressure                 0.69  0.33 0.59 0.413 1.4
## extracurricular_activities    0.68  0.35 0.59 0.413 1.5
## bullying                      0.78  0.29 0.69 0.310 1.3
## stress_level                  0.85  0.30 0.81 0.194 1.3
##
##                         ML2  ML1
## SS loadings           10.49 2.95
## Proportion Var         0.50 0.14
## Cumulative Var         0.50 0.64
## Proportion Explained   0.78 0.22
## Cumulative Proportion  0.78 1.00
##
## Mean item complexity =  1.3
```

```
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  210  with the objective function =  17.63
## df of  the model are 169  and the objective function was  0.82
##
## The root mean square of the residuals (RMSR) is  0.02
## The df corrected root mean square of the residuals is  0.03
##
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                  ML2  ML1
## Correlation of (regression) scores with factors   0.98 1.00
## Multiple R square of scores with factors          0.96 0.99
## Minimum correlation of possible factor scores     0.93 0.99
```

Nhận xét:

- Nhân tố 2: Ta nhận thấy nhân tố 2 sẽ được giải thích bởi 19 biến vì tỷ trọng của 19 biến này sẽ cao hơn tỷ trọng ở nhân tố 1, vậy trừ biến `blood_pressure` với tỷ trọng 0.99 ở nhân tố 1 cao hơn tỷ trọng 0.12 ở nhân tố 2, và biến `socialsupport` với tỷ trọng 0.69 ở nhân tố 1 cao hơn tỷ trọng 0.51 ở nhân tố 2, thì các biến còn lại sẽ giải thích cho nhân tố 2.
  - Với các biến `anxiety-level`, `mental-health-history`, `depression`, `headache`, `sleep-quality`, `breathing-problem`, `stress-level` liên quan đến vấn đề sức khoẻ và tinh nhần, các biến này đều có tỷ trọng dương trừ biến `sleep-quality` mang tỷ trong âm, vì ta thấy các biến mang giá trị dương này nghịch biến với biến `sleep-quality`, vì ta thấy các biến mang tỷ trọng dương này càng cao thì sức khoẻ và tinh thần đều càng không tốt điều đó sẽ dẫn đến chất lượng giấc ngủ sẽ không được tốt, vì vậy mà hai biến này sẽ nghịch biến với nhau.Điều này còn dẫn đến nhân tố 2 sẽ cao khi cá nhân đó có sức khoẻ và tinh thần không tốt.
  - Với các biến `self_esteem`, `noise_level`, `living_conditions`, `safety`, `basic_needs` sẽ liên quan đến quần đề đời sống xã hội, các biến này hầu như đều mang giá trị âm, trừ mức độ ồn mang giá trị dương, vì với các biến mang giá trị âm càng cao thì ta thấy đời sống xã hội sẽ càng

tốt mà đời dống xã hội tốt thì mức độ ồn sẽ càng giảm. Điều này dẫn đến nhân tố 1 sẽ thấp khi có mức sống và xã hội cao.

- Với các biến còn lại `academic_performance`, `study_load`, `teacher_student_relationship`, `future_career_concerns`, `peer_pressure`, `extracurricular_activities`, `bullying` sẽ liên quan đến học tập, các biến hầu như đêu mang tỷ trọng dương, trừ biến `academic_performance` và biến `teacher_student_relationship` thì mang giá trị âm, ta nhận xét được khi học sinh có thành tích học tâ tốt cũng sẽ có mối quan hệ tốt với giáo viên, nhưng điều này lại nghịc biến với các biến mang giá trị dương liên quan đến học tập không được tốt. Điều này dẫn đến khi một cá nhân có học tập không tốt thì nhân tố 2 càng cao
- Từ đó ta thấy nhân tố 2 sẽ giải thích cho mức độ sức khoẻ, tinh thần, đời sống, xã hội và học tập. Khi nhân tố 2 cao ta có thể nói sức khoẻ tinh thần sẽ không tốt, đời sống xa hội không tốt và học tập cũng không tốt.

- Nhân tố 1: Ta thấy nhânn tố 1 sẽ được giải thích bởi 2 biến còn lại là `blood_pressure` và `socialsupport`, vì 2 biến này có trọng số cao hơn ở trọng số ở nhân tố 2.

  - Ta thấy biến `blood_pressure` có trọng số dương, còn biến `social_support` có trọng số âm, khi đó ta nhận xét thấy nếu một người có sự trợ giúp từ bên ngoài cao thì huyết áp của người đó sẽ giảm xuống. Khi đó nhân tố 1 sẽ càng cao khi người đó có huyết áp cao và sự hỗ trợ bên ngoài sẽ giảm đi.

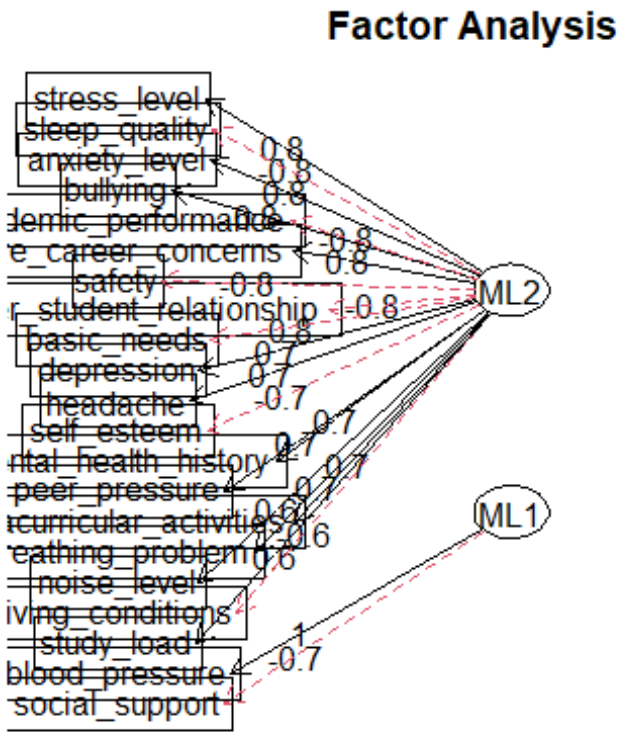
  - Nhân tố 1 sẽ chỉ sự đối lặp của sự hỗ trợ bên ngoài với huyết áp.

- Từ bảng kết quả, hai nhân tố *MR1* và *MR2* sẽ giải thích được 64% phương sai dữ liệu, trong đó *MR1* sẽ giải thích được 50% phương sai dữ liệu và *MR2* sẽ giải thích 14% còn lại.

- Với các trị riêng của *MR1* và *MR2* lần lượt là 10.57 và 2.84 thì theo tiêu chuẩn KAISER ta giữ lại hai biến này.
- Và bằng kiểm định Likelihood với prob < 6e-104, ta sẽ bác bỏ giả thuyết H_0 với H_0 là giả định mô hình 2 nhân tố là hợp lý. Vậy ta sẽ chấp nhận H_1 là giả định cho việc mô hình 2 nhân tố là không hợp lý.

Đồ thị mô tả cho mô hình:

```
fa.diagram(fa_stress)
```



❖ Kiểm tra mô hình thích hợp hơn:

Ta kiểm tra mô hình 8 nhân tố với Hệ số tương quan R.

```
fa_stress_corr = factanal(covmat = R,factors=8, rotation="varimax", n.obs=
1100)
print(fa_stress_corr)

##
## Call:
## factanal(factors = 8, covmat = R, n.obs = 1100, rotation = "varimax")
##
## Uniquenesses:
```

```
##                anxiety_level                     self_esteem
##                        0.285                           0.226
##        mental_health_history                      depression
##                        0.430                           0.287
##                     headache                 blood_pressure
##                        0.005                           0.005
##                sleep_quality              breathing_problem
##                        0.295                           0.533
##                  noise_level              living_conditions
##                        0.484                           0.525
##                       safety                    basic_needs
##                        0.309                           0.005
##         academic_performance                     study_load
##                        0.354                           0.499
## teacher_student_relationship       future_career_concerns
##                        0.278                           0.266
##               social_support                  peer_pressure
##                        0.048                           0.351
##   extracurricular_activities                      bullying
##                        0.367                           0.263
##                 stress_level
##                        0.078
##
## Loadings:
##                               Factor1 Factor2 Factor3 Factor4 Factor5 Fa
ctor6
## anxiety_level                  0.729   0.158  -0.277   0.208   0.186
## self_esteem                   -0.606  -0.362   0.275  -0.220  -0.183  -0
.118
## mental_health_history          0.623   0.141  -0.177   0.245   0.236
## depression                     0.671   0.274  -0.251   0.179   0.235
## headache                       0.488   0.186  -0.197   0.201   0.799
## blood_pressure                 0.187   0.973
## sleep_quality                 -0.716  -0.127   0.291  -0.189  -0.207
## breathing_problem              0.613          -0.184   0.185   0.103
## noise_level                    0.590   0.215  -0.133   0.240   0.166   0
.135
## living_conditions             -0.550  -0.135   0.253  -0.154  -0.193
## safety                        -0.593  -0.129   0.456  -0.223  -0.165  -0
.105
## basic_needs                   -0.473  -0.129   0.232  -0.811  -0.194
## academic_performance          -0.612           0.366  -0.248  -0.218  -0
.117
## study_load                     0.589   0.212  -0.122   0.166   0.189
## teacher_student_relationship -0.574  -0.195   0.496  -0.244  -0.205
## future_career_concerns         0.705   0.264  -0.219   0.207   0.246
## social_support                -0.301  -0.656   0.585  -0.224  -0.172
## peer_pressure                  0.699   0.246           0.207   0.229
## extracurricular_activities     0.699   0.274  -0.148           0.173
## bullying                       0.750   0.202  -0.223   0.228   0.145
## stress_level                   0.727   0.213  -0.233   0.261   0.243   0
```

```
.400
##                               Factor7 Factor8
## anxiety_level
## self_esteem                    0.323
## mental_health_history
## depression                    -0.146   0.125
## headache
## blood_pressure
## sleep_quality
## breathing_problem                     -0.101
## noise_level
## living_conditions                      0.147
## safety                                 0.152
## basic_needs
## academic_performance
## study_load                             0.123
## teacher_student_relationship
## future_career_concerns        -0.112
## social_support
## peer_pressure
## extracurricular_activities
## bullying
## stress_level
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
Factor8
## SS loadings      7.830   2.162   1.665   1.497   1.386   0.256   0.198
0.114
## Proportion Var   0.373   0.103   0.079   0.071   0.066   0.012   0.009
0.005
## Cumulative Var   0.373   0.476   0.555   0.626   0.692   0.705   0.714
0.719
##
## Test of the hypothesis that 8 factors are sufficient.
## The chi square statistic is 84.71 on 70 degrees of freedom.
## The p-value is 0.111
```

Nhận xét:

- Bằng kiểm định Likeliihood Chi Square, ta có giá trị p_value = 0.111 > 0.05 nên ta chưa đủ giả thuyết để bác bỏ H_0 là mô hình 8 nhân tố là phù hợp. Vậy nên ta có thể xem mô hình 8 nhân tố này với ma trận hiệp phương sai sẽ đưa ra kết quả tốt hơn.

2. BÀI TẬP

2.1. Bài 4.21 trang 205 của sách Johnson, 2013.

Cho mẫu ngẫu nhiên $\boldsymbol{X_1, \dots, X_{60}}$ kích thước 60 lấy từ tổng thể $\boldsymbol{\mathcal{N}_4(\mu, \Sigma)}$. Xác định:

a. Phân phối của $\boldsymbol{\bar{X}}$

b. Phân phối của $\boldsymbol{(X_1 - \mu)'\Sigma^{-1}(X_1 - \mu)}$

c. Phân phối của $\boldsymbol{n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu)}$

d. Phân phối gần đúng của $\boldsymbol{n(\bar{X} - \mu)'S^{-1}(\bar{X} - \mu)}$.

a. Xác định phân phối của $\bar{\boldsymbol{X}}$:

Theo kết quả (4-23) trang 174 sách Johnson, 2013: mẫu ngẫu nhiên $\mathbf{X_1, \dots, X_n}$ kích thước n lấy từ tổng thể $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ thì $\bar{\boldsymbol{X}}$ có phân phối $\mathcal{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$.

Theo đề bài, ta có: mẫu ngẫu nhiên $\mathbf{X_1, \dots, X_{60}}$ kích thước 60 lấy từ tổng thể $\mathcal{N}_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Do đó, $\bar{\boldsymbol{X}}$ có phân phối $\mathcal{N}_4(\boldsymbol{\mu}, \frac{1}{60}\boldsymbol{\Sigma})$.

b. Xác định phân phối của $\boldsymbol{(X_1 - \mu)'\Sigma^{-1}(X_1 - \mu)}$:

Theo định lý 2.1.4 trang 31 sách Thống Kê Nhiều Chiều, lưu hành nội bộ: Nếu $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, thì $(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2(p)$.

Theo đề bài, ta có: $\boldsymbol{X_1} \sim \mathcal{N}_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ do $\boldsymbol{X_i}$ độc lập và cùng phân phối.

Vì vậy, $(\boldsymbol{X_1} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X_1} - \boldsymbol{\mu}) \sim \chi^2(4)$.

c. Xác định phân phối của $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$:

Theo kết quả vừa tìm được ở câu (a) và định lý 2.1.4 trang 31 sách Thống Kê Nhiều Chiều, lưu hành nội bộ, ta có: $(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\left(\frac{1}{60}\boldsymbol{\Sigma}\right)^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi^2(4)$.

Mà $(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\left(\frac{1}{60}\boldsymbol{\Sigma}\right)^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) = (\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\left(\frac{1}{60}\right)^{-1}\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$

$= 60(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$

Do đó, $60(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi^2(4)$.

Với kích thước n, ta có: $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi^2(4)$.

d. Xác định phân phối của $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$:

Theo ghi chú 3 trang 43 sách Thống Kê Nhiều Chiều, lưu hành nội bộ: Khi $\bar{\boldsymbol{X}}$ có phân phối xấp xỉ chuẩn, phân phối mẫu $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$ được xấp xỉ bằng phân phối Chi-bình phương với p bậc tự do. Việc thay $\boldsymbol{\Sigma}^{-1}$ bằng $\boldsymbol{S}^{-1}$ không ảnh hưởng nhiều đến xấp xỉ này khi $n - p$ lớn.

Theo đề bài, $n = 60, p = 4$, khi đó $n - p = 60 - 4 = 56$ (lớn).

Vậy, theo kết quả câu (c), ta thay $\boldsymbol{\Sigma}^{-1}$ bằng $\boldsymbol{S}^{-1}$ ta được:

$60(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi^2(4)$.

Với kích thước n lớn, ta có: $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi^2(4)$.

2.2. Bài 6.33 trang 355-356 của sách Johnson, 2013.

❖ Thư viện:

```
library(tidyverse)
library(ggplot2)
library(rstatix)
library(ggpubr)
library(broom)
library(car)
```

❖ Nhập dữ liệu:

```
data = read.table("D:/TKNC/Datasets/Datasets/T6-18.dat")
names(data) = c("s_560nm","s_720nm","species","time","replication")
data = data |> mutate(species = factor(species), time = factor(time))
data
```

```
##    s_560nm s_720nm species time replication
## 1     9.33   19.14      SS    1           1
## 2     8.74   19.55      SS    1           2
## 3     9.31   19.24      SS    1           3
## 4     8.27   16.37      SS    1           4
## 5    10.22   25.00      SS    2           1
## 6    10.13   25.32      SS    2           2
## 7    10.42   27.12      SS    2           3
## 8    10.62   26.28      SS    2           4
## 9    15.25   38.89      SS    3           1
## 10   16.22   36.67      SS    3           2
## 11   17.24   40.74      SS    3           3
## 12   12.77   67.50      SS    3           4
## 13   12.07   33.03      JL    1           1
```

```
## 14   11.03   32.37      JL    1           2
## 15   12.48   31.31      JL    1           3
## 16   12.12   33.33      JL    1           4
## 17   15.38   40.00      JL    2           1
## 18   14.21   40.48      JL    2           2
## 19    9.69   33.90      JL    2           3
## 20   14.35   40.15      JL    2           4
## 21   38.71   77.14      JL    3           1
## 22   44.74   78.57      JL    3           2
## 23   36.67   71.43      JL    3           3
## 24   37.21   45.00      JL    3           4
## 25    8.73   23.27      LP    1           1
## 26    7.94   20.87      LP    1           2
## 27    8.37   22.16      LP    1           3
## 28    7.86   21.78      LP    1           4
## 29    8.45   26.32      LP    2           1
## 30    6.79   22.73      LP    2           2
## 31    8.34   26.67      LP    2           3
## 32    7.54   24.87      LP    2           4
## 33   14.04   44.44      LP    3           1
## 34   13.51   37.93      LP    3           2
## 35   13.33   37.93      LP    3           3
## 36   12.77   60.87      LP    3           4
```

```
dim(data)
```

```
## [1] 36  5
```

```
str(data)
```

```
## 'data.frame':    36 obs. of  5 variables:
##  $ s_560nm    : num  9.33 8.74 9.31 8.27 10.22 ...
##  $ s_720nm    : num  19.1 19.6 19.2 16.4 25 ...
##  $ species    : Factor w/ 3 levels "JL","LP","SS": 3 3 3 3 3 3 3 3 3 3
...
##  $ time       : Factor w/ 3 levels "1","2","3": 1 1 1 1 2 2 2 2 3 3 ...
##  $ replication: int  1 2 3 4 1 2 3 4 1 2 ...
```

❖ Mô tả dữ liệu:

Dữ liệu về phản xạ quang phổ gồm 36 quan trắc và 5 biến:

1. `s_560nm`: Phần trăm phản xạ quang phổ ở bước sóng 560nm (màu xanh lá cây)
2. `s_720nm`: Phần trăm phản xạ quang phổ ở bước sóng 720nm (gần hồng ngoại)
3. `species`: Giống loài (sitka spruce [SS], Japanese larch [JL], and lodgepole pine [LP])

4. `time`: thời gian của cây con 1 tuổi lấy tại 3 thời điểm khác nhau (Julian day 150 [1], Julian day 235 [2], and Julian day 320 [3]) trong mùa sinh trưởng.
5. `replication`: nhân rộng.

Trong đó, hai biến `species` và `time` là hai nhân tố.

a. Thực hiện MANOVA hai nhân tố. Kiểm định hiệu ứng species, hiệu ứng time và tương tác species – time. Sử dụng $\alpha = 0.05$.

❖ MANOVA hai nhân tố

```
fit_1 = manova(cbind(s_560nm,s_720nm)~species*time,data = data)
summary(fit_1,test = "Wilks")
```

```
##              Df    Wilks approx F num Df den Df    Pr(>F)
## species       2 0.068774   36.571      4     52 1.554e-14 ***
## time          2 0.049166   45.629      4     52 < 2.2e-16 ***
## species:time  4 0.087070   15.528      8     52 2.217e-11 ***
## Residuals    27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nhận xét: Từ bảng MANOVA với mức ý nghĩa 5%

- Kiểm định sự tương tác giữa 2 nhân tố: ta thấy p_value của tương tác 2 biến rất nhỏ (giá trị thống kê = 15.528 > 2.25 = $F_{4,52}^{5\%}$) nên ta bác bỏ giả thuyết $H_0: \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$. Nghĩa là có sự tương tác giữa hai nhân tố `species` và `time`.

- Kiểm định sự ảnh hưởng của nhân tố 1 (`species`) lên vector phụ thuộc: ta thấy p_value rất nhỏ (giá trị thống kê = 36.571 > 2.25 = $F_{4,52}^{5\%}$) nên ta bác bỏ giả thuyết $H_0: \tau_1 = \tau_2 = 0$, nghĩa là nhân tố 1 ảnh hưởng lên vector phụ thuộc mà ta đang xét.

- Kiểm định sự ảnh hưởng của nhân tố 2 (`time`) lên vector phụ thuộc: ta thấy p_value rất nhỏ (giá trị thống kê = 45.629 > 2.25 = $F_{4,52}^{5\%}$) nên ta bác bỏ giả thuyết $H_0: \beta_1 = \beta_2 = 0$, nghĩa là nhân tố 2 ảnh hưởng lên vector phụ thuộc mà ta đang xét.

b. Bạn có nghĩ rằng các giả định MANOVA thông thường được thỏa mãn đối với những dữ liệu này không? Thảo luận với sự tham khảo về phân tích phần dư và khả năng quan sát tương quan theo thời gian.

❖ Vẽ boxplot cho dữ liệu và kiểm tra giá trị ngoại lai

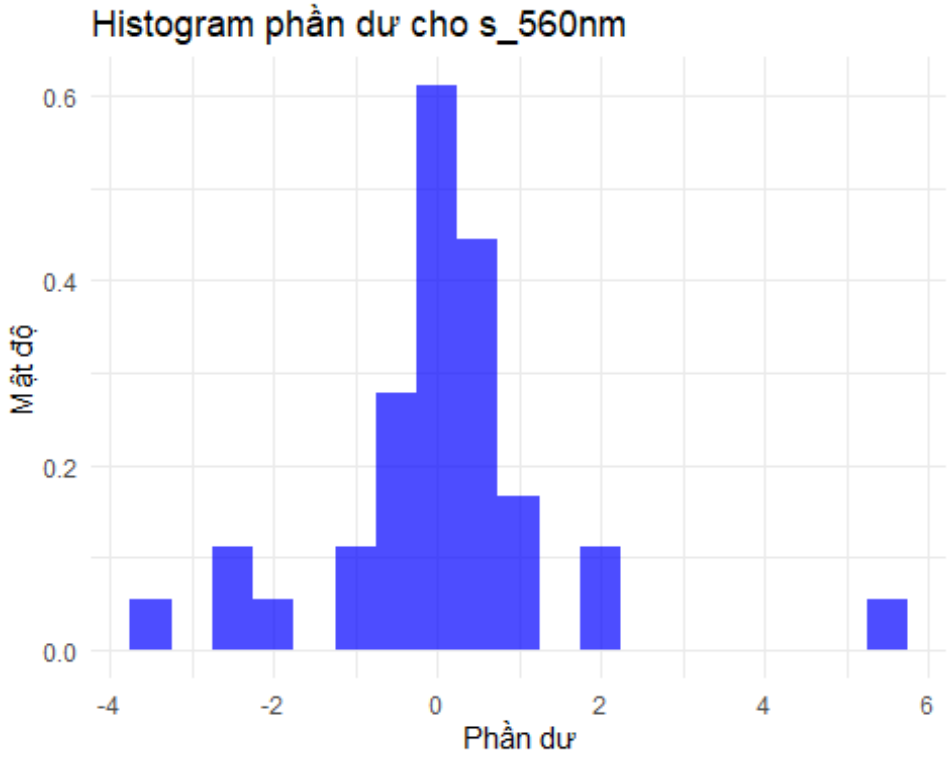
```
# Tính phần dư cho từng biến phản hồi
residuals <- residuals(fit_1)
residuals
```

```
##    s_560nm  s_720nm
## 1   0.4175   0.5650
## 2  -0.1725   0.9750
## 3   0.3975   0.6650
## 4  -0.6425  -2.2050
## 5  -0.1275  -0.9300
## 6  -0.2175  -0.6100
## 7   0.0725   1.1900
## 8   0.2725   0.3500
## 9  -0.1200  -7.0600
## 10  0.8500  -9.2800
## 11  1.8700  -5.2100
## 12 -2.6000  21.5500
## 13  0.1450   0.5200
## 14 -0.8950  -0.1400
## 15  0.5550  -1.2000
## 16  0.1950   0.8200
## 17  1.9725   1.3675
## 18  0.8025   1.8475
## 19 -3.7175  -4.7325
## 20  0.9425   1.5175
## 21 -0.6225   9.1050
## 22  5.4075  10.5350
## 23 -2.6625   3.3950
## 24 -2.1225 -23.0350
## 25  0.5050   1.2500
## 26 -0.2850  -1.1500
## 27  0.1450   0.1400
## 28 -0.3650  -0.2400
## 29  0.6700   1.1725
## 30 -0.9900  -2.4175
## 31  0.5600   1.5225
## 32 -0.2400  -0.2775
## 33  0.6275  -0.8525
## 34  0.0975  -7.3625
## 35 -0.0825  -7.3625
## 36 -0.6425  15.5775
```

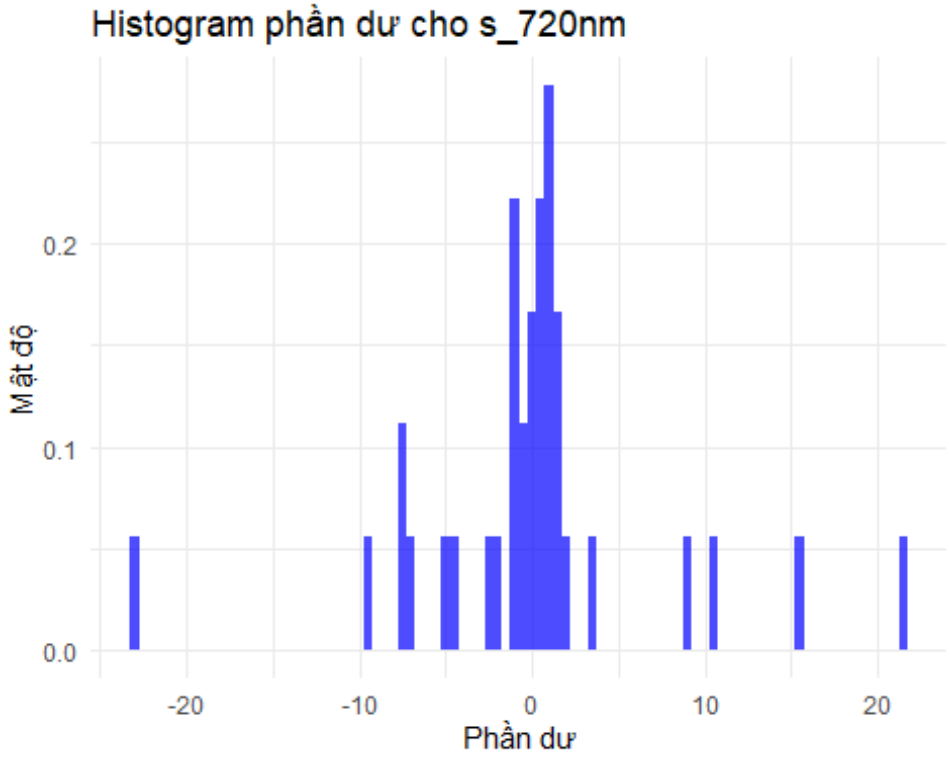
```
# Chuyển phần dư thành data frame
residuals_df <- data.frame(residuals)
```

```r
# Vẽ biểu đồ phần dư cho s_560nm
ggplot(residuals_df, aes(x = s_560nm)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.5, fill = "blue", alph
a = 0.7) +
  labs(title = "Histogram phần dư cho s_560nm", x = "Phần dư", y = "Mật độ
") +
  theme_minimal()
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2
3.4.0.
## ℹ Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning w
as
## generated.
```




```r
# Vẽ biểu đồ phần dư cho s_720nm
ggplot(residuals_df, aes(x = s_720nm)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.5, fill = "blue", alph
a = 0.7) +
  labs(title = "Histogram phần dư cho s_720nm", x = "Phần dư", y = "Mật độ
") +
  theme_minimal()
```



Nhận xét: Từ biểu đồ Histogram ta thấy dữ liệu vẫn còn một vài giá trị ngoại lai, nhưng đồ thị xấp xỉ về chuẩn nên các giả định MANOVA thông thường thoả mãn đối với dữ liệu này.

c. Người làm rừng đặc biệt quan tâm đến sự tương tác giữa loài và thời gian. Sự tương tác có hiển thị cho một biến nhưng không hiển thị cho biến kia không? Kiểm tra bằng cách chạy ANOVA hai yếu tố đơn biến cho mỗi câu trả lời trong số hai câu trả lời.

❖ Mô hình ANOVA 2 nhân tố cho biến s_560nm

```
model_2 = aov(s_560nm ~ species*time,data=data)
summary(model_2)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## species       2  965.2   482.6  169.97 5.03e-16 ***
## time          2 1275.2   637.6  224.58  < 2e-16 ***
## species:time  4  795.8   199.0   70.07 7.34e-14 ***
## Residuals    27   76.7     2.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nhận xét: Từ bảng ANOVA với mức ý nghĩa là 5%: kiểm định sự tương tác của 2 nhân tố `species` và `time` ở bước sóng 560nm, với p_value rất nhỏ, nên ta không đủ cơ sở bác bỏ giả thuyết $H_0: \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$. Nghĩa là không có sự tương tác giữa hai nhân tố `species` và `time` ở bước sóng 560nm.

❖ Mô hình ANOVA 2 nhân tố cho biến s_720nm

```
model_3 = aov(s_720nm ~ species*time,data=data)
summary(model_3)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## species       2   2027  1013.4  15.462 3.35e-05 ***
## time          2   5574  2786.9  42.521 4.54e-09 ***
## species:time  4    194    48.4   0.738    0.574
## Residuals    27   1770    65.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nhận xét: Từ bảng ANOVA với mức ý nghĩa là 5%: kiểm định sự tương tác của 2 nhân tố `species` và `time` ở bước sóng 720nm, với p_value = 0.574 > 0.05, nên ta bác bỏ giả thuyết $H_0: \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$. Nghĩa là có sự tương tác giữa hai nhân tố `species` và `time` ở bước sóng 720nm.

d. Bạn có thể nghĩ ra một phương pháp khác để phân tích những dữ liệu này (hoặc một thiết kế thử nghiệm khác) có thể cho phép xác định xu hướng thời gian tiềm năng của các số phản xạ quang phổ không?

Dữ liệu còn có thể phân tích bằng phương pháp đường cong tăng trưởng được thảo luận trong phần 6.4 sách Johnson. Dữ liệu cũng có thể được phân tích với giả định rằng các loài được "lồng nhau" trong ngày tháng. Từ đó ta đặt ra vấn đề: Độ phản xạ quang phổ có giống nhau đối với tất cả các loài trong mỗi ngày không?

❖ Mô hình tuyến tính hỗn hợp với loài lồng nhau trong ngày tháng

```
mixed_model_560 <- lmer(s_560nm ~ time * species + (1 | replication), data
= data)
```

```
## boundary (singular) fit: see help('isSingular')
```

```
summary(mixed_model_560)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: s_560nm ~ time * species + (1 | replication)
##    Data: data
##
## REML criterion at convergence: 117.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2062 -0.2548  0.0504  0.3301  3.2092
##
## Random effects:
##  Groups      Name        Variance  Std.Dev.
##  replication (Intercept) 2.485e-15 4.985e-08
##  Residual                2.839e+00 1.685e+00
## Number of obs: 36, groups:  replication, 4
##
## Fixed effects:
##                 Estimate Std. Error t value
## (Intercept)      11.9250     0.8425  14.154
## time2             1.4825     1.1915   1.244
## time3            27.4075     1.1915  23.003
## speciesLP        -3.7000     1.1915  -3.105
## speciesSS        -3.0125     1.1915  -2.528
## time2:speciesLP  -1.9275     1.6850  -1.144
## time3:speciesLP -22.2200     1.6850 -13.187
## time2:speciesSS  -0.0475     1.6850  -0.028
## time3:speciesSS -20.9500     1.6850 -12.433
##
## Correlation of Fixed Effects:
##             (Intr) time2  time3  spcsLP spcsSS tm2:LP tm3:LP tm2:SS
## time2       -0.707
## time3       -0.707  0.500
## speciesLP   -0.707  0.500  0.500
## speciesSS   -0.707  0.500  0.500  0.500
## tim2:spcsLP  0.500 -0.707 -0.354 -0.707 -0.354
## tim3:spcsLP  0.500 -0.354 -0.707 -0.707 -0.354  0.500
## tim2:spcsSS  0.500 -0.707 -0.354 -0.354 -0.707  0.500  0.250
## tim3:spcsSS  0.500 -0.354 -0.707 -0.354 -0.707  0.250  0.500  0.500
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```
mixed_model_720 <- lmer(s_720nm ~ time * species + (1 | replication), data
= data)
```

```
## boundary (singular) fit: see help('isSingular')
```

```
summary(mixed_model_720)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: s_720nm ~ time * species + (1 | replication)
```

```
##    Data: data
##
## REML criterion at convergence: 202
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.84530 -0.17926  0.03026  0.15803  2.66187
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  replication (Intercept)  0.00    0.000
##  Residual                65.54    8.096
## Number of obs: 36, groups:  replication, 4
##
## Fixed effects:
##                 Estimate Std. Error t value
## (Intercept)       32.510      4.048   8.031
## time2              6.123      5.725   1.070
## time3             35.525      5.725   6.206
## speciesLP        -10.490      5.725  -1.832
## speciesSS        -13.935      5.725  -2.434
## time2:speciesLP   -2.995      8.096  -0.370
## time3:speciesLP  -12.252      8.096  -1.513
## time2:speciesSS    1.232      8.096   0.152
## time3:speciesSS   -8.150      8.096  -1.007
##
## Correlation of Fixed Effects:
##             (Intr) time2  time3  spcsLP spcsSS tm2:LP tm3:LP tm2:SS
## time2       -0.707
## time3       -0.707  0.500
## speciesLP   -0.707  0.500  0.500
## speciesSS   -0.707  0.500  0.500  0.500
## tim2:spcsLP  0.500 -0.707 -0.354 -0.707 -0.354
## tim3:spcsLP  0.500 -0.354 -0.707 -0.707 -0.354  0.500
## tim2:spcsSS  0.500 -0.707 -0.354 -0.354 -0.707  0.500  0.250
## tim3:spcsSS  0.500 -0.354 -0.707 -0.354 -0.707  0.250  0.500  0.500
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

Nhận xét:

Biến `s_560nm`:

- Phương sai rất nhỏ (gần như bằng 0), điều này cho thấy rằng không có sự biến thiên lớn giữa các lần lặp lại.
- (Intercept): 11.9250, nghĩa là giá trị trung bình của `s_560nm` khi các yếu tố khác (`time` và `species`) bằng 0.

- `time2`: 1.4825, cho thấy rằng giá trị `s_560nm` tăng trung bình khoảng 1.4825 đơn vị từ thời điểm 1 đến thời điểm 2, nhưng không có ý nghĩa thống kê (t = 1.244).
- `time3`: 27.4075, cho thấy rằng giá trị `s_560nm` tăng trung bình khoảng 27.4075 đơn vị từ thời điểm 1 đến thời điểm 3, và có ý nghĩa thống kê (t = 23.003).
- `speciesLP`: -3.7000, cho thấy rằng giá trị `s_560nm` của loài LP thấp hơn trung bình khoảng 3.7000 đơn vị so với loài gốc (reference species), và có ý nghĩa thống kê (t = -3.105).
- `speciesSS`: -3.0125, cho thấy rằng giá trị `s_560nm` của loài SS thấp hơn trung bình khoảng 3.0125 đơn vị so với loài gốc (reference species), và có ý nghĩa thống kê (t = -2.528).
- Các hệ số tương tác giữa thời gian và loài:
  - `time2:speciesLP`: -1.9275 (t = -1.144)
  - `time3:speciesLP`: -22.2200 (t = -13.187)
  - `time2:speciesSS`: -0.0475 (t = -0.028)
  - `time3:speciesSS`: -20.9500 (t = -12.433)
- Nhìn vào các giá trị t-value và độ lớn của hệ số tương tác:
  - Ở thời điểm 2, giá trị phản xạ quang phổ của loài LP và SS không khác biệt nhiều so với loài gốc (các giá trị t-value không có ý nghĩa thống kê).
  - Ở thời điểm 3, giá trị phản xạ quang phổ của loài LP và SS khác biệt đáng kể so với loài gốc (các giá trị t-value rất lớn và có ý nghĩa thống kê).

Biến `s_720nm`:

- Phương sai bằng 0, cho thấy không có sự biến thiên giữa các lần lặp lại.
- (Intercept): 32.510, nghĩa là giá trị trung bình của `s_720nm` khi các yếu tố khác (`time` và `species`) bằng 0.

- `time2`: 6.123, cho thấy rằng giá trị `s_720nm` tăng trung bình khoảng 6.123 đơn vị từ thời điểm 1 đến thời điểm 2, nhưng không có ý nghĩa thống kê (t = 1.070).
- `time3`: 35.525, cho thấy rằng giá trị `s_720nm` tăng trung bình khoảng 35.525 đơn vị từ thời điểm 1 đến thời điểm 3, và có ý nghĩa thống kê (t = 6.206).
- `speciesLP`: -10.490, cho thấy rằng giá trị `s_720nm` của loài LP thấp hơn trung bình khoảng 10.490 đơn vị so với loài gốc, nhưng không có ý nghĩa thống kê (t = -1.832).
- `speciesSS`: -13.935, cho thấy rằng giá trị `s_720nm` của loài SS thấp hơn trung bình khoảng 13.935 đơn vị so với loài gốc, và có ý nghĩa thống kê (t = -2.434).
- Các hệ số tương tác giữa thời gian và loài:
  - `time2:speciesLP`: -2.995 (t = -0.370)
  - `time3:speciesLP`: -12.252 (t = -1.513)
  - `time2:speciesSS`: 1.232 (t = 0.152)
  - `time3:speciesSS`: -8.150 (t = -1.007)
- Nhìn vào các giá trị t-value và độ lớn của hệ số tương tác: Ở thời điểm 2 và 3, giá trị phản xạ quang phổ của loài LP và SS không khác biệt nhiều so với loài gốc (các giá trị t-value không có ý nghĩa thống kê).

Kết luận:

- Biến `s_560nm`:
  - Độ phản xạ quang phổ không giống nhau đối với tất cả các loài ở thời điểm 3 (có sự khác biệt đáng kể giữa các loài), nhưng ở thời điểm 2 thì không có sự khác biệt lớn giữa các loài.
  - Các loài LP và SS có giá trị `s_560nm` thấp hơn so với loài gốc, và sự khác biệt này có ý nghĩa thống kê.

  - Tương tác giữa thời gian và loài cũng cho thấy sự khác biệt đáng kể giữa các thời điểm và loài.

- Biến `s_720nm`:
  - Độ phản xạ quang phổ tương đối giống nhau đối với tất cả các loài ở cả thời điểm 2 và 3 (không có sự khác biệt lớn giữa các loài).
  - Loài SS có giá trị `s_720nm` thấp hơn so với loài gốc và sự khác biệt này có ý nghĩa thống kê.
  - Tương tác giữa thời gian và loài không có nhiều ý nghĩa thống kê.

Giải thích kết quả:

- Sự tăng trưởng của giá trị `s_560nm` và `s_720nm` qua các thời điểm cho thấy rằng thời gian có tác động lớn đến các giá trị đo lường này.
- Loài LP và SS có xu hướng có giá trị thấp hơn so với loài gốc, điều này có thể do các đặc điểm sinh học hoặc điều kiện môi trường khác nhau.
- Tương tác giữa thời gian và loài cũng cho thấy rằng sự thay đổi qua các thời điểm không giống nhau giữa các loài, đặc biệt là từ thời điểm 1 đến thời điểm 3.

Tóm lại: Độ phản xạ quang phổ có sự khác biệt đáng kể giữa các loài ở thời điểm 3 cho biến `s_560nm`. Tuy nhiên, với biến `s_720nm`, không có sự khác biệt lớn giữa các loài ở cả thời điểm 2 và 3. Vậy là ta đã trả lời được cho vấn đề đặt ra.

B. PHẦN 2

Vì bộ dữ liệu **dự đoán hóa đơn giả (fake bills)** khá tốt nên nhóm em sẽ sử dụng một bộ dữ liệu này thay thế cho hai bộ dữ liệu mà đề bài yêu cầu để phân tích theo hai phương pháp khác nhau là phương pháp phân tích thành phần chính và phương pháp phân tích nhân tố.

1. PHÂN TÍCH DỮ LIỆU DỰ ĐOÁN HÓA ĐƠN GIẢ BẰNG PHƯƠNG PHÁP PHÂN TÍCH THÀNH PHẦN CHÍNH

❖ Thư viện

```
library(janitor)
library(tidyverse)
library(FactoMineR)
library(ggplot2)
library(ggbiplot)
library(ggfortify)
library(dplyr)
library(psych)
library(factoextra)  #fviz
library(base)
library(naniar)      #gg_miss_var
```

```
## Warning: package 'naniar' was built under R version 4.4.1
```

❖ Dữ liệu

```
# Nguồn dữ liệu: https://www.kaggle.com/datasets/alexandrepetit881234/fake-bills

fake = read.csv("D:/TKNC/fake_bills.csv", header = TRUE, sep = ";")
head(fake)
```

```
##   is_genuine diagonal height_left height_right margin_low margin_up length
## 1       True   171.81      104.86       104.95       4.52      2.89 112.83
## 2       True   171.46      103.36       103.66       3.77      2.99 113.09
## 3       True   172.69      104.48       103.50       4.40      2.94 113.16
## 4       True   171.36      103.91       103.94       3.62      3.01 113.51
## 5       True   171.73      104.28       103.46       4.04      3.48 112.54
## 6       True   172.17      103.74       104.08       4.42      2.95 112.81
```

```
names(fake)
```

```
## [1] "is_genuine"   "diagonal"     "height_left"  "height_right" "margin_low"
## [6] "margin_up"    "length"
```

```
str(fake)
```

```
## 'data.frame':    1500 obs. of  7 variables:
##  $ is_genuine  : chr  "True" "True" "True" "True" ...
##  $ diagonal    : num  172 171 173 171 172 ...
##  $ height_left : num  105 103 104 104 104 ...
##  $ height_right: num  105 104 104 104 103 ...
##  $ margin_low  : num  4.52 3.77 4.4 3.62 4.04 4.42 4.58 3.98 4 4.04 ...
##  $ margin_up   : num  2.89 2.99 2.94 3.01 3.48 2.95 3.26 2.92 3.25 3.25
...
##  $ length      : num  113 113 113 114 113 ...
```

❖ Mô tả dữ liệu

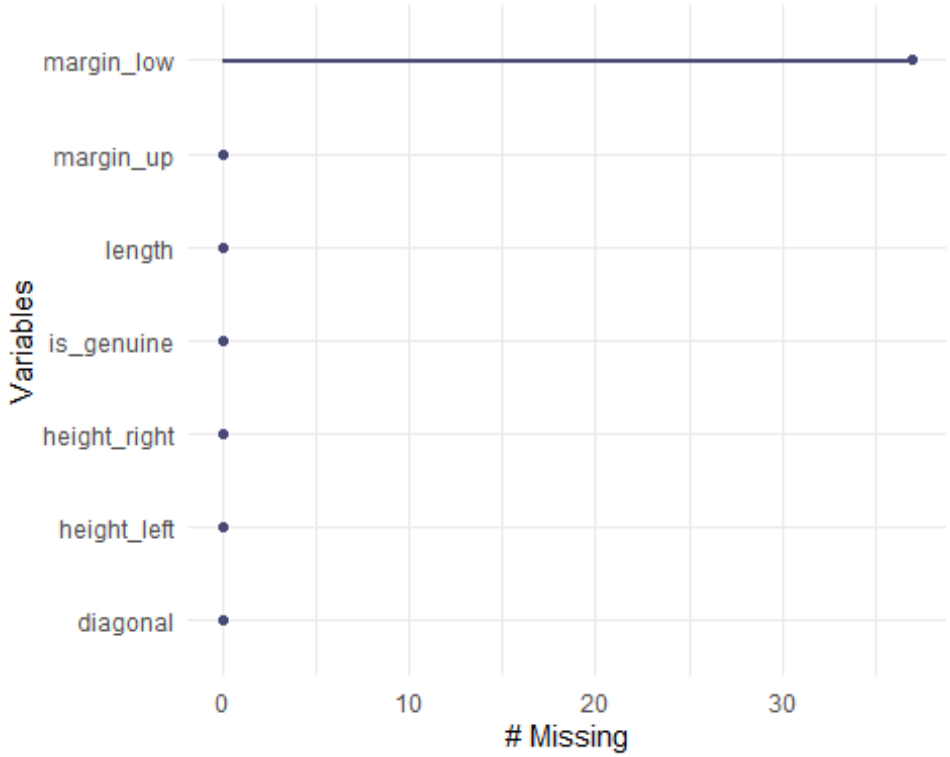
Dữ liệu dự đoán hóa đơn giả gồm 1500 quan trắc và 7 biến:

6. `is_genuine`: Hóa đơn có phải hàng thật không? (TRUE/FALSE)
7. `diagonal`: số đo đường chéo tính bằng mm
8. `height_left`: chiều cao của cạnh trái tính bằng mm
9. `height_right`: chiều cao của cạnh phải tính bằng mm
10. `margin_low`: lề dưới tính bằng mm
11. `margin_up`: lề trên tính bằng mm
12. `length`: chu vi tính bằng mm

Biến `is_genuine` được dùng để phân loại nhóm.

❖ Kiểm tra và xử lý giá trị khuyết

```
# Kiểm tra giá trị khuyết
gg_miss_var(fake)
```



Nhận xét: Có 1 biến có giá trị khuyết là `margin_low`.

```
# Loại bỏ giá trị khuyết
fake_new = na.omit(fake)
head(fake_new)
```

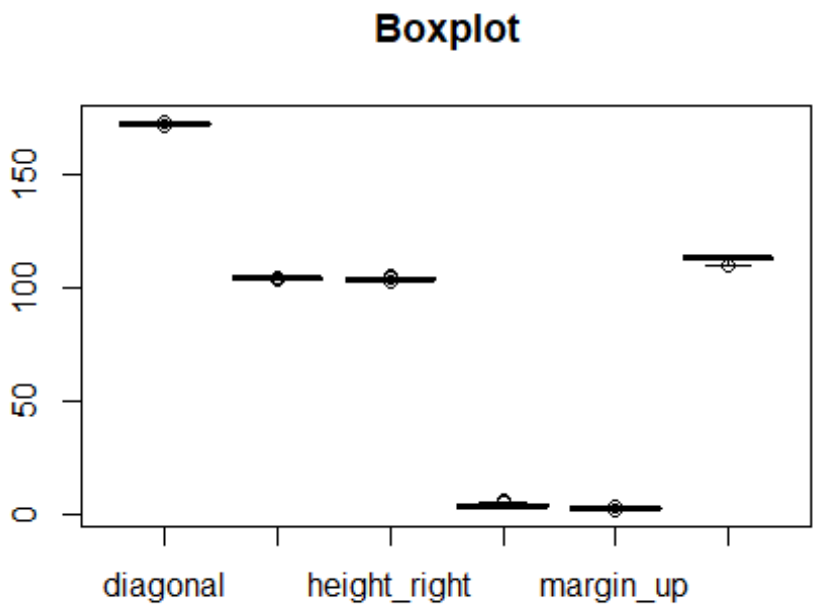
```
##   is_genuine diagonal height_left height_right margin_low margin_up len
gth
## 1       True   171.81      104.86       104.95       4.52       2.89 112
.83
## 2       True   171.46      103.36       103.66       3.77       2.99 113
.09
## 3       True   172.69      104.48       103.50       4.40       2.94 113
.16
## 4       True   171.36      103.91       103.94       3.62       3.01 113
.51
## 5       True   171.73      104.28       103.46       4.04       3.48 112
.54
## 6       True   172.17      103.74       104.08       4.42       2.95 112
.81
```

```
dim(fake_new)
```

```
## [1] 1463    7
```

❖ Kiểm tra và xử lý giá trị ngoại lai

```r
# Kiểm tra giá trị ngoại lai
dat_fake = fake_new[,-1]
boxplot(dat_fake, main = "Boxplot")
```



Nhận xét: Nhìn vào boxplot ta thấy có khá nhiều giá trị ngoại lai.

```r
Check_Outliers = function(Variable){
  Q1 = quantile(Variable,0.25)
  Q3 = quantile(Variable,0.75)
  IQR = Q3 - Q1
  return (Variable <= (Q1 - (1.5 * IQR)) | Variable >= (Q3 + (1.5 * IQR)))
}
matr_outlier = dat_fake %>%
  mutate(across(everything(), Check_Outliers))
data_outlier = dat_fake[rowSums(matr_outlier) > 0,]
head(data_outlier)
```

```
##     diagonal height_left height_right margin_low margin_up length
## 1     171.81      104.86       104.95       4.52      2.89 112.83
## 78    171.84      104.09       103.03       4.11      2.77 113.18
## 177   171.75      103.63       102.97       4.46      2.77 113.22
## 194   172.35      103.73       102.95       4.49      3.37 112.49
## 225   172.12      103.20       103.92       4.46      3.26 113.44
## 293   172.09      103.14       103.81       4.88      3.01 113.69
```

```r
dim(data_outlier)
```

```
## [1] 53  6
```

Nhận xét: Có 53 quan trắc có chứa giá trị ngoại lai, do đó ta sẽ loại bỏ các quan trắc đó.

```
# Loại bỏ giá trị ngoại lai
dele = which(rowSums(matr_outlier)>0)
dele
```

```
##    1   78  177  194  225  293  523  665  730  762  829  843 1023 1024 1
028 1030
##    1   77  174  191  221  286  510  650  713  743  808  822  994  995
999 1001
## 1032 1042 1054 1076 1083 1091 1093 1111 1125 1134 1135 1143 1151 1170 1
200 1255
## 1003 1013 1025 1047 1053 1061 1063 1081 1094 1103 1104 1112 1120 1139 1
168 1223
## 1271 1278 1291 1311 1322 1323 1332 1346 1349 1354 1356 1383 1389 1421 1
427 1442
## 1239 1246 1259 1278 1288 1289 1298 1312 1314 1319 1321 1348 1354 1386 1
392 1405
## 1454 1460 1465 1474 1485
## 1417 1423 1428 1437 1448
```

```
length(dele)
```

```
## [1] 53
```

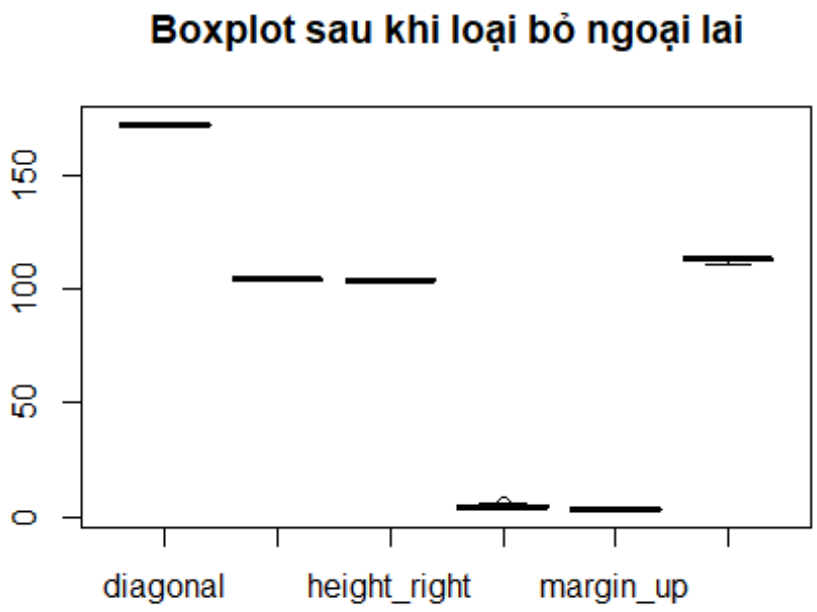
```
fake_clean = fake_new[-dele,]
head(fake_clean)
```

```
##   is_genuine diagonal height_left height_right margin_low margin_up len
gth
## 2       True   171.46      103.36       103.66       3.77      2.99 113
.09
## 3       True   172.69      104.48       103.50       4.40      2.94 113
.16
## 4       True   171.36      103.91       103.94       3.62      3.01 113
.51
## 5       True   171.73      104.28       103.46       4.04      3.48 112
.54
## 6       True   172.17      103.74       104.08       4.42      2.95 112
.81
## 7       True   172.34      104.18       103.85       4.58      3.26 112
.81
```

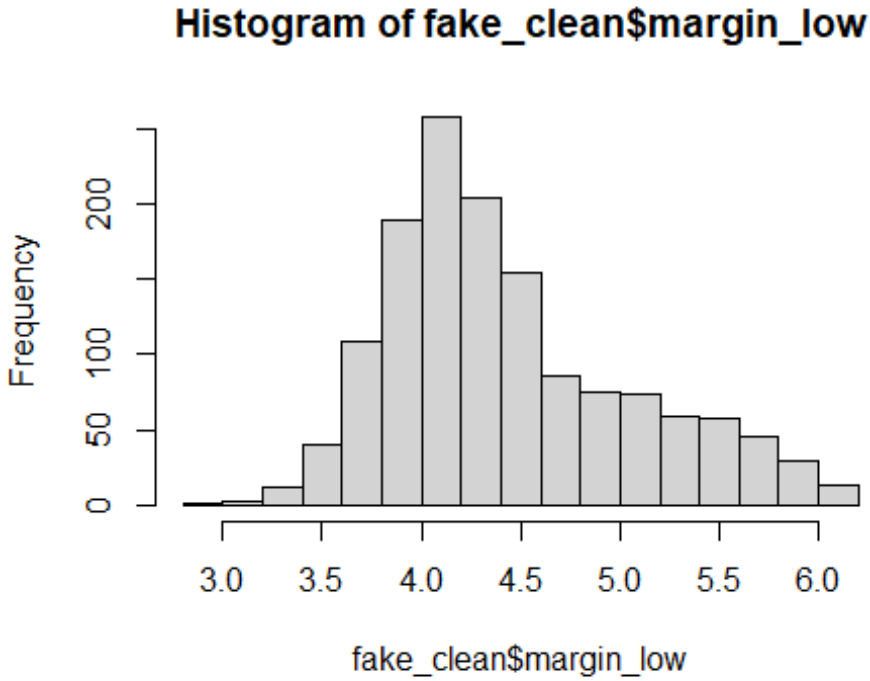
```
dim(fake_clean)
```

```
## [1] 1410    7
```

```
boxplot(fake_clean[,-1], main = "Boxplot sau khi loại bỏ ngoại lai")
```




```
hist(fake_clean$margin_low)
```



Nhận xét: Sau khi loại bỏ các quan trắc có chứa giá trị ngoại lai, vẽ lại boxplot ta thấy vẫn còn ngoại lai ở biến `margin_low`. Khi vẽ biểu đồ Histogram của `margin_low` ta thấy nó có dạng xấp xỉ dạng chuẩn nên ta sẽ không loại bỏ tiếp các giá trị ngoại lai đó.

❖ Chuẩn hóa dữ liệu

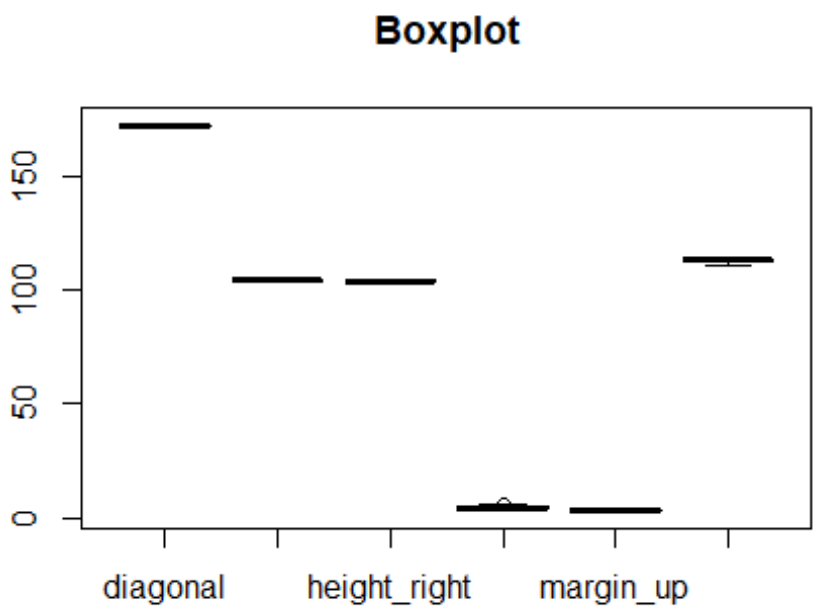
```
apply(fake_clean[,-1], 2, mean)
```

```
##     diagonal  height_left height_right   margin_low    margin_up
length
##   171.960645   104.025922   103.915546     4.444043     3.148560   112.
703759
```

```
apply(fake_clean[,-1], 2, var)
```

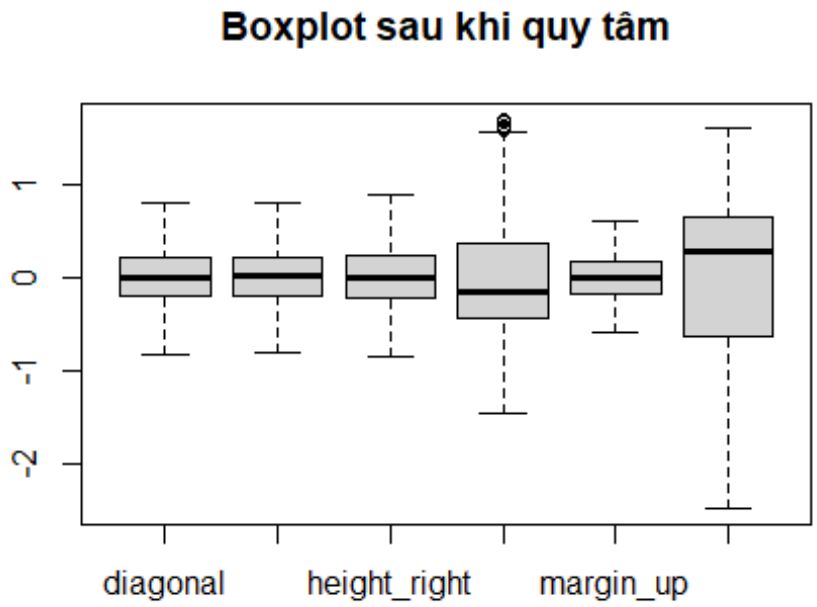
```
##     diagonal  height_left height_right   margin_low    margin_up
length
##   0.09004493   0.08648180   0.09918341   0.38061062   0.05167507   0.72
596627
```

```
boxplot(fake_clean[,-1], main = "Boxplot")
```



Nhận xét: Ta thấy phương sai của các biến nhỏ và không chênh lệch nhiều nên ta có thể phân tích dữ liệu này, không cần chuẩn hóa.

```
sc_fake = as.data.frame(scale(fake_clean[,-1],scale = F))
boxplot(sc_fake, main = "Boxplot sau khi quy tâm")
```

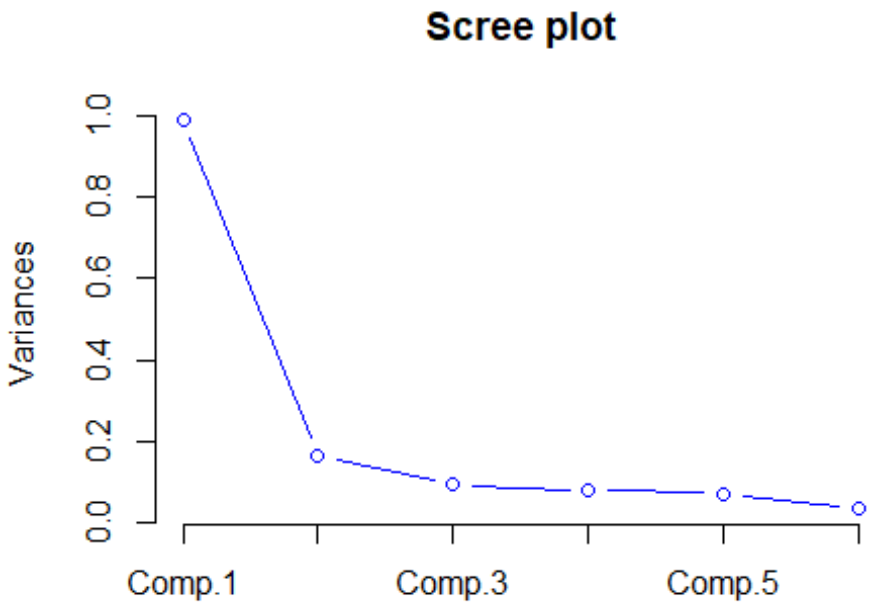


❖ Chọn thành phần chính giữ lại

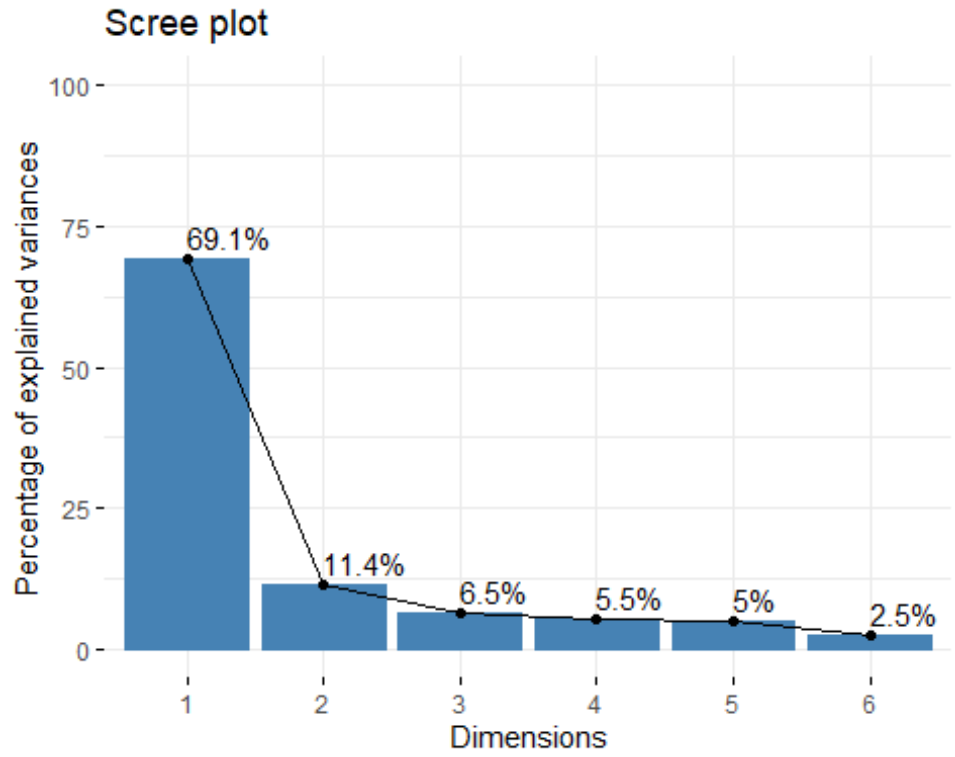
```
pca_fake = princomp(sc_fake)
summary(pca_fake)

## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4     Co
mp.5
## Standard deviation     0.9951361 0.4045469 0.30532383 0.28131806 0.2668
8002
## Proportion of Variance 0.6910912 0.1142111 0.06505667 0.05522881 0.0497
0528
## Cumulative Proportion  0.6910912 0.8053023 0.87035895 0.92558776 0.9752
9305
##                            Comp.6
## Standard deviation     0.18815874
## Proportion of Variance 0.02470695
## Cumulative Proportion  1.00000000

#screeplot
screeplot(pca_fake,type="lines",col="blue", main="Scree plot")
```




```
fviz_eig(pca_fake, addlabels = T, ylim = c(0,100))
```



Nhận xét: Qua đồ thị scree-plot ta thấy sự thay đổi rõ ràng nhất ở thành phần chính thứ nhất (PC1), PC1 đóng góp 69.1% vào phương sai suy rộng. Bên cạnh đó, thành phần

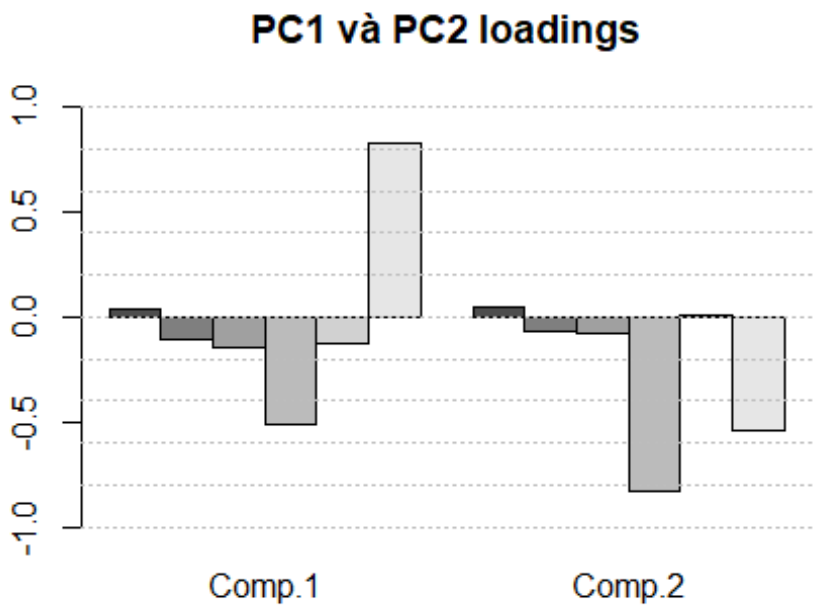
chính thứ hai (PC2) đóng góp 11.4% vào phương sai, cả PC1 và PC2 giải thích được 80.5%. Do đó, ta sẽ giữ lại hai thành phần chính đầu tiên.

❖ Loadings

```
Load_fake = pca_fake$loadings
Load_fake
```

```
##
## Loadings:
##              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## diagonal                    0.836  0.516  0.176
## height_left  -0.105         0.412 -0.375 -0.818
## height_right -0.150         0.349 -0.736  0.547
## margin_low   -0.516 -0.832         0.190
## margin_up    -0.129                              -0.987
## length        0.826 -0.543                       -0.101
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
## Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000
```

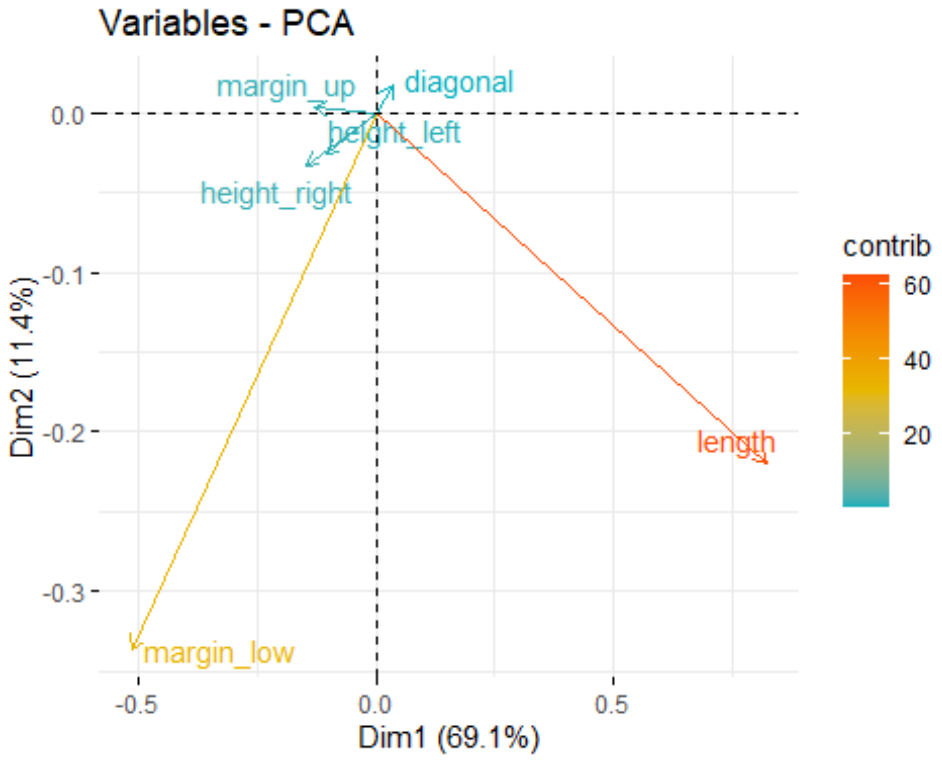
```
barplot(Load_fake[,1:2],beside = T, ylim = c(-1,1), main = "PC1 và PC2 loadings")
abline(h = seq(-1,1, by = 0.2), col = "gray",lty = "dotted")
```



Nhận xét:

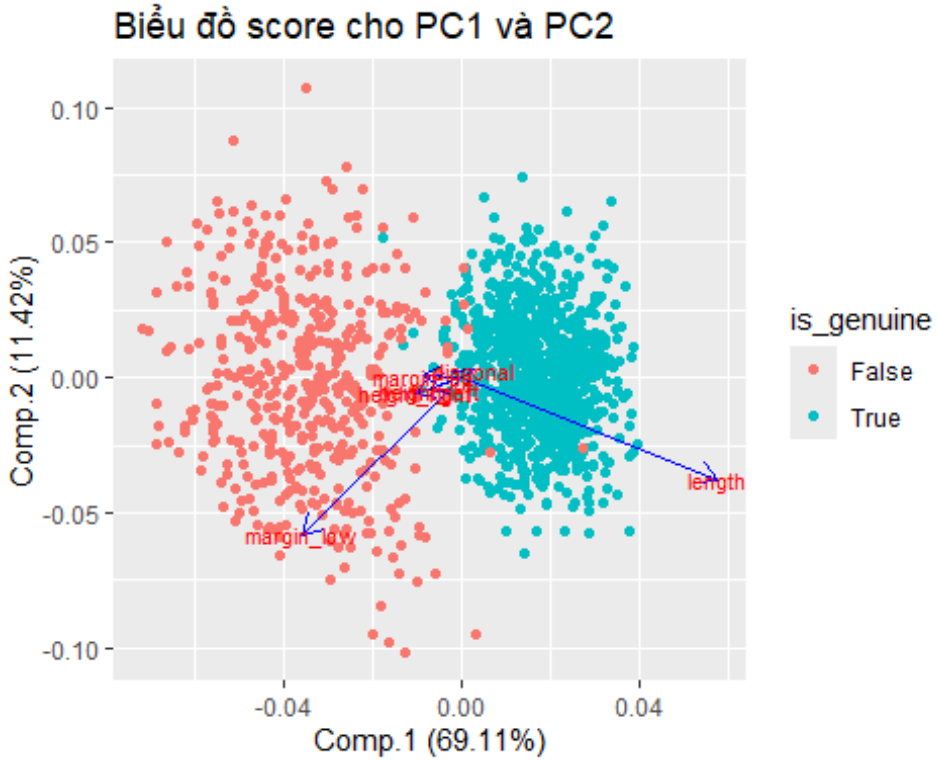
1. Thành phần chính thứ nhất (PC1): Có 2 biến có độ lớn trọng số khá lớn, lớn nhất là `length` (0.826), tiếp đến là `margin_low` (-0.516). Trong đó, 1 biến có trọng số dương và 1 biến có trọng số âm, thể hiện sự tương quan nghịch của hai biến này. PC1 có thể giải thích về kích thước lề và chu vi hóa đơn.
2. Thành phần chính thứ hai (PC2): Có biến `margin_low` có độ lớn trọng số lớn nhất và mang giá trị âm (-0.832), lớn hơn PC1. Do đó, PC2 giải thích rõ hơn về kích thước lề của hóa đơn.

```
fviz_pca_var(pca_fake,
             axes = c(1,2),
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
```



❖ Scores

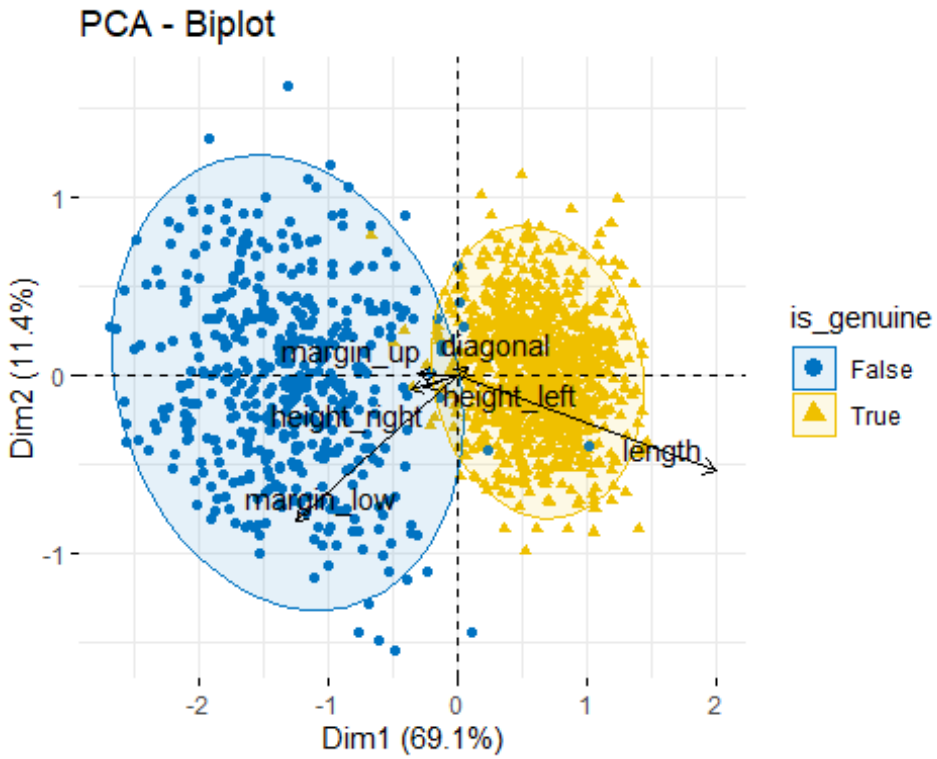
```
autoplot(pca_fake, loadings = TRUE, loadings.colour = 'blue', loadings.lab
el = TRUE, loadings.label.size = 3,
         colour = "is_genuine", data = fake_clean, main="Biểu đồ score cho
PC1 và PC2")
```




```
table(fake_clean$is_genuine)

##
## False  True
##   451   959

fviz_pca_biplot(pca_fake,
                col.ind = fake_clean$is_genuine, palette = "jco",
                addEllipses = TRUE, label = "var",
                col.var = "black", repel=TRUE,
                legend.title = "is_genuine")
```



Nhận xét: Nhìn vào biểu đồ ta thấy hóa đơn thật là những hóa đơn có PC1 dương, ta có thể dự đoán hóa đơn thật dựa trên chu vi. Hóa đơn giả là những hóa đơn có PC1 âm, có thể dự đoán được hóa đơn giả dựa vào kích thước lề.

2. PHÂN TÍCH DỮ LIỆU DỰ ĐOÁN HÓA ĐƠN GIẢ BẰNG PHƯƠNG PHÁP PHÂN TÍCH NHÂN TỐ

❖ Thư viện

```
library(janitor)
library(tidyverse)
library(FactoMineR)
library(ggplot2)
library(ggbiplot)
library(ggfortify)
library(dplyr)
library(psych)
library(factoextra)  #fviz
library(base)
library(naniar)      #gg_miss_var

## Warning: package 'naniar' was built under R version 4.4.1

library(EFA.dimensions)  #FACTORABILITY

## Warning: package 'EFA.dimensions' was built under R version 4.4.1
```

❖ Dữ liệu

```
# Nguồn dữ liệu: https://www.kaggle.com/datasets/alexandrepetit881234/fake-bills

fake = read.csv("D:/TKNC/fake_bills.csv", header = TRUE, sep = ";")
head(fake)
```

```
##   is_genuine diagonal height_left height_right margin_low margin_up length
## 1       True   171.81      104.86       104.95       4.52      2.89 112.83
## 2       True   171.46      103.36       103.66       3.77      2.99 113.09
## 3       True   172.69      104.48       103.50       4.40      2.94 113.16
## 4       True   171.36      103.91       103.94       3.62      3.01 113.51
## 5       True   171.73      104.28       103.46       4.04      3.48 112.54
## 6       True   172.17      103.74       104.08       4.42      2.95 112.81
```

```
dim(fake)
```

```
## [1] 1500    7
```

```
str(fake)
```

```
## 'data.frame':    1500 obs. of  7 variables:
##  $ is_genuine  : chr  "True" "True" "True" "True" ...
##  $ diagonal    : num  172 171 173 171 172 ...
##  $ height_left : num  105 103 104 104 104 ...
##  $ height_right: num  105 104 104 104 103 ...
##  $ margin_low  : num  4.52 3.77 4.4 3.62 4.04 4.42 4.58 3.98 4 4.04 ...
##  $ margin_up   : num  2.89 2.99 2.94 3.01 3.48 2.95 3.26 2.92 3.25 3.25 ...
##  $ length      : num  113 113 113 114 113 ...
```

❖ Mô tả dữ liệu

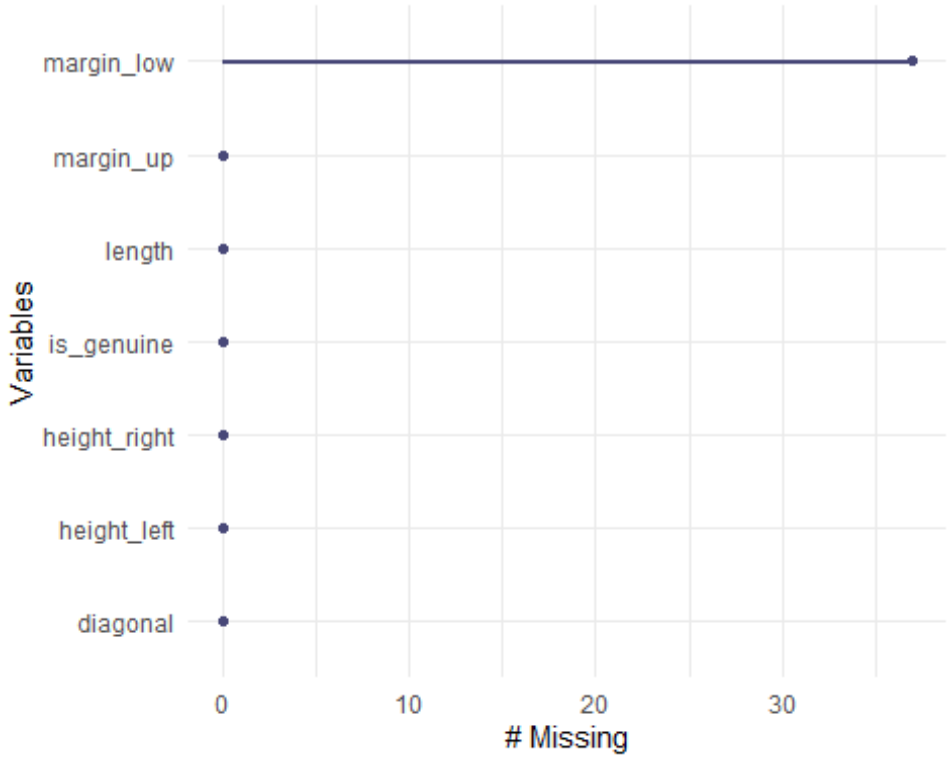
Dữ liệu nghiên cứu về hoá đơn giả và thật với 1500 quan trắc và 7 biến:

1. `is_genuine`: Hóa đơn có phải là hàng thật không? Đúng/sai
2. `diagonal`: Số đo đường chéo tính bằng mm
3. `height_left`: chiều cao của cạnh trái tính bằng mm
4. `height_right`: chiều cao của phía bên phải tính bằng mm

5. `margin_low`: lề dưới tính bằng mm
6. `margin_up`: lề trên tính bằng mm
7. `length`: tổng chiều dài tính bằng mm

❖ Kiểm tra và xử lý giá trị khuyết

```
# Kiểm tra giá trị khuyết
gg_miss_var(fake)
```



Nhận xét: Ta thấy biến có giá trị khuyết là `margin_low`.

```
# Loại bỏ các giá trị khuyết
fake_new = na.omit(fake)
head(fake_new)
```

```
##   is_genuine diagonal height_left height_right margin_low margin_up len
gth
## 1       True   171.81      104.86       104.95       4.52      2.89 112
.83
## 2       True   171.46      103.36       103.66       3.77      2.99 113
.09
## 3       True   172.69      104.48       103.50       4.40      2.94 113
.16
## 4       True   171.36      103.91       103.94       3.62      3.01 113
.51
## 5       True   171.73      104.28       103.46       4.04      3.48 112
```

```
.54
## 6       True    172.17      103.74       104.08       4.42       2.95 112
.81
```

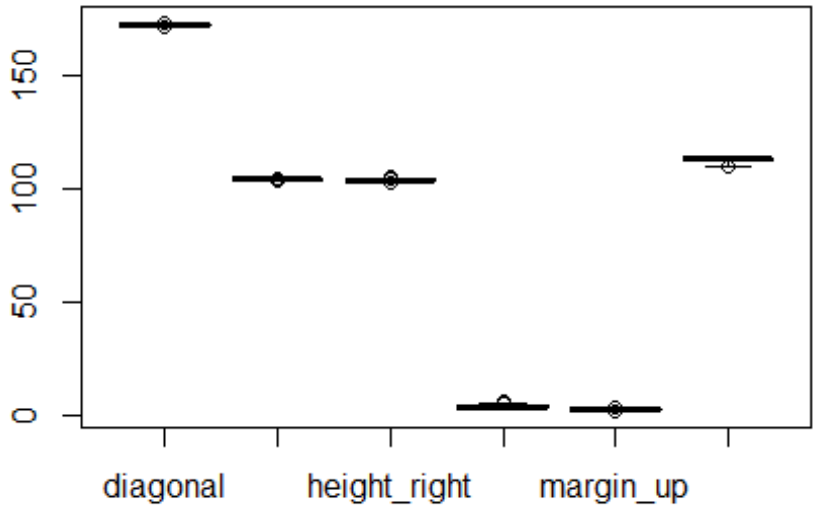
```
dim(fake_new)
```

```
## [1] 1463    7
```

Nhận xét: Sau khi loại bỏ các giá trị khuyết dữ liệu còn lại 1463 quan trắc và 7 biến.

❖ Kiểm tra giá trị ngoại lai

```
boxplot(fake_new[,-1])
```



Nhận xét: Ta thấy dữ liệu có nhiều điểm ngoại lai, vì vậy ta cần xử lý ngoại lai. Vì dữ liệu khá lớn nên ta sẽ xoá các ngoại lai.

❖ Kiểm tra và xử lý giá trị ngoại lai

```
# Kiểm tra giá trị ngoại lai

Check_Outliers = function(Variable){
  Q1 = quantile(Variable,0.25)
  Q3 = quantile(Variable,0.75)
  IQR = Q3 - Q1
  return (Variable <= (Q1 - (1.5 * IQR)) | Variable >= (Q3 + (1.5 * IQR)))
```

```
}
mat_outlier = (fake_new[,-1]) %>%
  mutate(across(everything(), Check_Outliers))
dat_outlier = (fake_new[,-1])[rowSums(mat_outlier) > 0,]
head(dat_outlier)
```

```
##     diagonal height_left height_right margin_low margin_up length
## 1     171.81      104.86       104.95       4.52      2.89 112.83
## 78    171.84      104.09       103.03       4.11      2.77 113.18
## 177   171.75      103.63       102.97       4.46      2.77 113.22
## 194   172.35      103.73       102.95       4.49      3.37 112.49
## 225   172.12      103.20       103.92       4.46      3.26 113.44
## 293   172.09      103.14       103.81       4.88      3.01 113.69
```

```
dim(dat_outlier)
```

```
## [1] 53  6
```

Nhận xét: Có 53 quan trắc có chứa giá trị ngoại lai, do đó ta sẽ loại bỏ các quan trắc đó.

```
# Loại bỏ giá trị ngoại lai
del = which(rowSums(mat_outlier)>0)
del
```

```
##    1   78  177  194  225  293  523  665  730  762  829  843 1023 1024 1
028 1030
##    1   77  174  191  221  286  510  650  713  743  808  822  994  995
999 1001
## 1032 1042 1054 1076 1083 1091 1093 1111 1125 1134 1135 1143 1151 1170 1
200 1255
## 1003 1013 1025 1047 1053 1061 1063 1081 1094 1103 1104 1112 1120 1139 1
168 1223
## 1271 1278 1291 1311 1322 1323 1332 1346 1349 1354 1356 1383 1389 1421 1
427 1442
## 1239 1246 1259 1278 1288 1289 1298 1312 1314 1319 1321 1348 1354 1386 1
392 1405
## 1454 1460 1465 1474 1485
## 1417 1423 1428 1437 1448
```

```
length(del)
```

```
## [1] 53
```

```
fake_cleand = fake_new[-del,]
head(fake_cleand)
```

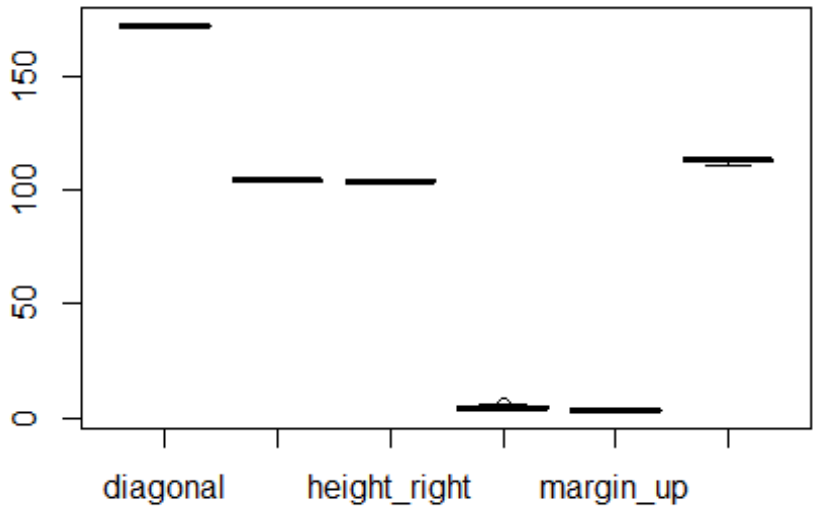
```
##   is_genuine diagonal height_left height_right margin_low margin_up len
gth
## 2       True   171.46      103.36       103.66       3.77      2.99 113
.09
## 3       True   172.69      104.48       103.50       4.40      2.94 113
```

```
.16
## 4       True   171.36      103.91       103.94       3.62       3.01 113
.51
## 5       True   171.73      104.28       103.46       4.04       3.48 112
.54
## 6       True   172.17      103.74       104.08       4.42       2.95 112
.81
## 7       True   172.34      104.18       103.85       4.58       3.26 112
.81
```

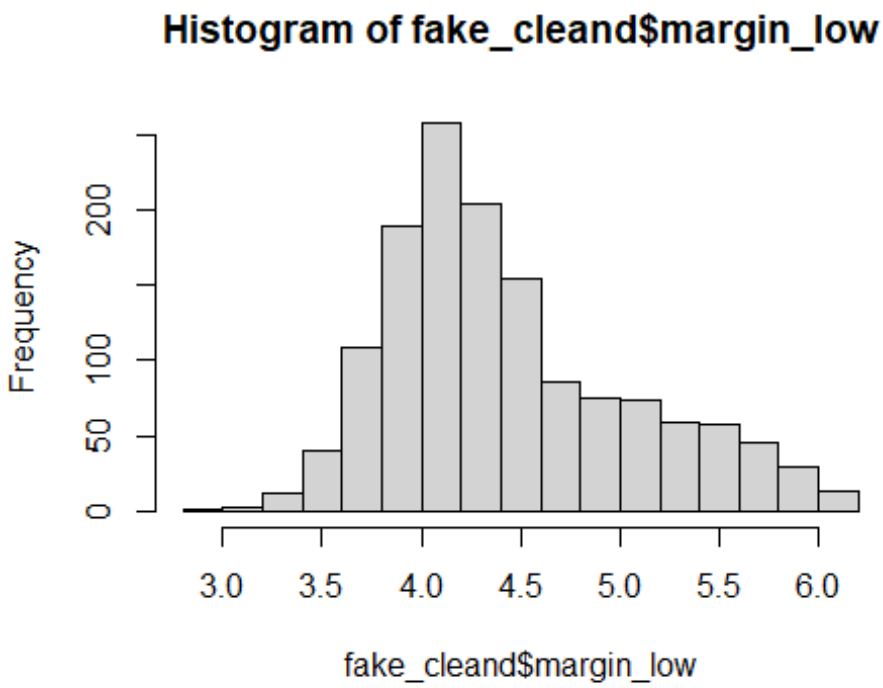
```
dim(fake_cleand)
```

```
## [1] 1410    7
```

```
boxplot(fake_cleand[,-1])
```




```
hist(fake_cleand$margin_low)
```



Nhận xét: Sau khi loại bỏ các quan trắc có chứa giá trị ngoại lai, vẽ lại boxplot ta thấy vẫn còn ngoại lai của biến `margin_low` nhưng biểu đồ Histogram của biến đó có dạng xấp xỉ dạng chuẩn, nên ta sẽ không loại bỏ tiếp các ngoại lai đó.

```
apply(fake_cleand,2, sd)
```

```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), n
a.rm =
## na.rm): NAs introduced by coercion
```

```
##   is_genuine     diagonal  height_left height_right   margin_low    mar
gin_up
##           NA    0.3000749    0.2940779    0.3149340    0.6169365    0.2
273215
##       length
##    0.8520365
```

Nhận xét: Dữ liệu có độ lệch chuẩn thấp, hay sự giao động của các biến dữ liệu không nhiều. Nên dữ liệu đã sạch và ta có thể phân tích thống kê.

- ❖ Ma trận hệ số tương quan

```
(R_2=cor(fake_cleand[,-1]))
```

```
##                diagonal height_left height_right margin_low   margin_u
p
## diagonal      1.00000000  0.02771492  -0.01749189 -0.1078682 -0.0482203
2
## height_left   0.02771492  1.00000000   0.21682712  0.2961057  0.2375650
6
## height_right -0.01749189  0.21682712   1.00000000  0.3893127  0.3028484
4
## margin_low   -0.10786816  0.29610570   0.38931271  1.0000000  0.4401619
2
## margin_up    -0.04822032  0.23756506   0.30284844  0.4401619  1.0000000
0
## length        0.09627867 -0.30230542  -0.40364691 -0.6665901 -0.5239451
0
##                   length
## diagonal      0.09627867
## height_left  -0.30230542
## height_right -0.40364691
## margin_low   -0.66659009
## margin_up    -0.52394510
## length        1.00000000
```

```r
corr.test(fake_cleand[,-1],use="complete.obs")
```

```
## Call:corr.test(x = fake_cleand[, -1], use = "complete.obs")
## Correlation matrix
##              diagonal height_left height_right margin_low margin_up len
gth
## diagonal         1.00        0.03        -0.02      -0.11     -0.05   0
.10
## height_left      0.03        1.00         0.22       0.30      0.24  -0
.30
## height_right    -0.02        0.22         1.00       0.39      0.30  -0
.40
## margin_low      -0.11        0.30         0.39       1.00      0.44  -0
.67
## margin_up       -0.05        0.24         0.30       0.44      1.00  -0
.52
## length           0.10       -0.30        -0.40      -0.67     -0.52   1
.00
## Sample Size
## [1] 1410
## Probability values (Entries above the diagonal are adjusted for multipl
e tests.)
##              diagonal height_left height_right margin_low margin_up len
gth
## diagonal         0.00         0.6          0.6          0      0.21
0
## height_left      0.30         0.0          0.0          0      0.00
0
## height_right     0.51         0.0          0.0          0      0.00
0
```

```
## margin_low       0.00         0.0          0.0          0      0.00
0
## margin_up        0.07         0.0          0.0          0      0.00
0
## length           0.00         0.0          0.0          0      0.00
0
##
##  To see confidence intervals of the correlations, print with the short=
FALSE option
```

```
FACTORABILITY(fake_cleand[,-1],)
```

```
##
##
## Three methods of assessing the factorability of a correlation matrix or
raw data set:

##
## Specified kind of correlations for this analysis: Pearson

##
##
## The determinant of the correlation matrix should be > 0.00001 for facto
rability.

##
## The determinant is 0.2745664 which is > 0.00001, indicating factorabili
ty.

##
##
## The Bartlett test of whether a correlation matrix is significantly diff
erent

## from an identity matrix (wherein all of the off-diagonal elements are z
ero):

##
## chisq = 1817.55760310169    df= 15     p = 0

##
## A significant difference is required for factorability.

##
##
## The Kaiser-Meyer-Olkin measure of sampling adequacy (MSA):

##              Variable MSA
## diagonal             0.66
## height_left          0.88
## height_right         0.88
## margin_low           0.74
## margin_up            0.83
## length               0.71
```

```
##
## The overall measure of sampling adequacy (MSA) = 0.78

##
## The Kaiser & Rice (1974) interpretation guidelines for MSA values:

##
##     KMO >= .9 is marvelous
##     KMO in the .80s is mertitorious
##     KMO in the .70s is middling
##     KMO in the .60s is mediocre
##     KMO in the .50s is miserable
##     KMO < .5 is unacceptable

##
## Consider excluding items with KMO values < .5 and then re-run the FACTO
RABILITY analyses.

##
## The overall KMO coefficient indicates the proportion of

## variance in the variables that might be caused by underlying

## factors. If the variables share common factors, then the

## overall KMO coefficient should be close to 1.0. The overall

## KMO indicates the extent to which there is at least one

## latent factor underlying the variables. The overall KMO

## index is considered particularly meaningful when the cases

## to variables ratio is less than 1:5. The KMO coefficient for

## a variable is a kind of summary index of how much a

## variable overlaps with the other variables.
```

Nhận xét: Với kết quả

```
The determinant of the correlation matrix should be > 0.00001 for factorability.

The determinant is 0.2732442 which is > 0.00001, indicating factorability.
```

Để các biến có tương quan với nhau thì giá trị định thức của ma trận tương quan phải > 0.000001, với kết quả định thức của ma trận hệ số tương quan là 0.2732442 > 0.000001, nên ta nói các biến có tương quan với nhau.

- ❖ Kiểm tra dữ liệu có đủ để phân tích nhân tố:

Sử dụng phương pháp kiểm định KMO

```
KMO(fake_cleand[,-1])
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = fake_cleand[, -1])
## Overall MSA =  0.78
## MSA for each item =
##     diagonal  height_left height_right   margin_low    margin_up
length
##         0.66         0.88         0.88         0.74         0.83
0.71
```

Nhận xét: Giá trị KMO của dữ liệu là 0.78 > 0.6, đây là một kết quả khá tốt cho dữ liệu, và các giá trị KMO các biến riêng lẻ cũng cao trên 0.6. Nên ta nói dữ liệu này đủ để phân tích nhân tố.

- ❖ Xác định số nhân tố:

Sử dụng kiểm định Kaiser để xác định số nhân tố:

```
ev = eigen(R_2)
print(ev$values)
```

```
## [1] 2.5753292 1.0227183 0.7997552 0.7129039 0.5665932 0.3227002
```

Nhận xét: Bằng kiểm định Kaiser, ta sẽ giữ lại các biến có giá trị riêng lớn hơn 1, vì vậy từ kết quả trên, ta sẽ giữ lại 2 nhân tố, và thực hiện phân tích 2 nhân tố này.

- ❖ Phân tích nhân tố với m = 2:

Ta phân tích 2 nhân tố

```
fa_fake = fa(fake_cleand[,-1],nfactors = 2, rotate = "varimax", residuals
= TRUE, fm ="ml")
fa_fake
```

```
## Factor Analysis using method =  ml
## Call: fa(r = fake_cleand[, -1], nfactors = 2, rotate = "varimax", resid
uals = TRUE,
##     fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                ML1   ML2   h2   u2 com
## diagonal     -0.03  0.24 0.06 0.94 1.0
## height_left   0.45  0.15 0.23 0.77 1.2
## height_right  0.51 -0.03 0.26 0.74 1.0
```

```
## margin_low    0.72 -0.27 0.59 0.41 1.3
## margin_up     0.58 -0.15 0.36 0.64 1.1
## length       -0.79  0.35 0.75 0.25 1.4
##
##                        ML1  ML2
## SS loadings           1.95 0.30
## Proportion Var        0.32 0.05
## Cumulative Var        0.32 0.37
## Proportion Explained  0.87 0.13
## Cumulative Proportion 0.87 1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  15  with the objective function =  1.29 with Chi Squar
e =  1817.56
## df of  the model are 4  and the objective function was  0.01
##
## The root mean square of the residuals (RMSR) is  0.01
## The df corrected root mean square of the residuals is  0.02
##
## The harmonic n.obs is  1410 with the empirical chi square  4.81  with p
rob <  0.31
## The total n.obs was  1410  with Likelihood Chi Square =  8.22  with pro
b <  0.084
##
## Tucker Lewis Index of factoring reliability =  0.991
## RMSEA index =  0.027  and the 90 % confidence intervals are  0 0.054
## BIC =  -20.79
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                ML1   ML2
## Correlation of (regression) scores with factors   0.88  0.50
## Multiple R square of scores with factors          0.77  0.25
## Minimum correlation of possible factor scores     0.55 -0.50
```

Nhận xét:

- Nhân tố 1, được giải thích bởi các biến có trọng số loadings > 0.3 gồm các biến `height_left`, `height_weight`, `margin_low`, `margin_up`, `lenght`. Trong đó biến `lenght` mang tỷ trọng âm và các biến còn lại mang tỷ trọng dương. Bốn biến `height_left`, `height_weight`, `margin_low`, `margin_up` liên quan đến *chiều cao và lề* nên ta có tể nói nhân tố 1 chỉ sự đối lặp giữa *chiều cao, lề* với *độ dài* của hoá đơn.

- Ở nhân tố 2, ta thấy trọng số loading > 0.3 gồm biến `margin_low` và `length`, mà 2 biến này đã được giải thích cho nhân tố 1 và tỷ trọng của 2 biến này ở nhân tố 1 cũng lớn hơn, nên ta không dùng lại 2 biến này cho nhân tố 2. Ta thấy còn biến `diagonal` chưa được đưa vào giải thích cho nhân tố 1, nhưng trọng số loadings của biến này ở cả hai nhân tố đều thấp, và có giá trị **h2** rất thấp, dẫn đến việc biến này không đóng góp nhiều vào việc giải thích phương sai. Từ đó ta nên loại biến `diagonal` khỏi mô hình.

- Bằng kiểm định Kaiser, ta sẽ nhận các nhân tố có trị riêng lớn hơn 1, vì vậy ta giữ lại nhân tố 1 đủ để giải thích cho phương sai tổng thể.

- Theo kiểm định Likelihood với giá trị p_value <0.056, thì ta sẽ không đủ cơ sở để bác bỏ H_0 với giả định mô hình 2 nhân tố là hợp lý và bac bỏ H_1 với giả định mô hình 2 nhân tố là không hợp lý. Vậy mô hình 2 nhân tố là hợp lý.

- Ở phân tích trên ta không dùng biến `is_genuine` vì biến này đã khảo xác tính thật giả của hoá đơn, nên không cần đưa vào mô hình phân tích. Thay vào đó biến này có thể dùng để thiết lập mô hình để kiểm tra tính thật giả của hoá đơn bằng các nhân tố đã phân tích như ở trên.

TÀI LIỆU THAM KHẢO

1. [1] Bác sĩ Chuyên khoa I Nguyễn Thị Cẩm Tú. (2021). *Căng thẳng là gì? nguyên nhân, triệu chứng và cách điều trị hiệu quả*. BookingCare. https://bookingcare.vn/cam-nang/cang-thang-la-gi-nguyen-nhan-trieu-chung-va-cach-dieu-tri-hieu-qua-p478.html
2. [2] *Sử dụng hóa đơn giả bị xử lý như thế nào?*. Thư Viện Pháp Luật. https://thuvienphapluat.vn/chinh-sach-phap-luat-moi/vn/ho-tro-phap-luat/tu-van-phap-luat/46767/su-dung-hoa-don-gia-bi-xu-ly-nhu-the-nao
3. Ung Dung Thong ke. (2021, September 8). *Xử lý dữ liệu khuyết với phân tích thành phần chính: Thực hành trên R* [Video]. YouTube. https://www.youtube.com/watch?v=2XcMzpeHNZU
4. Sai, C. (2020, August 20). *Sử dụng thống kê để xác định và loại bỏ dữ liệu ngoại lai cho machine learning trong R và Python*. Khoa Học Dữ Liệu. https://svcuong.github.io/post/remove-outliers/
5. Spencer Pao. (2021, March 1). *Understanding and applying factor analysis in R* [Video]. YouTube. https://www.youtube.com/watch?v=kbJMz0KzMnI
6. Mike Crowson. (2023, May 17). *Determining number of factors: Exploratory factor analysis (EFA) using RStudio and EFA.dimensions* [Video]. YouTube. https://www.youtube.com/watch?v=qy1psjlOqjA
7. Dhaval Maheta (DM). (2022, April 28). *18. Factor Analysis in R || Dr. Dhaval Maheta* [Video]. YouTube. https://www.youtube.com/watch?v=n4LtDet48UA
8. Johnson. (2013). *Applied Multivariate Statistical Analysis*. Hoa Kỳ.
9. Nguyễn Thị Mộng Ngọc, Đinh Ngọc Thanh, Đặng Đức Trọng. (2023). *Thống Kê Nhiều Chiều*. Lưu hành nội bộ.

BẢNG PHÂN CÔNG CÔNG VIỆC

MSSV	Họ và tên	Nhiệm vụ	Mức độ hoàn thành
21110280	Đoàn Thị Kỳ Duyên	+ Phần 1: Phân tích nhân tố dữ liệu nghiên cứu về mức độ căng thẳng của sinh viên. Giải bài 6.33 trang 355-356 của sách Johnson, 2013. + Phần 2: Phân tích nhân tố dữ liệu dự đoán hóa đơn giả.	100%
21110281	Võ Thị Hồng Gấm	+ Phần 1: Phân tích thành phần chính dữ liệu nghiên cứu về mức độ căng thẳng của sinh viên. Giải bài 4.21 trang 205 của sách Johnson, 2013. + Phần 2: Phân tích thành phần chính dữ liệu dự đoán hóa đơn giả. + Tìm dữ liệu và trình bày word.	100%