Ứng dụng phân tích dữ liệu trong nhận diện hành vi nghiện mạng xã hội của sinh viên

Vũ Đức Anh

2025-05-15

# Đề tài: Ứng dụng phân tích dữ liệu trong nhận diện hành vi nghiện mạng xã hội của sinh viên

Báo cáo này tập trung phân tích hành vi nghiện mạng xã hội ở ba nhóm đối tượng: học sinh trung học phổ thông, sinh viên đại học và sinh viên đã tốt nghiệp đại học. Để khám phá đặc điểm và xu hướng sử dụng mạng xã hội. Bài sẽ áp dụng ba phương pháp phân tích dữ liệu chính: Phân tích thành phần chính PCAđược sử dụng để rút gọn và tổng hợp thông tin từ các biến định lượng, giúp làm nổi bật các chiều chính trong hành vi người dùng. Phân tích tương quan đa chiều CA được áp dụng nhằm làm rõ mối liên hệ giữa các biến phân loại, qua đó phát hiện các mô hình tiềm ẩn trong dữ liệu danh mục. Phân cụm K-means được triển khai để phân nhóm các cá nhân theo đặc điểm hành vi sử dụng mạng xã hội, từ đó nhận diện các nhóm có mức độ sử dụng và ảnh hưởng tâm lý khác nhau. Phân tích này không chỉ hỗ trợ mô tả thực trạng nghiện mạng xã hội trong giới trẻ, mà còn góp phần đề xuất hướng tiếp cận phù hợp trong việc hỗ trợ sức khỏe tinh thần và nâng cao nhận thức về việc sử dụng mạng xã hội một cách lành mạnh.

# 1. Phương pháp phân tích thành phần chính PCA

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(FactoMineR)  
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(psych)

##   
## Attaching package: 'psych'  
##   
## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(dplyr)  
library(clusterSim)

## Loading required package: cluster  
## Loading required package: MASS  
##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(ggplot2)  
data <- read.csv("Students Social Media Addiction.csv")  
dim(data)

## [1] 705 13

view(data)  
str(data)

## 'data.frame': 705 obs. of 13 variables:  
## $ Student\_ID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Age : int 19 22 20 18 21 19 23 20 18 21 ...  
## $ Gender : chr "Female" "Male" "Female" "Male" ...  
## $ Academic\_Level : chr "Undergraduate" "Graduate" "Undergraduate" "High School" ...  
## $ Country : chr "Bangladesh" "India" "USA" "UK" ...  
## $ Avg\_Daily\_Usage\_Hours : num 5.2 2.1 6 3 4.5 7.2 1.5 5.8 4 3.3 ...  
## $ Most\_Used\_Platform : chr "Instagram" "Twitter" "TikTok" "YouTube" ...  
## $ Affects\_Academic\_Performance: chr "Yes" "No" "Yes" "No" ...  
## $ Sleep\_Hours\_Per\_Night : num 6.5 7.5 5 7 6 4.5 8 6 6.5 7 ...  
## $ Mental\_Health\_Score : int 6 8 5 7 6 4 9 6 7 7 ...  
## $ Relationship\_Status : chr "In Relationship" "Single" "Complicated" "Single" ...  
## $ Conflicts\_Over\_Social\_Media : int 3 0 4 1 2 5 0 2 1 1 ...  
## $ Addicted\_Score : int 8 3 9 4 7 9 2 8 5 4 ...

summary(data)

## Student\_ID Age Gender Academic\_Level   
## Min. : 1 Min. :18.00 Length:705 Length:705   
## 1st Qu.:177 1st Qu.:19.00 Class :character Class :character   
## Median :353 Median :21.00 Mode :character Mode :character   
## Mean :353 Mean :20.66   
## 3rd Qu.:529 3rd Qu.:22.00   
## Max. :705 Max. :24.00   
## Country Avg\_Daily\_Usage\_Hours Most\_Used\_Platform  
## Length:705 Min. :1.500 Length:705   
## Class :character 1st Qu.:4.100 Class :character   
## Mode :character Median :4.800 Mode :character   
## Mean :4.919   
## 3rd Qu.:5.800   
## Max. :8.500   
## Affects\_Academic\_Performance Sleep\_Hours\_Per\_Night Mental\_Health\_Score  
## Length:705 Min. :3.800 Min. :4.000   
## Class :character 1st Qu.:6.000 1st Qu.:5.000   
## Mode :character Median :6.900 Median :6.000   
## Mean :6.869 Mean :6.227   
## 3rd Qu.:7.700 3rd Qu.:7.000   
## Max. :9.600 Max. :9.000   
## Relationship\_Status Conflicts\_Over\_Social\_Media Addicted\_Score   
## Length:705 Min. :0.00 Min. :2.000   
## Class :character 1st Qu.:2.00 1st Qu.:5.000   
## Mode :character Median :3.00 Median :7.000   
## Mean :2.85 Mean :6.437   
## 3rd Qu.:4.00 3rd Qu.:8.000   
## Max. :5.00 Max. :9.000

##xoá các biến thiếu  
pca\_new <- na.omit(data)

#Chọn các biến định lượng  
library(dplyr)  
pca\_new <- data %>%  
 dplyr::select(Avg\_Daily\_Usage\_Hours, Sleep\_Hours\_Per\_Night,  
 Mental\_Health\_Score, Conflicts\_Over\_Social\_Media, Addicted\_Score)

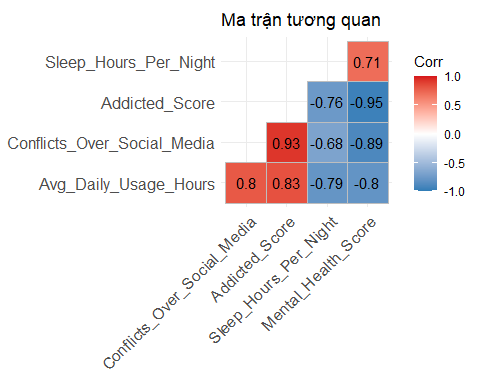
#tính ma trận tương quan giữa các biến trong pca\_new  
cor\_mat = cor(pca\_new, use = 'complete.obs')  
cor\_mat = cor(pca\_new, use = 'complete.obs')

Ma trận tương quan cho thấy các biến định lượng trong dữ liệu có mối liên hệ chặt chẽ. Điểm nghiện mạng xã hội (Addicted\_Score) có tương quan thuận mạnh với thời gian sử dụng và xung đột do mạng xã hội, nhưng tương quan âm rất mạnh với sức khỏe tâm thần và thời gian ngủ. Những kết quả này phản ánh rõ ràng rằng việc sử dụng mạng xã hội quá mức có thể liên quan đến các hậu quả tiêu cực về tâm lý và sinh hoạt cá nhân. Mức độ tương quan giữa các biến cũng cho thấy dữ liệu phù hợp để tiếp tục phân tích PCA.

library(ggplot2)  
library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 4.4.3

ggcorrplot(cor\_mat,   
 hc.order = TRUE,   
 type = "lower",  
 lab = TRUE,  
 colors = c("#2c7bb6", "white", "#d7191c"),  
 title = "Ma trận tương quan")

 Phân tích ma trận tương quan cho thấy thời gian sử dụng mạng xã hội trung bình mỗi ngày có mối tương quan thuận mạnh với mức độ nghiện và mâu thuẫn phát sinh từ mạng xã hội. Ngược lại, mức độ nghiện mạng xã hội có tương quan nghịch rất mạnh với sức khỏe tâm thần, đồng thời thời gian sử dụng mạng xã hội và mâu thuẫn xã hội đều liên quan âm với giấc ngủ và sức khỏe tâm thần. Bên cạnh đó, giấc ngủ đủ có liên hệ tích cực với sức khỏe tâm thần. Kết quả này cho thấy việc sử dụng mạng xã hội quá mức có thể ảnh hưởng tiêu cực đến giấc ngủ và sức khỏe tinh thần, đồng thời làm gia tăng các vấn đề xã hội liên quan.

#Kiểm định ba lét  
cortest.bartlett(pca\_new)

## R was not square, finding R from data

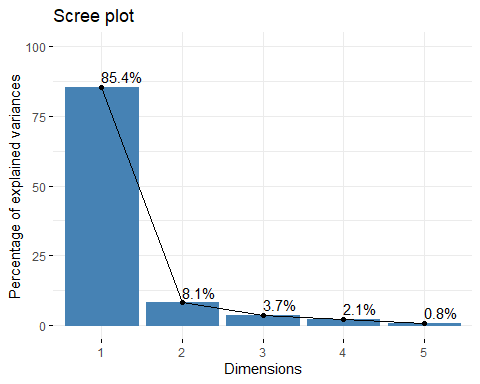
## $chisq  
## [1] 4671.597  
##   
## $p.value  
## [1] 0  
##   
## $df  
## [1] 10

Giả thuyết Kiểm định: H0: Ma trận hệ số tương quan của tổng thể là ma trận đơn vị H1: Ma trận hệ số tương quan của tổng thể không phải là ma trận đơn vị Ta có P-value= 0< 0,05 => Bác bỏ H0 => ma trận tương quan có ý nghĩa thống kê và phân tích thành phần chính là phù hợp với tập dữ liệu

#phân tích thành phần chính   
res.pca <- PCA(pca\_new, graph = FALSE)  
summary(res.pca)

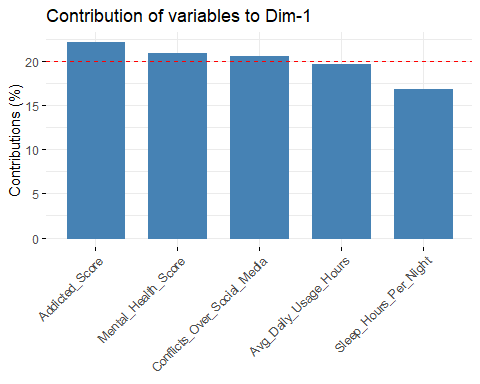
##   
## Call:  
## PCA(X = pca\_new, graph = FALSE)   
##   
##   
## Eigenvalues  
## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5  
## Variance 4.268 0.405 0.185 0.104 0.039  
## % of var. 85.363 8.098 3.692 2.072 0.776  
## Cumulative % of var. 85.363 93.461 97.152 99.224 100.000  
##   
## Individuals (the 10 first)  
## Dist Dim.1 ctr cos2 Dim.2 ctr  
## 1 | 1.093 | 0.862 0.025 0.622 | 0.039 0.001  
## 2 | 4.635 | -4.326 0.622 0.871 | -1.161 0.473  
## 3 | 2.964 | 2.875 0.275 0.941 | -0.307 0.033  
## 4 | 2.988 | -2.642 0.232 0.782 | -0.841 0.248  
## 5 | 1.289 | 0.028 0.000 0.000 | -0.711 0.177  
## 6 | 4.410 | 4.367 0.634 0.981 | -0.196 0.014  
## 7 | 5.604 | -5.430 0.980 0.939 | -1.130 0.447  
## 8 | 1.700 | 0.782 0.020 0.212 | -0.855 0.256  
## 9 | 2.385 | -1.812 0.109 0.577 | -1.263 0.559  
## 10 | 2.874 | -2.537 0.214 0.779 | -0.908 0.289  
## cos2 Dim.3 ctr cos2   
## 1 0.001 | -0.206 0.033 0.036 |  
## 2 0.063 | -0.893 0.613 0.037 |  
## 3 0.011 | -0.574 0.253 0.037 |  
## 4 0.079 | -0.798 0.489 0.071 |  
## 5 0.304 | -0.763 0.447 0.350 |  
## 6 0.002 | -0.145 0.016 0.001 |  
## 7 0.041 | -0.766 0.451 0.019 |  
## 8 0.253 | -0.025 0.000 0.000 |  
## 9 0.280 | -0.458 0.161 0.037 |  
## 10 0.100 | -0.597 0.274 0.043 |  
##   
## Variables  
## Dim.1 ctr cos2 Dim.2 ctr cos2   
## Avg\_Daily\_Usage\_Hours | 0.914 19.594 0.836 | -0.176 7.677 0.031 |  
## Sleep\_Hours\_Per\_Night | -0.847 16.818 0.718 | 0.493 60.075 0.243 |  
## Mental\_Health\_Score | -0.945 20.908 0.892 | -0.212 11.108 0.045 |  
## Conflicts\_Over\_Social\_Media | 0.937 20.553 0.877 | 0.254 15.972 0.065 |  
## Addicted\_Score | 0.972 22.127 0.944 | 0.145 5.167 0.021 |  
## Dim.3 ctr cos2   
## Avg\_Daily\_Usage\_Hours 0.362 70.828 0.131 |  
## Sleep\_Hours\_Per\_Night 0.192 19.892 0.037 |  
## Mental\_Health\_Score 0.094 4.829 0.009 |  
## Conflicts\_Over\_Social\_Media 0.009 0.045 0.000 |  
## Addicted\_Score -0.090 4.406 0.008 |

fviz\_eig(res.pca, addlabels = TRUE, ylim = c(0, 100))



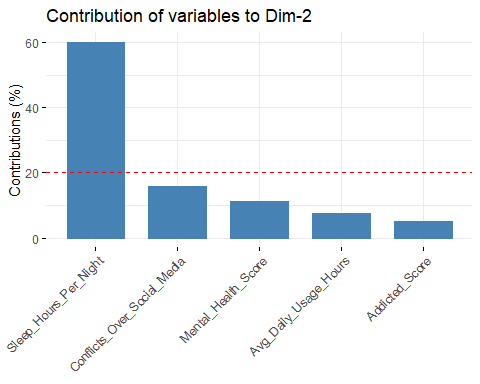
Biểu đồ Scree Plot cho thấy rằng thành phần chính đầu tiên Dim1 giải thích đến 85.4% phương sai của toàn bộ dữ liệu, trong khi thành phần chính thứ hai Dim2 chỉ giải thích 8.1% phương sai. Các thành phần còn lại giải thích lần lượt 3.7%, 2.1% và 0.8% phương sai. Như vậy, chỉ với hai thành phần chính đầu tiên, ta đã có thể giữ lại hơn 93.5% phương sai của dữ liệu gốc, điều này cho phép giảm số chiều từ 5 xuống còn 2 mà vẫn giữ lại hầu hết thông tin quan trọng. Việc giảm chiều như vậy không chỉ giúp đơn giản hóa mô hình mà còn giảm thiểu tính phức tạp tính toán, giúp quá trình phân tích dữ liệu trở nên hiệu quả hơn. Vậy trong bài sẽ quyết định sử dụng pc1 để phân tích thêm.

# Mức độ đóng góp của từng biến vào PC1  
fviz\_contrib(res.pca, choice = "var", axes = 1)



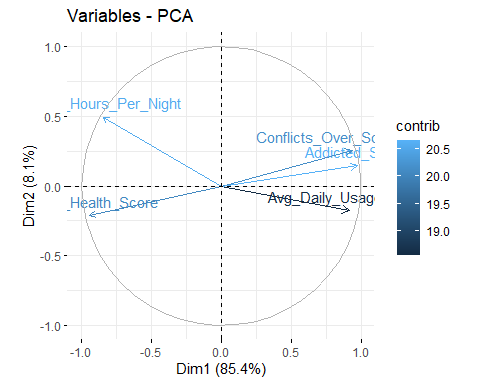
Biểu đồ cho thấy các biến Addicted\_Score, Mental\_Health\_Score, Conflicts\_Over\_Social\_Media, và Avg\_Daily\_Usage\_Hours có mức đóng góp đáng kể vào Dimension 1, đều trên mức 20%. Điều này cho thấy các yếu tố này là những nhân tố chính tạo nên trục Dimension 1, phản ánh rõ rệt về mức độ nghiện mạng xã hội và tác động của nó đến sức khỏe tâm lý. Ngược lại, biến Sleep\_Hours\_Per\_Night có mức đóng góp thấp hơn so với các biến còn lại, chứng tỏ rằng Dim 1 ít liên quan đến giấc ngủ mà chủ yếu phản ánh những khía cạnh liên quan đến mức độ sử dụng mạng xã hội và những xung đột từ việc lạm dụng mạng xã hội. Từ đó, có thể suy ra rằng Dimension 1 mang ý nghĩa chính là “Ảnh hưởng của nghiện mạng xã hội đến sức khỏe tâm lý”. Việc xác định đúng bản chất của Dimension 1 giúp phân tích chuyên sâu hơn về tác động của các yếu tố đến sức khỏe tinh thần và hành vi của người dùng mạng xã hội.

# Mức độ đóng góp của từng biến vào PC2  
fviz\_contrib(res.pca, choice = "var", axes = 2)



Biểu đồ đóng góp cho Dimension 2 cho thấy các biến Sleep\_hours\_per\_night là yếu tố chính định hình thành phần này, với mức đóng góp hơn 60%. Trong khi đó, các biến còn lại đều đóng góp ít hơn 20%.

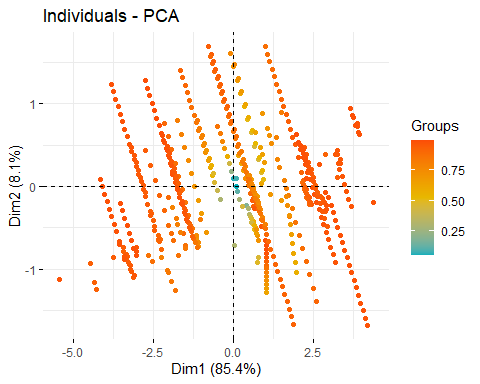
#Variables  
fviz\_pca\_var(res.pca, col.var = "contrib",)



gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")

Nhìn biểu đồ ta thấy Avg\_Daily\_Usage\_Hours cùng hướng với trục chính PC1, cho thấy đây là nhung biến chính chi phối mức độ nghiện. Ngược lại, Mental\_Health\_Score và Sleep\_Hours\_Per\_Night đi ngược chiều, phản ánh các yếu tố bảo vệ. Trong khi đó, Sleep\_Hours\_Per\_Night cũng góp phần mạnh cho PC2, giúp tách biệt yếu tố giấc ngủ với các yếu tố còn lại

#Biểu đồ pca ind dim1 dim2  
fviz\_pca\_ind(res.pca, geom.ind = "point", col.ind = "cos2",   
 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),  
 legend.title = "Groups", pvalueCutoff =1, qvalueCutoff = 1)



Biểu đồ cho thấy PCA đã mô tả tốt phần lớn cá thể trong không gian hai chiều đầu tiên. Dim1 chiếm tới 85.4% phương sai và rõ ràng là trục chính phân biệt hành vi.

# 2. Phân Tích CA

#làm crosstable  
library(gmodels)

## Warning: package 'gmodels' was built under R version 4.4.3

ca\_count<-CrossTable(data$Academic\_Level, data$Most\_Used\_Platform)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 705   
##   
##   
## | data$Most\_Used\_Platform   
## data$Academic\_Level | Facebook | Instagram | KakaoTalk | LINE | LinkedIn | Snapchat | TikTok | Twitter | VKontakte | WeChat | WhatsApp | YouTube | Row Total |   
## --------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Graduate | 87 | 82 | 0 | 0 | 21 | 0 | 51 | 17 | 12 | 12 | 43 | 0 | 325 |   
## | 16.189 | 9.365 | 5.532 | 5.532 | 13.235 | 5.993 | 5.630 | 0.727 | 7.563 | 3.740 | 13.170 | 4.610 | |   
## | 0.268 | 0.252 | 0.000 | 0.000 | 0.065 | 0.000 | 0.157 | 0.052 | 0.037 | 0.037 | 0.132 | 0.000 | 0.461 |   
## | 0.707 | 0.329 | 0.000 | 0.000 | 1.000 | 0.000 | 0.331 | 0.567 | 1.000 | 0.800 | 0.796 | 0.000 | |   
## | 0.123 | 0.116 | 0.000 | 0.000 | 0.030 | 0.000 | 0.072 | 0.024 | 0.017 | 0.017 | 0.061 | 0.000 | |   
## --------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## High School | 0 | 12 | 0 | 0 | 0 | 5 | 8 | 0 | 0 | 0 | 0 | 2 | 27 |   
## | 4.711 | 0.637 | 0.460 | 0.460 | 0.804 | 40.712 | 0.749 | 1.149 | 0.460 | 0.574 | 2.068 | 6.827 | |   
## | 0.000 | 0.444 | 0.000 | 0.000 | 0.000 | 0.185 | 0.296 | 0.000 | 0.000 | 0.000 | 0.000 | 0.074 | 0.038 |   
## | 0.000 | 0.048 | 0.000 | 0.000 | 0.000 | 0.385 | 0.052 | 0.000 | 0.000 | 0.000 | 0.000 | 0.200 | |   
## | 0.000 | 0.017 | 0.000 | 0.000 | 0.000 | 0.007 | 0.011 | 0.000 | 0.000 | 0.000 | 0.000 | 0.003 | |   
## --------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Undergraduate | 36 | 155 | 12 | 12 | 0 | 8 | 95 | 13 | 0 | 3 | 11 | 8 | 353 |   
## | 10.631 | 7.375 | 5.975 | 5.975 | 10.515 | 0.341 | 4.151 | 0.272 | 6.009 | 2.709 | 9.513 | 1.789 | |   
## | 0.102 | 0.439 | 0.034 | 0.034 | 0.000 | 0.023 | 0.269 | 0.037 | 0.000 | 0.008 | 0.031 | 0.023 | 0.501 |   
## | 0.293 | 0.622 | 1.000 | 1.000 | 0.000 | 0.615 | 0.617 | 0.433 | 0.000 | 0.200 | 0.204 | 0.800 | |   
## | 0.051 | 0.220 | 0.017 | 0.017 | 0.000 | 0.011 | 0.135 | 0.018 | 0.000 | 0.004 | 0.016 | 0.011 | |   
## --------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 123 | 249 | 12 | 12 | 21 | 13 | 154 | 30 | 12 | 15 | 54 | 10 | 705 |   
## | 0.174 | 0.353 | 0.017 | 0.017 | 0.030 | 0.018 | 0.218 | 0.043 | 0.017 | 0.021 | 0.077 | 0.014 | |   
## --------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

library(openxlsx)

## Warning: package 'openxlsx' was built under R version 4.4.3

openxlsx::write.xlsx(ca\_count, 'ca\_count.xlsx')  
#Đọc dữ liệu từ excel   
ca <- read.xlsx("ca\_count.xlsx", 1, colNames=TRUE,rowNames=TRUE)  
view(ca)

## kiểm định khi bình phương xem dữ liệu có phù hợp với phân tích tương ứng hay không  
 # Chisq test  
chisq <- chisq.test(ca)

## Warning in chisq.test(ca): Chi-squared approximation may be incorrect

chisq

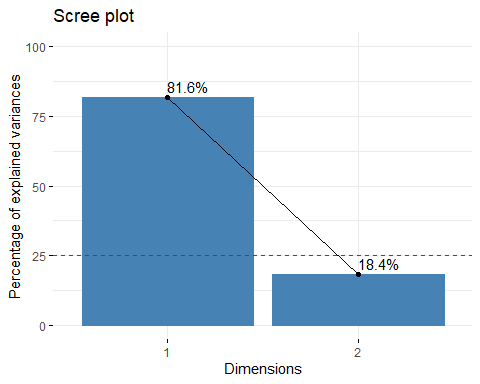
##   
## Pearson's Chi-squared test  
##   
## data: ca  
## X-squared = 216.15, df = 22, p-value < 2.2e-16

H₀: Hai biến là độc lập (không có mối liên hệ). H₁: Hai biến có liên hệ với nhau Ta có p-value <2.2e-16 và < 0.05 => Bác bỏ giả thuyết H0, hai biến có mối liên hệ với nhau

# Phân tích ca  
res.ca <- CA(ca, graph = FALSE)  
res.ca$eig

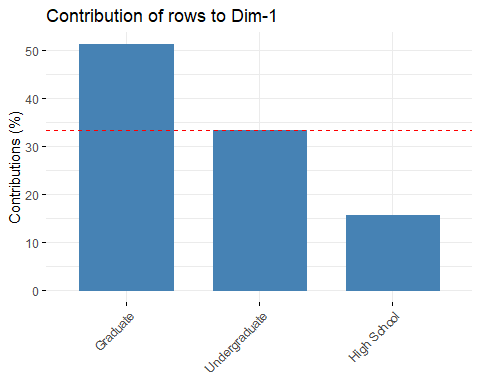
## eigenvalue percentage of variance cumulative percentage of variance  
## dim 1 0.25031291 81.64324 81.64324  
## dim 2 0.05628065 18.35676 100.00000

#vẽ biểu đồ scree plot  
library(ggplot2)  
fviz\_screeplot(res.ca, addlabels = TRUE, ylim = c(0, 100))+  
geom\_hline(yintercept=25, linetype=2, color="red")

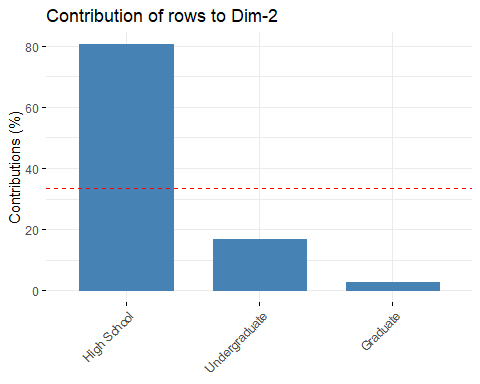


Kết quả phân tích Phân tích Thành phần Chính (CA) cùng với biểu đồ scree plot cho thấy hai thành phần đầu tiên đã giải thích được toàn bộ phương sai của bảng dữ liệu. Trong đó, riêng thành phần thứ nhất (Dim 1) chiếm đến 81.64%. Tuy nhiên, để phân tích CA đạt độ tin cậy cao hơn, thường yêu cầu Dim 1 chiếm ít nhất 90% phương sai. Vì vậy, trong bài, ta sẽ sử dụng cả hai thành phần Dim 1 và Dim 2 để đảm bảo bao phủ đầy đủ thông tin.

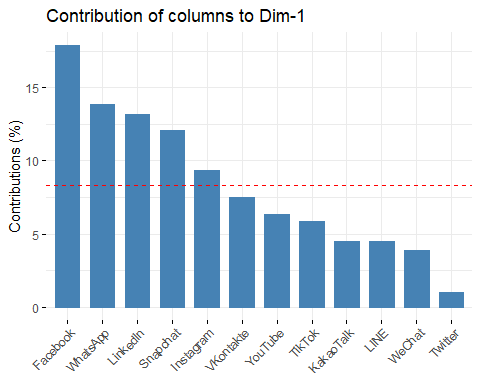
#Trực quan hoá xem cái nào đóng góp   
fviz\_contrib(res.ca, choice = "row", axes = 1)



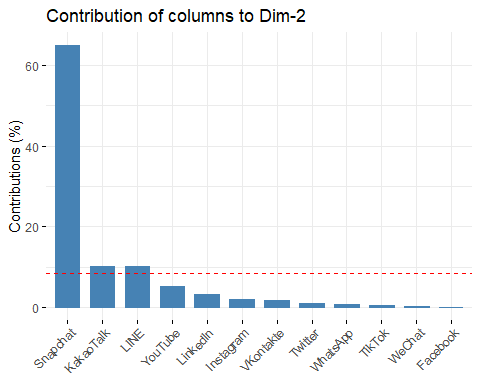
fviz\_contrib(res.ca, choice = "row", axes = 2)



fviz\_contrib(res.ca, choice = "col", axes = 1)

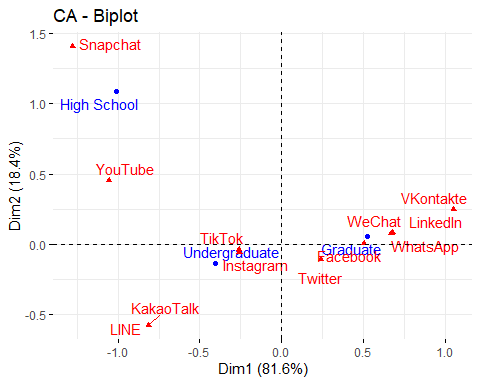


fviz\_contrib(res.ca, choice = "col", axes = 2)

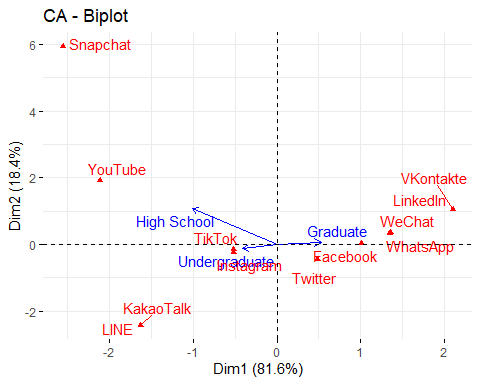
 Biểu đồ cho thấy Facebook, WhatsApp, và LinkedIn đóng góp nhiều nhất vào Dim-1, phản ánh sự phổ biến trong nhóm người trưởng thành và đã tốt nghiệp. Ngược lại, các nền tảng như Twitter, WeChat, và LINE có đóng góp thấp hơn, cho thấy ít đặc trưng hơn trong việc phân biệt nhóm đối tượng. Điều này cho thấy sự khác biệt trong cách sử dụng mạng xã hội giữa các nhóm người dùng có trình độ học vấn khác nhau.

Tuy nhiên, Biểu đồ cuối cùng lại cho thấy Snapchat là nền tảng có đóng góp lớn nhất vào Dim-2, nổi bật trong nhóm học sinh phổ thông. Ngược lại, các nền tảng như Facebook, LinkedIn, và WhatsApp có đóng góp nhỏ hơn, phù hợp với nhóm người đã tốt nghiệp. Từ đó ta có thể thấy rằng xu hướng người dùng đã tốt nghiệp, trường thành thường sẽ sử dụng các ứng dụng phục vụ cho công việc và đời sống, ngược lại với đối tượng trung học phổ thông, nhóm đối tượng có xu hướng sử dụng các ứng dụng có tính giải trí cao hơn.

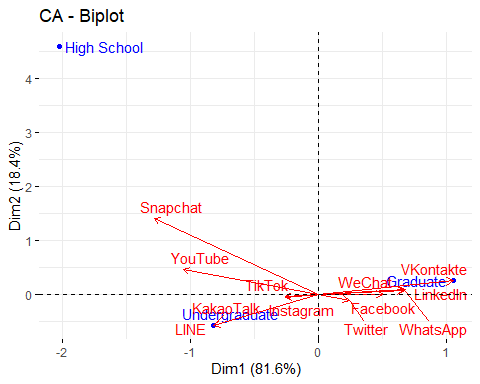
#Đồ thị bất đối xứng dòng   
fviz\_ca\_biplot(res.ca, repel = TRUE)



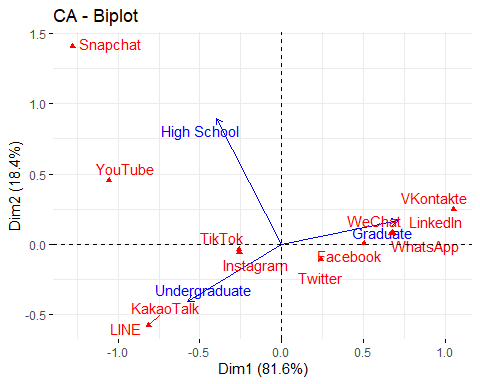
fviz\_ca\_biplot(res.ca,   
 map ="rowprincipal", arrow = c(TRUE, FALSE ),  
 repel = TRUE)



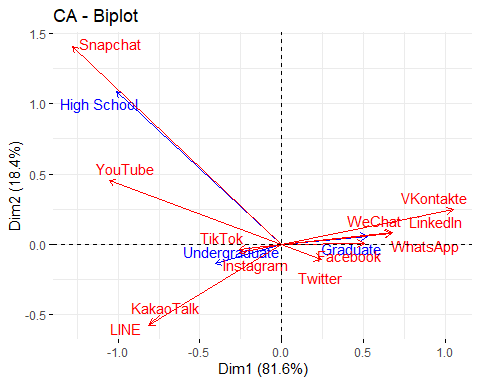
fviz\_ca\_biplot(res.ca,   
 map ="colprincipal", arrow = c(FALSE,TRUE ),  
 repel = TRUE)



fviz\_ca\_biplot(res.ca, map ="colgreen", arrow = c(TRUE, FALSE),  
 repel = TRUE)



fviz\_ca\_biplot(res.ca, arrow=c(TRUE, TRUE), repel=TRUE)



Nhìn vào biểu đồ trên, ta có thể nhận thấy các nền tảng như LinkedIn, VKontakte, WeChat, WhatsApp, Facebook, Twitter, Graduate nằm về phía bên phải của trục Dim1, cho thấy các biến này có mối liên hệ chặt chẽ với nhóm Graduate. Điều này có thể phản ánh xu hướng sử dụng các mạng xã hội và ứng dụng nhắm đến nhóm người đã tốt nghiệp đại học hoặc đang học sau đại học. Tuy nhiên linkedln có vẻ được ưa chuộng nhiều nhất so với các ứng dụng khác và có xu hướng được ít sử dụng twitter nhất Ngược lại, các ứng dụng như Snapchat, YouTube, TikTok, Instagram, LINE, KakaoTalk nằm về phía bên trái của trục Dim1, đồng thời gần với nhóm High School và Undergraduate, cho thấy các nền tảng này phổ biến hơn với học sinh trung học và sinh viên đại học. Snapchat được sử dụng rộng rãi nhất bởi nhóm highschool, nổi bật hơn hẳn các ứng dụng khác Trục Dim2 giúp phân biệt các nhóm và biến theo chiều dọc: ví dụ, Snapchat nằm cao hơn trên Dim2, có thể ám chỉ đặc điểm riêng biệt hoặc sự ưu tiên sử dụng của nhóm High School. Trong khi đó, LINE và KakaoTalk nằm thấp hơn trên Dim2, tương ứng với các nhóm và hành vi sử dụng khác nhau. Nhìn chung biểu đồ cho thấy sự khác biệt rõ ràng về xu hướng sử dụng mạng xã hội và ứng dụng giữa các nhóm học vấn khác nhau, với nhóm Graduate nghiêng về các nền tảng mạng xã hội chuyên nghiệp và truyền thống hơn, trong khi High School và Undergraduate ưa chuộng các ứng dụng giải trí và mạng xã hội phổ biến hơn.

# 3. Phương pháp phân cụm

library(factoextra)  
library(cluster)  
library(fpc)

## Warning: package 'fpc' was built under R version 4.4.3

library(dplyr)

#Chọn các biến định lượng  
cluster\_data <- pca\_new  
#Chuẩn hoá dữ liệu  
df.scaled <- scale(cluster\_data)

#Kiểm tra khả năng phân cụm  
hopkins <- get\_clust\_tendency(df.scaled, n=nrow(df.scaled)-1, graph = FALSE)  
hopkins

## $hopkins\_stat  
## [1] 0.9474291  
##   
## $plot  
## NULL

Ta có chỉ số hopkins = 0.9129166 > 0,5 => Vậy dữ liệu có xu hướng phân cụm rất mạnh.

#Phân cụm k-mean và trực quan hoá   
library(factoextra)  
kmean\_calc <- function(df, ...){  
 kmeans(df,scaled = ..., nstart = 30)  
}  
km2 <- kmean\_calc(df.scaled, 2)  
km3 <- kmean\_calc(df.scaled, 3)  
km4 <- kmean\_calc(df.scaled, 4)  
km5 <- kmean\_calc(df.scaled, 5)  
km6 <- kmean\_calc(df.scaled, 6)  
km7 <- kmean\_calc(df.scaled, 7)  
p1 <- fviz\_cluster(km2, data = df.scaled, ellipse.type = "convex") + theme\_minimal() + ggtitle("k = 2")   
p2 <- fviz\_cluster(km3, data = df.scaled, ellipse.type = "convex") + theme\_minimal() + ggtitle("k = 3")  
p3 <- fviz\_cluster(km4, data = df.scaled, ellipse.type= "convex") + theme\_minimal() + ggtitle("k = 4")  
p4 <- fviz\_cluster(km5, data = df.scaled, ellipse.type = "convex") + theme\_minimal() + ggtitle("k = 5")  
p5 <- fviz\_cluster(km6, data = df.scaled, ellipse.type = "convex") + theme\_minimal() + ggtitle("k = 6")  
p6 <- fviz\_cluster(km7, data = df.scaled, ellipse.type = "convex") + theme\_minimal() + ggtitle("k = 7")

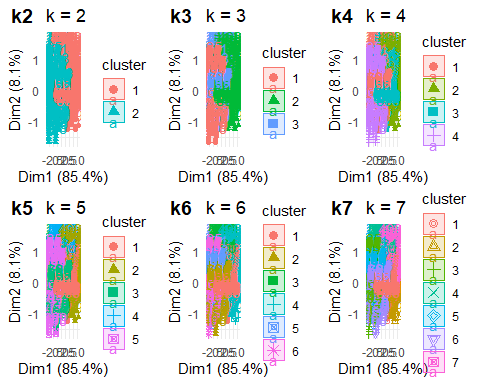
library(cowplot)

## Warning: package 'cowplot' was built under R version 4.4.3

##   
## Attaching package: 'cowplot'

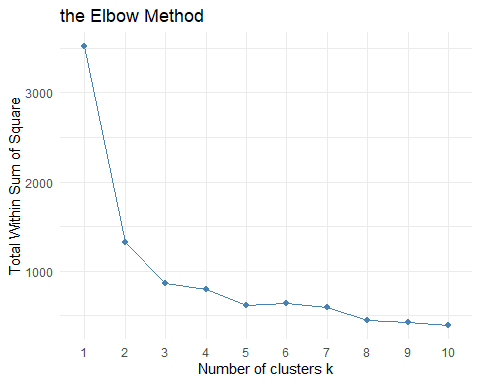
## The following object is masked from 'package:lubridate':  
##   
## stamp

library(ggplot2)  
print(plot\_grid(p1, p2, p3, p4, p5, p6, labels = c("k2", "k3", "k4", "k5", "k6", "k7")))



Các biểu đồ trên đã cho thấy Ở k=2, dữ liệu được chia thành hai nhóm khá rõ ràng, phân biệt chủ yếu theo chiều ngang (Dim1). Đây có thể là phân cụm đơn giản và dễ giải thích nhất, nhưng có thể chưa phản ánh đầy đủ tính đa dạng trong dữ liệu. Khi tăng lên k=3 và k=4, các nhóm trở nên phân tách tinh vi hơn với sự xuất hiện của nhiều nhóm nhỏ hơn, đồng thời có sự chồng lấn nhất định giữa các cụm. Ở k=5, k=6 và k=7, số lượng nhóm nhiều hơn, phân bố cụm khá phức tạp và các nhóm có xu hướng chồng lấn nhau nhiều hơn. Số cụm tối ưu là k = 2 hoặc 3, vì khi đó các nhóm phân tách rõ ràng và dễ giải thích. Số cụm lớn hơn gây chồng chéo và làm giảm độ chính xác phân loại.

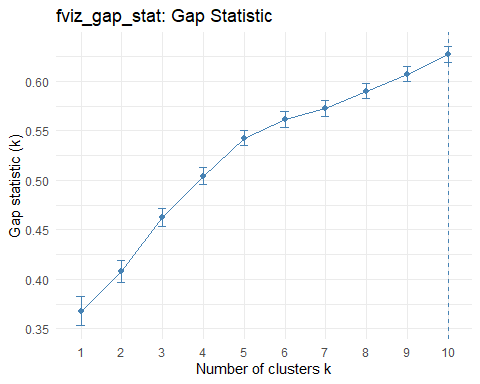
# vẽ đồ thị xác định số cụm  
## Dựa vào Elbow  
fviz\_nbclust(df.scaled, kmeans, method = "wss", k.max = 10) + theme\_minimal() + ggtitle("the Elbow Method")



## PP Gap  
gap\_stat <- clusGap(df.scaled, FUN = kmeans, nstart = 30, K.max = 10, B = 50)

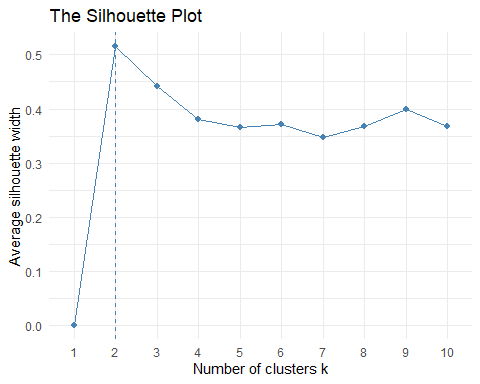
## Warning: did not converge in 10 iterations  
## Warning: did not converge in 10 iterations

fviz\_gap\_stat(gap\_stat) + theme\_minimal() + ggtitle("fviz\_gap\_stat: Gap Statistic")



Phương pháp khuỷu tay cho thấy điểm gãy rõ rệt ở k = 2 hoặc 3, trong khi Gap Statistic đạt giá trị cao nhất tại k = 10.

### pp Silhouette  
fviz\_nbclust(df.scaled, kmeans, method = "silhouette", k.max = 10) + theme\_minimal() + ggtitle("The Silhouette Plot")



kmean2<- kmeans (df.scaled, centers=2, nstart=10)# 2 cụm(nstart represents the number of random data sets used to run the algorithm)  
str(kmean2)

## List of 9  
## $ cluster : int [1:705] 2 1 2 1 2 2 1 2 1 1 ...  
## $ centers : num [1:2, 1:5] -0.838 0.658 0.854 -0.67 0.936 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : chr [1:2] "1" "2"  
## .. ..$ : chr [1:5] "Avg\_Daily\_Usage\_Hours" "Sleep\_Hours\_Per\_Night" "Mental\_Health\_Score" "Conflicts\_Over\_Social\_Media" ...  
## $ totss : num 3520  
## $ withinss : num [1:2] 546 772  
## $ tot.withinss: num 1318  
## $ betweenss : num 2202  
## $ size : int [1:2] 310 395  
## $ iter : int 1  
## $ ifault : int 0  
## - attr(\*, "class")= chr "kmeans"

kmean2

## K-means clustering with 2 clusters of sizes 310, 395  
##   
## Cluster means:  
## Avg\_Daily\_Usage\_Hours Sleep\_Hours\_Per\_Night Mental\_Health\_Score  
## 1 -0.8381496 0.8543093 0.9360073  
## 2 0.6577883 -0.6704706 -0.7345880  
## Conflicts\_Over\_Social\_Media Addicted\_Score  
## 1 -0.8768228 -0.9500255  
## 2 0.6881394 0.7455896  
##   
## Clustering vector:  
## [1] 2 1 2 1 2 2 1 2 1 1 2 2 1 2 1 2 1 2 1 2 2 1 2 2 1 2 1 2 1 2 1 2 2 1 2 2 1  
## [38] 2 1 2 1 2 1 2 2 1 2 2 1 2 1 2 1 2 1 2 2 1 2 2 1 2 2 2 1 2 1 2 1 2 2 1 2 2  
## [75] 1 2 2 1 2 2 1 2 1 1 2 2 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 1 2 1 2 2  
## [112] 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2  
## [149] 1 2 2 1 2 2 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2  
## [186] 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 2  
## [223] 1 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 2  
## [260] 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 1 1 2 2 1 1  
## [297] 1 2 1 1 2 2 1 1 1 2 1 1 2 2 1 1 1 2 1 1 2 2 1 1 2 1 1 2 1 2 2 2 1 2 2 1 1  
## [334] 2 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2  
## [371] 2 1 1 2 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2 2 1 1 2 1 2 2  
## [408] 2 1 2 2 1 1 2 1 2 2 2 1 2 1 1 1 2 1 2 2 2 1 1 1 1 1 2 1 2 2 2 1 1 1 1 1 2  
## [445] 1 1 2 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 2 1 1 2 2 1 1 1  
## [482] 1 1 2 1 1 2 2 1 1 1 1 1 2 1 1 2 2 1 1 2 1 1 1 2 2 2 1 2 1 2 1 1 1 2 2 2 1  
## [519] 2 1 2 1 1 1 2 2 2 1 2 1 2 1 1 1 2 2 2 1 2 1 2 1 1 1 2 2 2 1 2 1 2 1 1 1 2  
## [556] 2 2 1 2 1 2 1 1 1 2 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 1 2 1 2 2 1  
## [593] 2 2 2 2 1 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2  
## [630] 1 2 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2  
## [667] 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 2  
## [704] 1 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 545.6897 772.3138  
## (between\_SS / total\_SS = 62.6 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

print(kmean2)

## K-means clustering with 2 clusters of sizes 310, 395  
##   
## Cluster means:  
## Avg\_Daily\_Usage\_Hours Sleep\_Hours\_Per\_Night Mental\_Health\_Score  
## 1 -0.8381496 0.8543093 0.9360073  
## 2 0.6577883 -0.6704706 -0.7345880  
## Conflicts\_Over\_Social\_Media Addicted\_Score  
## 1 -0.8768228 -0.9500255  
## 2 0.6881394 0.7455896  
##   
## Clustering vector:  
## [1] 2 1 2 1 2 2 1 2 1 1 2 2 1 2 1 2 1 2 1 2 2 1 2 2 1 2 1 2 1 2 1 2 2 1 2 2 1  
## [38] 2 1 2 1 2 1 2 2 1 2 2 1 2 1 2 1 2 1 2 2 1 2 2 1 2 2 2 1 2 1 2 1 2 2 1 2 2  
## [75] 1 2 2 1 2 2 1 2 1 1 2 2 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 1 2 1 2 2  
## [112] 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2  
## [149] 1 2 2 1 2 2 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2  
## [186] 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 2  
## [223] 1 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 2  
## [260] 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 2 1 2 2 1 1 1 2 1 1 2 2 1 1  
## [297] 1 2 1 1 2 2 1 1 1 2 1 1 2 2 1 1 1 2 1 1 2 2 1 1 2 1 1 2 1 2 2 2 1 2 2 1 1  
## [334] 2 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2  
## [371] 2 1 1 2 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2 2 1 1 2 1 2 2  
## [408] 2 1 2 2 1 1 2 1 2 2 2 1 2 1 1 1 2 1 2 2 2 1 1 1 1 1 2 1 2 2 2 1 1 1 1 1 2  
## [445] 1 1 2 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 2 1 1 2 2 1 1 1  
## [482] 1 1 2 1 1 2 2 1 1 1 1 1 2 1 1 2 2 1 1 2 1 1 1 2 2 2 1 2 1 2 1 1 1 2 2 2 1  
## [519] 2 1 2 1 1 1 2 2 2 1 2 1 2 1 1 1 2 2 2 1 2 1 2 1 1 1 2 2 2 1 2 1 2 1 1 1 2  
## [556] 2 2 1 2 1 2 1 1 1 2 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 1 2 1 2 2 1  
## [593] 2 2 2 2 1 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2  
## [630] 1 2 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2  
## [667] 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 2  
## [704] 1 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 545.6897 772.3138  
## (between\_SS / total\_SS = 62.6 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

summary(kmean2)

## Length Class Mode   
## cluster 705 -none- numeric  
## centers 10 -none- numeric  
## totss 1 -none- numeric  
## withinss 2 -none- numeric  
## tot.withinss 1 -none- numeric  
## betweenss 1 -none- numeric  
## size 2 -none- numeric  
## iter 1 -none- numeric  
## ifault 1 -none- numeric

Theo phương pháp Silhouette, k = 2 là số cụm tối ưu nhất vì tạo ra các cụm tách biệt rõ ràng và có mức độ đồng nhất cao nhất bên trong từng cụm. Kết hợp với kết quả từ hai phương pháp còn lại, trong bài này sẽ sử dụng cả k = 2 và k = 3 để phân tích nhằm đảm bảo sự toàn diện và chính xác của kết quả.

#Thực hiện phương pháp k mean  
kmean2<- kmeans (df.scaled, centers=2, nstart=10)# 2 cụm(nstart represents the number of random data sets used to run the algorithm)  
str(kmean2)

## List of 9  
## $ cluster : int [1:705] 1 2 1 2 1 1 2 1 2 2 ...  
## $ centers : num [1:2, 1:5] 0.658 -0.838 -0.67 0.854 -0.735 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : chr [1:2] "1" "2"  
## .. ..$ : chr [1:5] "Avg\_Daily\_Usage\_Hours" "Sleep\_Hours\_Per\_Night" "Mental\_Health\_Score" "Conflicts\_Over\_Social\_Media" ...  
## $ totss : num 3520  
## $ withinss : num [1:2] 772 546  
## $ tot.withinss: num 1318  
## $ betweenss : num 2202  
## $ size : int [1:2] 395 310  
## $ iter : int 1  
## $ ifault : int 0  
## - attr(\*, "class")= chr "kmeans"

kmean2

## K-means clustering with 2 clusters of sizes 395, 310  
##   
## Cluster means:  
## Avg\_Daily\_Usage\_Hours Sleep\_Hours\_Per\_Night Mental\_Health\_Score  
## 1 0.6577883 -0.6704706 -0.7345880  
## 2 -0.8381496 0.8543093 0.9360073  
## Conflicts\_Over\_Social\_Media Addicted\_Score  
## 1 0.6881394 0.7455896  
## 2 -0.8768228 -0.9500255  
##   
## Clustering vector:  
## [1] 1 2 1 2 1 1 2 1 2 2 1 1 2 1 2 1 2 1 2 1 1 2 1 1 2 1 2 1 2 1 2 1 1 2 1 1 2  
## [38] 1 2 1 2 1 2 1 1 2 1 1 2 1 2 1 2 1 2 1 1 2 1 1 2 1 1 1 2 1 2 1 2 1 1 2 1 1  
## [75] 2 1 1 2 1 1 2 1 2 2 1 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 2 1 2 1 1  
## [112] 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1  
## [149] 2 1 1 2 1 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1  
## [186] 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 1  
## [223] 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 1  
## [260] 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 2 2 1 1 2 2  
## [297] 2 1 2 2 1 1 2 2 2 1 2 2 1 1 2 2 2 1 2 2 1 1 2 2 1 2 2 1 2 1 1 1 2 1 1 2 2  
## [334] 1 2 1 1 1 2 1 1 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1 1 2 1  
## [371] 1 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1  
## [408] 1 2 1 1 2 2 1 2 1 1 1 2 1 2 2 2 1 2 1 1 1 2 2 2 2 2 1 2 1 1 1 2 2 2 2 2 1  
## [445] 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 2 2  
## [482] 2 2 1 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 2 1 2 2 2 1 1 1 2 1 2 1 2 2 2 1 1 1 2  
## [519] 1 2 1 2 2 2 1 1 1 2 1 2 1 2 2 2 1 1 1 2 1 2 1 2 2 2 1 1 1 2 1 2 1 2 2 2 1  
## [556] 1 1 2 1 2 1 2 2 2 1 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 2 1 2 1 1 2  
## [593] 1 1 1 1 2 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1  
## [630] 2 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1  
## [667] 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1  
## [704] 2 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 772.3138 545.6897  
## (between\_SS / total\_SS = 62.6 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

print(kmean2)

## K-means clustering with 2 clusters of sizes 395, 310  
##   
## Cluster means:  
## Avg\_Daily\_Usage\_Hours Sleep\_Hours\_Per\_Night Mental\_Health\_Score  
## 1 0.6577883 -0.6704706 -0.7345880  
## 2 -0.8381496 0.8543093 0.9360073  
## Conflicts\_Over\_Social\_Media Addicted\_Score  
## 1 0.6881394 0.7455896  
## 2 -0.8768228 -0.9500255  
##   
## Clustering vector:  
## [1] 1 2 1 2 1 1 2 1 2 2 1 1 2 1 2 1 2 1 2 1 1 2 1 1 2 1 2 1 2 1 2 1 1 2 1 1 2  
## [38] 1 2 1 2 1 2 1 1 2 1 1 2 1 2 1 2 1 2 1 1 2 1 1 2 1 1 1 2 1 2 1 2 1 1 2 1 1  
## [75] 2 1 1 2 1 1 2 1 2 2 1 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 2 1 2 1 1  
## [112] 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1  
## [149] 2 1 1 2 1 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1  
## [186] 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 2 1 1 1  
## [223] 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 1  
## [260] 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1 2 2 1 1 2 2  
## [297] 2 1 2 2 1 1 2 2 2 1 2 2 1 1 2 2 2 1 2 2 1 1 2 2 1 2 2 1 2 1 1 1 2 1 1 2 2  
## [334] 1 2 1 1 1 2 1 1 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1 1 2 1  
## [371] 1 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1  
## [408] 1 2 1 1 2 2 1 2 1 1 1 2 1 2 2 2 1 2 1 1 1 2 2 2 2 2 1 2 1 1 1 2 2 2 2 2 1  
## [445] 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 2 2  
## [482] 2 2 1 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 2 1 2 2 2 1 1 1 2 1 2 1 2 2 2 1 1 1 2  
## [519] 1 2 1 2 2 2 1 1 1 2 1 2 1 2 2 2 1 1 1 2 1 2 1 2 2 2 1 1 1 2 1 2 1 2 2 2 1  
## [556] 1 1 2 1 2 1 2 2 2 1 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 2 1 2 1 1 2  
## [593] 1 1 1 1 2 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1  
## [630] 2 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1  
## [667] 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1  
## [704] 2 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 772.3138 545.6897  
## (between\_SS / total\_SS = 62.6 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

summary(kmean2)

## Length Class Mode   
## cluster 705 -none- numeric  
## centers 10 -none- numeric  
## totss 1 -none- numeric  
## withinss 2 -none- numeric  
## tot.withinss 1 -none- numeric  
## betweenss 1 -none- numeric  
## size 2 -none- numeric  
## iter 1 -none- numeric  
## ifault 1 -none- numeric

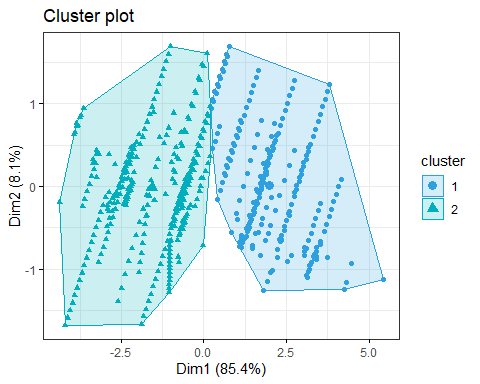
k\_stats2 <- cluster.stats(dist(df.scaled), kmean2$cluster)  
desc2 <- cluster.Description(cluster\_data,kmean2$cluster)  
k\_stats2$dunn

## [1] 0.02104322

k\_stats2

## $n  
## [1] 705  
##   
## $cluster.number  
## [1] 2  
##   
## $cluster.size  
## [1] 395 310  
##   
## $min.cluster.size  
## [1] 310  
##   
## $noisen  
## [1] 0  
##   
## $diameter  
## [1] 5.586668 5.662867  
##   
## $average.distance  
## [1] 1.782482 1.690633  
##   
## $median.distance  
## [1] 1.710812 1.635570  
##   
## $separation  
## [1] 0.119165 0.119165  
##   
## $average.toother  
## [1] 3.791722 3.791722  
##   
## $separation.matrix  
## [,1] [,2]  
## [1,] 0.000000 0.119165  
## [2,] 0.119165 0.000000  
##   
## $ave.between.matrix  
## [,1] [,2]  
## [1,] 0.000000 3.791722  
## [2,] 3.791722 0.000000  
##   
## $average.between  
## [1] 3.791722  
##   
## $average.within  
## [1] 1.742094  
##   
## $n.between  
## [1] 122450  
##   
## $n.within  
## [1] 125710  
##   
## $max.diameter  
## [1] 5.662867  
##   
## $min.separation  
## [1] 0.119165  
##   
## $within.cluster.ss  
## [1] 1318.003  
##   
## $clus.avg.silwidths  
## 1 2   
## 0.5021268 0.5323780   
##   
## $avg.silwidth  
## [1] 0.5154287  
##   
## $g2  
## NULL  
##   
## $g3  
## NULL  
##   
## $pearsongamma  
## [1] 0.6592383  
##   
## $dunn  
## [1] 0.02104322  
##   
## $dunn2  
## [1] 2.127215  
##   
## $entropy  
## [1] 0.6858612  
##   
## $wb.ratio  
## [1] 0.4594468  
##   
## $ch  
## [1] 1174.506  
##   
## $cwidegap  
## [1] 1.064917 1.185536  
##   
## $widestgap  
## [1] 1.185536  
##   
## $sindex  
## [1] 0.5411949  
##   
## $corrected.rand  
## NULL  
##   
## $vi  
## NULL

#Trực quan hoá  
fviz\_cluster(km2,   
 data = df.scaled,  
 palette = c("#2E9FDF", "#00AFBB"),   
 geom = "point",  
 ellipse.type = "convex",   
 ggtheme = theme\_bw())



Phân cụm K-means với K = 2 cho tỷ lệ phương sai giữa cụm đạt 62,6% vậy giữa 2 cụm có sự tách biệt. Ngoài ra,Việc sử dụng k = 2 trong phân cụm cho kết quả khá hiệu quả khi tạo ra hai nhóm phân biệt rõ ràng về hành vi sử dụng mạng xã hội và sức khỏe tâm thần. Cụm 1 gồm những người dùng mạng xã hội nhiều hơn(Avg\_Daily\_Usage\_Hours dương), có mức độ nghiện và xung đột liên quan đến mạng xã hội cao, các chỉ số khá đồng đều, đồng thời có số giờ ngủ và điểm sức khỏe tâm thần thấp hơn trung bình. Cụm 2 ngược lại, là nhóm người dùng mạng xã hội ít, giấc ngủ tốt hơn và sức khỏe tâm thần ổn định hơn.Vậy ta sẽ chia thành 2 nhóm là nhóm sử dụng mạng xã hội một cách hợp lý và nhóm có xu hướng nghiện mạng xã hội cao Chỉ số Silhouette trung bình lần lượt là 0.50 và 0.53, cho thấy phân cụm ở mức tương đối tốt. Chỉ số Dunn2 là 2.127215 cho thấy mức phân tách giữa cụm là tốt.

kmean3<- kmeans (df.scaled, centers=3, nstart=10)# 3 cụm(nstart represents the number of random data sets used to run the algorithm)  
str(kmean3)

## List of 9  
## $ cluster : int [1:705] 1 3 2 3 1 2 3 1 3 3 ...  
## $ centers : num [1:3, 1:5] 0.0513 1.1976 -0.8835 -0.0708 -1.1304 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : chr [1:3] "1" "2" "3"  
## .. ..$ : chr [1:5] "Avg\_Daily\_Usage\_Hours" "Sleep\_Hours\_Per\_Night" "Mental\_Health\_Score" "Conflicts\_Over\_Social\_Media" ...  
## $ totss : num 3520  
## $ withinss : num [1:3] 261 215 389  
## $ tot.withinss: num 865  
## $ betweenss : num 2655  
## $ size : int [1:3] 250 187 268  
## $ iter : int 2  
## $ ifault : int 0  
## - attr(\*, "class")= chr "kmeans"

kmean3

## K-means clustering with 3 clusters of sizes 250, 187, 268  
##   
## Cluster means:  
## Avg\_Daily\_Usage\_Hours Sleep\_Hours\_Per\_Night Mental\_Health\_Score  
## 1 0.05127794 -0.07076036 -0.3031073  
## 2 1.19761578 -1.13036448 -1.1151458  
## 3 -0.88348372 0.85473227 1.0608548  
## Conflicts\_Over\_Social\_Media Addicted\_Score  
## 1 0.1903557 0.3321146  
## 2 1.2343214 1.1432066  
## 3 -1.0388322 -1.1074936  
##   
## Clustering vector:  
## [1] 1 3 2 3 1 2 3 1 3 3 1 1 3 2 3 1 3 1 3 1 2 3 1 2 3 2 3 1 3 2 3 1 2 3 1 2 3  
## [38] 2 3 1 3 2 3 1 2 3 1 2 3 2 3 1 3 2 3 1 2 3 1 2 3 2 1 2 3 1 3 2 3 1 2 3 1 2  
## [75] 3 1 2 3 1 2 3 1 3 3 2 1 2 3 1 2 3 1 3 2 3 1 2 3 1 2 2 3 1 2 3 1 3 2 3 1 2  
## [112] 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 1 1 3 1 1 3 1 1 3 1 1 3 1 1  
## [149] 3 1 1 3 1 1 3 1 1 3 1 1 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1  
## [186] 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 2 1  
## [223] 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1  
## [260] 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3  
## [297] 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 1 3 1 2 3 1 2 1 3 1 1 3 1  
## [334] 2 3 1 2 1 3 1 1 3 1 2 3 1 2 1 3 1 1 3 1 2 3 1 2 1 3 1 1 3 1 2 3 1 2 1 3 1  
## [371] 1 3 1 2 3 1 2 1 3 1 1 3 1 2 3 1 2 1 3 1 1 3 3 2 3 1 2 1 3 1 1 3 3 2 3 1 2  
## [408] 1 3 1 1 3 3 2 3 1 2 1 3 1 1 3 3 2 3 1 2 2 3 1 1 3 3 2 3 1 2 2 3 1 1 3 3 2  
## [445] 3 1 2 2 3 1 1 3 3 2 3 1 2 2 3 1 1 3 3 2 3 1 2 2 3 1 1 3 3 2 3 1 2 2 3 1 1  
## [482] 3 3 2 3 1 2 2 3 1 1 3 3 2 3 1 2 2 3 1 1 3 3 1 1 1 2 3 2 1 1 3 3 1 1 1 2 3  
## [519] 2 1 1 3 3 1 1 1 2 3 2 1 1 3 3 1 1 1 2 3 2 1 1 3 3 3 1 1 2 3 2 1 1 3 3 3 1  
## [556] 1 2 3 2 3 1 3 3 3 1 1 2 3 2 3 1 3 1 2 3 2 1 2 3 1 2 3 2 1 3 2 3 2 3 1 2 3  
## [593] 2 1 2 1 3 2 3 2 1 3 2 3 2 3 1 2 3 2 1 2 3 1 2 3 2 1 3 2 3 2 3 1 2 3 2 1 2  
## [630] 3 1 2 3 2 1 3 2 3 2 3 1 2 3 2 1 2 3 1 2 3 2 1 3 2 3 2 3 1 2 3 2 1 2 3 1 2  
## [667] 3 2 1 3 2 3 2 3 1 2 3 2 1 2 3 1 2 3 2 1 3 2 3 2 3 1 2 3 2 1 2 3 1 2 3 2 1  
## [704] 3 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 261.0055 215.4353 388.5407  
## (between\_SS / total\_SS = 75.4 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

print(kmean3)

## K-means clustering with 3 clusters of sizes 250, 187, 268  
##   
## Cluster means:  
## Avg\_Daily\_Usage\_Hours Sleep\_Hours\_Per\_Night Mental\_Health\_Score  
## 1 0.05127794 -0.07076036 -0.3031073  
## 2 1.19761578 -1.13036448 -1.1151458  
## 3 -0.88348372 0.85473227 1.0608548  
## Conflicts\_Over\_Social\_Media Addicted\_Score  
## 1 0.1903557 0.3321146  
## 2 1.2343214 1.1432066  
## 3 -1.0388322 -1.1074936  
##   
## Clustering vector:  
## [1] 1 3 2 3 1 2 3 1 3 3 1 1 3 2 3 1 3 1 3 1 2 3 1 2 3 2 3 1 3 2 3 1 2 3 1 2 3  
## [38] 2 3 1 3 2 3 1 2 3 1 2 3 2 3 1 3 2 3 1 2 3 1 2 3 2 1 2 3 1 3 2 3 1 2 3 1 2  
## [75] 3 1 2 3 1 2 3 1 3 3 2 1 2 3 1 2 3 1 3 2 3 1 2 3 1 2 2 3 1 2 3 1 3 2 3 1 2  
## [112] 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 1 1 3 1 1 3 1 1 3 1 1 3 1 1  
## [149] 3 1 1 3 1 1 3 1 1 3 1 1 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1  
## [186] 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 1 3 2 2 1  
## [223] 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1  
## [260] 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3  
## [297] 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 3 1 1 3 2 1 3 3 1 3 1 2 3 1 2 1 3 1 1 3 1  
## [334] 2 3 1 2 1 3 1 1 3 1 2 3 1 2 1 3 1 1 3 1 2 3 1 2 1 3 1 1 3 1 2 3 1 2 1 3 1  
## [371] 1 3 1 2 3 1 2 1 3 1 1 3 1 2 3 1 2 1 3 1 1 3 3 2 3 1 2 1 3 1 1 3 3 2 3 1 2  
## [408] 1 3 1 1 3 3 2 3 1 2 1 3 1 1 3 3 2 3 1 2 2 3 1 1 3 3 2 3 1 2 2 3 1 1 3 3 2  
## [445] 3 1 2 2 3 1 1 3 3 2 3 1 2 2 3 1 1 3 3 2 3 1 2 2 3 1 1 3 3 2 3 1 2 2 3 1 1  
## [482] 3 3 2 3 1 2 2 3 1 1 3 3 2 3 1 2 2 3 1 1 3 3 1 1 1 2 3 2 1 1 3 3 1 1 1 2 3  
## [519] 2 1 1 3 3 1 1 1 2 3 2 1 1 3 3 1 1 1 2 3 2 1 1 3 3 3 1 1 2 3 2 1 1 3 3 3 1  
## [556] 1 2 3 2 3 1 3 3 3 1 1 2 3 2 3 1 3 1 2 3 2 1 2 3 1 2 3 2 1 3 2 3 2 3 1 2 3  
## [593] 2 1 2 1 3 2 3 2 1 3 2 3 2 3 1 2 3 2 1 2 3 1 2 3 2 1 3 2 3 2 3 1 2 3 2 1 2  
## [630] 3 1 2 3 2 1 3 2 3 2 3 1 2 3 2 1 2 3 1 2 3 2 1 3 2 3 2 3 1 2 3 2 1 2 3 1 2  
## [667] 3 2 1 3 2 3 2 3 1 2 3 2 1 2 3 1 2 3 2 1 3 2 3 2 3 1 2 3 2 1 2 3 1 2 3 2 1  
## [704] 3 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 261.0055 215.4353 388.5407  
## (between\_SS / total\_SS = 75.4 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

summary(kmean3)

## Length Class Mode   
## cluster 705 -none- numeric  
## centers 15 -none- numeric  
## totss 1 -none- numeric  
## withinss 3 -none- numeric  
## tot.withinss 1 -none- numeric  
## betweenss 1 -none- numeric  
## size 3 -none- numeric  
## iter 1 -none- numeric  
## ifault 1 -none- numeric

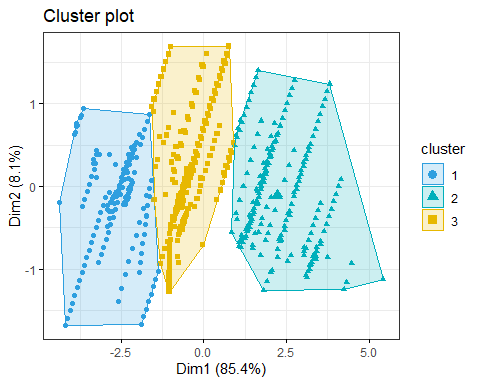
k\_stats3 <- cluster.stats(dist(df.scaled), kmean3$cluster)  
library(clusterSim)  
desc2 <- cluster.Description(cluster\_data,kmean3$cluster)  
k\_stats3$dunn

## [1] 0.02412383

k\_stats3

## $n  
## [1] 705  
##   
## $cluster.number  
## [1] 3  
##   
## $cluster.size  
## [1] 250 187 268  
##   
## $min.cluster.size  
## [1] 187  
##   
## $noisen  
## [1] 0  
##   
## $diameter  
## [1] 3.637724 3.784084 4.939720  
##   
## $average.distance  
## [1] 1.304973 1.365024 1.537050  
##   
## $median.distance  
## [1] 1.265105 1.336223 1.521449  
##   
## $separation  
## [1] 0.119165 0.119165 0.119165  
##   
## $average.toother  
## [1] 2.776397 3.785710 3.814391  
##   
## $separation.matrix  
## [,1] [,2] [,3]  
## [1,] 0.000000 0.119165 0.119165  
## [2,] 0.119165 0.000000 2.305903  
## [3,] 0.119165 2.305903 0.000000  
##   
## $ave.between.matrix  
## [,1] [,2] [,3]  
## [1,] 0.000000 2.521877 2.953992  
## [2,] 2.521877 0.000000 4.964658  
## [3,] 2.953992 4.964658 0.000000  
##   
## $average.between  
## [1] 3.445645  
##   
## $average.within  
## [1] 1.409124  
##   
## $n.between  
## [1] 163866  
##   
## $n.within  
## [1] 84294  
##   
## $max.diameter  
## [1] 4.93972  
##   
## $min.separation  
## [1] 0.119165  
##   
## $within.cluster.ss  
## [1] 864.9815  
##   
## $clus.avg.silwidths  
## 1 2 3   
## 0.4187654 0.4460615 0.4576857   
##   
## $avg.silwidth  
## [1] 0.4408009  
##   
## $g2  
## NULL  
##   
## $g3  
## NULL  
##   
## $pearsongamma  
## [1] 0.6200675  
##   
## $dunn  
## [1] 0.02412383  
##   
## $dunn2  
## [1] 1.640726  
##   
## $entropy  
## [1] 1.087323  
##   
## $wb.ratio  
## [1] 0.4089579  
##   
## $ch  
## [1] 1077.377  
##   
## $cwidegap  
## [1] 1.226639 1.359546 1.185536  
##   
## $widestgap  
## [1] 1.359546  
##   
## $sindex  
## [1] 0.474403  
##   
## $corrected.rand  
## NULL  
##   
## $vi  
## NULL

#Trực quan hoá  
fviz\_cluster(km3,   
 data = df.scaled,  
 palette = c("#2E9FDF", "#00AFBB","#E7B800"),   
 geom = "point",  
 ellipse.type = "convex",   
 ggtheme = theme\_bw())



Phân cụm K-means với K = 3 cho phương sai giữa cụm chiếm 75,4% tổng phương sai, khoảng cách quan sát trong nội bộ cụm khá nhỏ vậy kết quả phân cụm đảm bảo được tính chất đồng đều. Chỉ số Silhouette trung bình theo cụm dao động từ 0.419 đến 0.458, phản ánh chất lượng cụm ở mức trung bình, vẫn còn hiện tượng chồng lấn ở một số điểm. Chỉ số Dunn2 = 1.640726 cho thấy mức độ tách biệt giữa các cụm là khá tốt.

=>>So sánh giữa hai mô hình phân cụm cho thấy K = 2 cho kết quả tốt hơn K = 3. Cụ thể, mặc dù K = 3 có tỷ lệ phương sai = 75,4% giữa cụm cao hơn , nhưng các chỉ số đánh giá chất lượng cụm như Silhouette trung bình chỉ từ 0.419–0.458 và Dunn2= 1.6407 lại cho thấy mức độ đồng nhất và tách biệt chỉ ở mức trung bình. Ngược lại, với K = 2, tuy phương sai giữa cụm = 62,6% thấp hơn so với k = 33 nhưng chất lượng cụm lại tốt hơn rõ rệt, với chỉ số Silhouette trung bình đạt 0.50–0.53 và Dunn2 lên đến 2.1272. Ngoài ra, kết quả phân cụm với K = 2 cũng cho ra hai nhóm người dùng có sự phân biệt rõ ràng về hành vi sử dụng mạng xã hội và sức khỏe tâm thần, giúp dễ dàng diễn giải và ứng dụng trong thực tiễn. Do đó, mô hình với K = 2, chia ra làm 2 nhóm là nhóm sử dụng mạng xã hội một cách hợp lý và nhóm có xu hướng nghiện mạng xã hội cao, được đánh giá là phù hợp và hiệu quả hơn.

# 4. Kết luận

Trong thời đại số, mạng xã hội trở thành một phần thiết yếu trong cuộc sống của sinh viên. Tuy nhiên, sử dụng mạng xã hội quá mức có thể ảnh hưởng tiêu cực đến sức khỏe tinh thần và sự cân bằng cuộc sống.

Trong nghiên cứu này, chúng tôi đã sử dụng ba phương pháp phân tích: PCA, CA và K-means, để tìm hiểu về hành vi nghiện mạng xã hội của sinh viên. Phân tích PCA cho thấy các yếu tố như thời gian sử dụng mạng xã hội, sức khỏe tâm lý và xung đột qua mạng xã hội có ảnh hưởng lớn nhất đến hành vi nghiện. Điều này cho thấy việc sử dụng mạng xã hội không chỉ đơn thuần là về thời gian mà còn liên quan đến tâm lý và các mối quan hệ cá nhân.

Phân tích CA giúp làm rõ mối liên hệ giữa các yếu tố như trình độ học vấn và loại ứng dụng được sử dụng nhiều nhất. Điều này giúp nhận diện rõ hơn nhóm đối tượng dễ bị tác động từ mạng xã hội.

Đặc biệt, khi sử dụng phân cụm K-means với k = 2, chúng tôi đã chia sinh viên thành hai nhóm: nhóm cân bằng và nhóm có nguy cơ nghiện cao. Nhóm cân bằng kiểm soát tốt thời gian sử dụng mạng xã hội và ít gặp vấn đề về sức khỏe tâm lý. Trong khi đó, nhóm có nguy cơ nghiện cao thường sử dụng mạng xã hội quá mức, dễ bị xáo trộn về tâm lý.

Qua những kết quả này, có thể thấy rằng việc nghiện mạng xã hội không chỉ là vấn đề về tần suất sử dụng mà còn liên quan đến sự mất cân bằng trong cuộc sống. Những phát hiện này góp phần xây dựng các chương trình hỗ trợ giúp sinh viên quản lý hành vi sử dụng mạng xã hội một cách lành mạnh hơn.