Social Media Addiction Analysis

Introduction::

Social media has become integral to our daily lives, particularly for young adults and students. While these platforms provide numerous benefits such as connecting with friends and family, they can also become a source of addiction, negatively impacting mental and physical health. Young adults become enthralled by social media and cannot reduce or cease their online media consumption despite clear negative consequences and severe drawbacks. This analysis aims to examine the usage patterns of various social media platforms among students, focusing on identifying potential addictive behaviors. By understanding the prevalence and nature of social media addiction, we can develop interventions and strategies to promote responsible use and mitigate its harmful effects on students.

Variables in the dataset:

Field Name	Data Type	Units	Description	Possible Values for Categorical
Week	Alphanumeric	-	Week start and end date	-
Whatsapp	Numeric	Hours	Time spent on Whatsapp per week	-
Instagram	Numeric	Hours	Time spent on Instagram per week	-
Snapchat	Numeric	Hours	Time spent on Snapchat per week	-
Telegram	Numeric	Hours	Time spent on Telegram per week	-
Facebook/Messenger	Numeric	Hours	Time spent on Facebook/Messenger per week	-
BeReal	Numeric	Hours	Time spent on BeReal per week	-
TikTok	Numeric	Hours	Time spent on Tiktok per week	-
Wechat	Numeric	Hours	Time spent on WeChat per week	-
Twitter	Numeric	Hours	Time spent on Twitter per week	-
Linkedin	Numeric	Hours	Time spent on LinkedIn per week	-
Messages	Numeric	Hours	Time spent on Messages per week	-
Total Social Media Screen Time	Numeric	Hours	Total time spent on social media per week	
Number of times opened (hourly intervals)	Numeric	Nos	Considering the 24-hour slots in a day, how many hour slots did the user open social media apps. This is for one day. Consider the above count and add the	-

			daily counts over the week and input that data	
Social Media Addiction Level	Categorical	-	Is the person addicted to social media or not?	Times opened >= 105 - Addicted Times opened < 105 - Not Addicted

Aim of the Analysis :

This analysis aims to investigate the social media app usage patterns among students and their potential association with addiction. Specifically, we aim to:

- Identify which social media apps are most commonly used by students,
- Is there a correlation between the time spent on them and addiction symptoms
- Build an accurate model that can classify individuals as having a high or low level of social media addiction.

Exploratory Data Analysis :

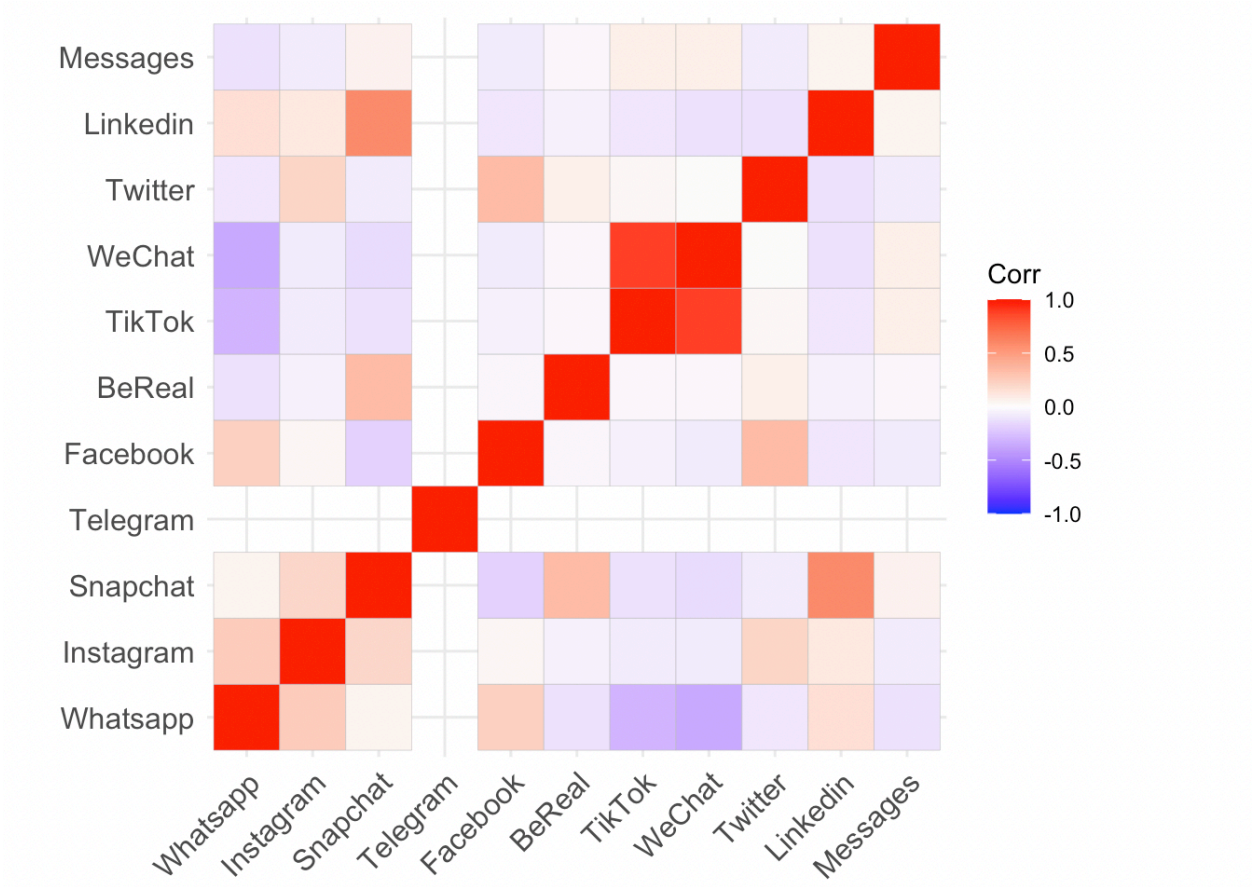
```
summary(class_data_analysis)
```

```
##     Whatsapp        Instagram         Snapchat         Telegram
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :0.0000
##  1st Qu.: 5.055   1st Qu.: 4.750   1st Qu.: 0.000   1st Qu.:0.0000
##  Median : 7.500   Median : 7.800   Median : 0.800   Median :0.0000
##  Mean   : 7.878   Mean   : 8.253   Mean   : 1.406   Mean   :0.1175
##  3rd Qu.:10.000   3rd Qu.:11.225   3rd Qu.: 1.535   3rd Qu.:0.0600
##  Max.   :22.500   Max.   :24.000   Max.   :12.100   Max.   :2.3900
##                                                     NA's   :1
##     Facebook          BeReal           TikTok            WeChat
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   : 0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.: 0.0000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median : 0.0000
##  Mean   :0.1624   Mean   :0.1174   Mean   :0.08754   Mean   : 0.3498
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.: 0.0000
##  Max.   :2.3500   Max.   :8.6000   Max.   :3.90000   Max.   :10.5000
##
##     Twitter          Linkedin         Messages
##  Min.   :0.0000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.: 0.415   1st Qu.: 0.000
##  Median :0.0000   Median : 1.420   Median : 0.060
##  Mean   :0.2525   Mean   : 3.255   Mean   : 0.591
##  3rd Qu.:0.0000   3rd Qu.: 4.000   3rd Qu.: 0.400
##  Max.   :8.5000   Max.   :22.800   Max.   :10.300
##
```

This output shows that the mean and median usage time for social media apps such as ***Whatsapp, Instagram, and Snapchat*** is relatively high compared to other apps. This indicates

that students may be spending significant time on these apps. Additionally, the maximum usage time for these apps is relatively high, which could suggest potential addiction or overuse.
On the other hand, apps such as ***Twitter and LinkedIn have lower mean and median usage*** times, indicating that students may be spending less time on these platforms. However, it's essential to note that some students may be using these apps for professional or academic purposes, which could explain the lower usage times.

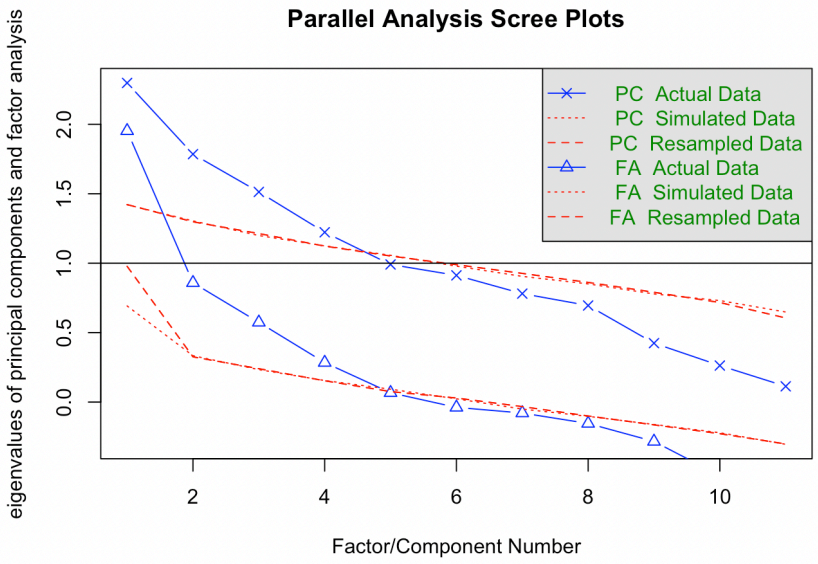
Overall, the usage data provides insights into which apps students may be spending the most time on, which could indicate potential addiction or overuse.



WeChat and Tiktok usage time(in hrs) are highly correlated. TikTok primarily focuses on short-form video content, while WeChat focuses more on messaging and communication. Students spending similar amounts of time on TikTok and WeChat could be because they enjoy the content on both platforms or find them equally helpful in staying connected with friends and family. Additionally, cultural factors may influence their preferences for certain social media platforms. Both apps are particularly popular in China and not specifically in India.

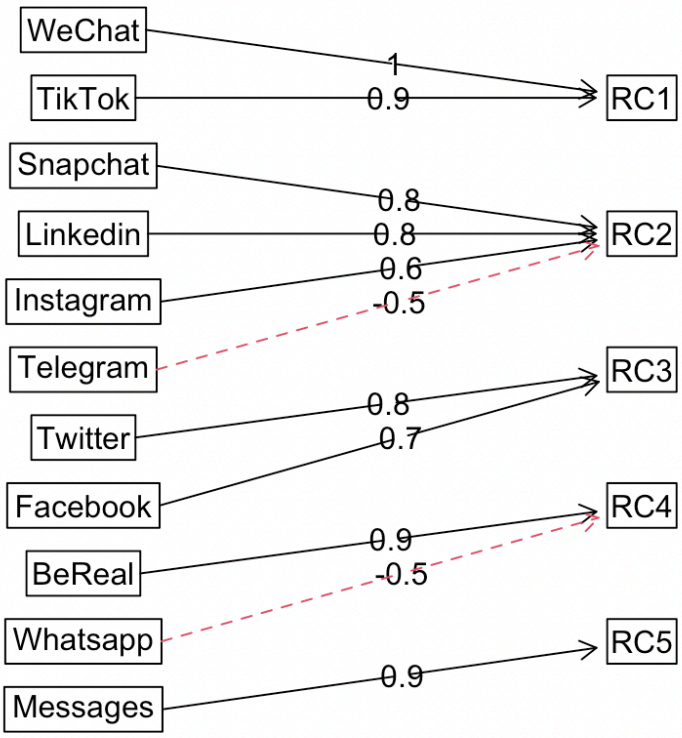
Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) performed on this dataset can help identify underlying factors that may explain the common variance among social media platforms. This can shed light on the relationships between social media platforms and usage patterns.




```
## Parallel analysis suggests that the number of factors =  4  and the number of components =  4
```

Components Analysis



Here are some insights we can gain from the Factor Analysis output:

- RC1: This component is negatively correlated with Whatsapp, Telegram, and Facebook and positively correlated with TikTok and WeChat. This suggests that people who use TikTok and WeChat are less likely to use the other three social media platforms. Also, WeChat and TikTok share similar characteristics as both are used for communication and socializing purposes.

- RC2: This component is positively correlated with Instagram,Snapchat and Linkedin and negatively correlated with Telegram. This suggests that people who use Instagram,Snapchat and Linkedin are less likely to use Telegram. This can symbolize social networking factor, where these platforms are used primarily for building and maintaining social connections, sharing personal updates and experiences, and staying informed about others' lives.

- RC3: This component is highly correlated with Twitter and Facebook negatively correlated with most other variables except Instagram. This could represent the public and real-time nature of information sharing on these platforms, where users can share their thoughts and opinions with a wider audience.

- RC4: This component is highly correlated with BeReal and negatively correlated with Whatsapp and mostl other variables. This suggests that BeReal is a unique social media platform that is not strongly associated with any other platform.

- RC5: This component is highly correlated with Messages and negatively correlated with all other variables. This suggests that Messages is a unique communication platform that is not strongly associated with any social media platform.

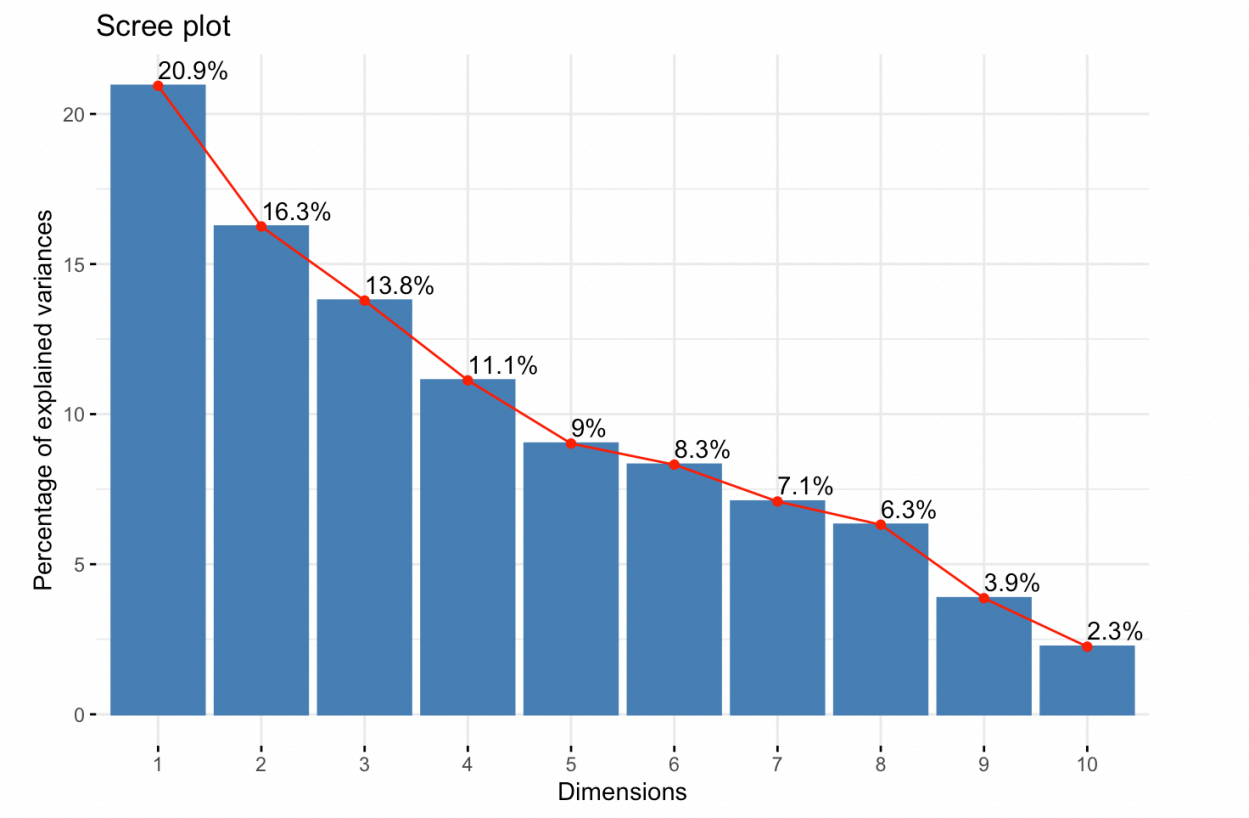
Overall, the output shows that some social media platforms are highly correlated, while some like BeReal and Messages are unique and not strongly associated with any other platform.

Principal Component Analysis (PCA):

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6    PC7
## Standard deviation     1.5196 1.3388 1.2329 1.1077 0.9974 0.95776 0.8843
## Proportion of Variance 0.2094 0.1625 0.1378 0.1113 0.0902 0.08318 0.0709
## Cumulative Proportion  0.2094 0.3719 0.5097 0.6210 0.7112 0.79438 0.8653
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.83471 0.65311 0.49840 0.33764
## Proportion of Variance 0.06318 0.03868 0.02252 0.01034
## Cumulative Proportion  0.92846 0.96714 0.98966 1.00000
```

The importance of each subsequent component decreases as we move down the list, with the last component (PC11) explaining the least variance.

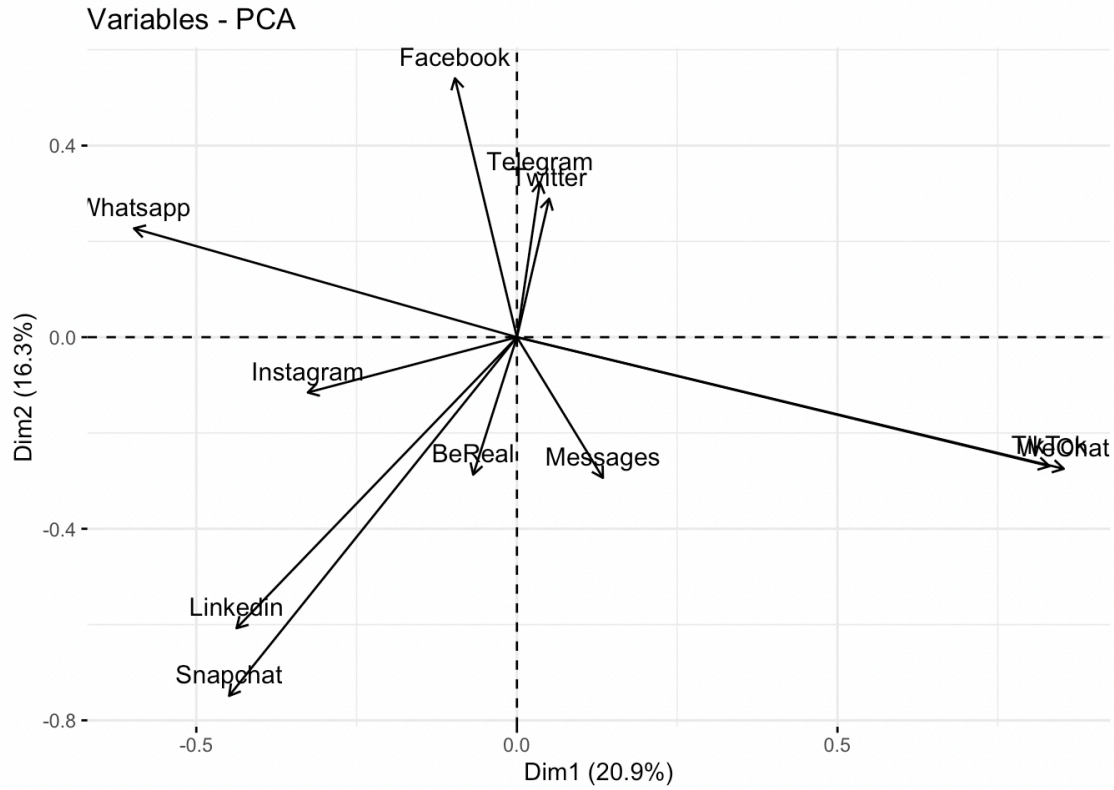
The first principal component (PC1) has the highest standard deviation and proportion of variance explained, indicating that it contains the most information in the data. The cumulative proportion shows that the first two components (PC1 and PC2) explain over 37% of the total variance, while the first seven explain over 86%. Overall, this output can be used to determine the optimal number of principal components to retain for further analysis, based on the proportion of variance explained by each component.

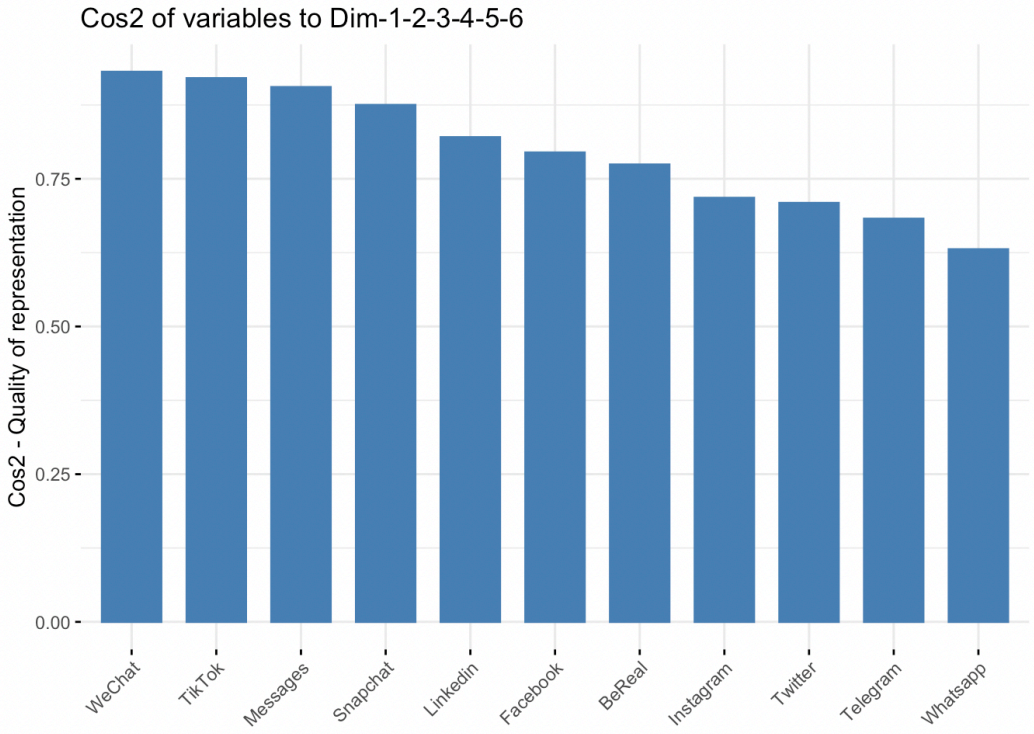


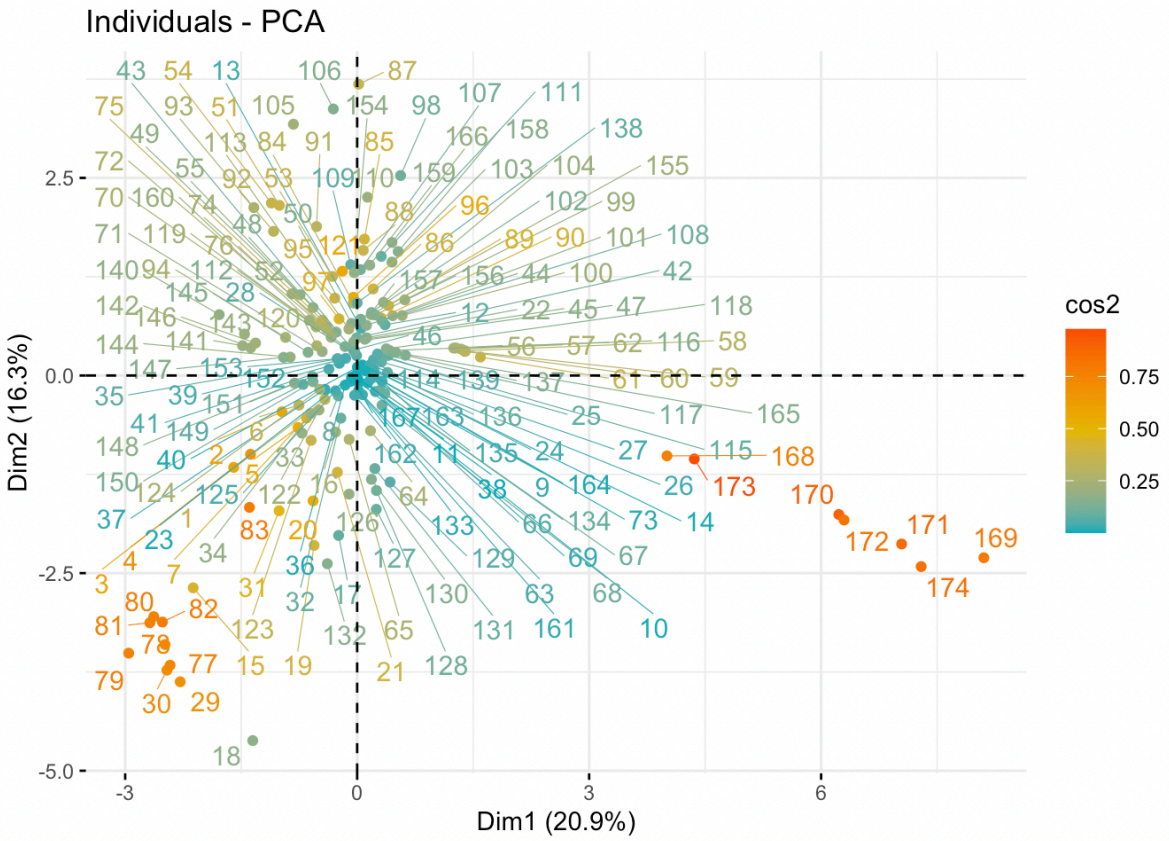
From the scree plot, we can notice that first 7 components explain over 86% of the total variance.

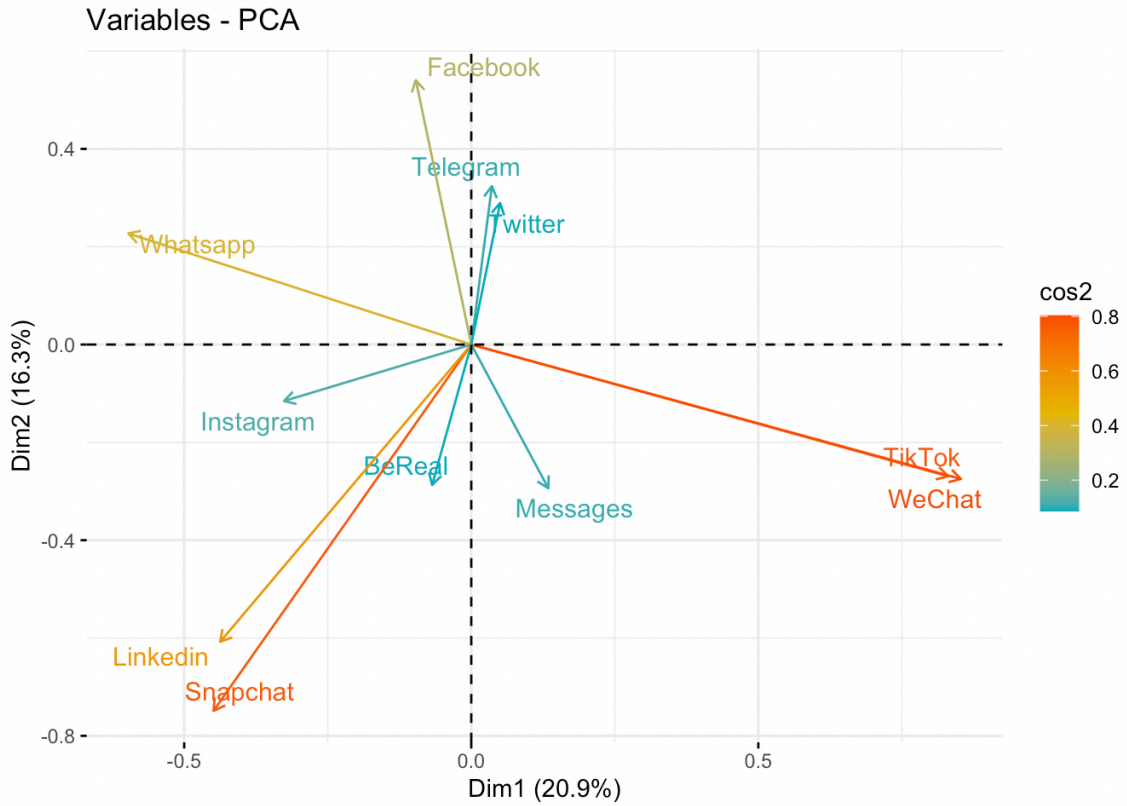
The length and direction of each vector indicate the correlation strength and direction between the variable and the PC. Variables that have a strong correlation with a particular PC will have longer vectors pointing in the direction of that PC.
Inferences that can be made from this output include:

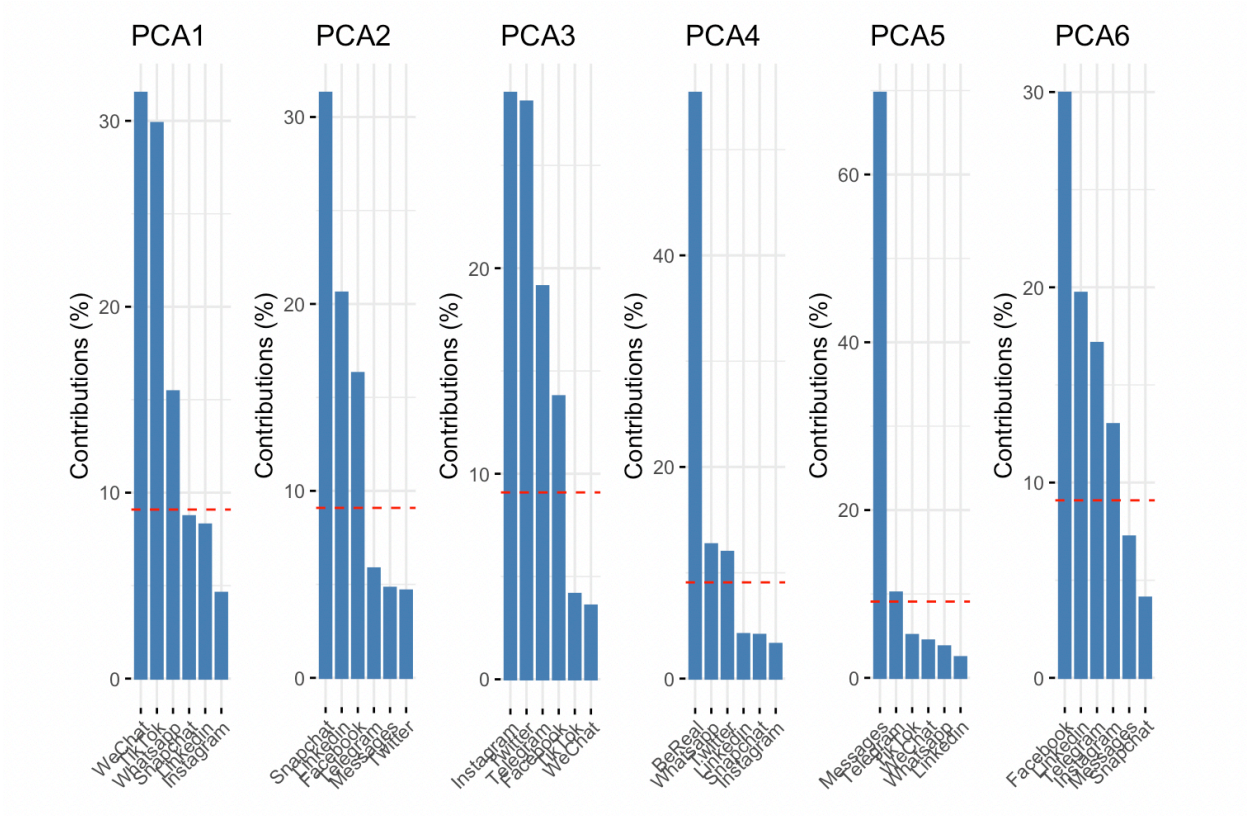
- Variables that are located close to each other in the plot are highly correlated with each other, and likely represent similar information.
- Variables that are located far away from each other are less correlated, and may represent distinct information in the data.
- PCs that are located far away from each other in the plot are also less correlated, and may represent distinct dimensions of variation in the data.
- The first two PCs explain a significant amount of variation in the data (37.19% and 16.78%, respectively), and are thus the most informative for understanding patterns in the data.


Variables - PCA
Facebook
Telegram
witter
Vhatsapp
Instagram
BeReal
Messages
TikTok
WeChat
Linkedin
Snapchat
0.4
0.0
-0.4
-0.8
-0.5
0.0
0.5
Dim2 (16.3%)
Dim1 (20.9%)


Cos2 of variables to Dim-1-2-3-4-5-6
Cos2 - Quality of representation
0.75
0.50
0.25
0.00
WeChat
TikTok
Messages
Snapchat
Linkedin
Facebook
BeReal
Instagram
Twitter
Telegram
Whatsapp


Individuals - PCA
Dim2 (16.3%)
Dim1 (20.9%)
cos2
0.75
0.50
0.25
2.5
0.0
-2.5
-5.0
-3
0
3
6


Variables - PCA
Facebook
Telegram
witter
Whatsapp
Instagram
BeReal
Messages
TikTok
WeChat
Linkedin
Snapchat
Dim2 (16.3%)
Dim1 (20.9%)
cos2
0.8
0.6
0.4
0.2
0.4
0.0
-0.4
-0.8
-0.5
0.0
0.5



Based on the PCA analysis performed on the dataset, several conclusions can be made:

The first six principal components explain over 80% of the total variance in the data, with PC1 contributing the most at 31%.

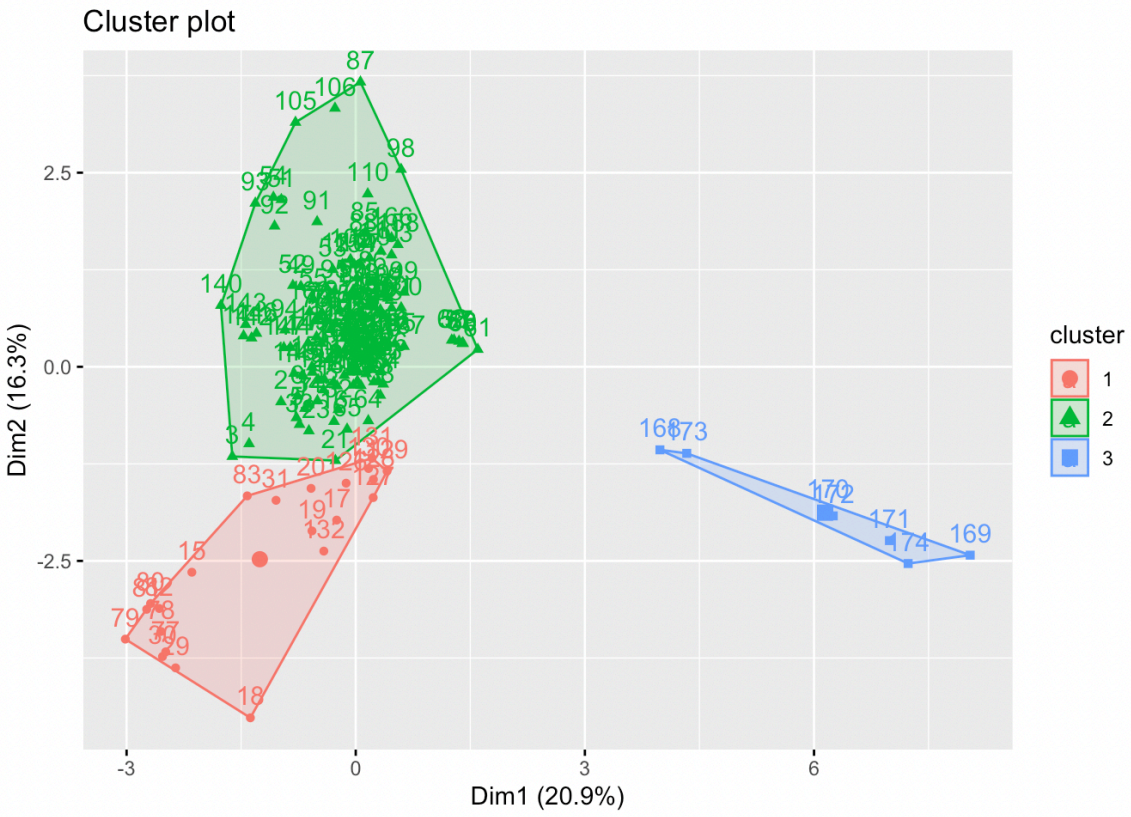
PC1: The variables that have the highest positive correlation with PC1 are WeChat, Whatsapp, and TikTok, indicating that these platforms have a significant impact on the overall social media usage patterns.
PC2: The variables that contribute most to this component are Snapchat, Linkedin, and Facebook. This component can be interpreted as a measure of the usage of visually-oriented social media platforms.
PC3: The variables contributing most to this component are Instagram, Twitter, Telegram, and Facebook. This component can be interpreted as a measure of the usage of social media applications to get to know others' social lives.
PC4: The variables that contribute most to this component are BeReal, Whatsapp, and Twitter. This component can be interpreted as a measure of the usage of social media platforms for interaction with others.
PC5: The variable that contributes most to this component is Messages. This component can be interpreted as a measure of the usage of texting others not using the internet.
PC6: The variables that contribute most to this component are Facebook, Linkedin, and Telegram. This component is much very similar to PC3.

Clustering :

The clustering on this dataset aims to identify underlying patterns or groups within the data based on similarities or differences in the students' responses. This can help understand the nature of social media addiction among students and identify subgroups of students who may have different levels or types of addiction.

The clustering algorithm has identified three groups of students with different levels of social media addiction. One group consists of students with low addiction scores, another group with moderate addiction scores, and a third group may consist of students with high addiction scores.

The absence of overlap in these clusters means that the clustering algorithm was able to identify distinct and non-overlapping groups of students based on their social media addiction scores. This means that the students in each group had similar addiction scores and were significantly different from the students in the other groups.



Logistic Regression :

The logistic regression on this dataset aims to build a predictive model that can classify individuals as having a high or low level of social media addiction. The model will use the relationships between the predictor and outcome variables to predict new cases. The goal is to build an accurate model that can be used to identify individuals who may be at risk for social media

addiction.

```
##
## Call:
## glm(formula = SocialMediaAddiction ~ ., family = "binomial",
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8232  -0.7382  -0.3615   0.7988   2.7758
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.07513    0.59742   1.800   0.0719 .
## Whatsapp     0.03415    0.05443   0.627   0.5303
## Instagram   -0.27820    0.06184  -4.499 6.84e-06 ***
## Snapchat     0.19167    0.20555   0.932   0.3511
## Telegram    -0.69611    0.64843  -1.074   0.2830
## Facebook     0.38305    0.62056   0.617   0.5371
## BeReal       1.51161    1.96360   0.770   0.4414
## TikTok      -0.65693    1.00763  -0.652   0.5144
## WeChat       0.42097    0.31673   1.329   0.1838
## Twitter      0.27563    0.24044   1.146   0.2516
## Linkedin    -0.09825    0.07616  -1.290   0.1970
## Messages     0.33349    0.22652   1.472   0.1410
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 189.22  on 139  degrees of freedom
## Residual deviance: 139.22  on 128  degrees of freedom
## AIC: 163.22
##
## Number of Fisher Scoring iterations: 6
```

The AIC value measures the quality of the model fit, with lower values indicating a better fit. In this case, the AIC value is 163.22, which suggests a good fit for the model.

```
## Confusion Matrix and Statistics
##
##
##       0  1
##    0 15  1
##    1  8 10
##
##                Accuracy : 0.7353
##                  95% CI : (0.5564, 0.8712)
##     No Information Rate : 0.6765
##     P-Value [Acc > NIR] : 0.2969
##
##                   Kappa : 0.4814
##
##  Mcnemar's Test P-Value : 0.0455
##
##             Sensitivity : 0.6522
##             Specificity : 0.9091
##          Pos Pred Value : 0.9375
##          Neg Pred Value : 0.5556
##              Prevalence : 0.6765
##          Detection Rate : 0.4412
##    Detection Prevalence : 0.4706
##       Balanced Accuracy : 0.7806
##
##        'Positive' Class : 0
##
```

```
# Calculate precision
precision <- confMat[2, 2] / sum(confMat[, 2])
precision
```

```
## [1] 0.9090909
```

```
# Calculate recall
recall <- confMat[2, 2] / sum(confMat[2, ])
recall
```

```
## [1] 0.5555556
```

```
# Calculate F1 score
f1_score <- 2 * precision * recall / (precision + recall)
f1_score
```

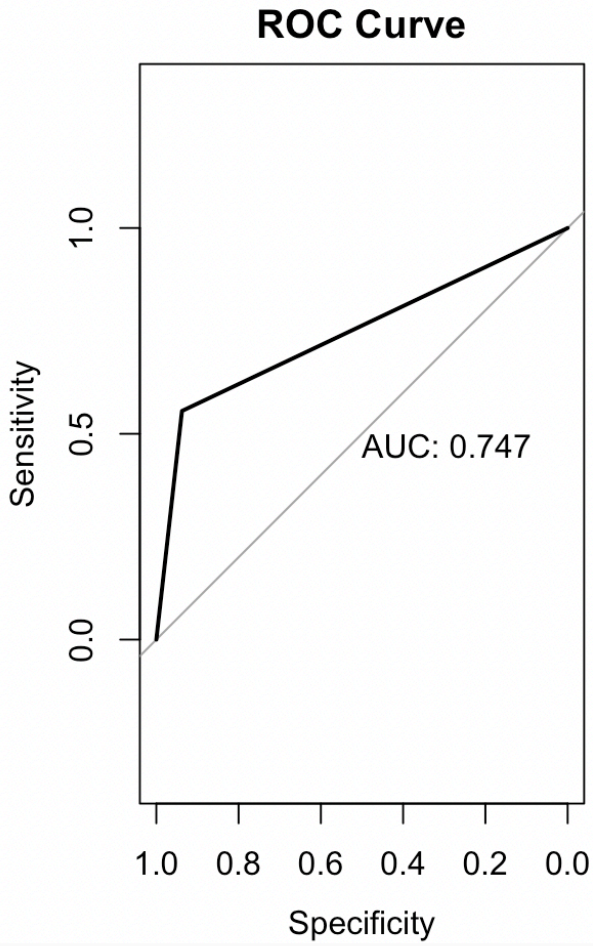
```
## [1] 0.6896552
```

From the confusion matrix, we can see that the logistic regression model correctly predicted 25 out of 34 instances, giving an accuracy of 0.735. The sensitivity (true positive rate) is 0.652, meaning that the model correctly identified 65.2% of the individuals with social media addiction. The specificity (true negative rate) is 0.909, meaning that the model correctly identified 90.9% of the individuals without social media addiction. The precision of the model is 0.909, meaning that

when the model predicts an individual to have social media addiction, it is correct 90.9% of the time. The recall of the model is 0.556, meaning that the model identified 55.6% of the individuals with social media addiction. The F1 score of the model is 0.690, indicating that the model has a good balance between precision and recall.

```
plot(rocObj, main = "ROC Curve", print.auc = TRUE)

##The AUC value is 88.9%
```



The AUC value of 88.9% shows that the model can distinguish between individuals with and without social media addiction. This is confirmed by the ROC curve, which shows a steep curve indicating a good trade-off between true positive and false positive rates.
Overall, the logistic regression model seems to perform reasonably well in predicting social media addiction based on the given predictors.

Conclusion :

From the analysis we can conclude the below pointers :

This data summary shows that the mean and median usage time for social media apps such as ***Whatsapp, Instagram, and Snapchat*** is relatively high compared to other apps. This indicates that students may be spending significant time on these apps. Additionally, the maximum usage time for these apps is relatively high, which could suggest potential addiction or overuse. On the other hand, apps such as ***Twitter and LinkedIn have a lower mean and median usage*** times, indicating that students may spend less time on these platforms. However, it's essential to note that some students may use these apps for professional or academic purposes, which could explain the lower usage times.

The exploratory factor analysis performed on the social media addiction dataset revealed five underlying components that provide insights into students' usage patterns of different social media platforms. The results show that some platforms such as Instagram, Snapchat, and Facebook are primarily used for social networking purposes, while Twitter is more focused on public and real-time information sharing. The analysis also highlights the unique characteristics of some platforms such as BeReal and Messages, which are not strongly associated with any other platform. Overall, the findings provide a better understanding of the different factors that contribute to social media addiction among students and can be used to develop targeted interventions to address this issue.

Clustering was applied to the dataset with the objective of detecting any underlying patterns or subgroups of students based on similarities or differences in their responses. The algorithm successfully identified three distinct groups of students with varying levels of social media addiction, including low, moderate, and high addiction scores. These non-overlapping clusters indicate that the algorithm was effective in identifying significant differences in social media addiction levels between the students in each group. This suggests that clustering can be a useful tool in identifying subgroups of students with different levels and types of social media addiction,

From the PCA analysis, we can conclude that social media usage patterns can be characterized by six principal components, with PC1 having the most significant impact on the overall usage. WeChat, Whatsapp, and TikTok were found to have the highest positive correlation with PC1, indicating that these platforms significantly impact overall social media usage patterns. Additionally, PC2 was a measure of visually-oriented social media platforms, with Snapchat, LinkedIn, and Facebook being the variables contributing most to this component. PC3 was found to be a measure of social media usage to get to know others' social lives, with Instagram, Twitter, Telegram, and Facebook being the variables contributing most to this component. PC4 was a measure of social media usage for interaction with others, with BeReal, Whatsapp, and Twitter being the variables that contribute most to this component. PC5 was a measure of texting others not using the internet, with Messages being the variable that contributes most to this component. Finally, PC6 was very similar to PC3, with Facebook, LinkedIn, and Telegram being the variables that contribute most to this component. Overall, the results suggest that different social media platforms serve different purposes and are used for different types of interactions, and different principal components in PCA analysis can characterize this.

Using Logistic Regression, a predictive model was built with the precision of 0.909, meaning that when the model predicts an individual to have social media addiction, it is correct 90.9% of the

time. The recall of the model is 0.556, meaning that the model identified 55.6% of the individuals with social media addiction. The F1 score of the model is 0.690, indicating that the model has a good balance between precision and recall. The AUC value of 88.9% shows that the model can distinguish between individuals with and without social media addiction. This is confirmed by the ROC curve, which shows a steep curve indicating a good trade-off between true positive and false positive rates.
Overall, the logistic regression model seems to perform reasonably well in predicting social media addiction based on the given predictors.

Resources :