REPORT ON
MULTIVARIATE ANALYSIS
Social Media Class Project

Source of dataset:

The dataset was collectively generated by the entire class and comprises information on social media app usage on a weekly basis, including the frequency of app openings. The data collection process relied on manual tracking, as mobile phones maintain records of app usage duration. The ultimate goal of the dataset was to determine whether individuals are addicted to social media or not, based on the frequency of app openings. A threshold of 105 app openings was established as the criterion for making the determination of addiction/non-addiction.

Link:
https://docs.google.com/spreadsheets/d/1acMUlnd25q6ShvsT7xXkUGurXKCDO4MnKu99ftDQ-DI/edit#gid=0

Data Dictionary:

Name of Variable	Description	Type of Value
Student	The name of the student who provided the data	String
Week	The timeframe of the data collection, typically a one-week period.	Date
Whatsapp (hrs)	The number of hours spent on Whatsapp during the specified week.	Integer (hrs)
Instagram (hrs)	The number of hours spent on Instagram during the specified week.	Integer (hrs)
Snapchat (hrs)	The number of hours spent on Snapchat during the specified week.	Integer (hrs)
Telegram (hrs)	The number of hours spent on Telegram during the specified week.	Integer (hrs)
Facebook/Messenger (hrs)	The number of hours spent on Facebook/Messenger during the specified week.	Integer (hrs)
BeReal (hrs)	The number of hours spent on BeReal during the specified week.	Integer (hrs)
TikTok (hrs)	The number of hours spent on TikTok during the specified week.	Integer (hrs)
WeChat (hrs)	The number of hours spent on WeChat during the specified week.	Integer (hrs)
Twitter (hrs)	The number of hours spent on Twitter during the specified week.	Integer (hrs)

Linkedin (hrs)	The number of hours spent on Linkedin during the specified week.	Integer (hrs)
Messages (hrs)	The number of hours spent on Messages during the specified week.	Integer (hrs)
Total Social Media Screen Time (hrs)	The total number of hours spent on all social media apps combined during the specified week.	Integer (hrs)
Number of times opened (hourly intervals)	The number of times the social media apps were opened in hourly intervals during the specified week.	Integer (count)
Social Media Addiction	A binary indicator indicating whether the individual is addicted or not addicted to social media, based on a predefined threshold of app openings (e.g., 105 times per week).	Binary

Questions that I tried to answer (Hypothesis):

1. How time spent on social media reveals underlying pattern on addiction?

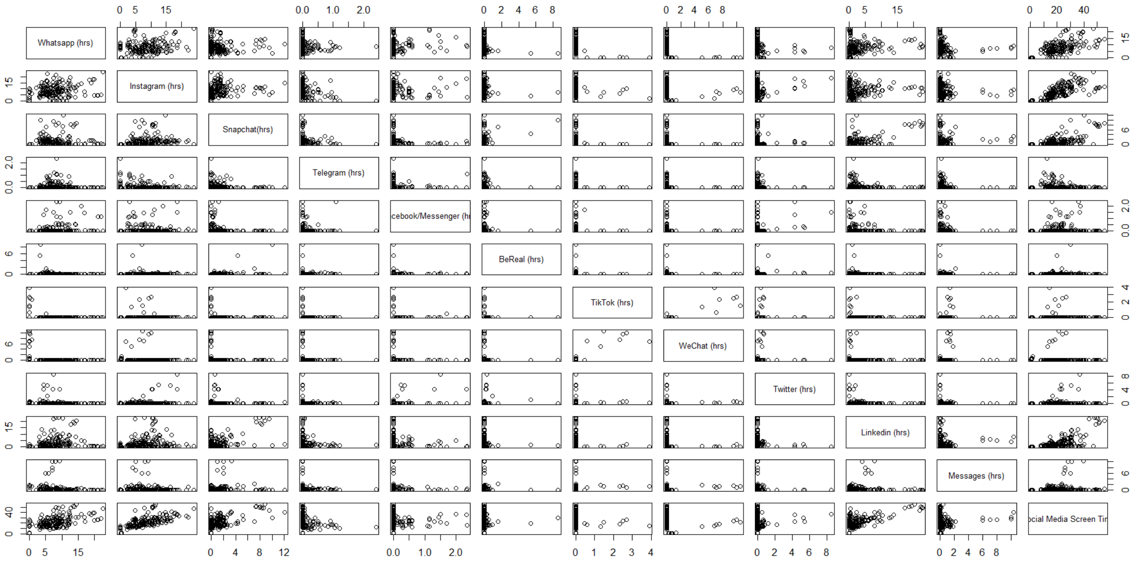
2. Are we able to distinguish between different people's choices based on the time spent on app?

3. Does time spent on each app really counts towards person's being addicted?

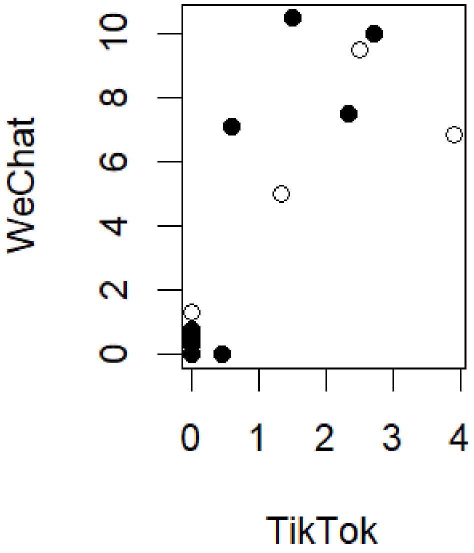
Note: Brief conclusion on last page

Conclusion:

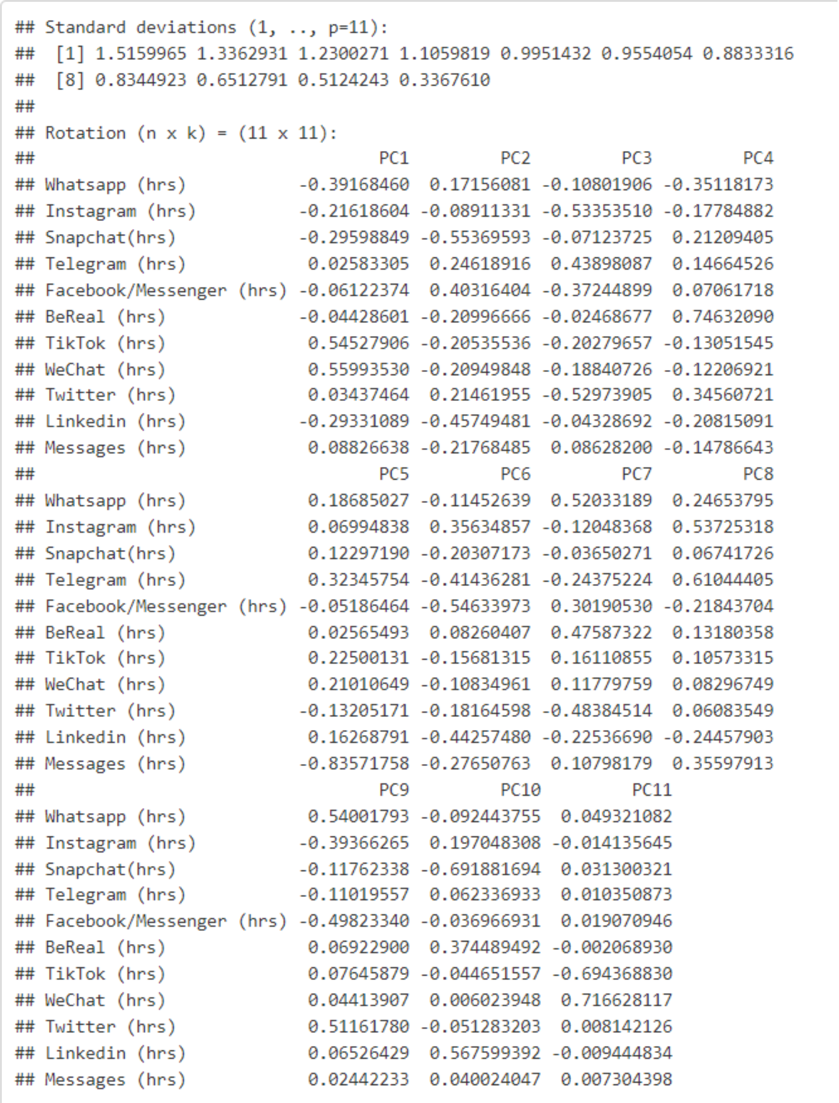
➢ Principal Component Analysis (PCA):



- Whatsapp and Total Social Media Screen Time is positively correlated.
- Instagram and Total Social Media Screen Time is positively correlated.
- Snapchat and Total Social Media Screen Time is positively correlated, but not as Whatsapp and Instagram.
- Even LinkedIn is quiet correlated positively with Total Social Media Screen Time.
- **From this we can derive that these four apps are responsible for making a person socially additive, but to prove our point we need some analysis.**

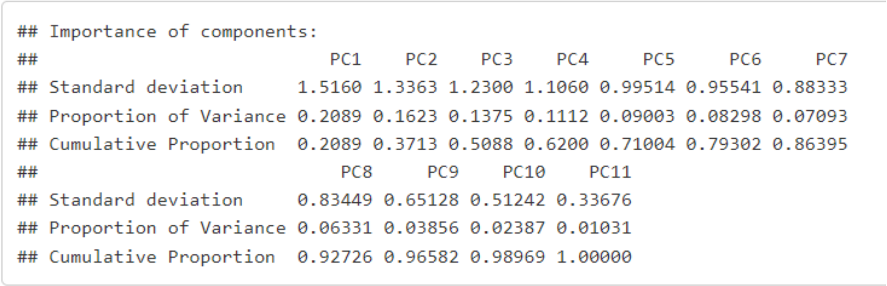


- TikTok and WeChat also look positively correlated but due to lack of data, we can't guarantee.



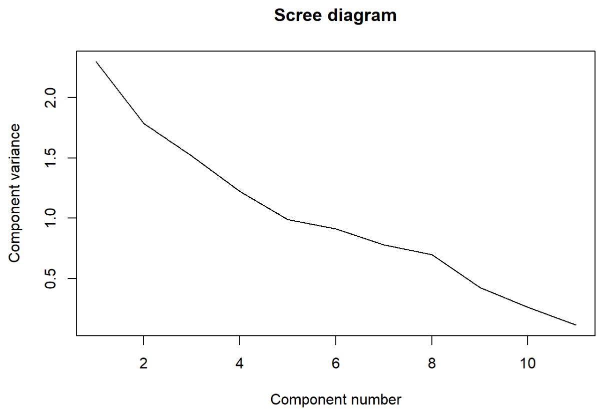
```
## Standard deviations (1, .., p=11):
##  [1] 1.5159965 1.3362931 1.2300271 1.1059819 0.9951432 0.9554054 0.8833316
##  [8] 0.8344923 0.6512791 0.5124243 0.3367610
##
## Rotation (n x k) = (11 x 11):
##                                PC1         PC2         PC3         PC4
## Whatsapp (hrs)          -0.39168460  0.17156081 -0.10801906 -0.35118173
## Instagram (hrs)         -0.21618604 -0.08911331 -0.53353510 -0.17784882
## Snapchat(hrs)           -0.29598849 -0.55369593 -0.07123725  0.21209405
## Telegram (hrs)           0.02583305  0.24618916  0.43898087  0.14664526
## Facebook/Messenger (hrs) -0.06122374  0.40316404 -0.37244899  0.07061718
## BeReal (hrs)            -0.04428601 -0.20996666 -0.02468677  0.74632090
## TikTok (hrs)             0.54527906 -0.20535536 -0.20279657 -0.13051545
## WeChat (hrs)             0.55993530 -0.20949848 -0.18840726 -0.12206921
## Twitter (hrs)            0.03437464  0.21461955 -0.52973905  0.34560721
## Linkedin (hrs)          -0.29331089 -0.45749481 -0.04328692 -0.20815091
## Messages (hrs)           0.08826638 -0.21768485  0.08628200 -0.14786643
##                                PC5         PC6         PC7         PC8
## Whatsapp (hrs)           0.18685027 -0.11452639  0.52033189  0.24653795
## Instagram (hrs)          0.06994838  0.35634857 -0.12048368  0.53725318
## Snapchat(hrs)            0.12297190 -0.20307173 -0.03650271  0.06741726
## Telegram (hrs)           0.32345754 -0.41436281 -0.24375224  0.61044405
## Facebook/Messenger (hrs) -0.05186464 -0.54633973  0.30190530 -0.21843704
## BeReal (hrs)             0.02565493  0.08260407  0.47587322  0.13180358
## TikTok (hrs)             0.22500131 -0.15681315  0.16110855  0.10573315
## WeChat (hrs)             0.21010649 -0.10834961  0.11779759  0.08296749
## Twitter (hrs)           -0.13205171 -0.18164598 -0.48384514  0.06083549
## Linkedin (hrs)           0.16268791 -0.44257480 -0.22536690 -0.24457903
## Messages (hrs)          -0.83571758 -0.27650763  0.10798179  0.35597913
##                                PC9         PC10         PC11
## Whatsapp (hrs)           0.54001793 -0.092443755  0.049321082
## Instagram (hrs)         -0.39366265  0.197048308 -0.014135645
## Snapchat(hrs)           -0.11762338 -0.691881694  0.031300321
## Telegram (hrs)          -0.11019557  0.062336933  0.010350873
## Facebook/Messenger (hrs) -0.49823340 -0.036966931  0.019070946
## BeReal (hrs)             0.06922900  0.374489492 -0.002068930
## TikTok (hrs)             0.07645879 -0.044651557 -0.694368830
## WeChat (hrs)             0.04413907  0.006023948  0.716628117
## Twitter (hrs)            0.51161780 -0.051283203  0.008142126
## Linkedin (hrs)           0.06526429  0.567599392 -0.009444834
## Messages (hrs)           0.02442233  0.040024047  0.007304398
```

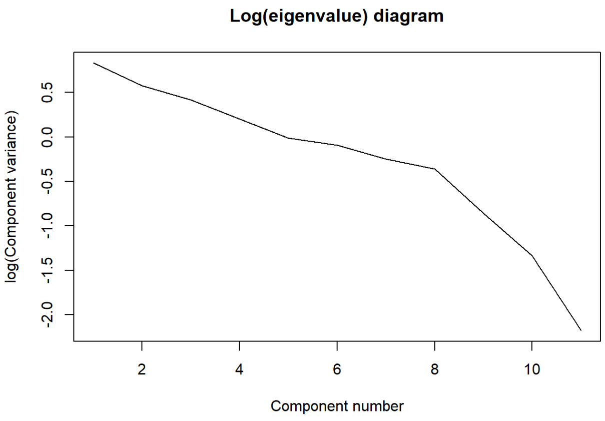
- We got an 11x11 rotation matrix.
- These PCs are ordered in the order of importance, with PC1 being most important and so on.
- The numbers represented as a list shows the loadings/weights of each original variable.
- The larger the value of the loading, the more important the variable is in determining the value of that principal component.
- For example, for PC1, the variables with the higher loading are TikTok (0.5452) and WeChat (0.5599).
- It indicates that these variables are making large contributions to PC1.
- Note: Negative sign indicates the inverse relationship or correlation of that variable to the corresponding principal component.



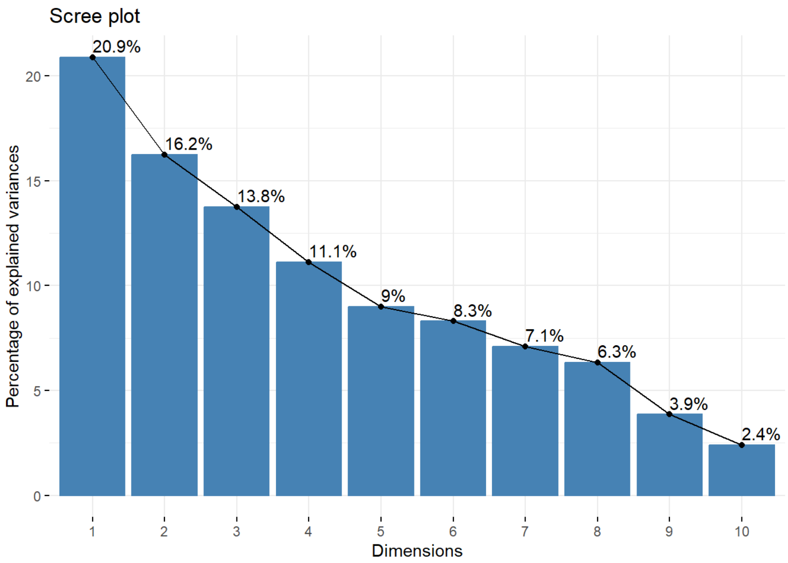
```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.5160 1.3363 1.2300 1.1060 0.99514 0.95541 0.88333
## Proportion of Variance 0.2089 0.1623 0.1375 0.1112 0.09003 0.08298 0.07093
## Cumulative Proportion  0.2089 0.3713 0.5088 0.6200 0.71004 0.79302 0.86395
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.83449 0.65128 0.51242 0.33676
## Proportion of Variance 0.06331 0.03856 0.02387 0.01031
## Cumulative Proportion  0.92726 0.96582 0.98969 1.00000
```

- Standard Deviation indicates the amount of variability or information PC captures from the original variable.
- The proportion of Variance explains the variance explained by each PC.
- Cumulative Proportion is just the combination of the current PC and the PCs before it.
- **Before we go further from here, we can infer that the first 2 components are not able to provide more than 37% of the explained variance.**
- **Therefore, we can say that Principal Component Analysis fails here.**
- **Below provided is a little evidence, why we are saying PCA fails.**





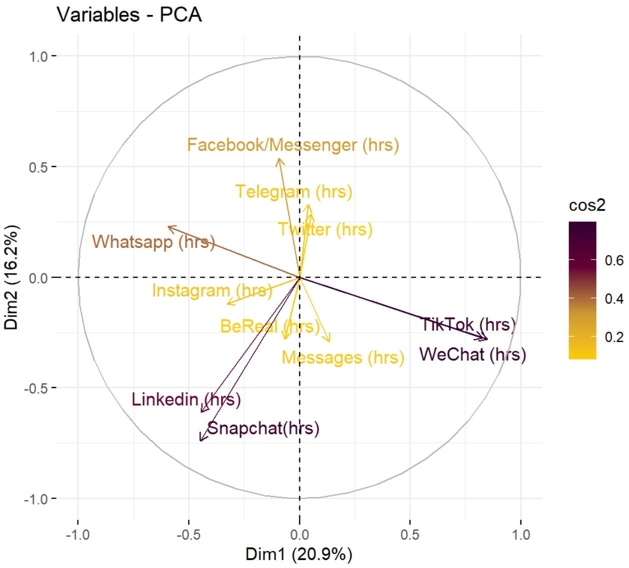
- From the Scree Diagram, we can't say where the elbow is formed.
- It is difficult to observe it from the scree diagram.
- Even looking at the log(eigenvalue) diagram, we can see that the elbow is formed at component number 7.
- From summary we can see that we are able to explain around 86.3% of the total variance using the first 7 components.



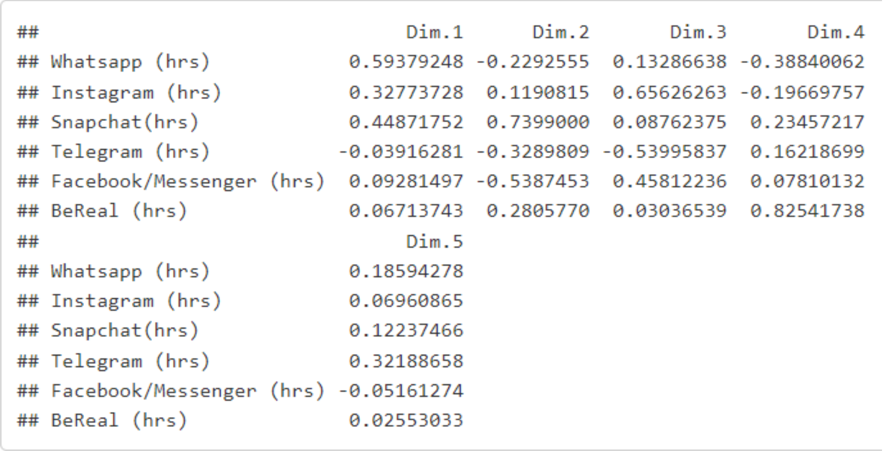
- PC1 explains around 20.9% of variance.
- PC2 explains around 16.2% of variance.
- PC3 explains around 13.8% of variance.
- PC4 explains around 11.1% of variance.
- PC5 explains around 9% of variance.
- PC6 explains around 8.3% of variance.
- PC7 explains around 7.1% of variance.
- PC8 explains around 6.3% of variance.
- PC9 explains around 3.9% of variance.
- PC10 explains around 2.4% of variance.

OPTIONAL ANALYSIS

- If we agree to proceed with the above PCA results, we would be getting the following analysis.

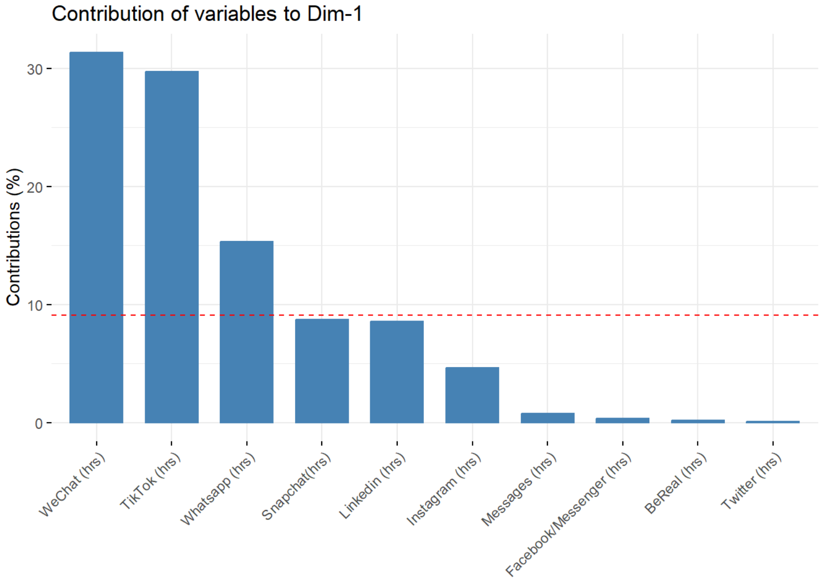


- We can see that TikTok and WeChat are close to each other and can infer that both have similar patterns and are correlated as well.
- Similarly, LinkedIn and SnapChat follow the same trend.

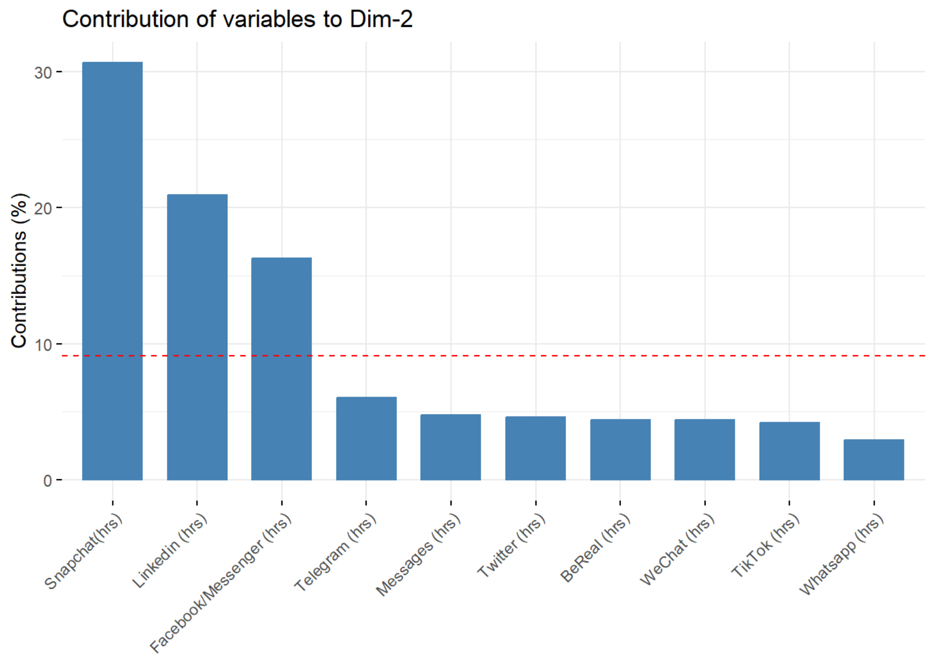


```
##                                Dim.1      Dim.2       Dim.3       Dim.4
## Whatsapp (hrs)            0.59379248 -0.2292555  0.13286638 -0.38840062
## Instagram (hrs)           0.32773728  0.1190815  0.65626263 -0.19669757
## Snapchat(hrs)             0.44871752  0.7399000  0.08762375  0.23457217
## Telegram (hrs)           -0.03916281 -0.3289809 -0.53995837  0.16218699
## Facebook/Messenger (hrs)  0.09281497 -0.5387453  0.45812236  0.07810132
## BeReal (hrs)              0.06713743  0.2805770  0.03036539  0.82541738
##                                Dim.5
## Whatsapp (hrs)            0.18594278
## Instagram (hrs)           0.06960865
## Snapchat(hrs)             0.12237466
## Telegram (hrs)            0.32188658
## Facebook/Messenger (hrs) -0.05161274
## BeReal (hrs)              0.02553033
```

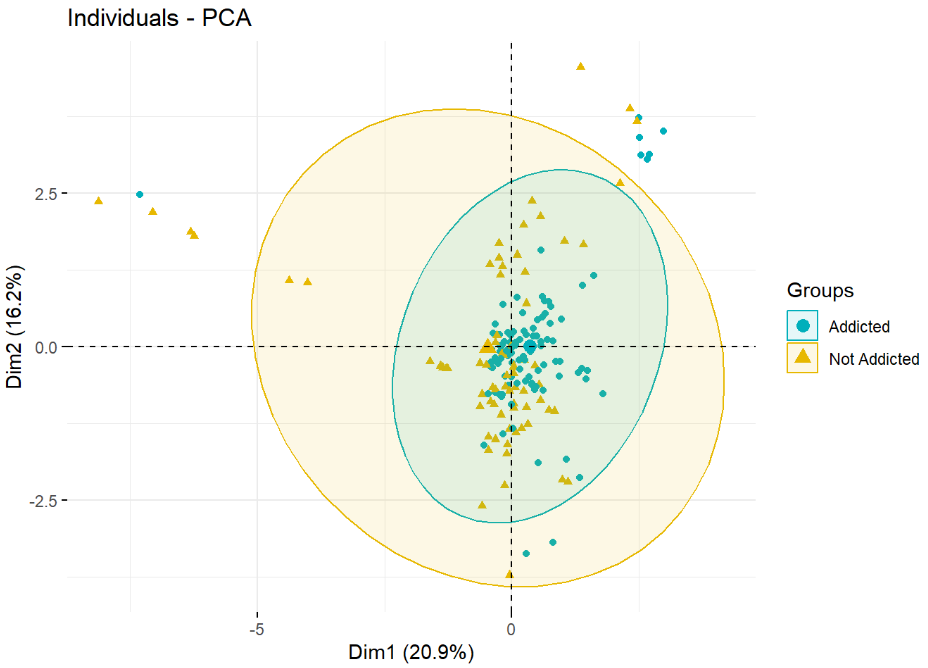
- It shows how each variable contributes to Principal Components and how they are positioned relative to each other in multi-dimensional space.



- We can see that WeChat, TikTok, and Whatsapp are the variables that contribute to Dim. 1 based on this plot.



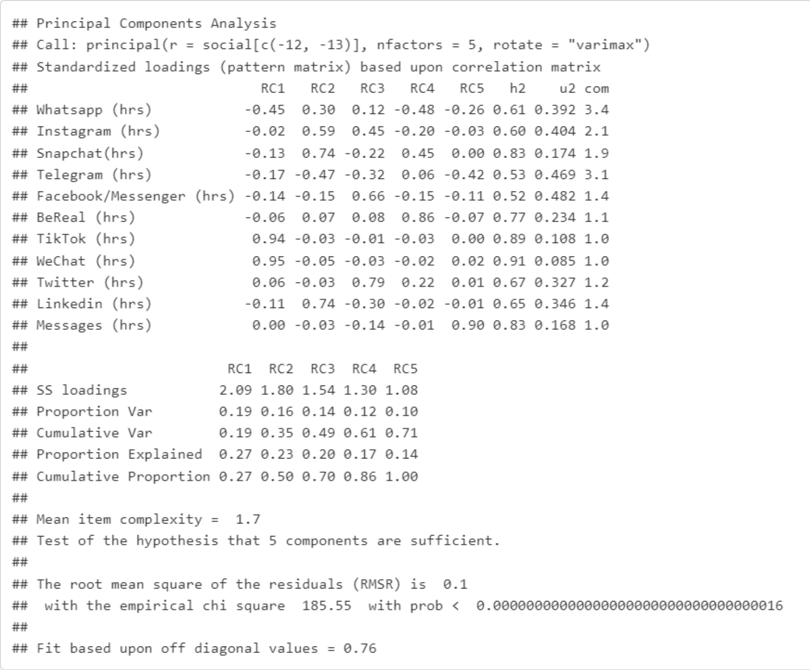
- We can see that SnapChat, Linkedin, and Facebook/Messenger are the variables that contributes to Dim. 2 based on this plot.



- We can make some conclusion from this plot based on our Principal Component that, the blue cluster shows us that, the people who spends more time on Whatsapp, TikTok, WeChat, Snapchat, Linkedin, and Facebook/Messenger are considered as Addictive.
- We can see that the blue clusters are close to the origin.

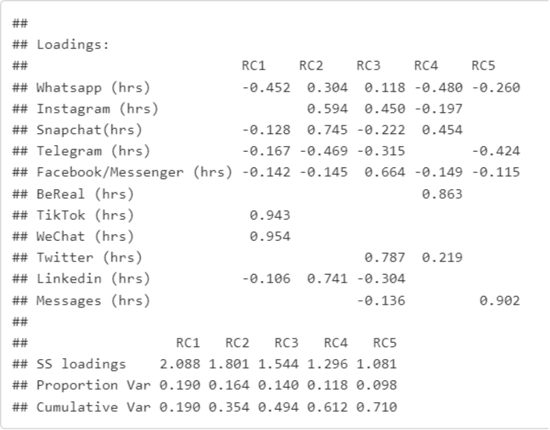
- **But since our first two PCA components are not able to explain more than 70% of the variance, we can't rely on this method to answer our hypothesis.**
- **Therefore, we can say that PCA is not able to explain us about the relationship between the app usage other than discussed above on our dataset.**

- Exploratory Factor Analysis (EFA):



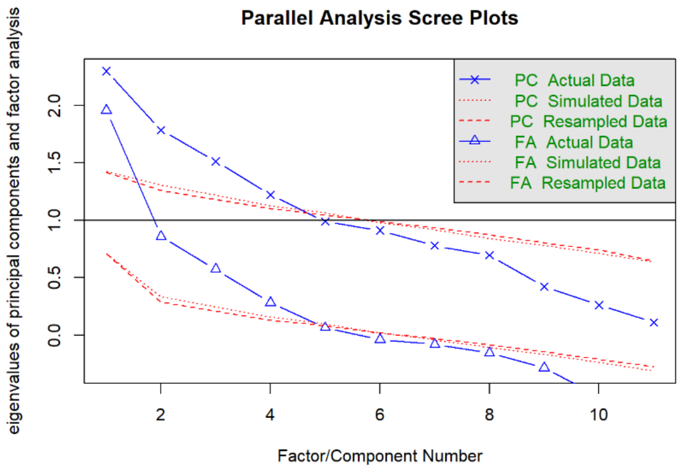
```
## Principal Components Analysis
## Call: principal(r = social[c(-12, -13)], nfactors = 5, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                           RC1   RC2   RC3   RC4   RC5   h2    u2 com
## Whatsapp (hrs)          -0.45  0.30  0.12 -0.48 -0.26 0.61 0.392 3.4
## Instagram (hrs)         -0.02  0.59  0.45 -0.20 -0.03 0.60 0.404 2.1
## Snapchat(hrs)           -0.13  0.74 -0.22  0.45  0.00 0.83 0.174 1.9
## Telegram (hrs)          -0.17 -0.47 -0.32  0.06 -0.42 0.53 0.469 3.1
## Facebook/Messenger (hrs) -0.14 -0.15  0.66 -0.15 -0.11 0.52 0.482 1.4
## BeReal (hrs)            -0.06  0.07  0.08  0.86 -0.07 0.77 0.234 1.1
## TikTok (hrs)             0.94 -0.03 -0.01 -0.03  0.00 0.89 0.108 1.0
## WeChat (hrs)             0.95 -0.05 -0.03 -0.02  0.02 0.91 0.085 1.0
## Twitter (hrs)            0.06 -0.03  0.79  0.22  0.01 0.67 0.327 1.2
## Linkedin (hrs)          -0.11  0.74 -0.30 -0.02 -0.01 0.65 0.346 1.4
## Messages (hrs)           0.00 -0.03 -0.14 -0.01  0.90 0.83 0.168 1.0
##
##                        RC1  RC2  RC3  RC4  RC5
## SS loadings           2.09 1.80 1.54 1.30 1.08
## Proportion Var        0.19 0.16 0.14 0.12 0.10
## Cumulative Var        0.19 0.35 0.49 0.61 0.71
## Proportion Explained  0.27 0.23 0.20 0.17 0.14
## Cumulative Proportion 0.27 0.50 0.70 0.86 1.00
##
## Mean item complexity =  1.7
## Test of the hypothesis that 5 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.1
##  with the empirical chi square  185.55  with prob <  0.000000000000000000000000000000000016
##
## Fit based upon off diagonal values = 0.76
```

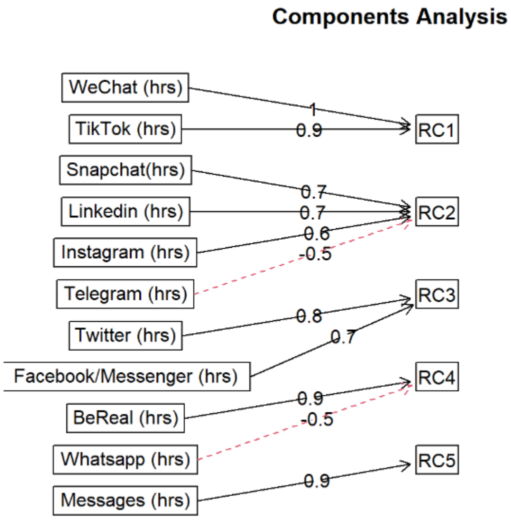
- We passed nfactors = 5 which refers to the number of factors or PCs.
- RC1, RC2, RC3, RC4, and RC5 and so on have some values which represent the loadings of each variables.
- Positive loading indicates that there is a positive correlation between the original variable and extracted Principal Components.
- While negative loading indicates a negative relationship.
- h2 shows the proportion of variance in each variable explained by the principal components.
- u2 shows the proportion of variance in each variable unexplained by the principal components.
- com shows the complexity to explain the variance.
- SS Loadings shows the squared loading of each principal components.
- Proportion Variance shows the proportion of total variance explained by each principal component.
- Cumulative Variance shows the cumulative proportion of total variance explained by each principal component.
- Proportion Explained shows the proportion of total variance explained by each principal component, expressed as a percentage.
- Cumulative Proportion shows the cumulative proportion of total variance explained by each principal component, expressed as a percentage.



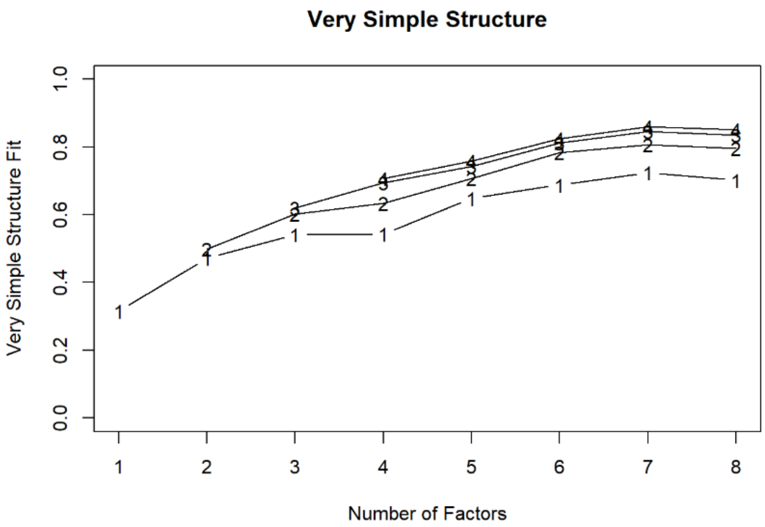
```
##
## Loadings:
##                            RC1    RC2    RC3    RC4    RC5
## Whatsapp (hrs)           -0.452  0.304  0.118 -0.480 -0.260
## Instagram (hrs)                  0.594  0.450 -0.197
## Snapchat(hrs)            -0.128  0.745 -0.222  0.454
## Telegram (hrs)           -0.167 -0.469 -0.315        -0.424
## Facebook/Messenger (hrs) -0.142 -0.145  0.664 -0.149 -0.115
## BeReal (hrs)                                   0.863
## TikTok (hrs)              0.943
## WeChat (hrs)              0.954
## Twitter (hrs)                           0.787  0.219
## Linkedin (hrs)           -0.106  0.741 -0.304
## Messages (hrs)                         -0.136         0.902
##
##                    RC1   RC2   RC3   RC4   RC5
## SS loadings    2.088 1.801 1.544 1.296 1.081
## Proportion Var 0.190 0.164 0.140 0.118 0.098
## Cumulative Var 0.190 0.354 0.494 0.612 0.710
```

- From the result we can make following conclusions:
- RC1 shows the relation with WeChat, and TikTok.
- RC2 shows the relation with Snapchat, LinkedIn, Instagram, and Telegram.
- RC3 shows the relation with Twitter, and Facebook/Messenger.
- RC4 shows the relation with BeReal, and Whatsapp.
- RC5 shows the relation with only Messages. Therefore, we can just take Messages as a features and exclude RC5 since it only has one variable in it that is contribution or showing relationship.



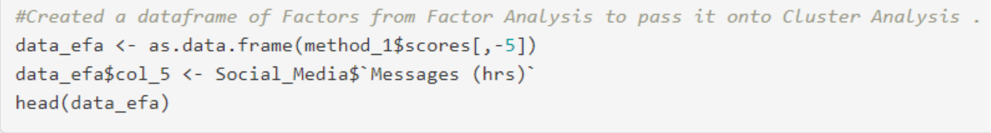


- Parallel analysis suggests that the number of factors = 4 and the number of components = 4.
- We can say that we reduced the factors from 8 to 4.
- It can be seen from the Components Analysis Plot.

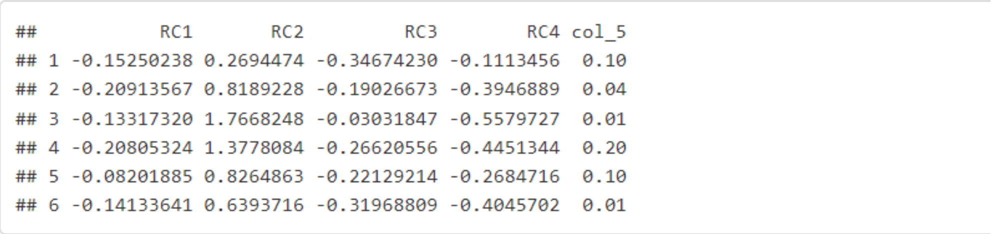


- We can also see from Very Simple Structure that number of factors = 4 is a good fit.

- ➢ Clustering Analysis:
- We are using the factors for clustering, because our original data was getting overlapped when forming the clusters.

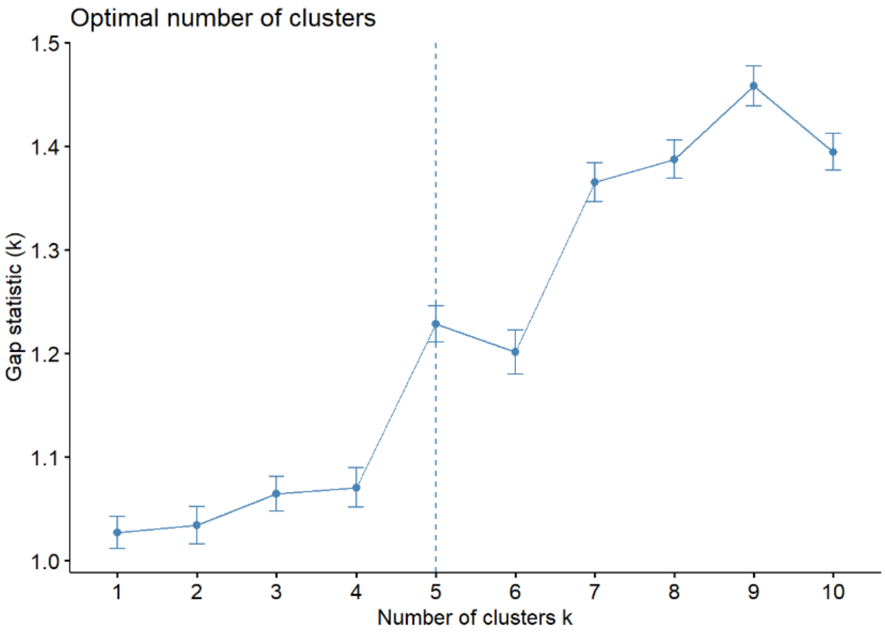


```
#Created a dataframe of Factors from Factor Analysis to pass it onto Cluster Analysis .
data_efa <- as.data.frame(method_1$scores[,-5])
data_efa$col_5 <- Social_Media$`Messages (hrs)`
head(data_efa)
```

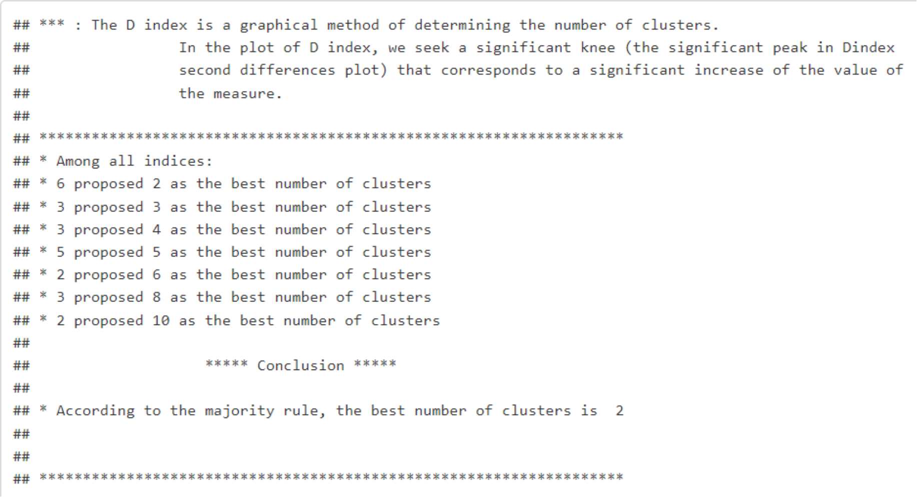


```
##           RC1       RC2         RC3        RC4 col_5
## 1 -0.15250238 0.2694474 -0.34674230 -0.1113456  0.10
## 2 -0.20913567 0.8189228 -0.19026673 -0.3946889  0.04
## 3 -0.13317320 1.7668248 -0.03031847 -0.5579727  0.01
## 4 -0.20805324 1.3778084 -0.26620556 -0.4451344  0.20
## 5 -0.08201885 0.8264863 -0.22129214 -0.2684716  0.10
## 6 -0.14133641 0.6393716 -0.31968809 -0.4045702  0.01
```

- Since, clustering was not useful for our original dataset and therefore, we are using the Factors we got from our factor analysis.

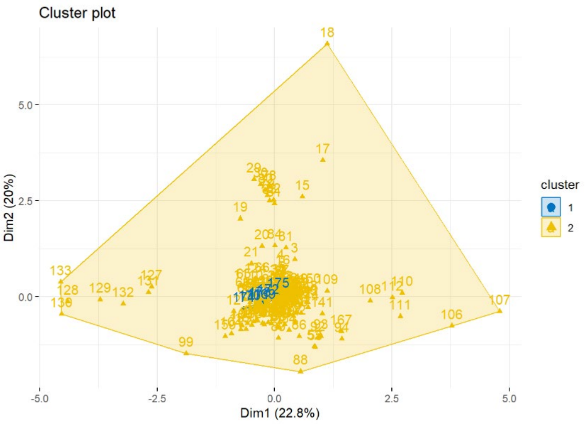


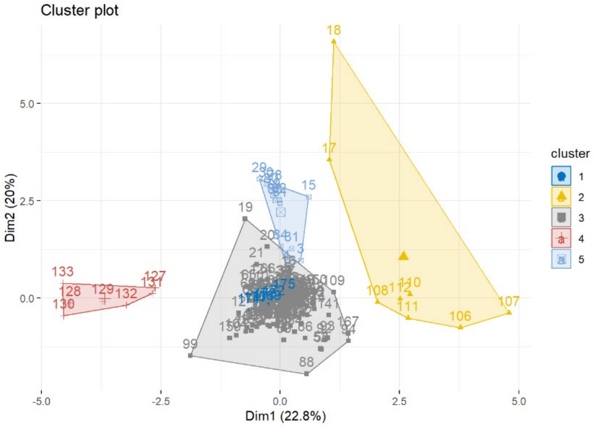
- We used fviz_nbclust to find the optimal number of clusters. Here, we got our k = 5.

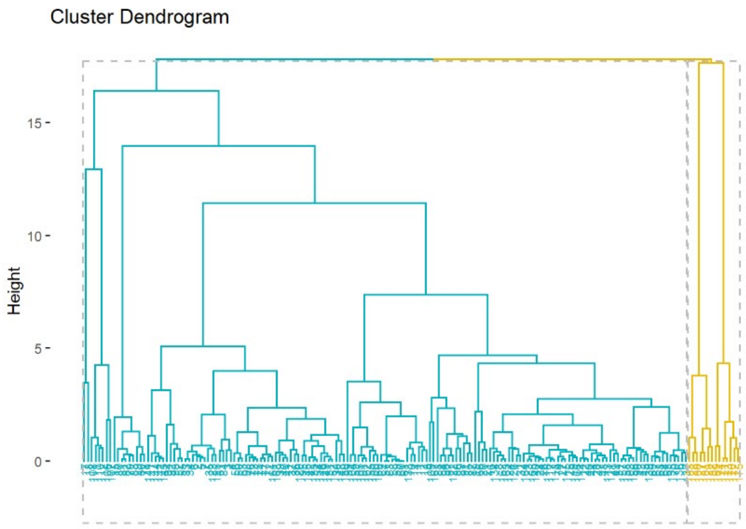


```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 6 proposed 2 as the best number of clusters
## * 3 proposed 3 as the best number of clusters
## * 3 proposed 4 as the best number of clusters
## * 5 proposed 5 as the best number of clusters
## * 2 proposed 6 as the best number of clusters
## * 3 proposed 8 as the best number of clusters
## * 2 proposed 10 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```

- When using Hubert Index and D-index method, we got the optimal number of cluster as 2.
- We tried to plot the graph for k = 2 and 5.
- We can see that; both the cluster plots have overlapping.

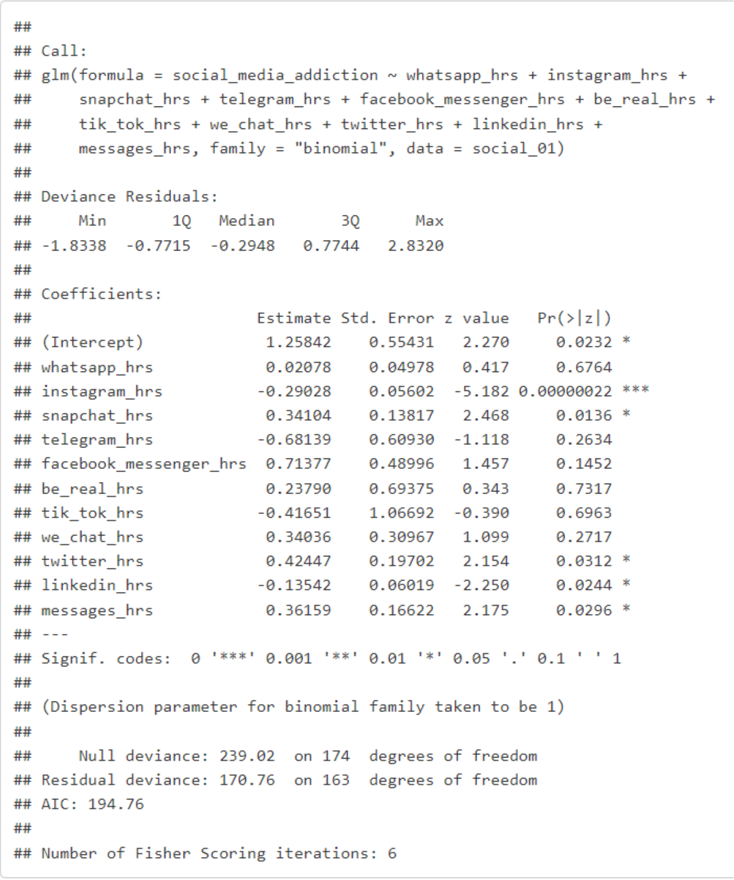






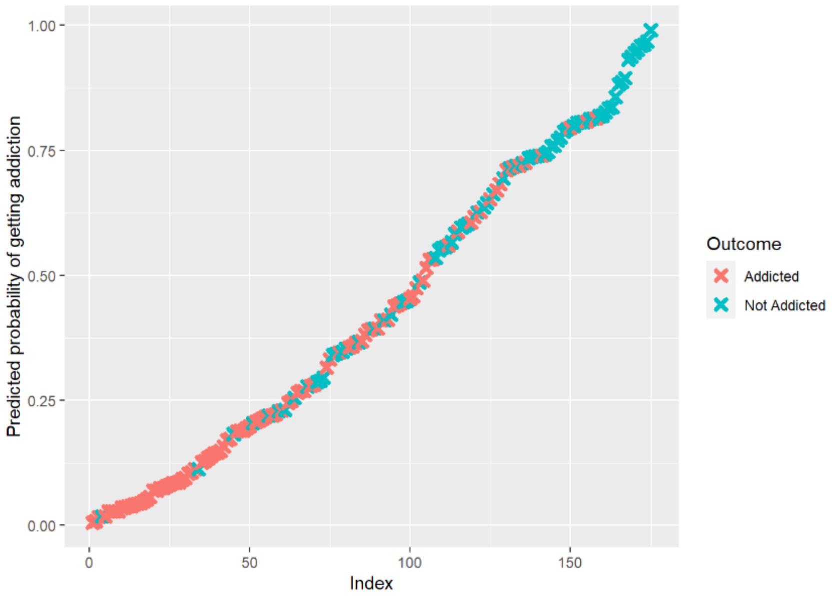
- We can say that clustering can't be as useful to classify between addicted and not addicted.
- We can say this because, when we plotted our clusters, we got overlapped clusters.

- Logistic Regression:

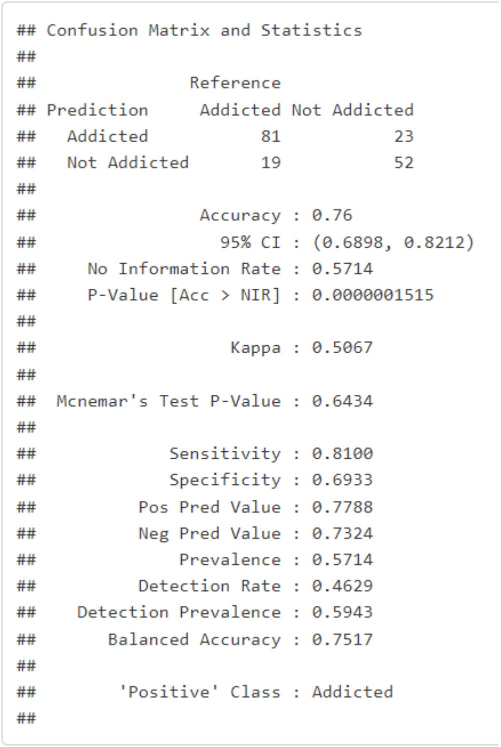


```
##
## Call:
## glm(formula = social_media_addiction ~ whatsapp_hrs + instagram_hrs +
##     snapchat_hrs + telegram_hrs + facebook_messenger_hrs + be_real_hrs +
##     tik_tok_hrs + we_chat_hrs + twitter_hrs + linkedin_hrs +
##     messages_hrs, family = "binomial", data = social_01)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8338  -0.7715  -0.2948   0.7744   2.8320
##
## Coefficients:
##                        Estimate Std. Error z value   Pr(>|z|)
## (Intercept)             1.25842    0.55431   2.270     0.0232 *
## whatsapp_hrs            0.02078    0.04978   0.417     0.6764
## instagram_hrs          -0.29028    0.05602  -5.182 0.00000022 ***
## snapchat_hrs            0.34104    0.13817   2.468     0.0136 *
## telegram_hrs           -0.68139    0.60930  -1.118     0.2634
## facebook_messenger_hrs  0.71377    0.48996   1.457     0.1452
## be_real_hrs             0.23790    0.69375   0.343     0.7317
## tik_tok_hrs            -0.41651    1.06692  -0.390     0.6963
## we_chat_hrs             0.34036    0.30967   1.099     0.2717
## twitter_hrs             0.42447    0.19702   2.154     0.0312 *
## linkedin_hrs           -0.13542    0.06019  -2.250     0.0244 *
## messages_hrs            0.36159    0.16622   2.175     0.0296 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 239.02  on 174  degrees of freedom
## Residual deviance: 170.76  on 163  degrees of freedom
## AIC: 194.76
##
## Number of Fisher Scoring iterations: 6
```

- This is the logistic regression with all variables.
- The intercept is the log(odds) a person will be Addicted.
- Each one-unit change in the number of snapchat_hrs will increase the log odds of getting addicted by 0.34104.
- Each one-unit change in the whatsapp_hrs will increase the log odds of getting addicted by 0.02078.
- Even the p-value states that both these attributes are quite significant in determining addiction.
- The interpretation of telegram_hrs is different.
- Getting telegram_hrs changed by one unit will decrease the log odds of getting addicted by -0.68139.
- Difference between Null deviance and Residual deviance tells us that the model is a good fit.
- The greater the difference means better the model.

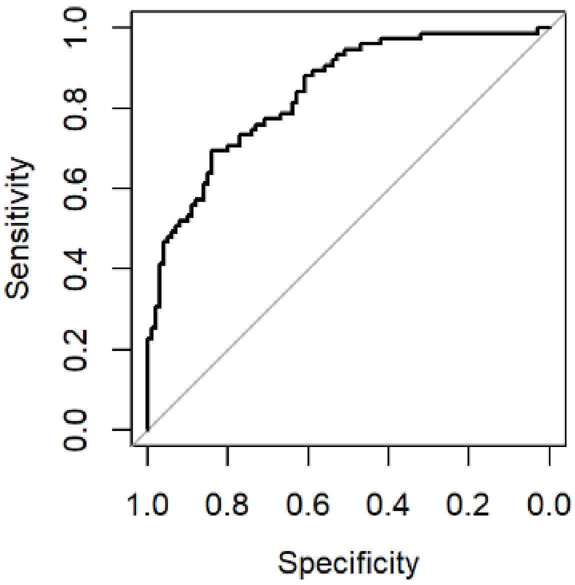


- We know that as probability of getting addiction increases, a person has greater chance of not having addiction and it is showed with blue color.
- Similarly, a person having less probability of addiction has high chance of being addicted.
- In our case, 0.50 (approximately from the graph), can be identified as the cut-off or threshold of having addiction and not having addiction.



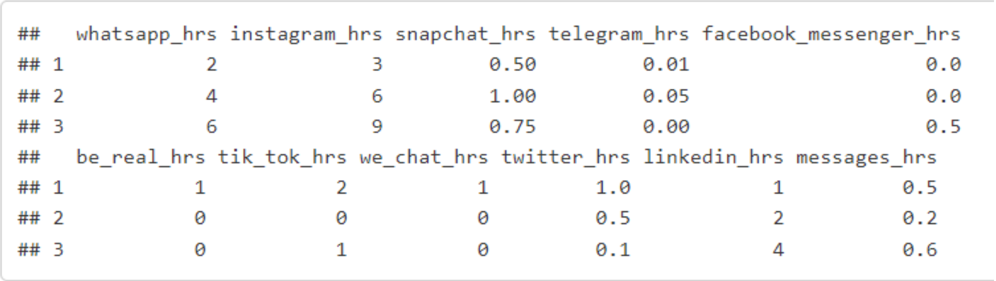
```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     Addicted Not Addicted
##   Addicted           81           23
##   Not Addicted       19           52
##
##                Accuracy : 0.76
##                  95% CI : (0.6898, 0.8212)
##     No Information Rate : 0.5714
##     P-Value [Acc > NIR] : 0.0000001515
##
##                   Kappa : 0.5067
##
##  Mcnemar's Test P-Value : 0.6434
##
##             Sensitivity : 0.8100
##             Specificity : 0.6933
##          Pos Pred Value : 0.7788
##          Neg Pred Value : 0.7324
##              Prevalence : 0.5714
##          Detection Rate : 0.4629
##    Detection Prevalence : 0.5943
##       Balanced Accuracy : 0.7517
##
##        'Positive' Class : Addicted
##
```

- 81 + 52 = 133 were predicted as Addicted and Not Addicted correctly (diagonal)
- 19 + 23 = 42 were predicted incorrectly (off-diagonal).
- The accuracy is reported as 0.76, which means that the model correctly predicted around 76% of the instances.
- The NIR is reported as 0.5714, which is the accuracy that would be achieved by always predicting Addicted Class.
- A higher kappa value indicates a better agreement between the model's predictions and the actual values.
- In our case, the kappa is reported as 0.5067, which indicates moderate agreement between model's prediction and actual prediction.
- Sensitivity is reported as 0.8100, indicating that the model correctly predicted around 81.00% of the instances of class Addicted.
- Specificity is reported as 0.6833, indicating that the model correctly predicted around 69.33% of the instances of class Not Addicted.
- Balanced Accuracy is the average of Sensitivity and Specificity.
- We got balanced accuracy as 0.7517.



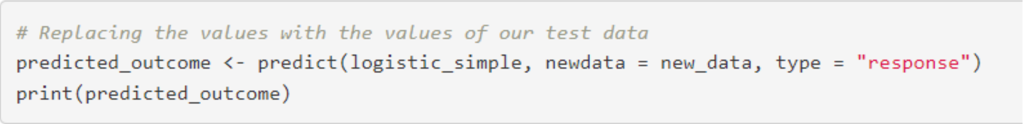
- ROC curve is a graphical plot which shows the trade-off between the true positive rate (Sensitivity) and the false positive rate (Specificity).
- Higher value indicates better overall model performance.
- The AUC is reported as 0.8396, indicating that the model has good discriminatory power with an AUC value close to 1, suggesting that it can effectively distinguish between the two classes (Addicted and Not Addicted) based on the predicted probabilities from the logistic regression model.

- We created a new dataframe to predict on our logistic regression model.

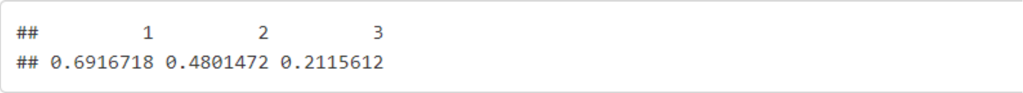


```
##   whatsapp_hrs instagram_hrs snapchat_hrs telegram_hrs facebook_messenger_hrs
## 1            2             3         0.50         0.01                    0.0
## 2            4             6         1.00         0.05                    0.0
## 3            6             9         0.75         0.00                    0.5
##   be_real_hrs tik_tok_hrs we_chat_hrs twitter_hrs linkedin_hrs messages_hrs
## 1           1           2           1         1.0            1          0.5
## 2           0           0           0         0.5            2          0.2
## 3           0           1           0         0.1            4          0.6
```

- The below is the predicted probability we got from our logistic regression model.



```
# Replacing the values with the values of our test data
predicted_outcome <- predict(logistic_simple, newdata = new_data, type = "response")
print(predicted_outcome)
```



```
##         1         2         3
## 0.6916718 0.4801472 0.2115612
```

- For 1st, we got probability of 0.6916, which means that the person was predicted as **Not-Addictive.**
- For 2nd, we got probability of 0.4801, which means that the person was predicted as **Addictive** (Moderately).
- For 3rd, we got probability of 0.2511, which means that the person was predicted as **Addictive** (Highly).

- Brief Conclusion:

- After performing Principal Component Analysis (PCA), we observed that the first two components did not account for more than 70% of the variance. Based on this, we concluded that PCA was not useful in our analysis.
- We identified correlations and insights among variables such as WhatsApp, Instagram, Snapchat, and Linkedin with Total social media screen time, suggesting that these variables contribute to social media addiction.
- Through Exploratory Factor Analysis (EFA), we reduced our initial set of 11 variables to 5, including 4 factors and one variable (Messages). This was supported by Scree Plot and Very Simple Structure analyses, providing evidence for our decision.
- Subsequently, we conducted Clustering Analysis on the original data, but found that the clusters were overlapping, leading to ambiguous results.
- To address this, we performed clustering analysis on our EFA factors and identified two optimal clusters with k values of 2 and 5. However, even with this approach, we observed overlapping clusters, which made it challenging to clearly determine addiction status.
- Finally, we utilized Logistic Regression and built a model using all the variables. The resulting accuracy was 76%. We also created a dataframe to predict addiction status for random values as a validation step.