Social Media Usage Analysis

Introduction:

We collected the social media usage data of 25 students from the class. The data points include Time spent on social media apps such as WhatsApp, Instagram, Facebook, etc., per week, Number of times opened (hourly intervals), and Social Media Addiction Level. The data set has 175 observations of 16 variables.

The description of variables is as follows:

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?

Questions:

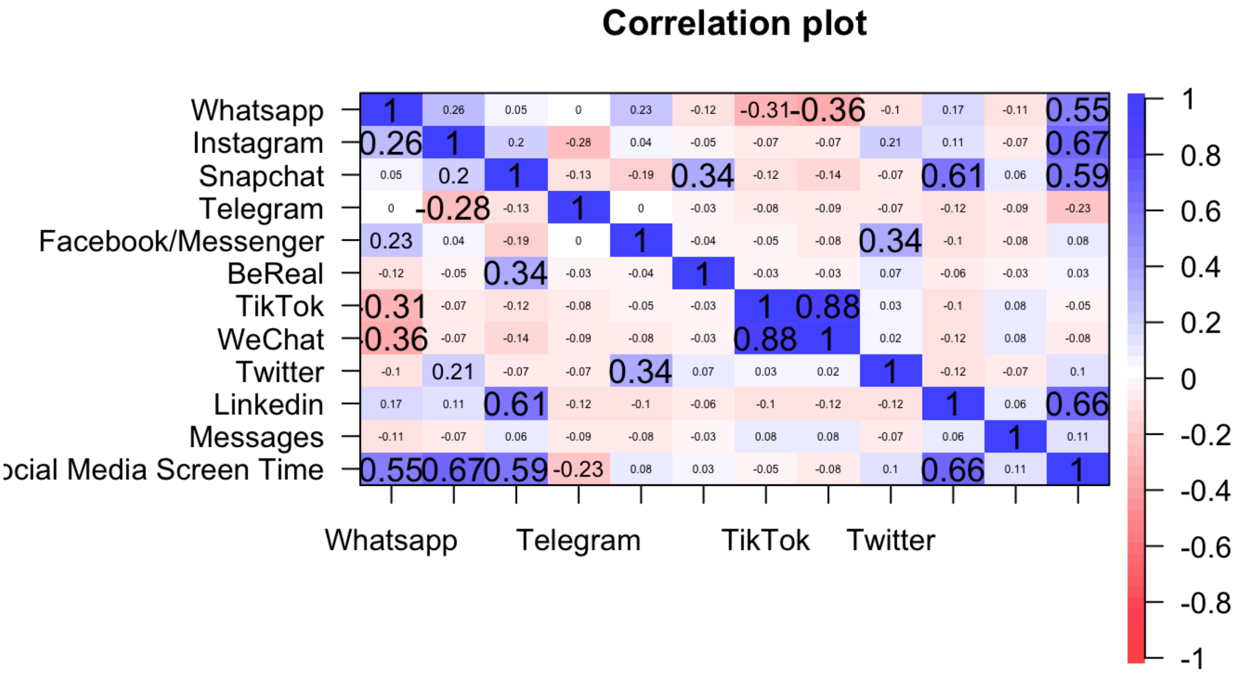
1. Use principal components analysis to investigate the relationships between the number of hours spent and addiction.
2. Carry out cluster analysis to study the relation between the number of hours spent and addiction.
3. Identify the important factors underlying the observed variables and examine the relationships between the addiction with respect to these factors.
4. Can you predict social media addiction from the number of hours spent on these aforementioned apps as input?

Hypothesis:

1. There may be a relationship between the number of hours spent on social media apps and addiction. This hypothesis could be tested by conducting a principal components analysis to explore the relationships between these variables.
2. There may be underlying factors that influence both social media addiction and the number of hours spent on these apps. This hypothesis could be tested by identifying these factors using factor analysis or other similar techniques and then examining the relationship.
3. That there may be distinct groups or clusters of individuals based on their level of addiction and the number of hours spent on social media apps. This hypothesis could be tested by carrying out cluster analysis to identify these groups or clusters.
4. Predicting social media addiction from the number of hours spent on these apps may be possible. This hypothesis could be tested by developing a logistic regression predictive model, with the number of hours spent as the input variable and addiction as the outcome variable.

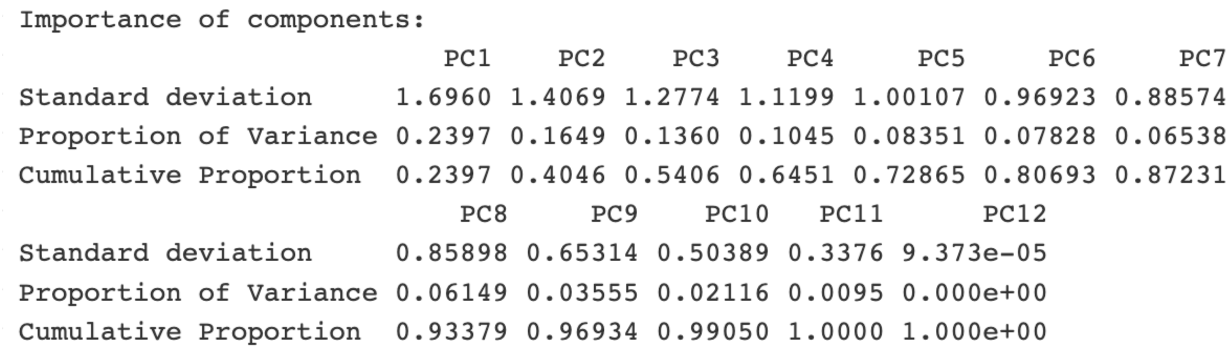
Working:

First, let's look at how the variables are correlated to each other by plotting a correlation plot.



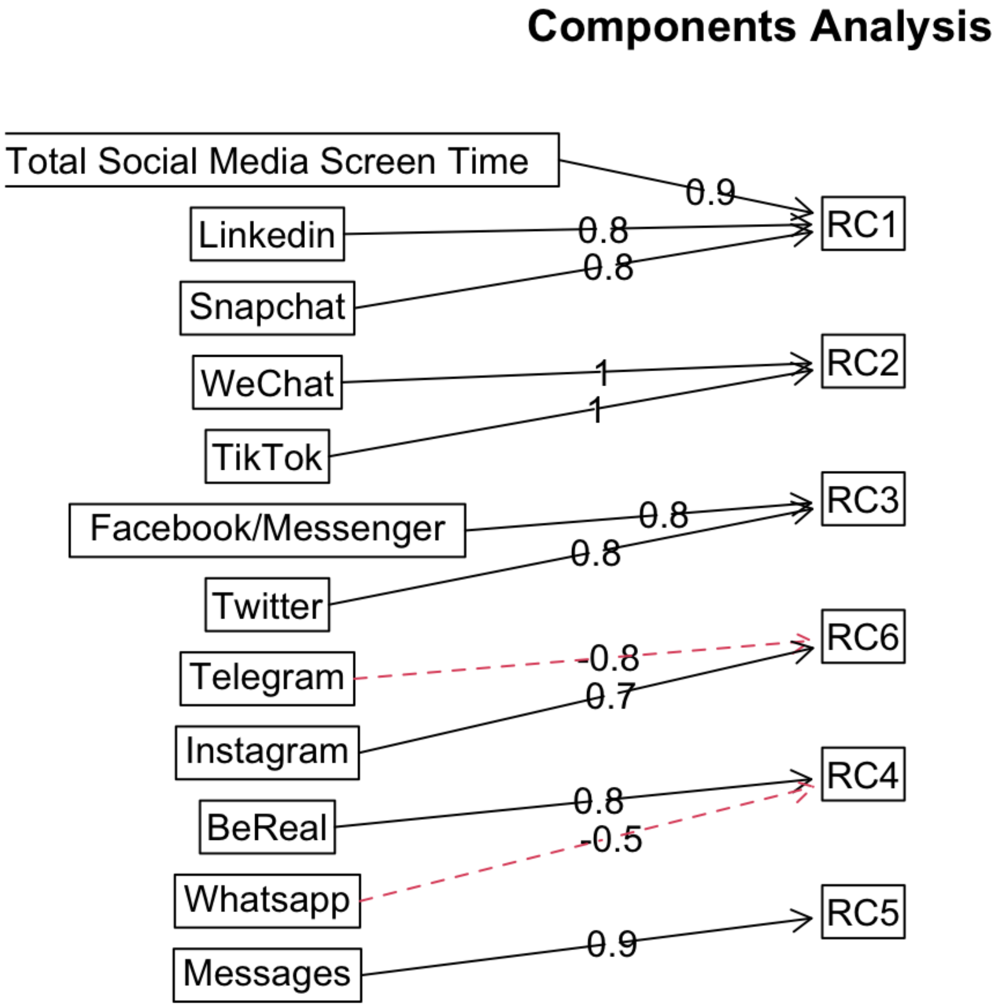
The correlation plot shows us that "Social media screen time" is highly correlated with WhatsApp, Instagram, Snapchat, and Linkedin.

- Let's do PCA on the given variables to check whether there is a any relationship between the number of hours spent on social media apps and addiction:
  > After performing PCA on the given data, we can see that the proportion of variance is very low, and there isn't any benefit in performing the Principal Component Analysis on this data.

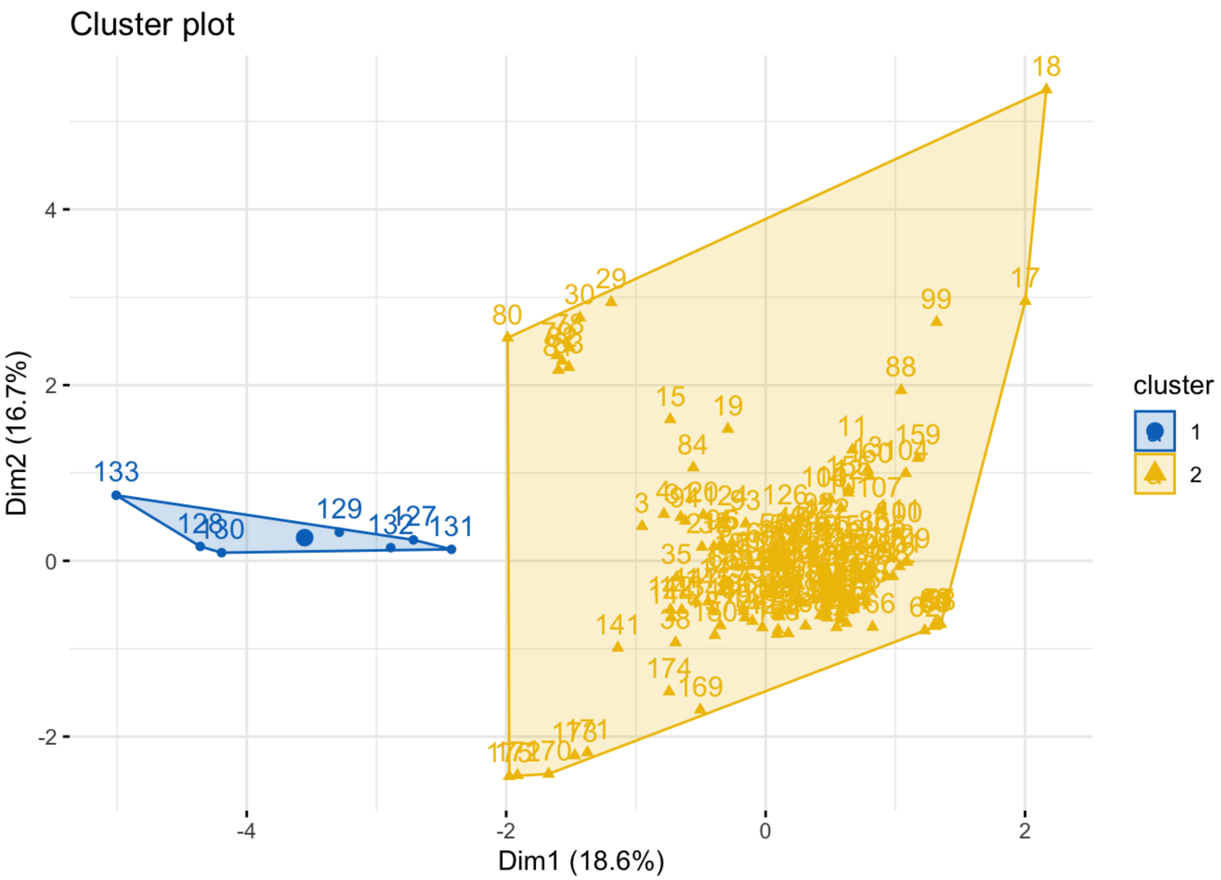


```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7
Standard deviation     1.6960 1.4069 1.2774 1.1199 1.00107 0.96923 0.88574
Proportion of Variance 0.2397 0.1649 0.1360 0.1045 0.08351 0.07828 0.06538
Cumulative Proportion  0.2397 0.4046 0.5406 0.6451 0.72865 0.80693 0.87231
                           PC8     PC9    PC10   PC11      PC12
Standard deviation     0.85898 0.65314 0.50389 0.3376 9.373e-05
Proportion of Variance 0.06149 0.03555 0.02116 0.0095 0.000e+00
Cumulative Proportion  0.93379 0.96934 0.99050 1.0000 1.000e+00
```

- Let’s see if we get any results from Factor analysis and Clustering:
  > Analyzing the factors gives us the following diagram.



We can see that the columns are divided into 6 factors, and in RC5, only the Messages column is considered. So we replaced RC5 with messages. Let’s perform cluster analysis on given factors. We used the K-means algorithm for plotting the Cluster Plot with 2 as the number of clusters as we are trying to determine Addiction.

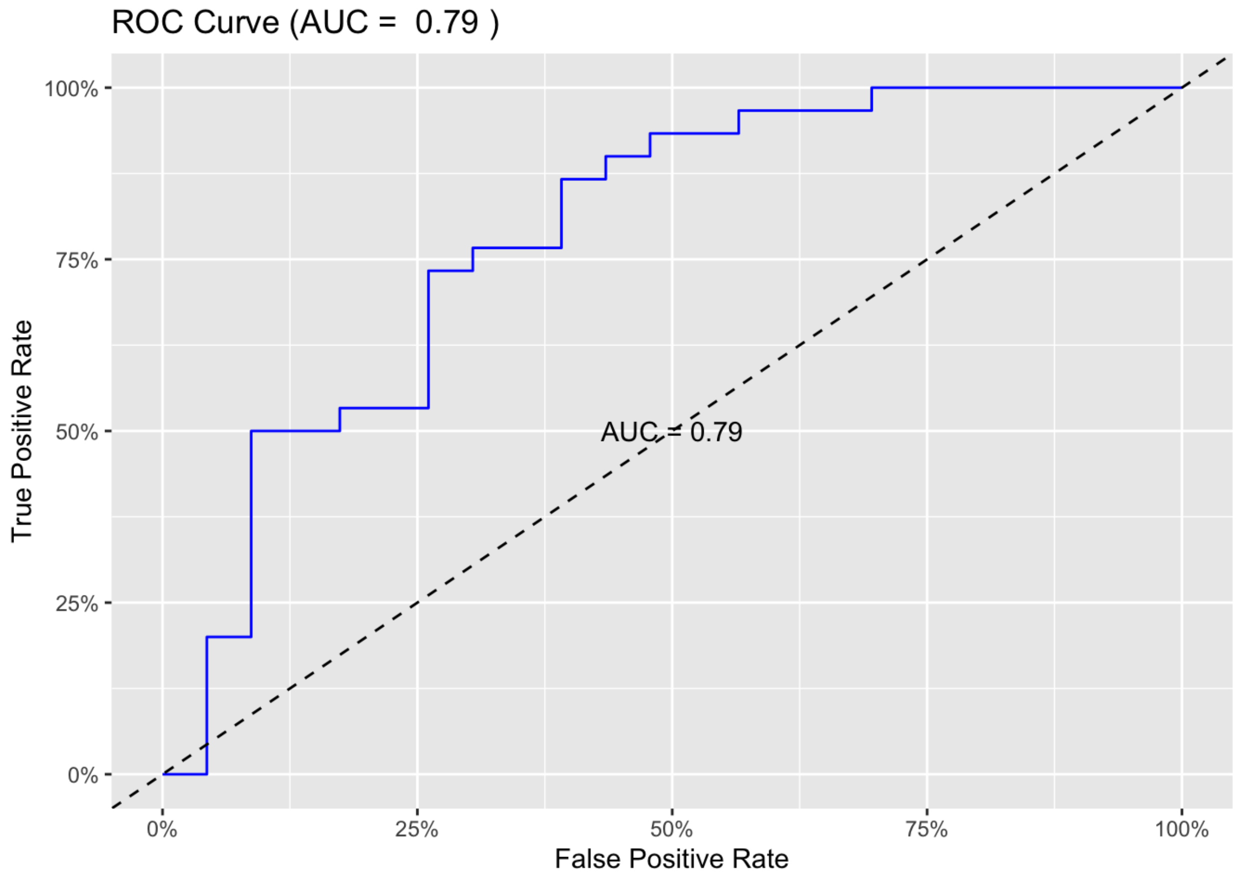


From the above Cluster plot, we are not able to determine the addiction with given factors, so the hypothesis is wrong.

- Let’s see if it is possible to predict social media addiction from the number of hours spent on these apps.

>
We are using logistic regression as we need the output in the form of either Addicted or Not Addicted. We created two models, one with all variables and one with only those variables which are significant. We get better output from the first model where all variables are considered with the following outputs:

Accuracy: 0.717
Precision: 0.767
Recall: 0.742



We can see AUC is 0.79, indicating that the predictive model used to generate the ROC curve can reasonably distinguish between positive and negative cases. Specifically, an AUC of 0.79 means that the model correctly identifies a random positive case more often than a random negative case about 79% of the time. We can say that our model has fair discrimination ability. Hence we can say the hypothesis is true that we can predict social media addiction from the number of hours spent on these apps.

The whole project with the dataset, rmd file, and the HTML file is available on my **Github account**.