Multivariate Analysis – Class Final Project

Kireeti Mantrala – sm2594

About the Data:

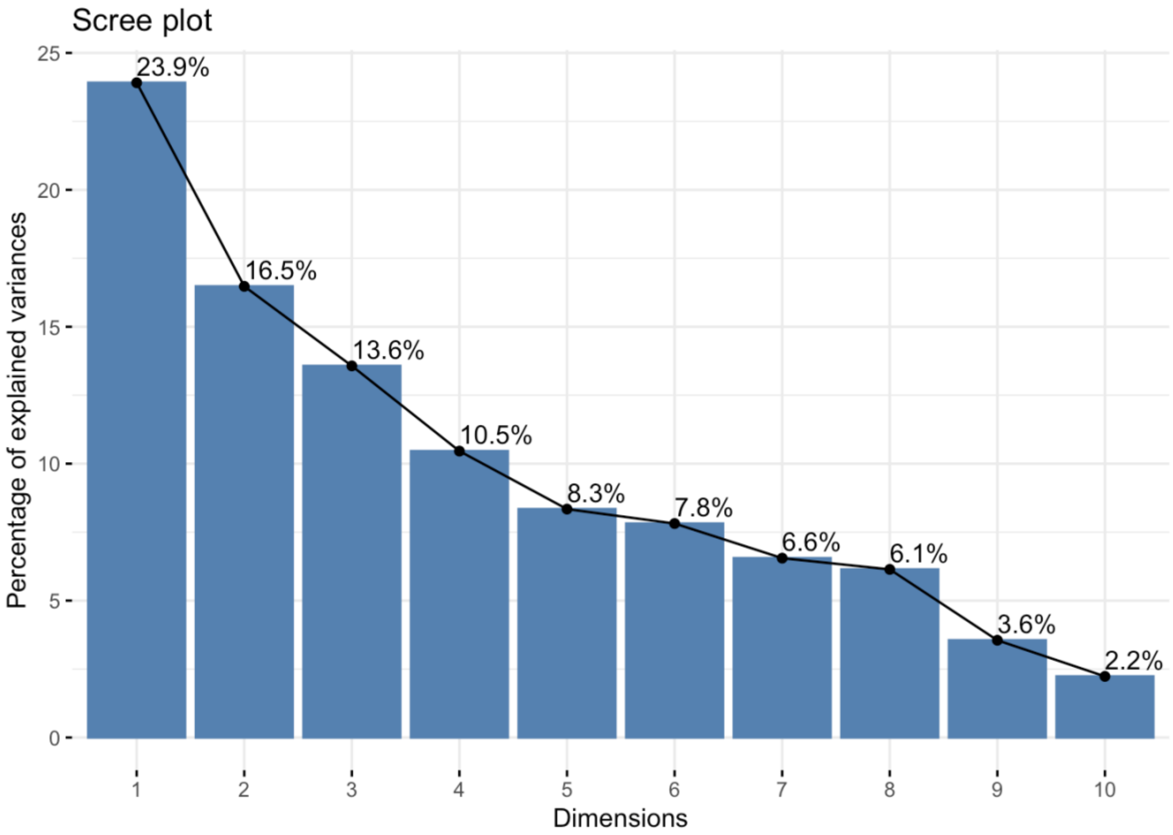
This dataset is self-made by a survey organized within the class of 25 students among themselves. They survey is based on number of hours spent per week on individual social media platforms. The dataset consists of social media platforms, WhatsApp, Instagram, Snapchat, Telegram, Facebook/Messenger, BeReal, TikTok, WeChat, Twitter, LinkedIn, and Messages. We have also considered the number times each individual has opened the applications per week too. And the last Column Talks about Addicted – 1 and Not Addicted – 0. We will be performing required Multivariate Analysis on the data and create a model to predict if a student/person is addicted to social media or not.

Data Dictionary:

Field Name	Data Type	Units	Description
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	No.s	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?

Execution:

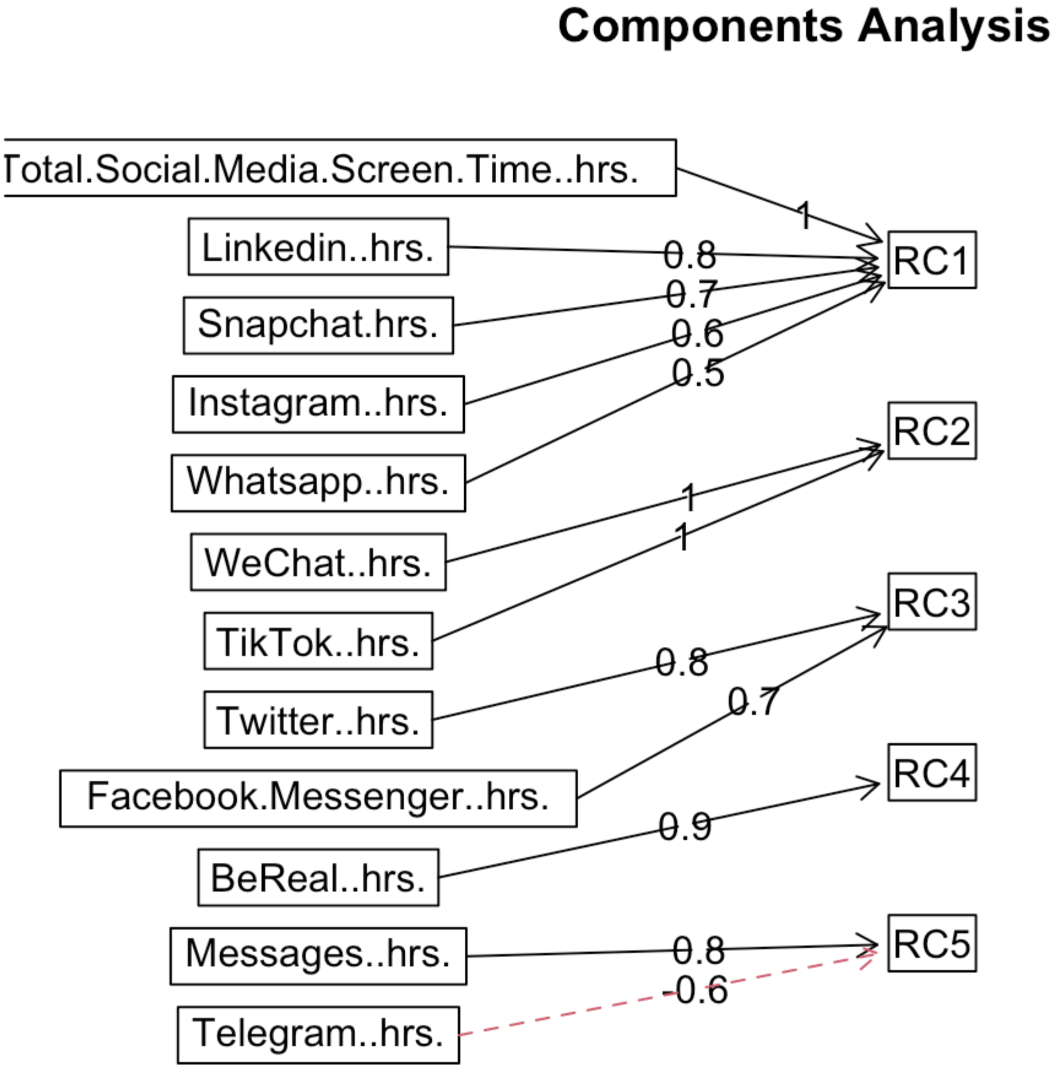
We have first loaded the dataset, Class_survey.csv. Next, we can find the Scree Plot to understand if we can perform PCA on the given dataset or not.



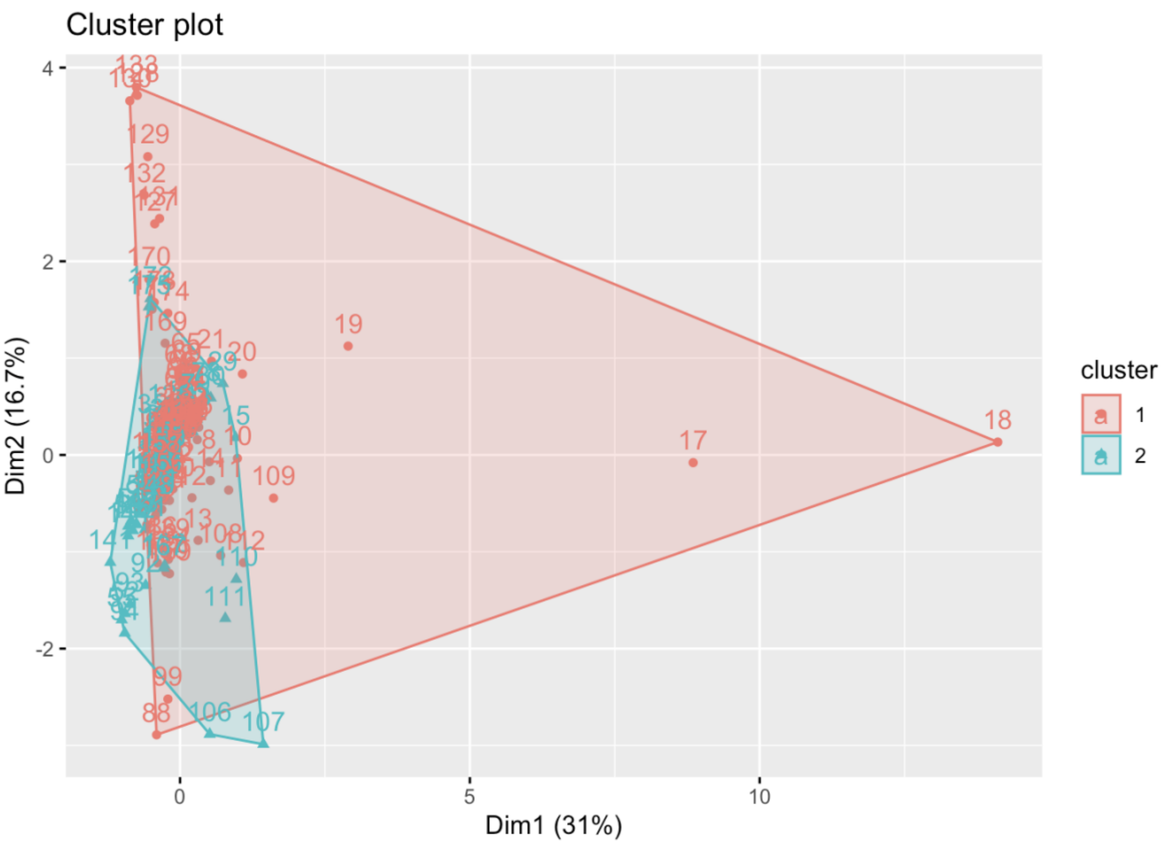
We get the correlation between the factors and through Scree Plot, we concluded that, since the percentage is not above 75, we shall not proceed further with PCA and now perform the Factor Analysis to reduce and explore the underlying structure of these set of variables to provide insights into the relationships between different aspects or the Social Media Platforms of the dataset.

Next, we perform the Factor Analysis and obtain the Components Analysis Diagram; we can identify the underlying factors that are driving the correlation between the set of variables determining the optimal number of factors to extract from the data. The subsets are:

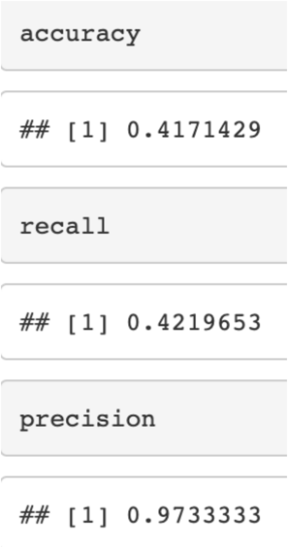
- RC1 has Total Social Media Screen Time, LinkedIn, Snapchat, Instagram, WhatsApp
- RC2 has WeChat, TikTok
- RC3 Twitter, Facebook/Messenger
- RC4 has only BeReal
- RC5 Messages and Telegram with Telegram being inversely related.



Next, we performed the Cluster Analysis. Before which, we see that ‘BeReal’ has been separated as a single component (RC4) and I have added the column to the new data. This will be used for the Clustering. We obtain the Cluster plot.

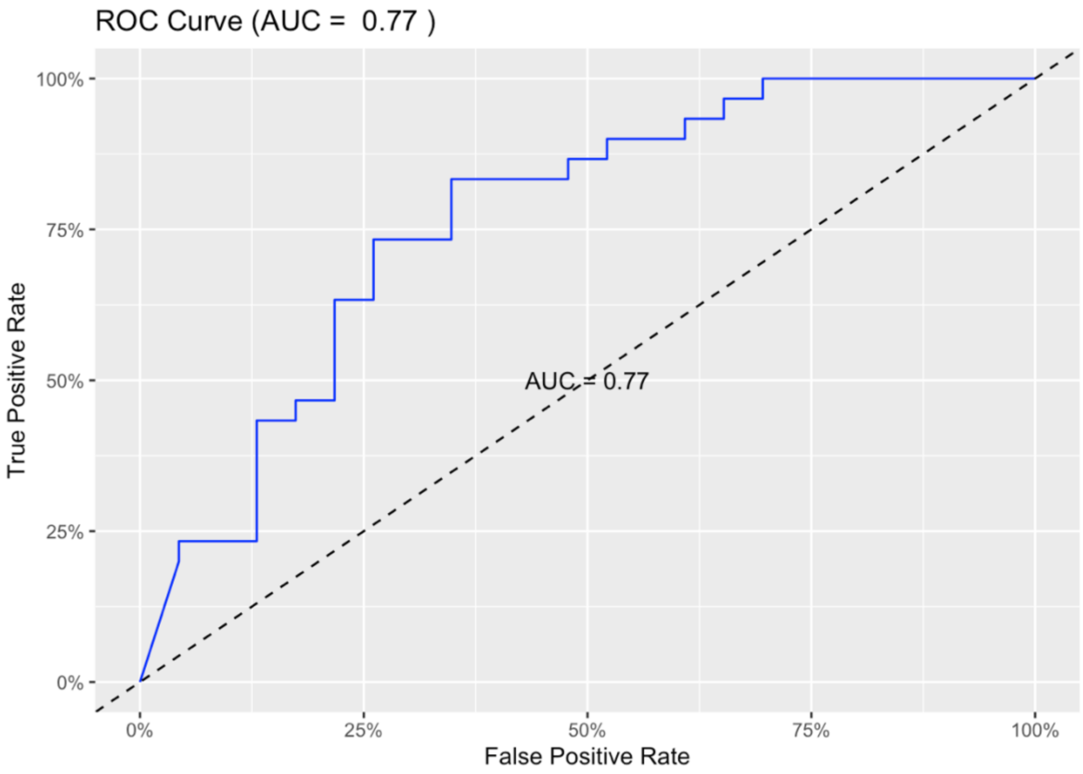


We could not classify insights for the given elements in the dataset because these clusters are overlapping. We drew the Confusion Matrix and calculated the Accuracy, Precision and Recall Metrics. With the confusion matrix, we have performed Accuracy, Precision and Recall and obtained results: **Accuracy: 41%, Precision: 42%, Recall: 97%**



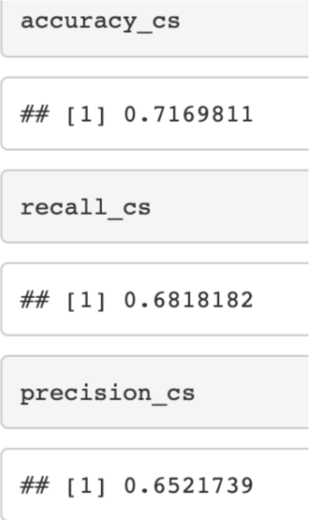
```
accuracy

## [1] 0.4171429

recall

## [1] 0.4219653

precision

## [1] 0.9733333
```

The Cluster Analysis has a low accuracy of 41%, this indicates that the clusters obtained are not accurately separated and will be having an overlap. The precision of the Cluster Analysis is 42%, this indicates that there may be a high number of false positives in the clusters, and it may be incorrectly classifying some instances as belonging to the cluster. The recall of the model is high at 97%, which means that out of all the actual instances belonging to the cluster, the analysis was able to correctly identify 97% of them. Next, we performed the Logistic Regression.



Performed Logistic Regression and obtained errors and z-value table. Now, we drew the Confusion Matrix and calculated the Accuracy, Precision and Recall Metrics. With the confusion matrix on the Logistic Regression, we have performed Accuracy, Precision and Recall and obtained results:

Accuracy: 71%, Precision: 65%, Recall: 68%.



```
accuracy_cs
```

```
## [1] 0.7169811
```

```
recall_cs
```

```
## [1] 0.6818182
```

```
precision_cs
```

```
## [1] 0.6521739
```

Question:

With the data can we predict if a person is addicted to social media or not addicted?

Answer:

Yes. The model has an overall accuracy of 71%, which means that it correctly predicts social media addiction 71% of the time. This indicates that the model may be useful, but there is still room for improvement. The precision of the model is 65%, which means that out of all the instances the model predicted as social media addiction, only 65% were correct. This indicates that there may be some false positives in the model's predictions, and it may be flagging some students as addicted to social media when they are not. The recall of the model is 68%, which means that out of all the actual cases from the survey, the model was able to correctly identify 68% of them. This indicates that there may be some cases of social media addiction that the model might be missing. Based on these results, we can conclude that the model may be useful for predicting social media addiction in a multivariate analysis, but it could benefit from further refinement to improve its precision and recall values. Additionally, it is important to consider the cost of false positives and false negatives for the model.