Social_Media

Dataset -
The dataset comprises social media usage reported by 21 students, with a total of 12 columns and 22 rows. Each row represents cumulative data for an individual student.

Data Dictionary -

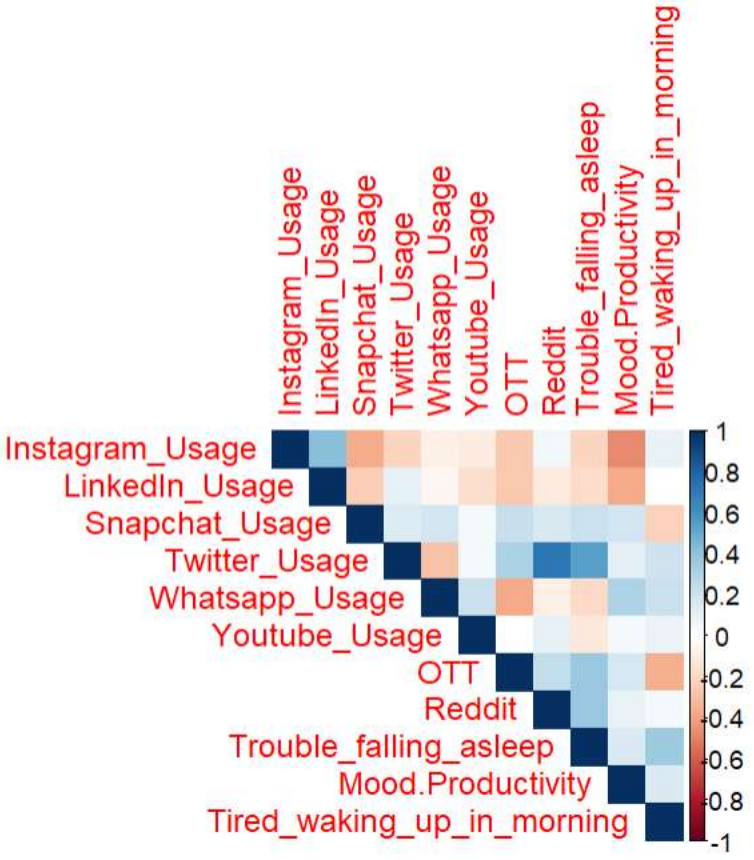
Dataset ID: Unique identifier assigned to each student's data entry.
Instagram Usage: Duration of Instagram app usage.
LinkedIn Usage: Duration of LinkedIn app usage.
Snapchat Usage: Duration of Snapchat app usage.
Twitter Usage: Duration of Twitter app usage.
Whatsapp Usage: Duration of WhatsApp app usage.
Youtube Usage: Duration of YouTube app usage.
OTT Usage: Duration of Over-the-Top media services usage.
Reddit Usage: Duration of Reddit app usage.
Trouble Falling Asleep: Indicates if the student reported difficulty falling asleep (0: No, 1: Yes).
Mood Productivity: Subjective measure of the student's mood and productivity level (0: Bad, 1: Good).
Tiredness upon Waking Up in the Morning: Indicates the level of tiredness reported by the student upon waking up in the morning (0: Low, 1: High).

QUESTIONS and HYPOTHESIS

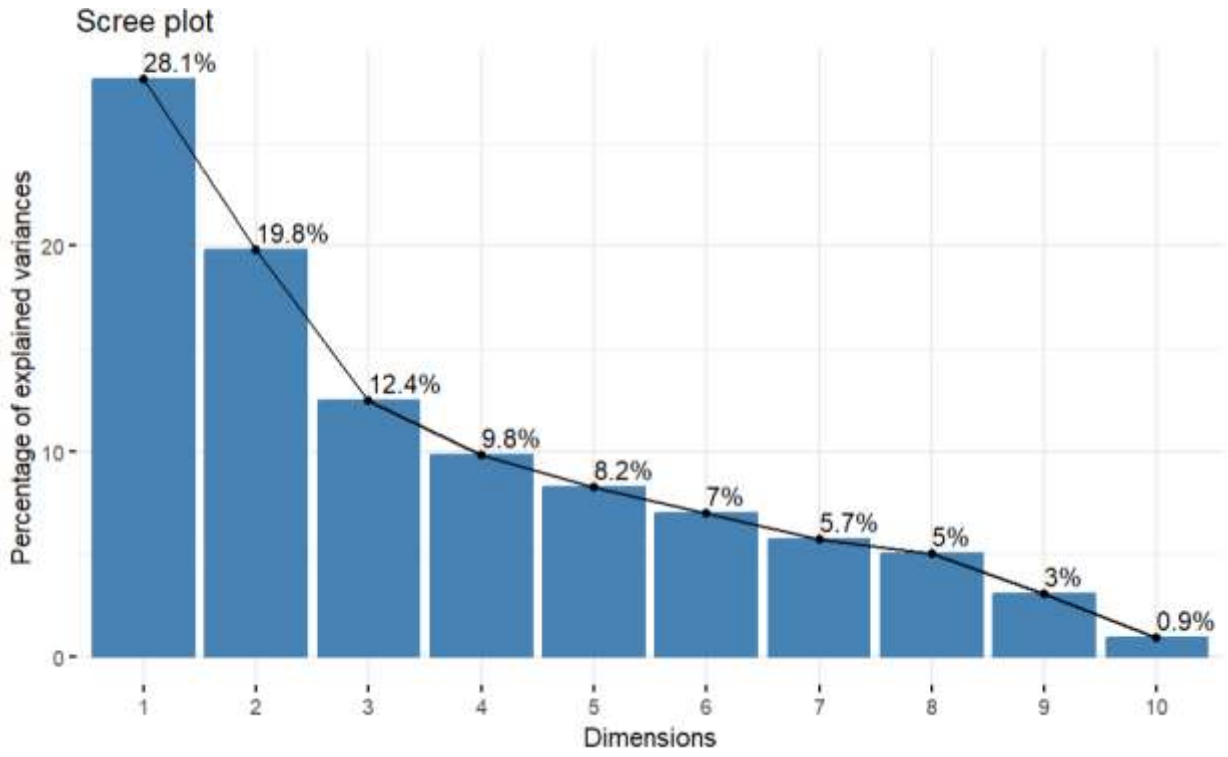
Questions :

1. Using the provided information, can we determine if students are experiencing tiredness upon waking up in the morning due to social media usage.
2. Using the provided data, can we predict whether students are experiencing tiredness upon waking up in the morning due to social media usage.

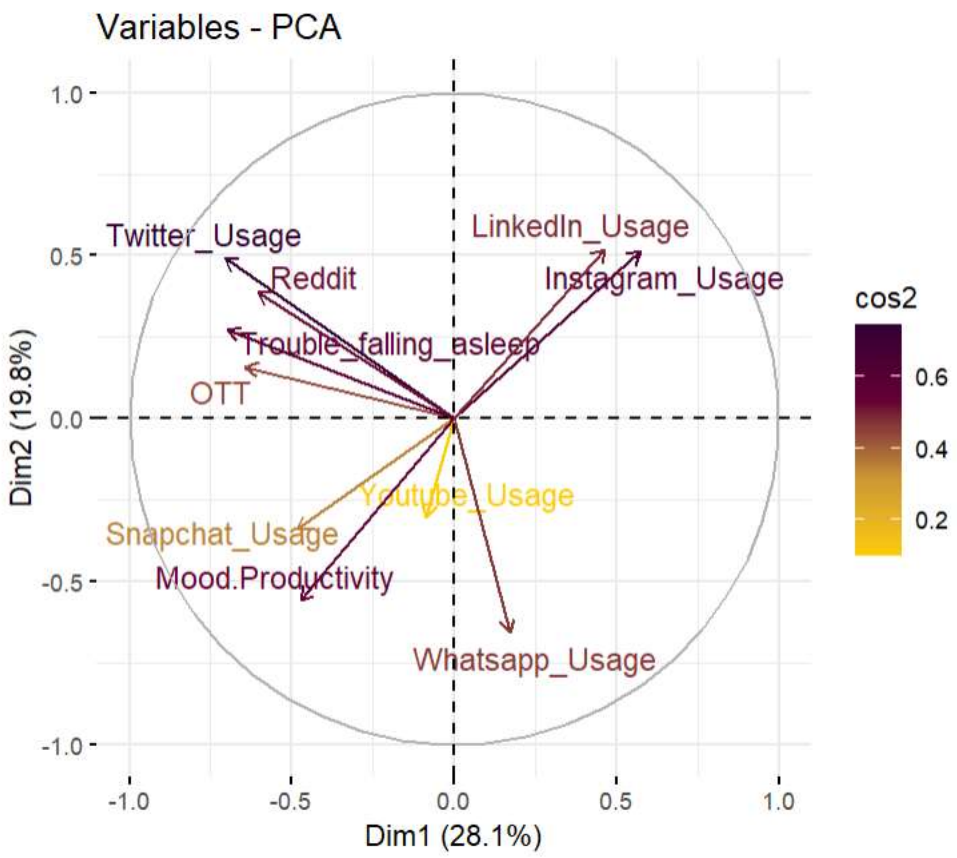
Hypothesis :
We can use the amount of time students spend on different social media apps to predict if they're affected by tiredness when waking up in the morning.



The correlation matrix shows us a correlation between the columns in both cases. So, Principal Component Analysis (PCA) can help in reducing the number of columns used for analysis.



The scree plot indicates that the total variance explained by the first 2 principal components is less than 70%. Therefore, using PCA to reduce the number of columns may not be suitable. Next, we'll explore Exploratory Factor Analysis (EFA) for this dataset.

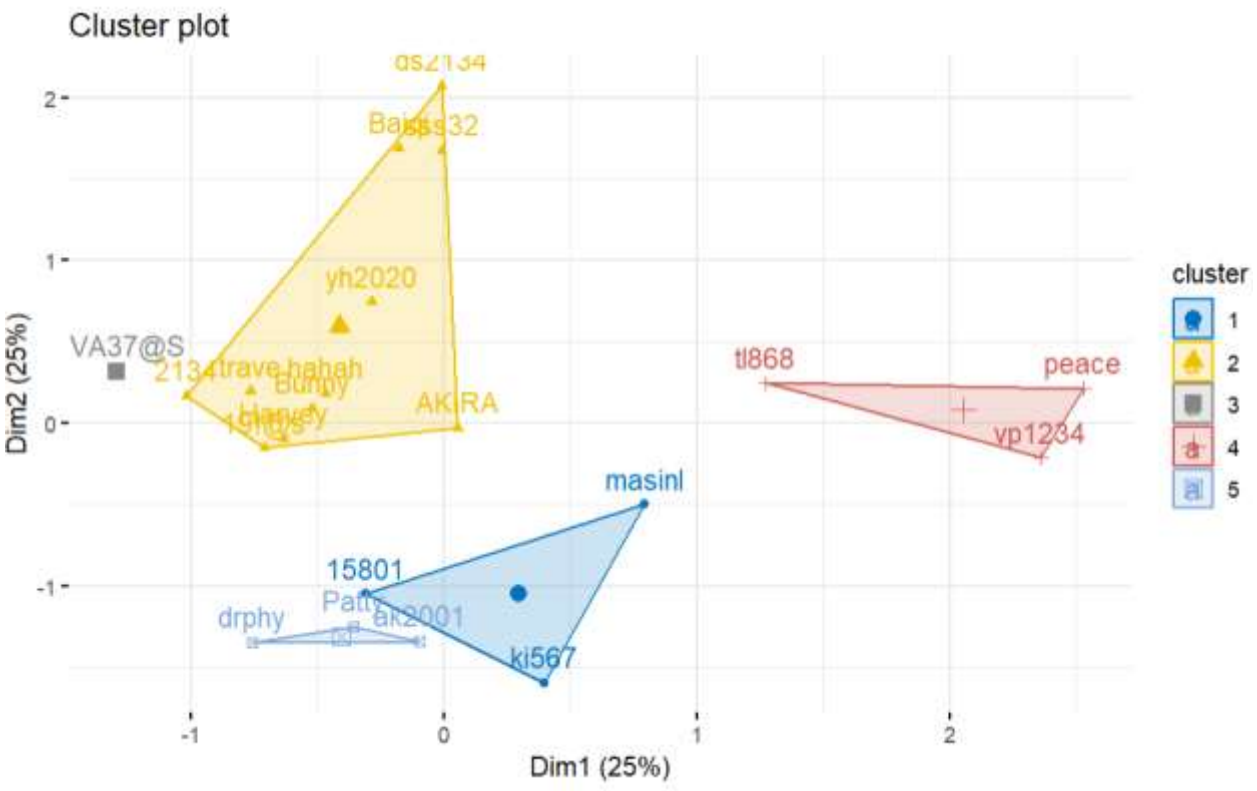


RC1: Content Engagement: Reflects user activity and interaction with various content sharing platforms like Twitter and Reddit.

RC2: Social Media Usage: Represents the extent of engagement with popular social media platforms such as Snapchat, LinkedIn, and Instagram.

RC3: Communication and Entertainment: Indicates usage patterns focused on communication and entertainment through platforms like Whatsapp and OTT services.

RC4: Information Consumption: Signifies engagement with content consumption and learning activities primarily through platforms like YouTube.

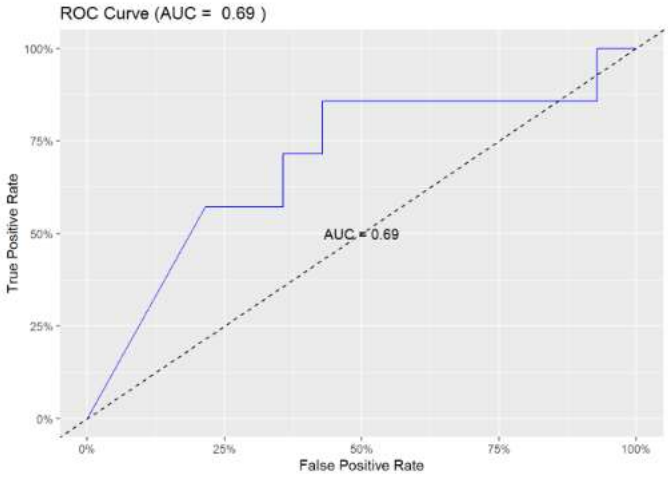


	Actual	
Clustered	No trouble waking up	Trouble waking up
NO Trouble waking up	13	5
Trouble waking up	1	2

The confusion matrix indicates that the clustering predominantly categorizes users as not experiencing any difficulty falling asleep. The accuracy is 71%, which is not that good.

Logistic regression is a statistical approach used to classify data into two categories. It estimates the likelihood that an observation belongs to one of two classes. It establishes the connection between a binary outcome and one or multiple independent variables.

```
##
## Call:
## glm(formula = Tired_waking_up_in_morning ~ ., family = "binomial",
##     data = data)
##
## Coefficients: (10 not defined because of singularities)
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.557e+01  2.160e+05       0        1
## ID19!@s                5.113e+01  3.055e+05       0        1
## ID2134                -1.813e-14  3.055e+05       0        1
## IDak2001              -3.196e-14  3.055e+05       0        1
## IDAKIRA               -2.711e-14  3.055e+05       0        1
## IDBaiqi                5.113e+01  3.055e+05       0        1
## IDBunny                5.555e-14  3.055e+05       0        1
## IDdrphy                1.947e-14  3.055e+05       0        1
## IDds2134              -2.398e-15  3.055e+05       0        1
## IDhahah               -1.143e-14  3.055e+05       0        1
## IDHarvey               5.113e+01  3.055e+05       0        1
## IDki567                2.003e-14  3.055e+05       0        1
## IDmasinl              -6.943e-15  3.055e+05       0        1
## IDMVA37@S              5.113e+01  3.055e+05       0        1
## IDPatty                2.624e-14  3.055e+05       0        1
## IDpeace               -5.489e-14  3.055e+05       0        1
## IDsss32               -3.527e-14  3.055e+05       0        1
## IDtl868                5.113e+01  3.055e+05       0        1
## IDtrave                5.113e+01  3.055e+05       0        1
## IDvp1234               5.113e+01  3.055e+05       0        1
## IDyh2020              -1.159e-13  3.055e+05       0        1
## Instagram_Usage               NA         NA      NA       NA
## LinkedIn_Usage                NA         NA      NA       NA
## Snapchat_Usage                NA         NA      NA       NA
## Twitter_Usage                 NA         NA      NA       NA
## Whatsapp_Usage                NA         NA      NA       NA
## Youtube_Usage                 NA         NA      NA       NA
## OTT                           NA         NA      NA       NA
## Reddit                        NA         NA      NA       NA
## Trouble_falling_asleep        NA         NA      NA       NA
## Mood.Productivity             NA         NA      NA       NA
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.6734e+01  on 20  degrees of freedom
## Residual deviance: 3.3117e-10  on  0  degrees of freedom
## AIC: 42
##
## Number of Fisher Scoring iterations: 24
```



These visualizations provide valuable insights into the logistic regression model's performance by showcasing Accuracy and ROC Curve of the model.

CONCLUSION :

The logistic regression model demonstrated high accuracy and superior performance, as evident from its ROC curve. It particularly excelled in effectively predicting outcomes, specifically in distinguishing between individuals experiencing sleep disturbances and those who are not, using their social media usage patterns as predictors. This suggests that we can utilize the time spent on individual social media apps to predict whether a student is affected by or facing difficulties waking up early.