SOCIAL MEDIA DATA
MULTIVARIATE ANALYSIS

Ashita Shetty

Questions Explored

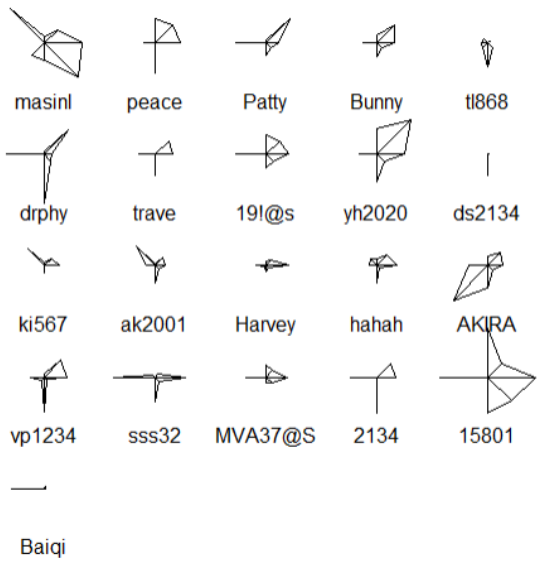
1. **Investigate the correlation between different social media applications with Productivity, Sleep Patterns, and Energy throughout the week**
2. **Recognize underlying patterns between variables to group them together**
3. **Identify clusters of students based on social media usage**
4. **Pin down the best model to predict the outcome of Trouble falling asleep and factors that are statistically significant in the prediction**
5. **Where do I as a student stand compared to my fellow students**

Data Dictionary:

- id: Unique identifier for each student
- Instagram_Hours: Number of hours spent on Instagram
- LinkedIn_Hours: Number of hours spent on LinkedIn
- Snapchat_Hours: Number of hours spent on Snapchat
- Twitter_Hours: Number of hours spent on Twitter
- Whatsapp_Wechat_hours: Number of hours spent on WhatsApp or WeChat
- Reddit_hours: Number of hours spent on Reddit
- Youtube_hours: Number of hours spent on YouTube
- OTT_hours: Number of hours spent on over-the-top (OTT) platforms
- Application type: Type of application used (Social media, OTT, Learning)
- Interview_calls: Number of interview calls received
- Networking: Indicates whether networking was conducted (1 for yes, 0 for no)
- Items_learned: Number of items learned
- Mood_Productivity: Mood and productivity level (e.g., Yes, No)
- Tired_Morning: Indicates if the individual felt tired in the morning (Yes or No)
- Trouble_Sleeping: Indicates if the individual had trouble sleeping (Yes or No)
- Entire_Week_Feeling: Overall feeling throughout the entire week

Exploratory Data Analysis (EDA)

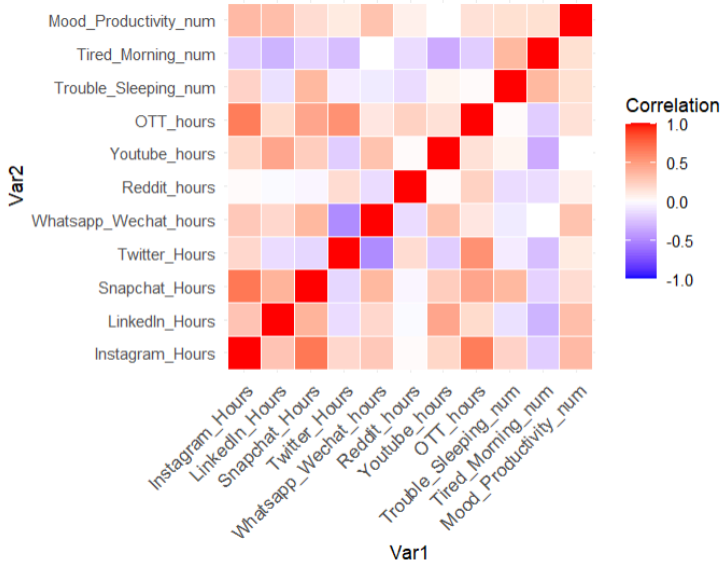
Similar students based on the social media application usage



- Stars diagram is a great way to observe students with similar application usage. While not everybody has the exact same pattern, similar students can be spotted based on the applications used or average time of use.

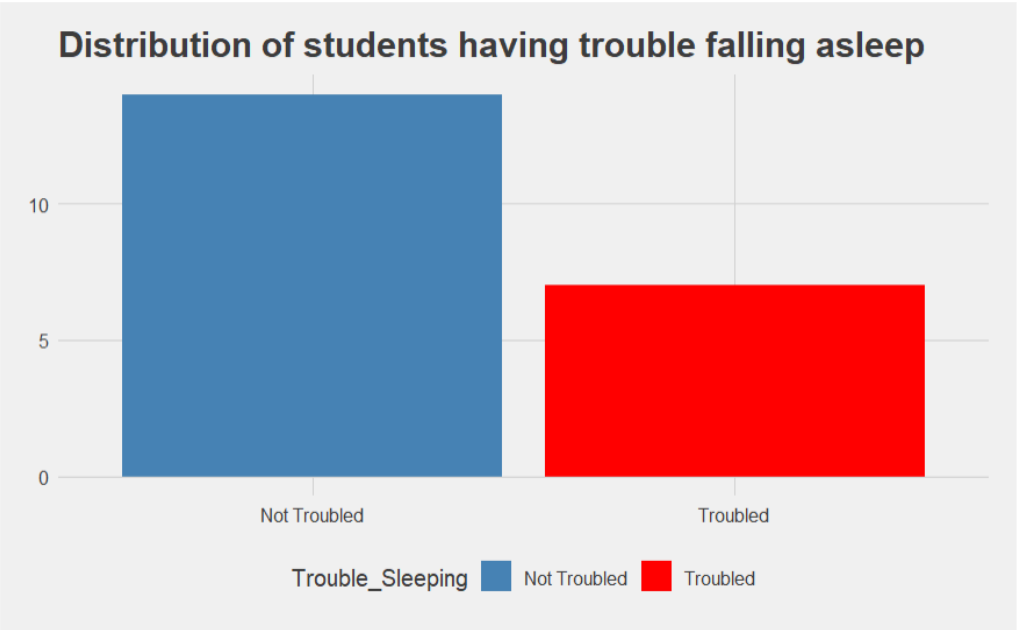
- For example, 19!@s, MVA37@S, yh2020, Bunny, 15801 are students that have used similar apps, however the average consumption differed. While MVA37@S is not an extensive social media user, 15801 seems to have been using social media quite a lot.

Correlation between different applications and the target variables

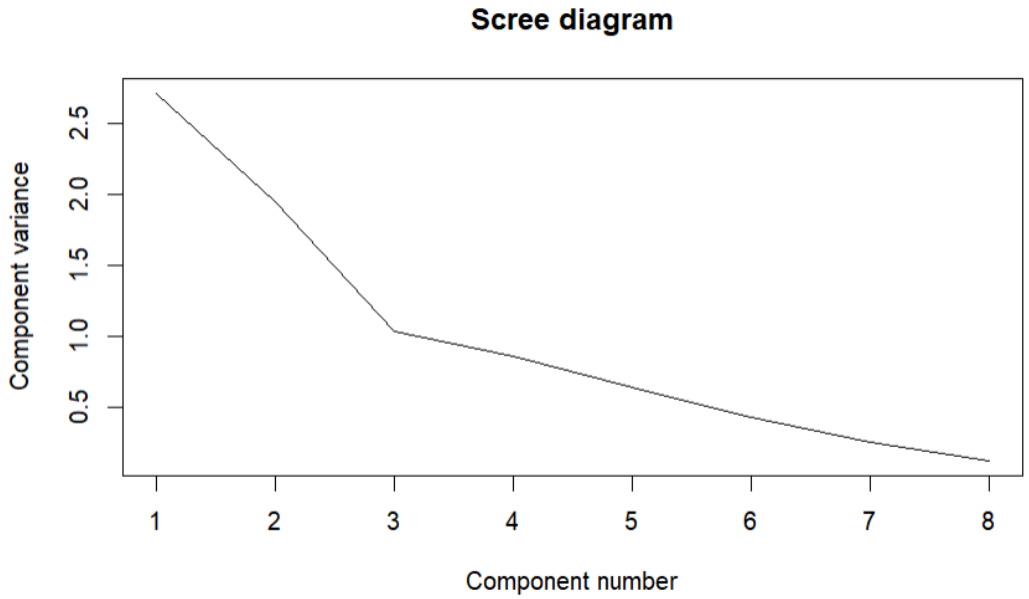


- As it may be observed that Snapchat and Instagram are highly correlated to having trouble falling asleep.
- For Tired_Morning, most applications are negatively correlated.
- Whereas applications such as WhatsApp/WeChat, LinkedIn, Instagram, Snapchat, and OTT correlate with the Mood Productivity of the students.

Distribution of data in the Target Class



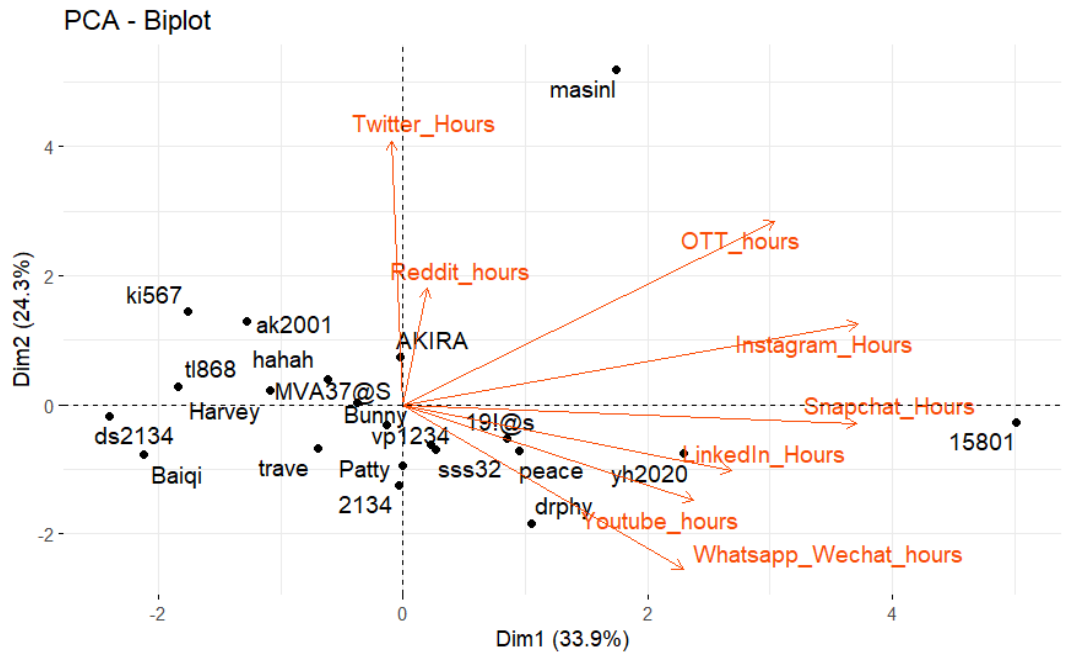
Principal Component Analysis



After having found **3 as the ideal number of components.** We conducted the Principal Component Analysis and compared the cumulative variance in the 3 components

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8
Standard deviation     1.6463 1.3947 1.0176 0.9265 0.80136 0.65798 0.50370 0.34905
Proportion of Variance 0.3388 0.2431 0.1294 0.1073 0.08027 0.05412 0.03171 0.01523
Cumulative Proportion  0.3388 0.5819 0.7114 0.8187 0.89894 0.95306 0.98477 1.00000
```

For 3 PCs, we can observe the cumulative variance to be **71.14%**

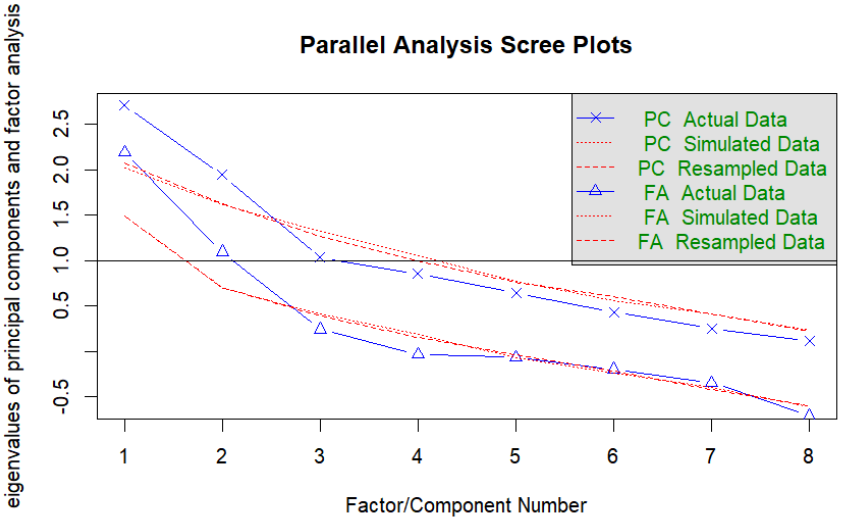


- The Bi-plot helps to understand to clearly see the penetration of the applications among students.
- The ranking of applications based on the usage (high to low): **WhatsApp/Wechat > Twitter > OTT platforms > Instagram > Snapchat > LinkedIn > Youtube > Reddit**.
- Student "15801" in general spends more time on social media, especially Snapchat. Whereas Student "masinl" spends most of his/her time on Twitter and OTT.
- The club of students on the left have been grouped due to their similar activities of social media usage. These students are not as extensive users as the ones to the right highlight modularity.

Factor Analysis

FA helps to group variables based on underlying patterns and to understand the relationship between them.

Using Parallel Analysis Scree Plot to visualize the ideal number of factors

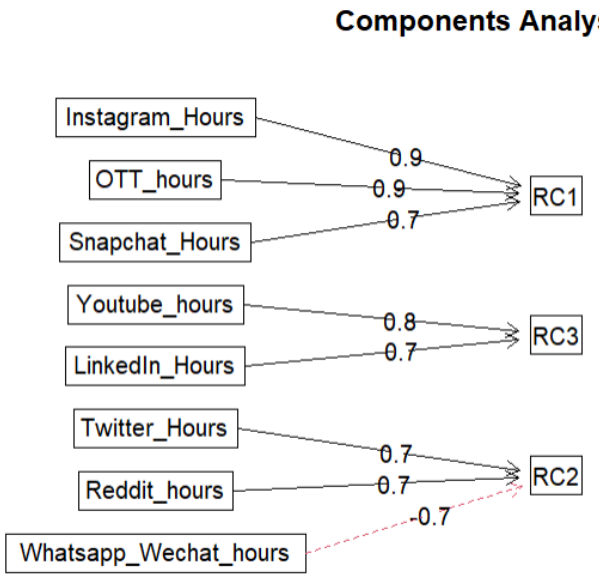


- Based on the 'FA Actual Data', we can infer that **3 is the ideal number of Components.**

Carrying out Factor Analysis using 3 Factors

```
Loadings:
                      RC1    RC3    RC2
Instagram_Hours        0.896  0.145
LinkedIn_Hours         0.233  0.726
Snapchat_Hours         0.731  0.306 -0.329
Twitter_Hours          0.386 -0.388  0.725
Whatsapp_Wechat_hours  0.239  0.364 -0.658
Reddit_hours                  0.329  0.682
Youtube_hours                 0.808
OTT_hours              0.853         0.353

                 RC1   RC3   RC2
SS loadings    2.334 1.691 1.665
Proportion Var 0.292 0.211 0.208
Cumulative Var 0.292 0.503 0.711
```

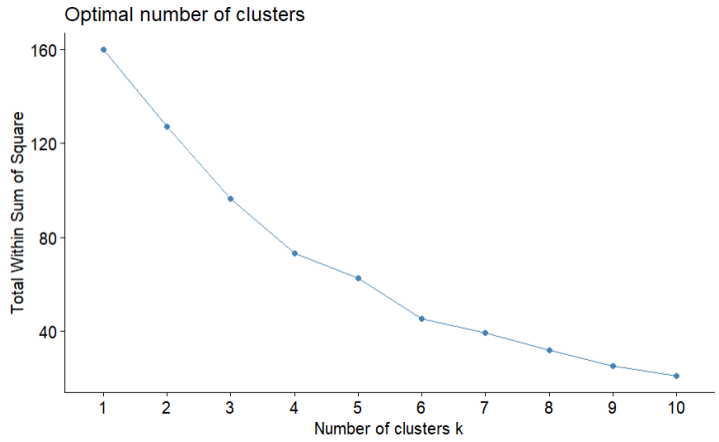


- Based on the **Loadings,** and **Component Analysis,** we can observe that there are 3 groups to be considered: RC1, RC2, and RC3
- We can divide these Factors into:
  - **RC1: Visual Social Media Engagement**
  - **RC2: Professional Networking and Content Consumption**
  - **RC3: Social Interaction and Messaging Platforms**

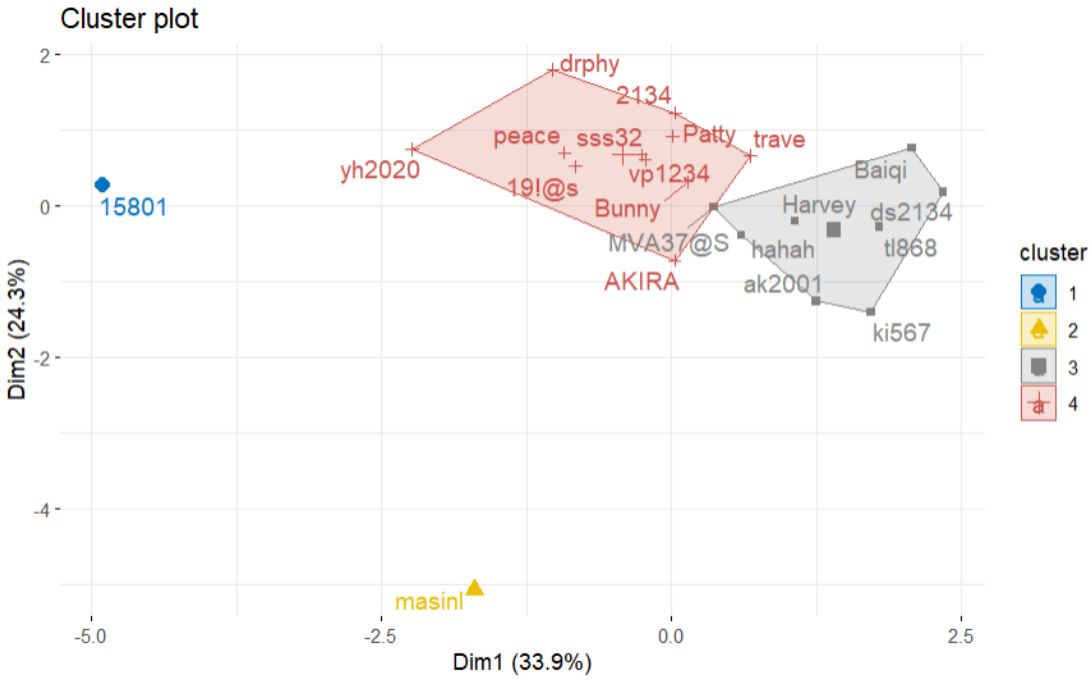
Cluster Analysis

The aim was to group students based on their social media consumption together

To find the optimal number of clusters -



As **4 is the ideal number of clusters.** We create 4 clusters to group students.



- Majority of the class can be majorly divided into two groups of social media users: **Heavy Users (Cluster 4)** and **Moderate to Low (Cluster 3)** (except the outlier users, "15801", and "masinl")

Models to Predict Trouble Falling Asleep

Multiple Regression

```
Call:
lm(formula = Trouble_Sleeping_num ~ Instagram_Hours + LinkedIn_Hours +
    Snapchat_Hours + Twitter_Hours + Whatsapp_Wechat_hours +
```

```
    Reddit_hours + Youtube_hours + OTT_hours, data = social_media)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60834 -0.24317 -0.05131  0.20427  0.66610

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.49305    0.38624   1.277    0.226
Instagram_Hours        0.01744    0.04981   0.350    0.732
LinkedIn_Hours        -0.07144    0.05424  -1.317    0.212
Snapchat_Hours         0.15215    0.10242   1.486    0.163
Twitter_Hours         -0.01944    0.17061  -0.114    0.911
Whatsapp_Wechat_hours -0.03855    0.03918  -0.984    0.345
Reddit_hours          -0.03560    0.07633  -0.466    0.649
Youtube_hours          0.05029    0.07772   0.647    0.530
OTT_hours             -0.02886    0.06448  -0.448    0.662

Residual standard error: 0.508 on 12 degrees of freedom
Multiple R-squared:  0.3363,     Adjusted R-squared:  -0.1061
F-statistic: 0.7602 on 8 and 12 DF,  p-value: 0.6429
```

- None of the predictor variables (Instagram_Hours, LinkedIn_Hours, etc.) show a statistically significant relationship with the Trouble_Sleeping_num variable, as indicated by their high p-values.
- The model's adjusted R-squared value of -0.1061 suggests that the model does not explain much of the variance in the Trouble_Sleeping_num variable.
- None of the social media usage variables significantly predict trouble sleeping, and the overall model is not statistically significant in explaining the variance in trouble sleeping behavior.

Logistic Regression

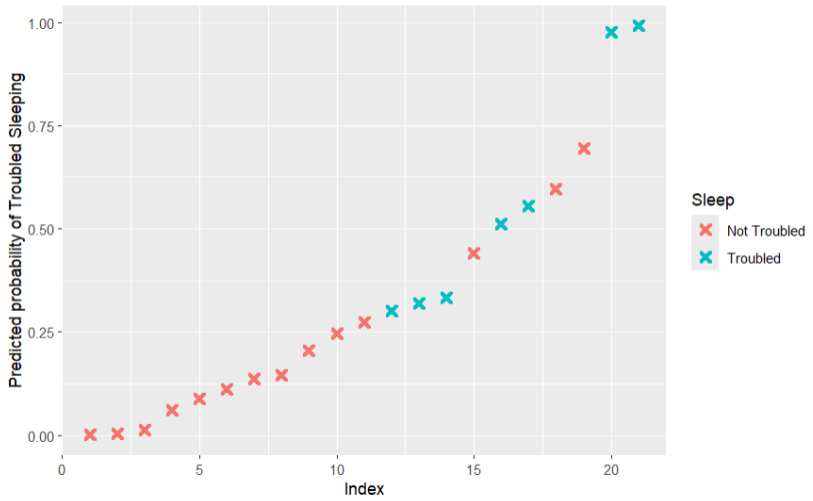
```
Call:
glm(formula = Trouble_Sleeping ~ Instagram_Hours + LinkedIn_Hours +
    Snapchat_Hours + Twitter_Hours + Whatsapp_Wechat_hours +
    Reddit_hours + Youtube_hours + OTT_hours, family = "binomial",
    data = social_media)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.49668    2.15669  -0.230    0.818
Instagram_Hours        0.30097    0.34407   0.875    0.382
LinkedIn_Hours        -0.69902    0.52211  -1.339    0.181
Snapchat_Hours         1.26462    0.98630   1.282    0.200
Twitter_Hours         -0.03163    0.77375  -0.041    0.967
Whatsapp_Wechat_hours -0.33578    0.27323  -1.229    0.219
Reddit_hours          -0.11034    0.44628  -0.247    0.805
Youtube_hours          0.53381    0.66481   0.803    0.422
OTT_hours             -0.33704    0.41879  -0.805    0.421

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.734  on 20  degrees of freedom
Residual deviance: 17.665  on 12  degrees of freedom
AIC: 35.665
```

```
Number of Fisher Scoring iterations: 6
```




```
Confusion Matrix and Statistics

              Reference
Prediction     Not Troubled Troubled
  Not Troubled            2        4
  Troubled               12        3

               Accuracy : 0.2381
                 95% CI : (0.0822, 0.4717)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.99999

                  Kappa : -0.3333

 Mcnemar's Test P-Value : 0.08012

            Sensitivity : 0.14286
            Specificity : 0.42857
         Pos Pred Value : 0.33333
         Neg Pred Value : 0.20000
             Prevalence : 0.66667
         Detection Rate : 0.09524
   Detection Prevalence : 0.28571
      Balanced Accuracy : 0.28571

       'Positive' Class : Not Troubled
```

- The model performs poorly with an accuracy of 23.81%.

- The 95% confidence interval is between 0.0822 and 0.4717 for our logistic regression model.
- Sensitivity (true positive rates), and Specificity (true negative rates) are 0.14 and 0.428 respectively.
- Balanced Accuracy provides a balanced measure of model performance by taking into account the imbalance in class distribution, which is calculated by taking the average of Sensitivity and Specificity which is 0.286.

Linear Discriminant Analysis (LDA)

```
Call:
lda(train_raw.df$Trouble_Sleeping ~ ., data = train_raw.df)

Prior probabilities of groups:
Not Troubled     Troubled
   0.6666667    0.3333333

Group means:
                        Instagram_Hours  LinkedIn_Hours  Snapchat_Hours  Twitter_Hours
Whatsapp_Wechat_hours
Not  Troubled                 5.600              3.748           0.847           0.527
7.147
Troubled                       6.864             3.172           2.884           0.618
6.706
             Reddit_hours Youtube_hours OTT_hours
Not Troubled         0.25         3.098     2.664
Troubled             0.20         3.182     3.000

Coefficients of linear discriminants:
                             LD1
Instagram_Hours       -0.2708934
LinkedIn_Hours        -0.3893767
Snapchat_Hours         1.9876064
Twitter_Hours          1.9938117
Whatsapp_Wechat_hours  0.2163040
Reddit_hours           2.3767141
Youtube_hours          0.3759605
OTT_hours             -0.9303420
```

- The prior probability of being Troubled is 0.33, indicating that approximately one-third of the observations in the training dataset are labeled as Troubled.
- The means of the predictor variables for each group (Not Troubled and Troubled) show differences between the two groups. For example, Troubled individuals tend to have higher mean values for "Snapchat_Hours", "Instagram_Hours", "Twitter_Hours", "YouTube_Hours", and "OTT_Hours" compared to Not Troubled individuals.
- The coefficients of the linear discriminants (LD1) indicate the contribution of each predictor variable to the separation between the two groups. Positive coefficients indicate that higher values of the predictor variable are associated with the Troubled group, while negative coefficients indicate the opposite.
- **Snapchat_Hours, Twitter_Hours, Reddit_hours** have positive coefficients, indicating that spending more time on these platforms is associated with a **higher likelihood** of experiencing trouble falling asleep.
- **Instagram_Hours, LinkedIn_Hours, and OTT_hours** have negative coefficients, suggesting that spending more time on these platforms is associated with a **lower likelihood** of experiencing trouble falling asleep.

- **Whatsapp_Wechat_hours and Youtube_hours** have coefficients close to zero, indicating that they have relatively **weaker associations** with trouble falling asleep compared to the other variables.

```
Accuracy: 0.3333333
```

Evaluation

Models	Evaluation Metrics
Multiple Regression	**Multiple R-squared: 0.3363**
Logistic Regression	**23.81%**
Linear Discriminant Analysis	**33.33%**

Understanding where I stand -

```
     Instagram_Hours        LinkedIn_Hours        Snapchat_Hours         Twitter_Hours
           0.19179943            -0.68112674            0.31592211            -0.44978498
Whatsapp_Wechat_hours          Reddit_hours         Youtube_hours             OTT_hours
           0.09836697            -0.33347426           -1.27756549             0.02379202
```

-
- The Z-scores can help me identify how my **Youtube**, **LinkedIn**, and **Twitter** usage are way below my fellow classmates.
- Whereas **Instagram**, **Snapchat** are slightly higher than the average among my classmates.