# MultiVariate Analysis Group Project

Name: Sailesh Potturi

Email : sp2245@scarletmail.rutgers.edu

Github link: https://github.com/SaileshRBS/Multi-Variate-Analysis.git

## Dataset: Class_Survey

The provided data appears to be a tabular dataset with various columns representing different attributes for everyone. The dataset consists of **176** Rows and **15** Columns. The 15$^{th}$ column (Social Media Addiction) is the one where we need to determine whether the person has heart disease or not using the 14 variables.

The dataset appears to be a record of social media usage for number of individuals, over a period of several weeks. The dataset includes the following columns:

1)Student: Represents the name of the student whose social media usage is being recorded.

2)Week: Represents the time period for which the social media usage is recorded, usually spanning a week.

3)Whatsapp (hrs): Represents the number of hours spent on WhatsApp during the given week.

4)Instagram (hrs): Represents the number of hours spent on Instagram during the given week.

5)Snapchat (hrs): Represents the number of hours spent on Snapchat during the given week.

6)Telegram (hrs): Represents the number of hours spent on Telegram during the given week.

7)Facebook/Messenger (hrs): Represents the number of hours spent on Facebook/Messenger during the given week.

8)BeReal (hrs): Represents the number of hours spent on BeReal during the given week.

9)TikTok (hrs): Represents the number of hours spent on TikTok during the given week.

10)WeChat (hrs): Represents the number of hours spent on WeChat during the given week.

11)Twitter (hrs): Represents the number of hours spent on Twitter during the given week.

12)LinkedIn (hrs): Represents the number of hours spent on LinkedIn during the given week.

13)Messages (hrs): Represents the number of hours spent on messaging apps other than the ones mentioned above during the given week.

14)Total Social Media Screen Time (hrs): Represents the total number of hours spent on all social media platforms during the given week.

15)SocialMediaAddiction: Represents a classification of whether the student is considered "addicted" to social media based on their total social media screen time.

The dataset provides information on the social media usage patterns of the two individuals over time, including the total hours spent on various social media platforms and whether they are classified as "addicted" or  based on their social media screen time.

***Data Cleaning:**

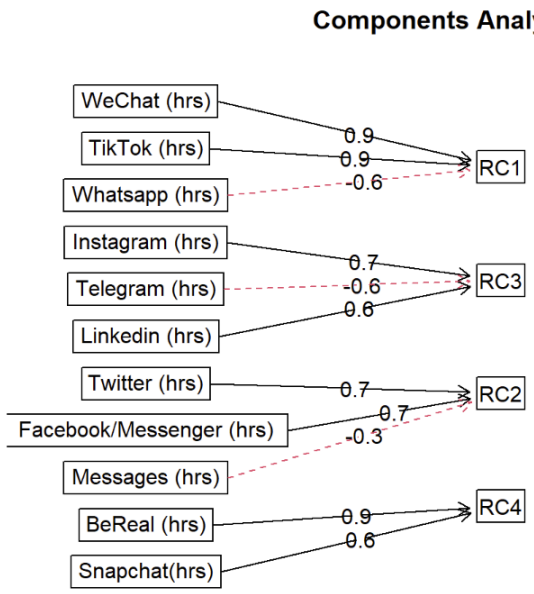
- No of times opened column is being removed as that column is not required for predicted of Addicted or Not Addicted.
- In Raw Data all the float values have being changed to numbers to avoid the as x should be numeric error.
- There is no null values in the raw data so na.omit function() is not required.

1st Question:
Identify the important factors underlying the observed variables and examine the relationships between the Social Media Addiction with respect to these factors.

Answer:

- Firstly, we will use Factor Analysis to reduce the number of columns by which we can see that the columns will get in to multiple factors. We can also observe the visualization of the relationship.

```
fa.diagram(fit.pc) # Visualize the relationship
```



All the Columns are categorized in to RC1, RC2 , RC3 and  RC4.
RC1 contains  WeChat ,TikTok , Whatsapp .
RC2 contains Instgram, Telegram and Linkedlin.
RC3 contains Twitter , Facebook/Messenger, Messages.
RC4 contains BeReal, Snapchat.

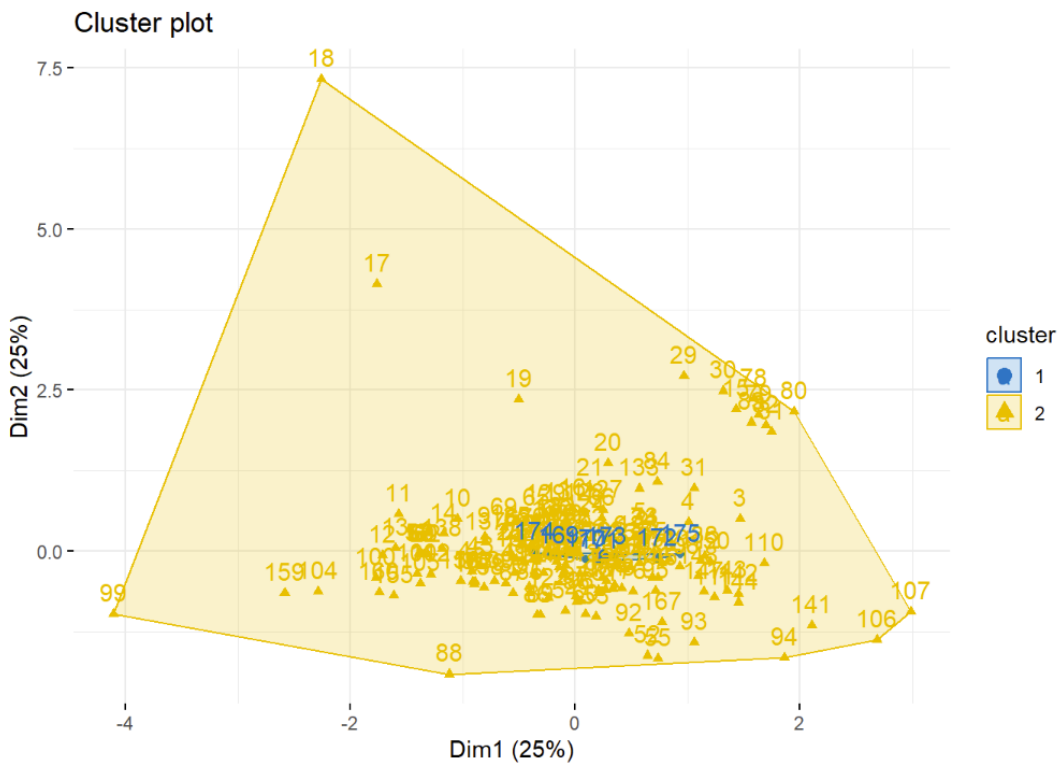
Now, we see the fit.pc$score values by the columns RC1, RC2, RC3 and RC4. Since the dataset has many columns in it. The below code shows random columns of the RC1, RC2, RC3 and RC4

```
# Rotated factor scores, Notice the columns ordering: RC1, RC3, RC2,RC4
fit.pc$scores
```

```
##                RC1          RC3         RC2           RC4
##   [1,] -0.240343094  0.106089671 -0.36857902 -5.665768e-02
##   [2,] -0.439794597  0.660342489 -0.34343847 -2.341429e-01
##   [3,] -0.518544040  1.598851361 -0.48226547 -2.348562e-01
##   [4,] -0.526441804  1.148041876 -0.58623950 -1.891017e-01
##   [5,] -0.274463815  0.673648084 -0.41433287 -1.175293e-01
##   [6,] -0.366212150  0.453404052 -0.40109816 -2.777271e-01
##   [7,] -0.223185408  0.622596643 -0.31775962 -1.398095e-01
##   [8,] -0.148084283 -0.054301064 -0.10052544  4.076200e-01
##   [9,] -0.086892972 -0.689453109 -0.31743105  3.485784e-01
##  [10,] -0.221740347 -0.863757257 -0.26890814  7.533271e-01
##  [11,] -0.391851789 -1.331233119 -0.51593389  8.777212e-01
##  [12,] -0.324374358 -1.653698957 -0.34336691  3.272058e-01
##  [13,] -0.481341343 -1.512788631 -0.26841703  4.572363e-01
##  [14,] -0.278050301 -1.034943134 -0.31657176  5.322682e-01
##  [15,] -0.543458044  2.022395464 -0.67297485  1.536015e+00
##  [16,] -0.057359002  0.178221782 -0.28398336  5.775437e-01
##  [17,]  0.045332070 -0.724060981  1.02144378  5.713688e+00
##  [18,] -0.247908912 -0.342655536  0.61768281  9.146890e+00
##  [19,] -0.060046222  0.180706417 -0.63675560  2.394270e+00
##  [20,] -0.008421984  0.683140358 -0.50282766  1.110946e+00
##  [21,]  0.083213490  0.431385986 -0.48777105  7.479913e-01
##  [22,] -0.107735117 -0.448657570 -0.01019298 -2.758570e-01
##  [23,]  0.007848939 -0.064431723  0.04392913 -2.281278e-02
##  [24,]  0.089948620 -0.188991032  0.03932552  8.200637e-02
##  [25,]  0.029569360 -0.554262830 -0.11555054  1.023651e-01
```

```
## [150,] -0.132066030  1.094843752  0.30464441 -4.282167e-01
## [151,] -0.186494859  0.667924316  0.06565398 -3.747871e-01
## [152,] -0.190036434  0.686994042  0.13812364 -3.918119e-01
## [153,] -0.052791024  0.391598930  0.20481121 -3.258631e-01
## [154,] -0.098491689  0.333852359  0.18092998 -3.709164e-01
## [155,] -0.701321602 -1.726718729 -0.25711310 -3.432171e-01
## [156,] -0.300060277 -1.089444850 -0.19947690 -2.674079e-01
## [157,] -0.235451749 -0.993774568 -0.16657286 -3.492799e-01
## [158,] -0.249397634 -0.920637972 -0.13060163 -4.139401e-01
## [159,] -0.615954858 -2.645780200 -0.37128035 -3.142190e-02
## [160,] -0.597939215 -1.848509911 -0.08123107 -1.482324e-01
## [161,] -0.503101435  0.016521449  0.11273823 -7.237844e-01
## [162,]  0.045430940  0.020514165 -0.02810872  2.215043e-01
## [163,]  0.001054569 -0.024632408 -0.10406219  2.687682e-01
## [164,]  0.191982134 -0.156240035  0.04477130  2.434200e-01
## [165,]  0.043681656 -0.085027056  0.04548100  3.176094e-02
## [166,]  0.152961693 -0.546456612  0.03716740  6.707498e-02
## [167,]  0.302289500  0.299671151  1.80412012 -4.658256e-01
## [168,]  0.033656663  0.237868029  0.15014858 -6.041263e-03
## [169,]  2.790222908 -0.135739232 -0.24607550 -1.674672e-01
## [170,]  5.777881200  0.105614536 -0.53781915 -4.750526e-01
## [171,]  4.477792651  0.202253125 -0.34330270 -4.286439e-01
## [172,]  5.228734482  0.692485323 -0.14881946 -4.588452e-01
## [173,]  4.556366347  0.271664544 -0.30219033 -3.658883e-01
## [174,]  2.980922237 -0.327189515 -0.27002591 -4.914429e-02
## [175,]  5.492415046  0.915492308 -0.23659703 -5.081354e-01
```

- The 2nd part is that we implement  Cluster Analysis using the fit.pc$score values of the RC's columns.
- Here , we scale the data derived from EFA with respect to the RC'S columns and clustering is performed.




- From the above plot we can see that there are many points formed in to 2 clusters 1 and cluster 2.
- But we can't which cluster and its points determine Social Media Addiction
- For a better understanding we perform confusion matrix for the above plot.

```
predicted_cs <- ifelse(kmeans.EFA_class$cluster > 1.5, "Not Addicted", "Addicted")
actual_cs <- ifelse(class$SocialMediaAddiction == "Addicted" , "Addicted", "Not Addicted")
confusion_cs <- table(predicted_cs, actual_cs)
confusion_cs
```

```
##               actual_cs
## predicted_cs   Addicted Not Addicted
##   Addicted            1            6
##   Not Addicted       99           69
```

- We can see that out of 100 it has predicted 1 as Addicted and 99 as Not Addicted.

2ND Question:

Using the Class_Survey dataset predict the number of individuals who are Addicted and Not Addicted and the accuracy of the derived result

Answer:

- The Class_Survey dataset is a type of dataset where we need to predict '1' or '0' or "Yes" or" No" or 'Addicted' and 'Not Addicted'. We use Logistic Regression for this.

- Firstly, We Split the dataset into Train and Test data with a Split ratio of 0.70.

```
# Let's try to predict the Social Media Addiction using the prediction

set.seed(123)
split <- sample.split(class$SocialMediaAddiction, SplitRatio = 0.70)
train_cs <- subset(class, split == TRUE)
test_cs <- subset(class, split == FALSE)

Xtrain_cs <- train_cs[, 1:14]
Ytrain_cs <- train_cs[, 15]
Ytrain_cs <- unlist(Ytrain_cs)
#Ytrain_cs <- as.integer(Ytrain_cs)


Xtest_cs <- test_cs[, 1:14]
x_cs <- cbind(Xtrain_cs, Ytrain_cs)
logistic_cs <- glm(Ytrain_cs ~ ., data = x_cs, family = 'binomial')
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logistic_cs)
```

- We use the confusion matrix to predict the values of the individuals who are Addicted and Not Addicted.

```
##              actual_cs
## predicted_cs   Addicted Not Addicted
##   Addicted            7           15
##   Not Addicted       23            8
```

```
TP <- sum(actual_cs == 'Addicted' & predicted_cs == 'Addicted')
FP <- sum(actual_cs == 'Not Addicted' & predicted_cs == 'Addicted')
TN <- sum(actual_cs == 'Not Addicted' & predicted_cs == 'Not Addicted')
FN <- sum(actual_cs == 'Addicted' & predicted_cs == 'Not Addicted')

recall <- TP / (TP + FN)
precision <- TP / (TP + FP)

recall
```

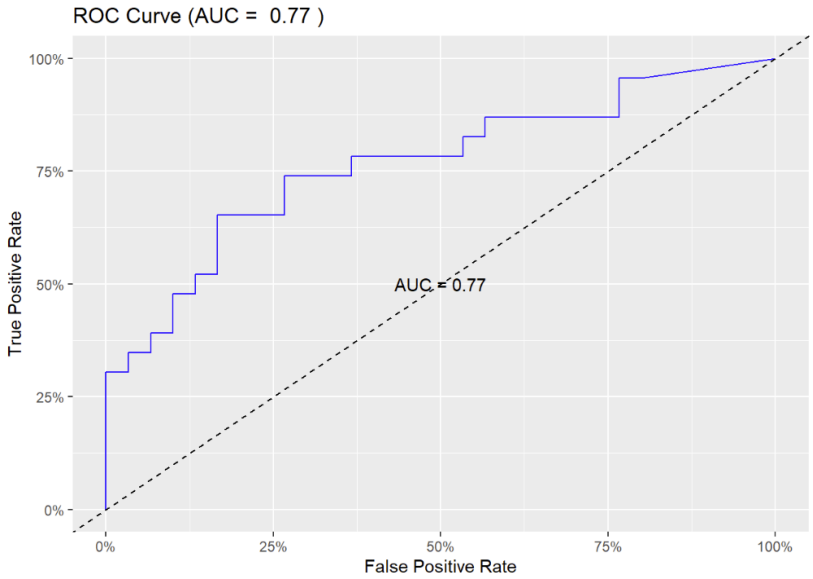
```
## [1] 0.2333333
```

```
precision
```

```
## [1] 0.3181818
```

- **We can see that the prediction shows 7 as Addicted and 23 as Not Addicted.**
- **The Recall shows 23% and Precision as 31%. This shows us the prediction model is not up to the mark.**

The accuracy of the dataset can be shown using the ROC Curve




- **From the above plot we see the dataset accuracy is 0.77(77%). We can say that the predicted was not accurate to the mark basing on the confusion matrix mentioned above and needs more improvement in prediction.**