Social Media

Social Media Usage Analysis

1. About the data

This dataset captures the patterns of social media usage along with associated metrics such as mood, productivity, tiredness, and how many interview calls an individual gets, etc. It provides insights into how individuals allocate their time across various social media platforms and their subjective experiences and well-being indicators. We collected this data over weeks from all the students in the class where each student was recording the time spent on each of the apps and the qualities such as mood productivity, tiredness waking up, etc.

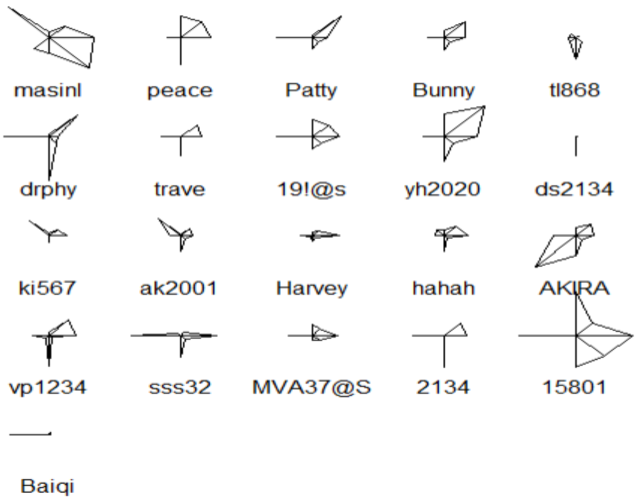
Data Dictionary

- Eight apps are considered for which the number of hours used by everyone has been recorded. The apps are:
- Instagram, LinkedIn, Snapchat, Twitter, WhatsApp/WeChat, YouTube, Reddit and OTT apps.
- There are also columns such as how many interview calls, how many coffee chats were done over networking, how many learning items were created, mood productivity, Tired waking up, trouble falling asleep, and how an individual felt throughout the week. All these were recorded to see how the usage of social media is affecting these activities.

Question: Wanted to observe how much time is spent on social media apps, which apps are used together, and what can be the reason these apps are used together. And also how the time spent on these apps is affecting the activities such as mood productivity etc.

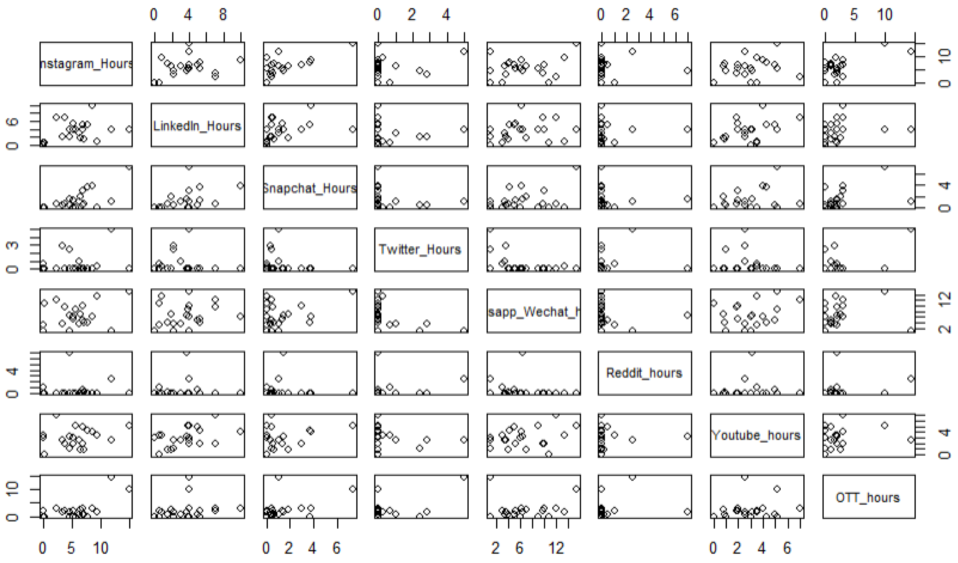
2. Analyzing the data

Star Plot



- Star plot helps us to check the similarity between the individuals.
- We can observe that none of them are very much related, except for a few have a little similarity.

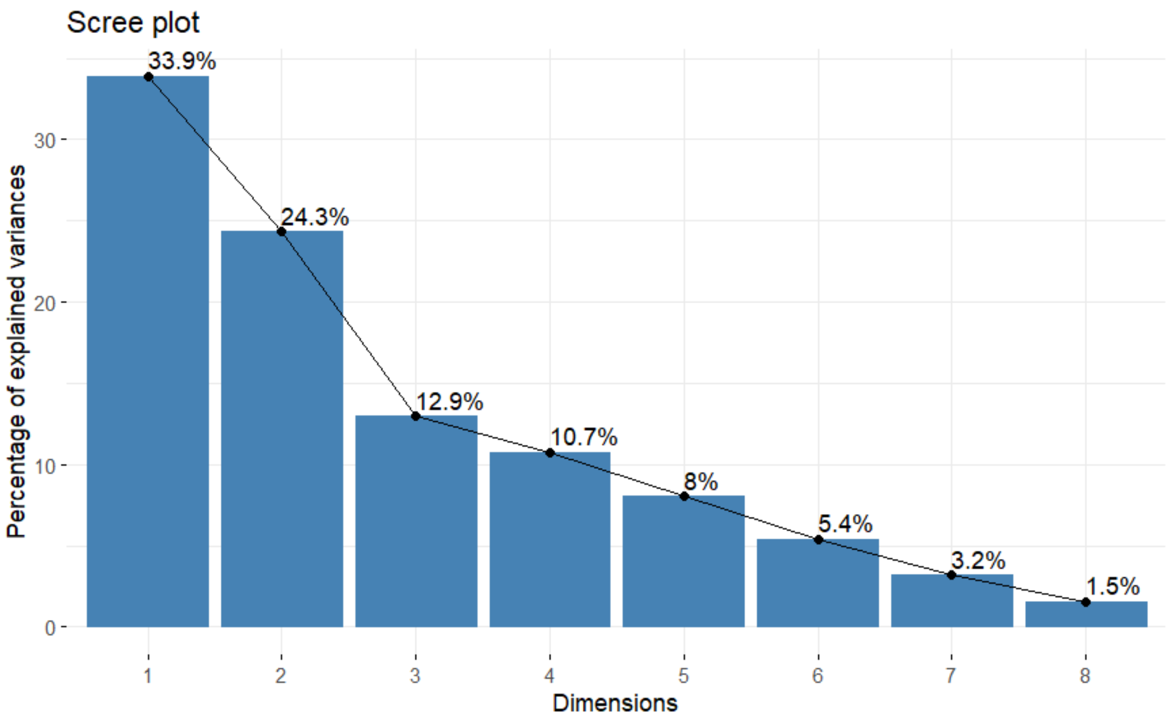
Correlation plot



- From the above plot we can observe that none of the columns are linearly correlated, they all are mostly in the form of clusters.

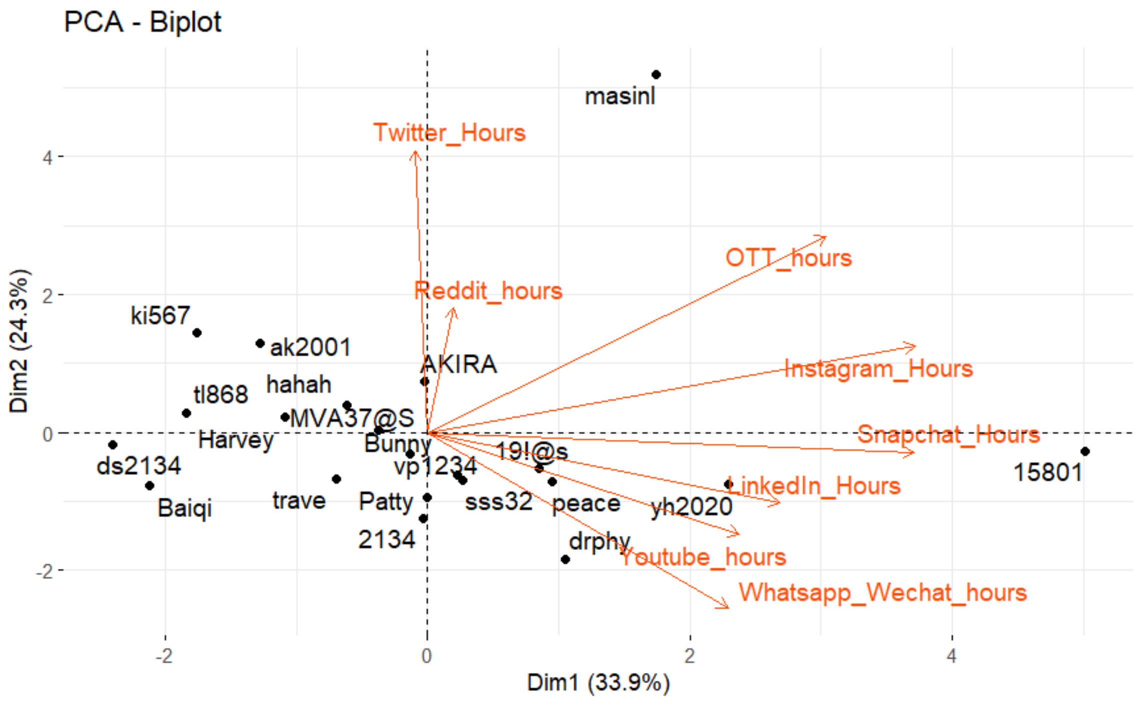
PCA Analysis

Performed PCA analysis for doing dimensionality reduction and understanding on what basis the reduction is happening.



- The scree diagram shows us that the sum of the first 3 principal components is 71.1% and tells us 3 PCs should be considered.
- So, we can use PCA for column reduction as well.
- And we can also observe a significant curve shift after the third dimension.

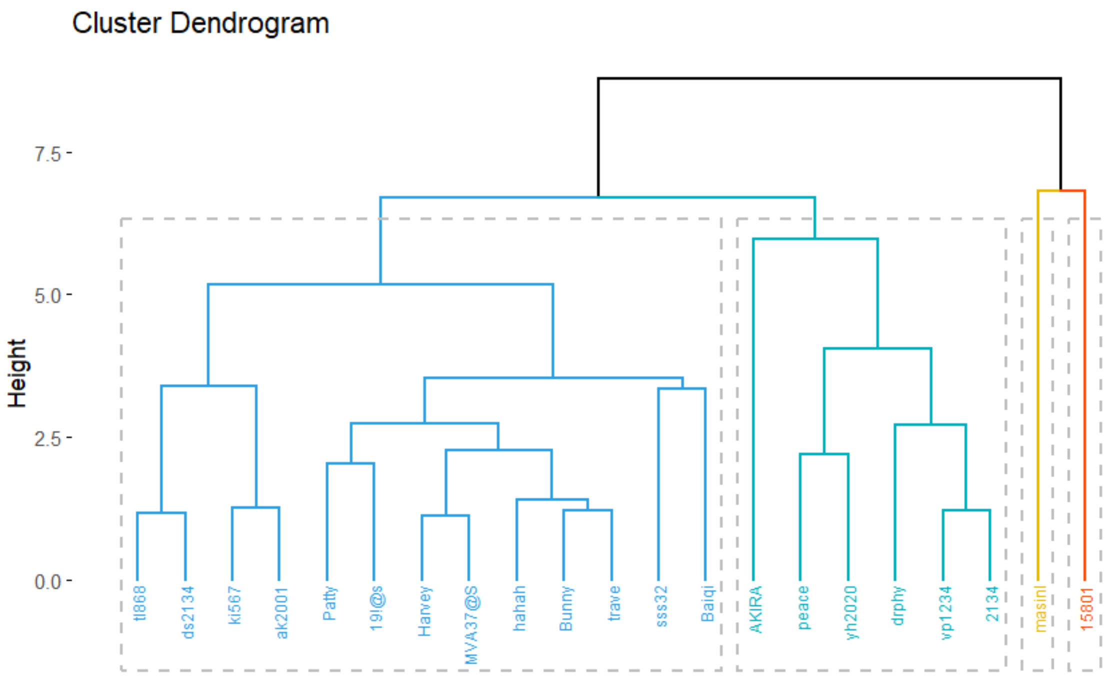
PCA Biplot



- We can observe how one set of students uses the apps more than the other set of students.
- We can also observe which apps are used commonly by which set of students.
- It also shows us which apps are strongly correlated. We can observe that YouTube and LinkedIn are strongly correlated as they have very small angles between them.
- We can also observe the two outliers in the class, 15801(which was me! :) and masinl are ones, this can be because they spent more hours on a certain app alone.

Cluster Analysis

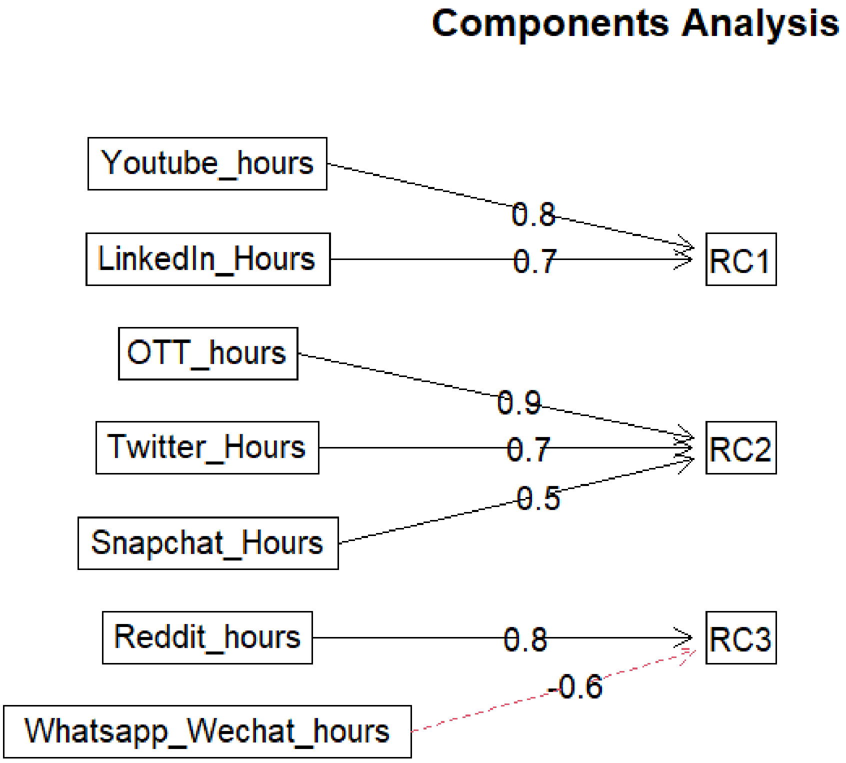
Cluster Dendrogram



- We can observe from that, the outliers we observed in PCA Analysis are formed into two separate clusters.
- The two sets that we observed, one set consisting of people who used the apps comparatively more than the other set are formed into two separate clusters.

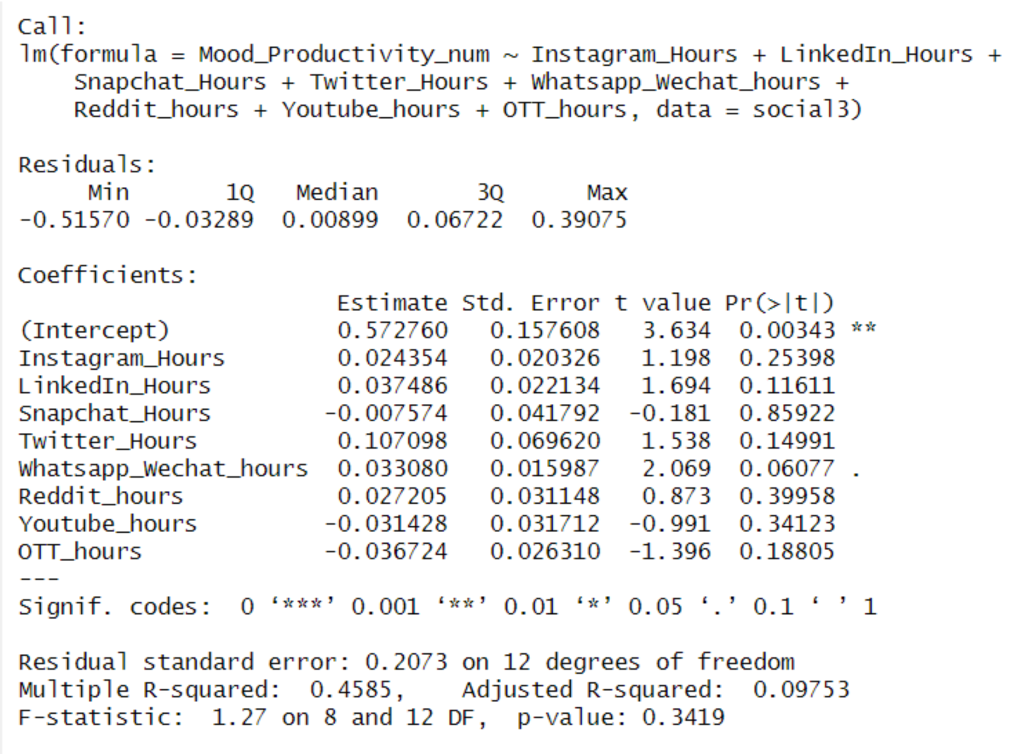
Factor Analysis

Component Analysis



- From factor analysis we can see which apps have similar usage.
- We can see that Youtube_Hours and LinkedIn_Hours are combined maybe because people might be learning professional experiences on both platforms to learn something in both concepts, or maybe LinkedIn for learning and YouTube for leisure.
- We can observe that OTT, Twitter and Snapchat are combined to one, because they are all giving social media entertainment.
- We can observe that whatsapp_wechat are reddit are inversely correlated, as in, if one is being used more the other is being used less.

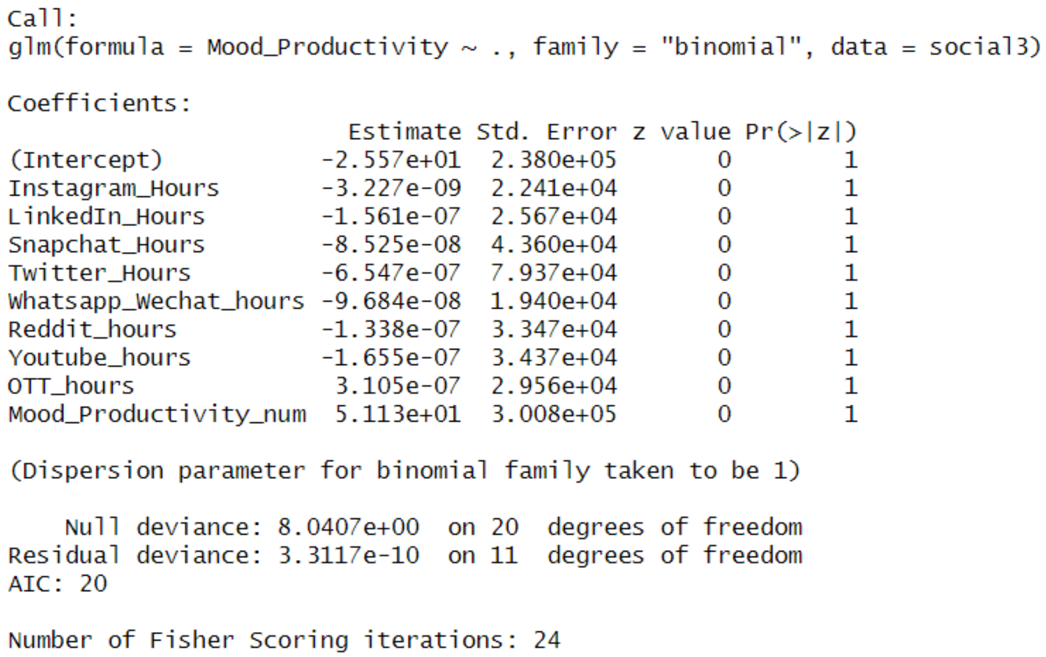
Multiple Regression



```
Call:
lm(formula = Mood_Productivity_num ~ Instagram_Hours + LinkedIn_Hours +
    Snapchat_Hours + Twitter_Hours + Whatsapp_Wechat_hours +
    Reddit_hours + Youtube_hours + OTT_hours, data = social3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51570 -0.03289  0.00899  0.06722  0.39075

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.572760   0.157608   3.634  0.00343 **
Instagram_Hours        0.024354   0.020326   1.198  0.25398
LinkedIn_Hours         0.037486   0.022134   1.694  0.11611
Snapchat_Hours        -0.007574   0.041792  -0.181  0.85922
Twitter_Hours          0.107098   0.069620   1.538  0.14991
Whatsapp_Wechat_hours  0.033080   0.015987   2.069  0.06077 .
Reddit_hours           0.027205   0.031148   0.873  0.39958
Youtube_hours         -0.031428   0.031712  -0.991  0.34123
OTT_hours             -0.036724   0.026310  -1.396  0.18805
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2073 on 12 degrees of freedom
Multiple R-squared:  0.4585,	Adjusted R-squared:  0.09753
F-statistic:  1.27 on 8 and 12 DF,  p-value: 0.3419
```

- For my analysis I considered the predicting column to be the mood productivity of an individual based on the social media app usage.
- The Median being close to 0 showcases that the model can predict perfectly.
- Furthermore, it can be observed that most columns have P-values > 0.05, which means that none of them significantly affect the target variable.
- We can tell that the model is not that great because the significance of each column on the predicting column is not that high.

Logistic Regression



```
Call:
glm(formula = Mood_Productivity ~ ., family = "binomial", data = social3)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.557e+01  2.380e+05       0        1
Instagram_Hours       -3.227e-09  2.241e+04       0        1
LinkedIn_Hours        -1.561e-07  2.567e+04       0        1
Snapchat_Hours        -8.525e-08  4.360e+04       0        1
Twitter_Hours         -6.547e-07  7.937e+04       0        1
Whatsapp_Wechat_hours -9.684e-08  1.940e+04       0        1
Reddit_hours          -1.338e-07  3.347e+04       0        1
Youtube_hours         -1.655e-07  3.437e+04       0        1
OTT_hours              3.105e-07  2.956e+04       0        1
Mood_Productivity_num  5.113e+01  3.008e+05       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.0407e+00  on 20  degrees of freedom
Residual deviance: 3.3117e-10  on 11  degrees of freedom
AIC: 20

Number of Fisher Scoring iterations: 24
```

- As we can observe the p-value is 1 for all the outcome variables, which means that there is a weak relationship between the predictor and outcome variable.

Takeaway from the overall analysis

- There are two sets of students in the class which were clearly clustered and differentiated. One has students who spent comparatively more time on these apps and the other which comparatively spent less time on the apps.
- There are few outliers, which means they spent a lot more time on certain types of apps than others.
- We could also see which apps are being used the most together, which can be used for better and further analysis.
- By performing the regression models we can see that the models performance is not that great, the variables have very less significance on the predicting variable.