Final Project

Naby Diop

12/16/2019

Introduction

On August 5, 2015, Facebook launched Lives. At first it was only destinated to celebrities, but after the success that it has encounter it was soon oppened to every Facebook user. The data Live collect the reactions of Facebook users from "Facebook pages of 10 Thai fashion and cosmetics retail sellers Posts of a different nature (video, photos, statuses, and links). Engagement metrics consist of comments, shares, and reactions." More information about the data can be found the on the UCI machine learning. The goal of this study is to do an comparative analysis between two distinct period: before and after August 5, 2015. Therefore, we will be using the technics of unsupervised learnling such as principal componant analysis, a tool used for data visualization or data pre-processing before supervised techniques are applied, and clustering, a broad class of methods for discovering unknown subgroups in data. We seek to visualize the importance of viodeo by secking how important is the variables (num_wows, num_loves, num_sade) that were introduce after Facebook live videos was launched.

Data

```
## 'data.frame':    7050 obs. of  16 variables:
##  $ status_id       : Factor w/ 6997 levels "1050855161656896_1050858841656528",..: 2832 2831 2830 283
##  $ status_type     : Factor w/ 4 levels "link","photo",..: 4 2 4 2 2 2 4 4 2 2 ...
##  $ status_published: Factor w/ 6913 levels "1/1/2018 1:39",..: 3950 3920 3922 3918 3833 3831 3827 381
##  $ num_reactions   : int  529 150 227 111 213 217 503 295 203 170 ...
##  $ num_comments    : int  512 0 236 0 0 6 614 453 1 9 ...
##  $ num_shares      : int  262 0 57 0 0 0 72 53 0 1 ...
##  $ num_likes       : int  432 150 204 111 204 211 418 260 198 167 ...
##  $ num_loves       : int  92 0 21 0 9 5 70 32 5 3 ...
##  $ num_wows        : int  3 0 1 0 0 1 10 1 0 0 ...
##  $ num_hahas       : int  1 0 1 0 0 0 2 1 0 0 ...
##  $ num_sads        : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ num_angrys      : int  0 0 0 0 0 0 3 1 0 0 ...
##  $ Column1         : logi  NA NA NA NA NA NA ...
##  $ Column2         : logi  NA NA NA NA NA NA ...
##  $ Column3         : logi  NA NA NA NA NA NA ...
##  $ Column4         : logi  NA NA NA NA NA NA ...
```

The dataset Live was collected by Nassim Dehouche, a Ph.D researcher at Mahidol University International College. It has 7050 observations, 16 variables. We first examine the data by looking at its structure. The first thing that we realized is that the data has four additional variables that should be removed. The second thing that we noticed is that the data is a mix of numerical and categorical variables. However, the variable **status_published** is recorded as factor type where it should be date type. We convert this variable to the apopriate type (date type), than we will look at the summary of the data and analyze it before forwarder analysis.

After we remove the four last colones we can look at the names of the varibles in the data using the function **names** in R.

```
##  [1] "status_id"        "status_type"      "status_published"
##  [4] "num_reactions"    "num_comments"     "num_shares"
##  [7] "num_likes"        "num_loves"        "num_wows"
```

```
## [10] "num_hahas"        "num_sads"         "num_angrys"
```

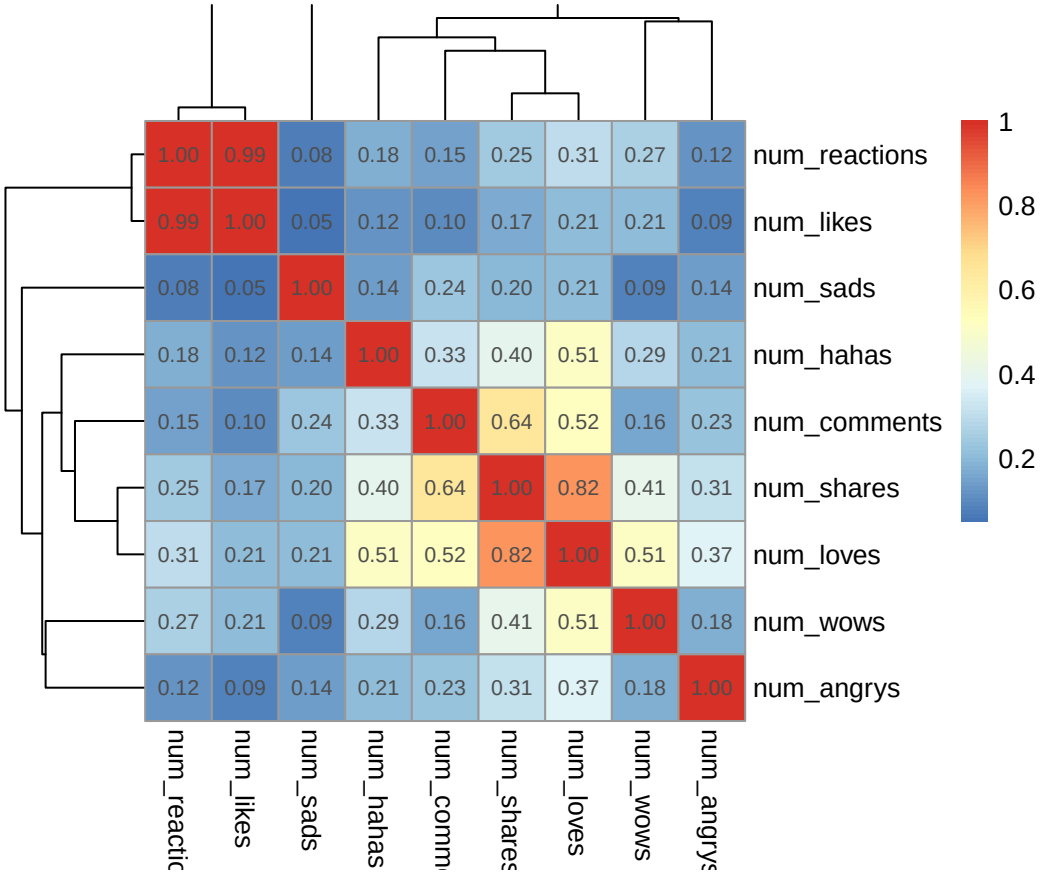
Before we look at the summary, we first convert the variable **status_published** to date type.

Here is the summary

```
##                              status_id    status_type
##  246675545449582_326883450762124 :   2   link  :  63
##  246675545449582_429583263825475 :   2   photo :4288
##  819700534875473_1000607730118085:   2   status: 365
##  819700534875473_1001982519980606:   2   video :2334
##  819700534875473_1002372733274918:   2
##  819700534875473_951614605017398 :   2
##  (Other)                         :7038
##  status_published     num_reactions     num_comments       num_shares
##  Min.   :2012-07-15   Min.   :   0.0   Min.   :    0.0   Min.   :   0.00
##  1st Qu.:2016-03-15   1st Qu.:  17.0   1st Qu.:    0.0   1st Qu.:   0.00
##  Median :2017-11-18   Median :  59.5   Median :    4.0   Median :   0.00
##  Mean   :2016-11-19   Mean   : 230.1   Mean   :  224.4   Mean   :  40.02
##  3rd Qu.:2018-03-09   3rd Qu.: 219.0   3rd Qu.:   23.0   3rd Qu.:   4.00
##  Max.   :2018-06-13   Max.   :4710.0   Max.   :20990.0   Max.   :3424.00
##
##    num_likes        num_loves         num_wows         num_hahas
##  Min.   :   0.0   Min.   :  0.00   Min.   :  0.000   Min.   :  0.0000
##  1st Qu.:  17.0   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.0000
##  Median :  58.0   Median :  0.00   Median :  0.000   Median :  0.0000
##  Mean   : 215.0   Mean   : 12.73   Mean   :  1.289   Mean   :  0.6965
##  3rd Qu.: 184.8   3rd Qu.:  3.00   3rd Qu.:  0.000   3rd Qu.:  0.0000
##  Max.   :4710.0   Max.   :657.00   Max.   :278.000   Max.   :157.0000
##
##     num_sads         num_angrys
##  Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000
##  Mean   : 0.2437   Mean   : 0.1132
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :51.0000   Max.   :31.0000
##
```

We first noticed from the summary that non of colones has missing values. We can see that the median status_published is 2017-11-18 mainning more data was collected after Facebook lives was launched. Also the status_id should be removed from the data because it won't have any effect on forwarder analysis.

Now we can look at the corrolation of the reactions (num_comments, num_shares, num_likes, num_loves, num_wows, num_hahas, num_sads, num_angrys) from users to the different types of posts (video, link, photo).



This heatmap of the corrolation matrix indicates a strong positive relationship between the variables num_likes and num_reactions (0.99) meaning almost all the users who likes the videos also react to it. num_shares and num_loves are also high corrolated (0.82). This means that more than 80% of users who love a given post also shared it. The variables num_likes and num_angrys are weakly corrolated (0.09) which makes sens because someone who is angry about a post certainly doesn't like it.

Analysis

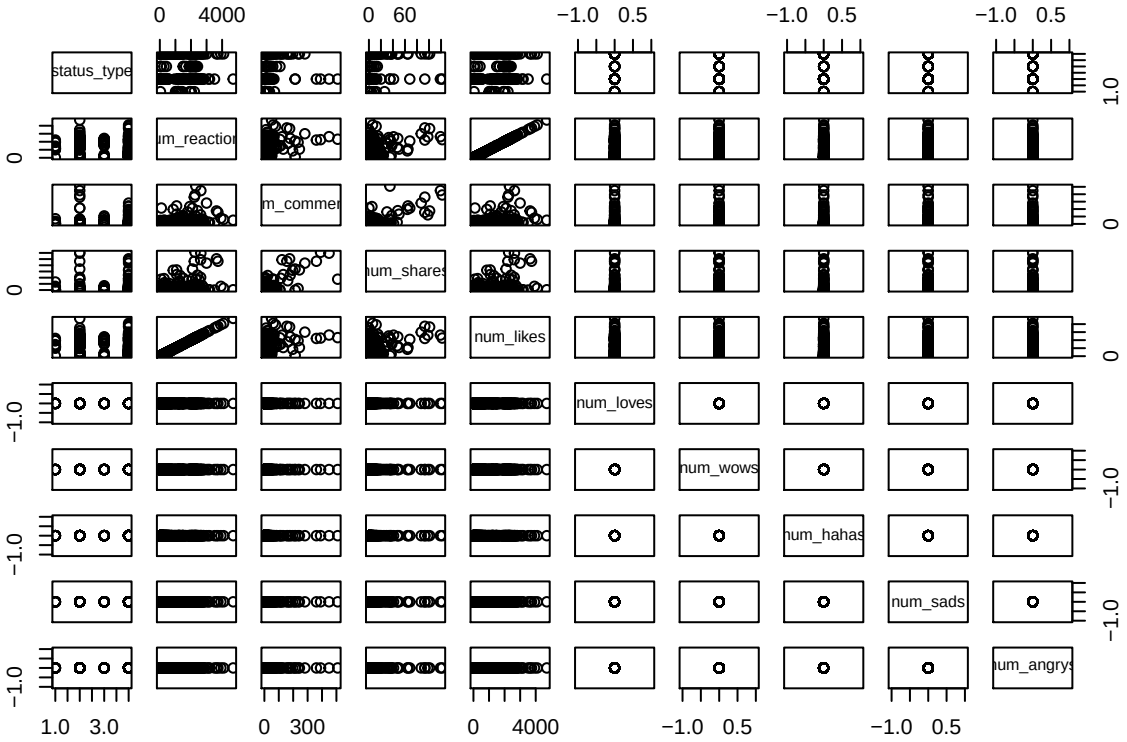
As stated in the introduction, we seek to compare users reactions before and after August 5, 2015. To do such analysis we first split the data into two parts. One part recording the reactions before August 5, 2015 and another part after that. We firs use the cluster method and end up with principal componant analysis method (PCA).

Clustering

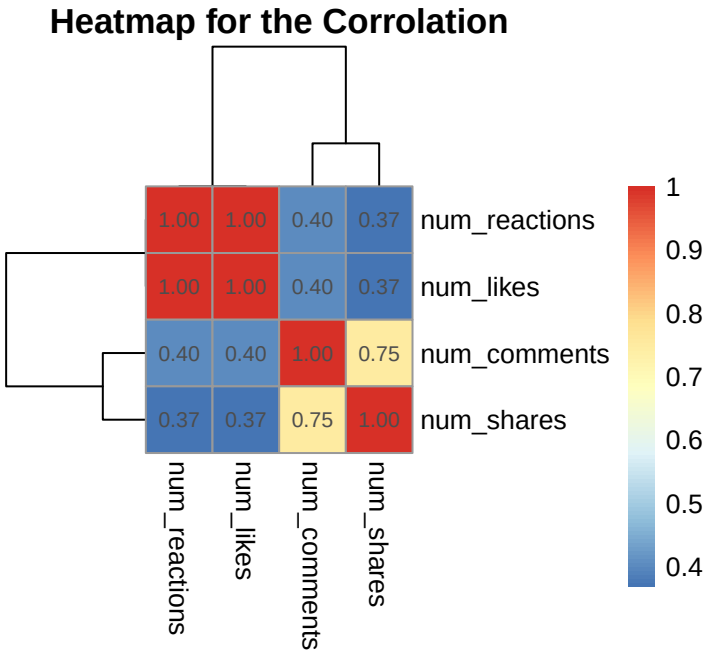
K-means method

Before August 5, 2015.

Lets first look at the scartteplot of the data.



The scartteplot shows that only the first quarter of the plots are maininful here. This fisrt quarter takes in count only the variables rum_reactions, num_likes, num_comments and num_shares. This makes sens since the others reactions do not exist at that time. It also indicate the presence of two single cluster with variables like num_shares, num_reactions, num_comments and num_likes. The other variables indicates only single cluster. Now before we fit a clustering model let look at first the correlation matrix of the first portion of the data.



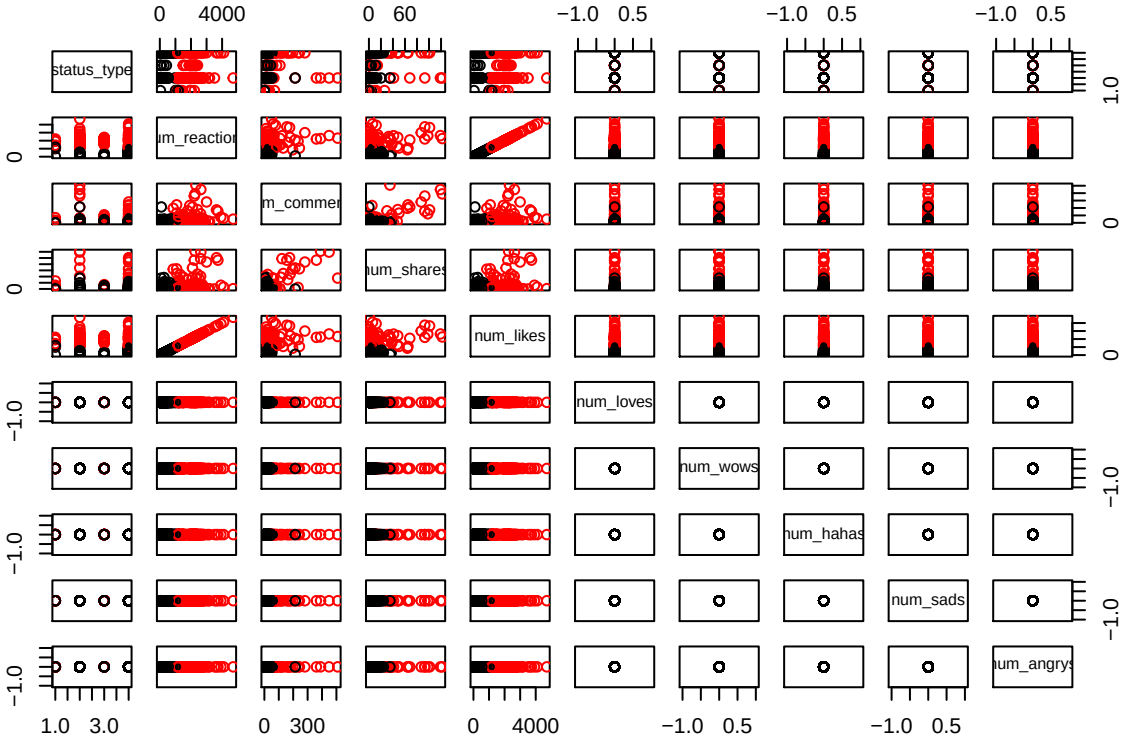
Reactions and shares are collinear (corrolation = 1). The variables num_shares and num_comments are also high corrolated (0.75) wheras num_shares, num_likes and num_reactions are weakly corrolated.

Fitting k-means

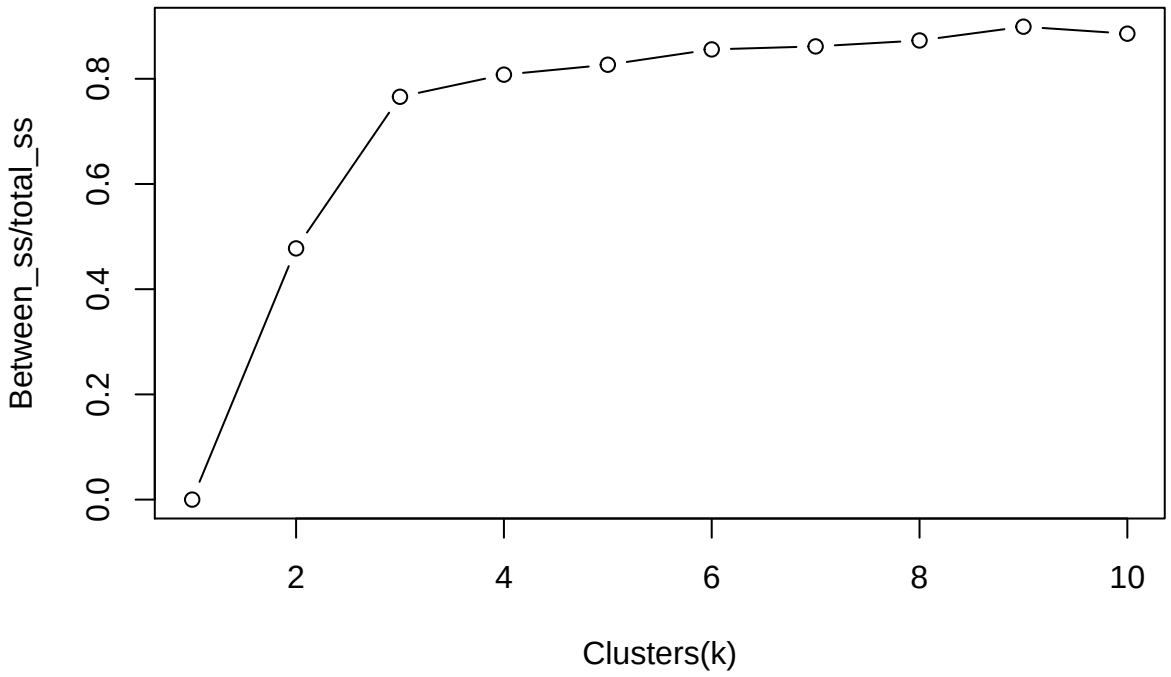
We have intuitively dirive from the scatteplot two clusters, we will be at first fitting the data with initialy k=2 (number of clusters).

```
##                  status_type
## km_Live$cluster link photo status video
##               1    5  1027      5   270
##               2    9   101     20    63
```

This table shows that first cluster contains only the 5links, 1027 photos, 5 status and 270 videos. The second cluster contains the 9 links, 101 photos, 20 status and 63 videos. A visualisation of the clusters is shown bellow.



We notice that the model does a very pore work. Some videos are being classified as photos and links as status vis versa. Therefore it is important to look for the optimal value of k. To do so we first fit kmeans for 10 different value ok k from 1 to 10. Then we calculate for each of this 10 models the percent variability in the data based on the model.Finally we plot the percent variability against it correspondind number or cluster.



This plot indicates that four cluster is the optimal number of cluster.

```
##               status_type
## k[[4]]$cluster link photo status video
##              1    1  1022      5   207
##              2    9    40      2    87
##              3    0     5      0     7
##              4    4    61     18    32
```

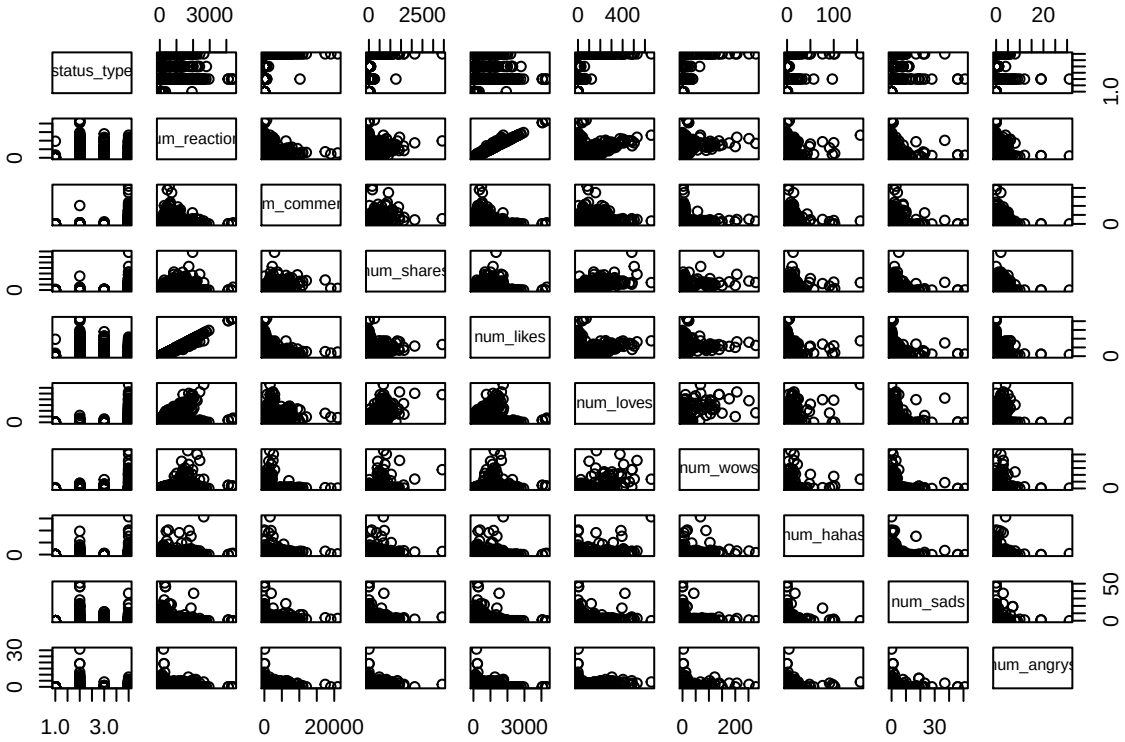
The table shows that first cluster has no links and status ant it contains only 5 photos and 7 videos. The second cluster contains 9 links, 40 photos, 2 status and 87 videos. The third cluster contains 1 link, 1022 photos, 5 status and 207 videos. The last cluster contains 4 links, 61 photos, 18 status and 32 videos. The accuracy of the model is given bellow.

```
## [1] 0.807847
```

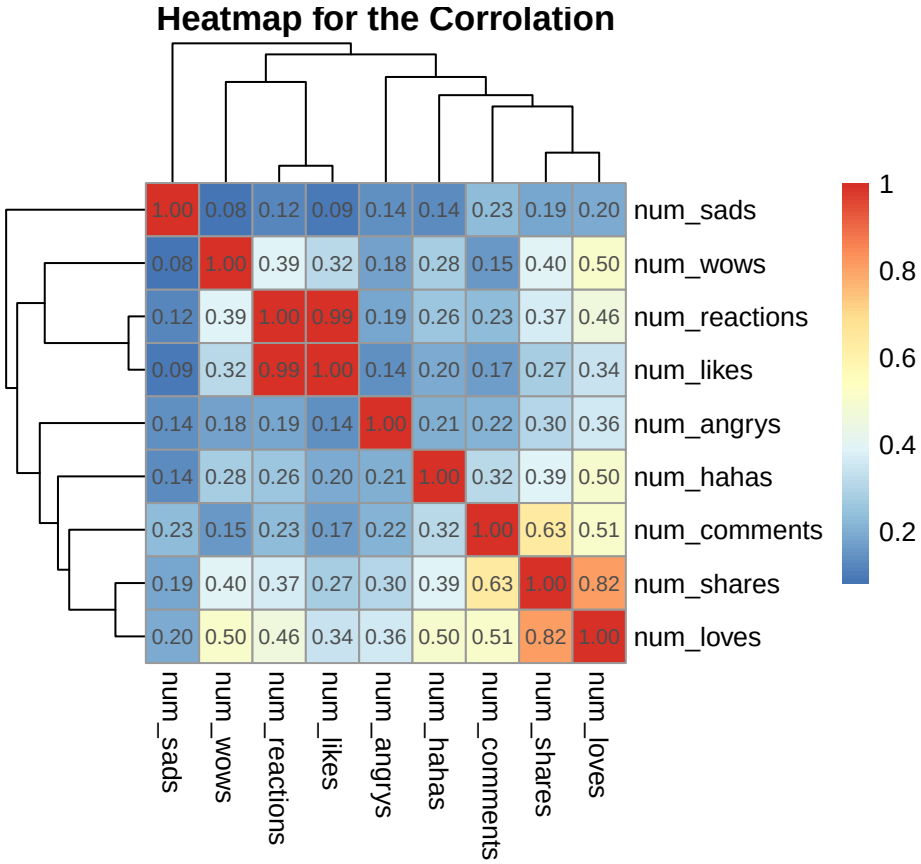
80% of the variability of the data is explained by the model with is a good indicator.

After August 5, 2015.

Lets look at the scartterplot first.



Over here all the variables are present since now Facebook lives are being used. The scatteplot roughly discribes that many variables have linear relationship. To have more insight about that we can look at the heatmap of the correlation matrix.



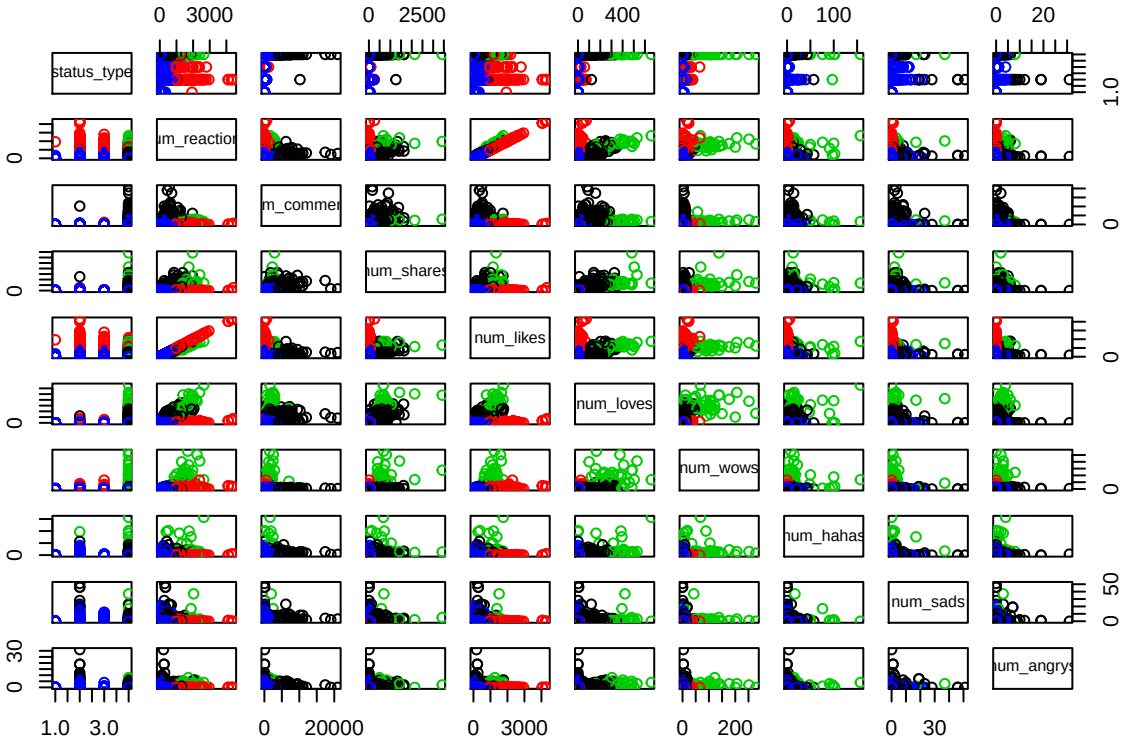
Here likes and reactions are colinear (0.99). shares and loves are very high corrolated. Wows and sades are not corrolated same as reactions and sades.

Fitting k-means

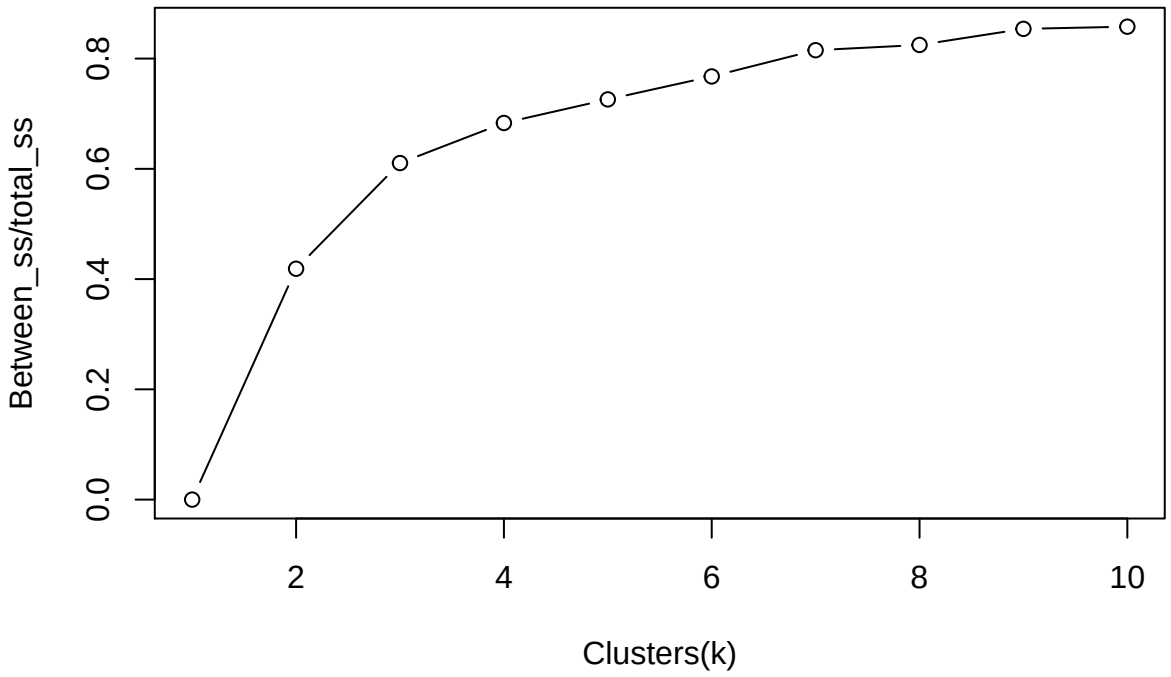
Here we initially fit kmeans with four cluster.

```
##                   status_type
## km_Live1$cluster link photo status video
##                1    0    21      0   341
##                2    1   130     63    10
##                3    0     1      0    37
##                4   48  3008    277  1613
```

The table shows, as for the first portion of the data, that first cluster has no links and status ant it contains only 1 photos and 37 videos. The second cluster contains 48 links, 3008 photos, 277 status and 1613 videos. The third cluster contains 1 link, 130 photos, 63 status and 10 videos. The last cluster contains no link and status 21 photos, and 341 videos.



The presence of miss classification is present, but the model does identify four acceptable clusters. We must, notice that in some case the clusters are overlapping.



As indicated above, here again four clusters seems to be optimal number of clusters.

```
## [1] 0.8578628
```

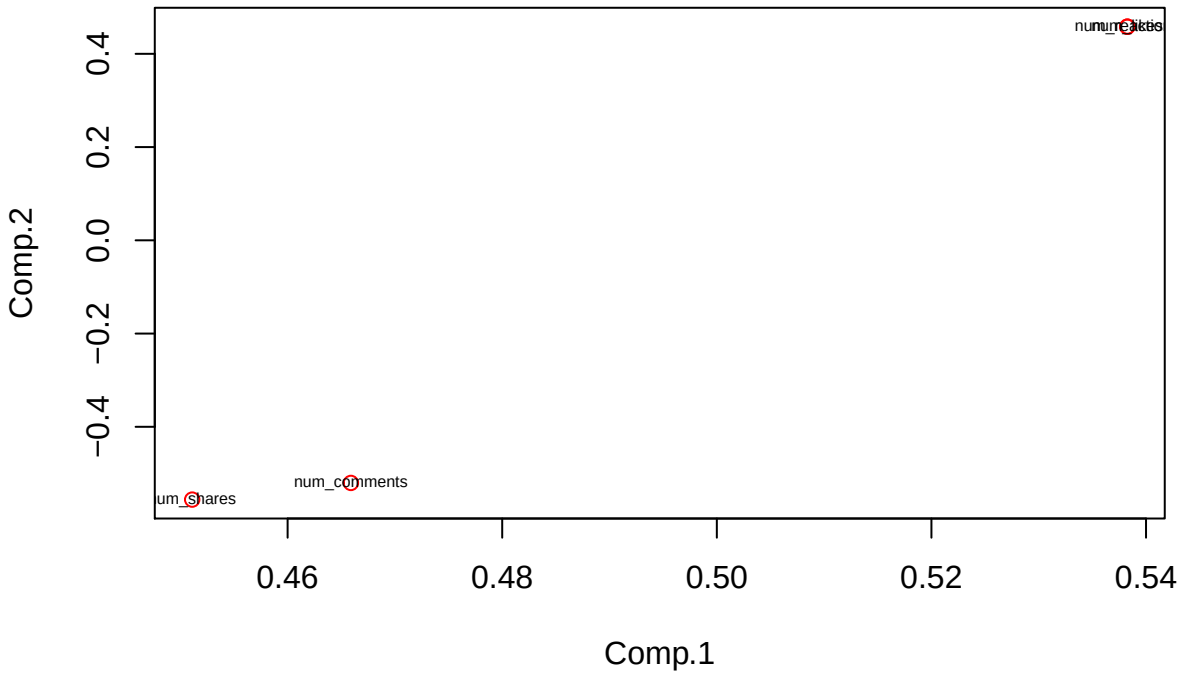
For the second portion of the data the model have better performance then for the first part (85%).

PCA

before August 5, 2015.

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3       Comp.4
## Standard deviation     1.6299976 1.0469628 0.49696733 1.214980e-08
## Proportion of Variance 0.6642231 0.2740328 0.06174413 3.690441e-17
## Cumulative Proportion  0.6642231 0.9382559 1.00000000 1.000000e+00
##
## Loadings:
##               Comp.1 Comp.2 Comp.3 Comp.4
## num_reactions  0.538  0.458         0.707
## num_comments   0.466 -0.520 -0.716
## num_shares     0.451 -0.556  0.698
## num_likes      0.538  0.458        -0.707
```

Two componants are enough for reduicing the dimension of this data, roughly 94% of the variability in data is explained by these two componants (cumulative Proportion = 0.938). We plot the loadings of this two componants in oder to visualize the most important variable for the two componants.



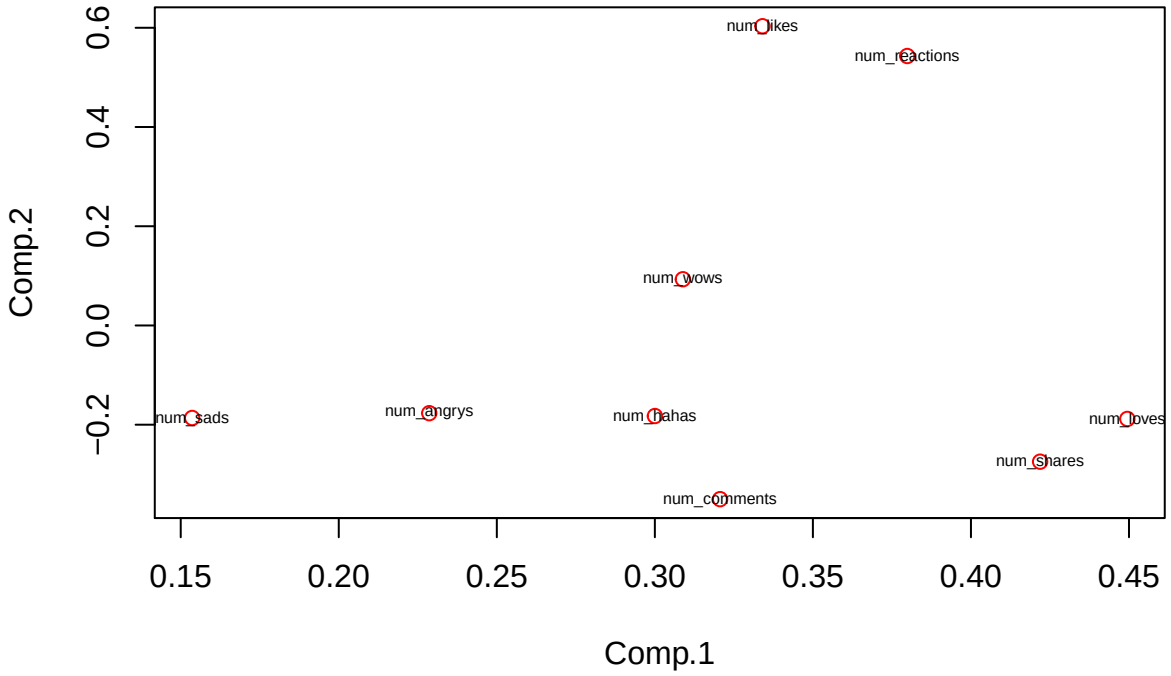
As indicated by the numerical values of the loadings, from this plot we can see that the variables num_likes and reactions are overlapping on the top right corner of the plot. The position of these two variables indicates that they are the most important for the first and second componant, Comments and shares are the least important for the two first componants. Also we notice that likes and reactions are equaly important for both componants. Whereas commantes are more important then shares for both componants.

After August 5, 2015.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.9264342 1.2165961 0.9771820 0.91687689 0.88312161
## Proportion of Variance 0.4123498 0.1644562 0.1060983 0.09340703 0.08665598
## Cumulative Proportion  0.4123498 0.5768061 0.6829044 0.77631140 0.86296737
##                           Comp.6     Comp.7     Comp.8       Comp.9
## Standard deviation     0.83333115 0.62212847 0.38962671 1.810855e-04
## Proportion of Variance 0.07716009 0.04300487 0.01686766 3.643550e-09
## Cumulative Proportion  0.94012746 0.98313233 1.00000000 1.000000e+00
##
## Loadings:
##               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## num_reactions  0.380  0.543  0.119         0.120
## num_comments   0.321 -0.350  0.144  0.336  0.448        -0.637  0.168
## num_shares     0.422 -0.275 -0.107  0.168  0.169 -0.238  0.397 -0.681
## num_likes      0.334  0.603  0.155         0.154
## num_loves      0.449 -0.188 -0.165                       0.482  0.696
## num_wows       0.309        -0.357        -0.622 -0.440 -0.421
## num_hahas      0.300 -0.182 -0.192        -0.302  0.848        -0.118
## num_sads       0.154 -0.186  0.857        -0.446
## num_angrys     0.229 -0.177        -0.918  0.232        -0.114
##               Comp.9
## num_reactions  0.726
## num_comments
## num_shares
```

```
## num_likes     -0.682
## num_loves
## num_wows
## num_hahas
## num_sads
## num_angrys
```

A visualization of the loadings is shown bellow.



From this plot we can see that the most important feature for the first componant is num_love follow by num_shares, num_reactions and num_likes. This basicly implying that videos incators are the most capture by componant one. The most important variables for the second componant are num_likes and um_reactions. These two variables are indicators of photos, links and status.

Conclusion

The introduction of facbook lives has a lot of impact. It increases the numbers of comments and shares. In fact for the PCA visualisation we notice before the introction of Facebook lives only comments and shares was important, but after that loves becomes more important which is a video indicator. Also the number of shares reactions double from before August 2015 to after the same date (sugestesd by the clusters). Finally we can conclude that Facebook videos has given more reactions in general to the Thai sellers posts.

Appendix

Data Preparation

```
library("PerformanceAnalytics")
library(ggplot2)
library(pheatmap)
```

```
Live = read.csv("Live.csv")
attach(Live)
```

```
str(Live)
```

```
Live = Live[ , -c(13, 14, 15, 16)]
```

```
names(Live)
```

```
time= as.character(Live$status_published)
Live$status_published = as.Date(time, format="%m/%d/%Y %H:%M")
Live$status_type = as.factor(Live$status_type)
```

```
summary(Live)
```

```
Live  = Live[-1]
```

```
Reactions  = Live[, c(3,4, 5, 6, 7, 8, 9, 10, 11)]
```

Preliminary analysis

```
S = cor(Reactions)
```

```
pheatmap(as.matrix(S),cellwidth = 25,main = "Heatmap for the
Corrolation", cellheight = 25, display_numbers = TRUE, Rowv =
TRUE, Colv = FALSE)
```

```
Lives <- split(Live,Live$status_published<as.Date("2015-08-05"))
```

```
Live1 = Lives[1]$"FALSE"
Live2 = Lives[2]$"TRUE"
```

```
plot(Live2[-2])
```

```
Reactions1  = Live1[, c(3,4, 5, 6, 7, 8, 9, 10, 11)]
Reactions2  = Live2[, c(3,4, 5, 6, 7, 8, 9, 10, 11)]
```

```
S1 = cor(Reactions1)
S2 = cor(Reactions2[, c(1, 2, 3, 4)])
```

```
pheatmap(as.matrix(S2),cellwidth = 25,main = "Heatmap for
the Corrolation", cellheight = 25, display_numbers = TRUE,
Rowv = TRUE, Colv = FALSE)
```

Kmeans model

```
km_Live = kmeans(scale(Live2[,c(3,4, 5, 6)]), 2)
```

```
xtabs(~ km_Live$cluster + status_type, data = Live2[-2])
```

```
plot(Live2[-2], col = km_Live$cluster)
```

```
k = list()
for (i in 1:10) {
  k[[i]] = kmeans(scale(Live2[,c(3,4, 5, 6)]), i)
}
```

```
betweenss_totss = list()
for (i in 1:10) {
  betweenss_totss[[i]] = k[[i]]$betweenss/k[[i]]$totss
}
```

```
plot(1:10, betweenss_totss, type = "b", ylab = "Between_ss/total_ss", xlab = "Clusters(k)")
```

```
betweenss_totss[[4]] = k[[i]]$betweenss/k[[4]]$totss
betweenss_totss[[4]]
```

```
plot(Live1[-2])
```

```
pheatmap(as.matrix(S1),cellwidth = 20,main = "Heatmap for
the Corrolation", cellheight = 20, display_numbers = TRUE,
Rowv = TRUE, Colv = FALSE)
```

```
km_Live1 = kmeans(scale(Live1[,c(3,4, 5, 6, 7, 8, 9,10,11)]), centers = 4)
```

```
xtabs(~ km_Live1$cluster + status_type, data = Live1[-2])
```

```
plot(Live1[-2], col = km_Live1$cluster)
```

```
plot(1:10, betweenss_totss, type = "b", ylab = "Between_ss/total_ss", xlab = "Clusters(k)")
```

PCA

```
pca_Live2 = princomp(covmat = S2)
summary(pca_Live2, loadings = T)
```

```
plot(pca_Live2$loadings, type = "p", col = "red")
text(pca_Live2$loadings, row.names(pca_Live2$loadings), cex = 0.5)
```

```
pca_Live1 = princomp(covmat = S1)
summary(pca_Live1, loadings = T)
```

```
plot(pca_Live1$loadings, col = "red")
text(pca_Live1$loadings, row.names(pca_Live1$loadings), cex = 0.5)
```

Reference

Alvin C. Rencher and William F. Christensen Method of Multivariate Analysis, John Wiley & Sons, 3^{th} edition, 2012.

Robert Gentleman, Kurt Hornik, and Giovanni Parmigiani, *An Introduction to Applied Multivariate Analysis with R*, Springer, 2011.

Gareth James, Daniela Witten, and Trevor Hastie *An Introduction to Statistical Learning*, Springer+Business Media, 2013.

Jay L. Devore, and Kenneth N. Berk, *Modern Mathematical Statistics With Applications*, Springer, Second Edition, 2012.

Julian J. Faraway *Extending the Linear Model with R*: Generalized Linear, Mixed Effect and Nonparametric Regression Models, CRC Press, Second edition, 2016.

Kutner, Nachtsheim, Neter and Li, *Applied Linear Statistical Models*, McGraw-Hill!Irwin, 5th edition, 2015.