Analysis of Facebook Metrics

MT5758 Applied Multivariate Analysis

Group: **03**

Student ID: **180025784**

Contents

Abstract

The growth of the Internet User Index has led to the global spread of social media, which may make the latter become an essential medium for major brands to reach customers. In this report, the author focuses on the relationship between the three response variables of type of one Facebook post, category of one post and if the post is paid respectively and other interactions attributes by mainly applying PCA (Principal Component Analysis) to the Facebook Metrics dataset obtained from UCI Machine Learning Repository (Moro, Rita and Vala, 2016). The number of people engaged with the post is found influential on type of one post while the number of people who liked the post is found important to the category of the post. Also, the number of total interactions of the post is affected by paid promotion.

Introduction

An increasing number of enterprises realise the potential of the utilisation of Internet-based social networks, such as Facebook, Twitter, Instagram etc., can attract customers. Lariscy et al. (2009) consider that measuring the impact of advertising is an important issue that should be included in the global social media strategy.The report aims at illustrating the relationship between three different response variables of type and category of Facebook post as well as the paid promotion of the post and some other interactions variables. The target audience of this report are people who have the basic knowledge of statistics, especially PCA.

The dataset used in this report is acquired from UCI Machine Learning Repository. The raw dataset contains 500 observations and 19 variables including page total likes, type, category, post month, post weekday, post hour, paid, lifetime post total reach, lifetime post total impressions, lifetime engaged users, lifetime post consumers, lifetime post consumptions, lifetime post impressions by people who have liked your page, lifetime post reach by people who like your page, lifetime people who have liked your page and engaged with your post and total interactions. Among all the variables, type and category are categorical variables, paid is a binary variable and all the others are numerical variables. When doing the data exploration, first let type be the response variable and other numericals be explanatory variables. Then the same procedure was conducted to category and paid variables. The software used in this analysis is R Studio 3.5.2 (R Core Team, 2019).

The multivariate method used to analyse the dataset is PCA. My analysis primarily found that in terms of type, covariates related to people engaged have higher loadings in the first principal component, which indicates the number of people who clicked the post has some relationship with the type of the post. In addition, in terms of category, covariates related to the number of people liked the post have high loadings in the first principal component, which demonstrates that the category of the post is relevant with people show interests to the post. Last but not least, the post with paid promotion can gain more interactions in total.

Methods

-Data Manipulation

First we did data cleaning to drop five missing values of the dataset, so the cleaned dataset has 495 observations for further analysis. In order to research the relationship between categorical variables and numerical variables, we enable non-numerical variables, that is, type, category and paid to be factors. Also, to make the dataset more clearly to be read, I reordered the variables and drop three variables, post hour, post month and post weekday, which clearly will not affect the output. Also, I change the name of the columns for easier read in the plots.

-Model Fitting and PCA

When fitting the model with type and category as response variables respectively, I used multinomial logistic regression to only retain the significant covariates for further PCA. However, when fitting the model with paid as the response variable, I used generalized linear model with binomial as family, but I realise that the significant covariates remaining cannot form a multivariate dataset. Thus, I give up doing PCA to this.

Because the datasets for all three analysis are multivariate variables, PCA is a suitable method for this. Two main approaches of choosing the number of principal components are Kaiser's Criterion and Scree plot. Because the use of correlation matrix in the analysis, we simply choose the components with eigenvalues that are greater than one. Plus, the scree plot shows the variances of all the components in the descending order, of which the components with high variances should be chosen.

After choosing the principal components, we can look at the loadings that indicate the percentages of contributions of every variable that comprises the data cloud. It is clear that the variables with higher proportion in the first principal component are believed to be influential. Furthermore, it is essential to check the scores of the variables in the second principal components which are also retained in PCA. When it comes to the plots, I used biplot to see the directions of the combination of each variable in the axes of first and second principal components and autoplot to see the performance of response variable in the axes of first and second principal components. Also, the network graph can show the correlation of the variables. After doing the PCA, I tried to fit the model with multinomial logistic regression again with the first and second principal component as the covariates to see the output, which undoubtfully shows the strong significance of principal component 1 and 2 in the model.

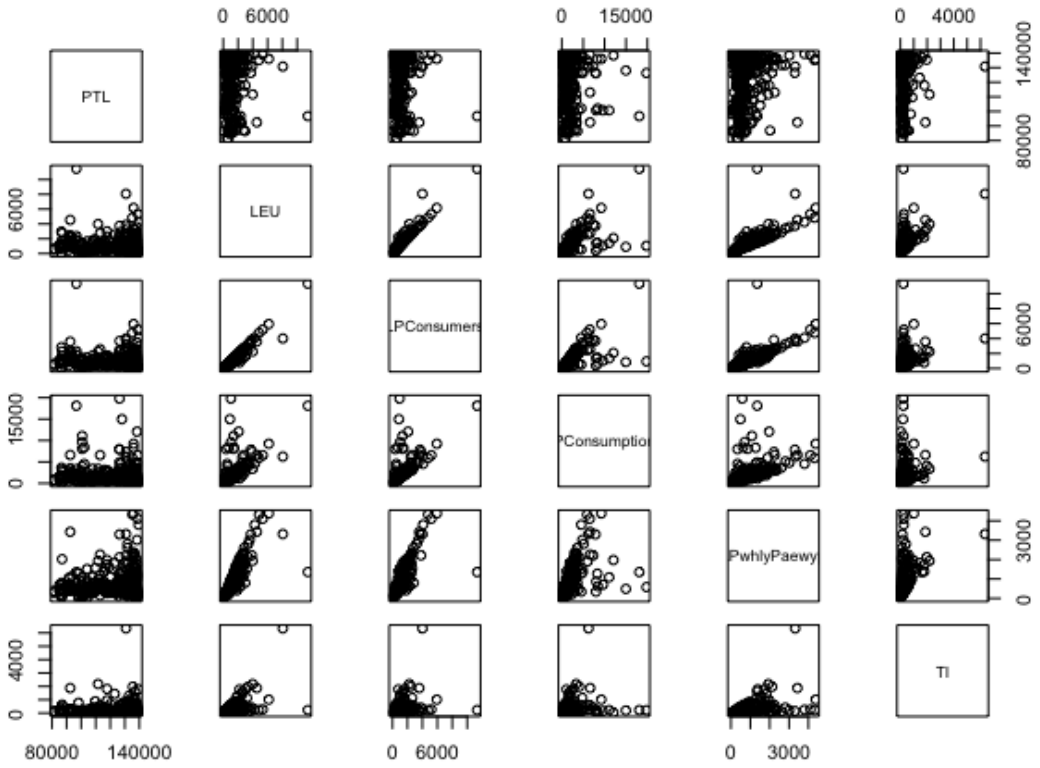
Results

Type as Response Variable

Multinomial Logistic Regression and Exploratory Analysis

The output of Anova of multinomial logistic regression indicates that page total likes, lifetime engaged users, lifetime post consumers, lifetime post consumptions, lifetime people who have liked your page and engaged with your post and total interactions are the six covariables that are significant and should be contained in the new dataset to conduct PCA. The scatter plot of every two variables of this new dataset are shown in Figure 1, from which we can clearly see the collinearity may exist between every two of lifetime engaged users, lifetime post consumers, lifetime post consumptions, lifetime people who have liked your page and engaged with your post. Furthermore, multinomial logistic regression was fitted again with the first and second principal component as the covariates to see the output, which undoubtfully shows the strong significance of principal component 1 and 2 in the model.

Figure 1 Scatter plot of dataset related to TYPE



Summary of PCA output

Figure 2 displays the summary of output of PCA. The first principal component has the largest standard deviation of about 1.89 and comprises 59.76% of the variance in the data cloud whilst the second principal component has the second largest standard deviation of 1.03 and proportion of 17.94%. And these two approximately contains 77.70% of the data cloud.

Figure 2 Output of PCA related to TYPE

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	1.893604	1.0373599	0.8689072	0.62650208	0.43507123	0.03681052
Proportion of Variance	0.597623	0.1793526	0.1258333	0.06541748	0.03154783	0.00022584
Cumulative Proportion	0.597623	0.7769756	0.9028089	0.96822633	0.99977416	1

Number of components to retain

-Kaiser's Criterion

In light of the criterion of Kaiser, the component with eigenvalues that are higher than 1 should be retained when using a correlation matrix. We can see from Figure 3 that the first two components have eigenvalues of 3.59 and 1.08 that should be retained.

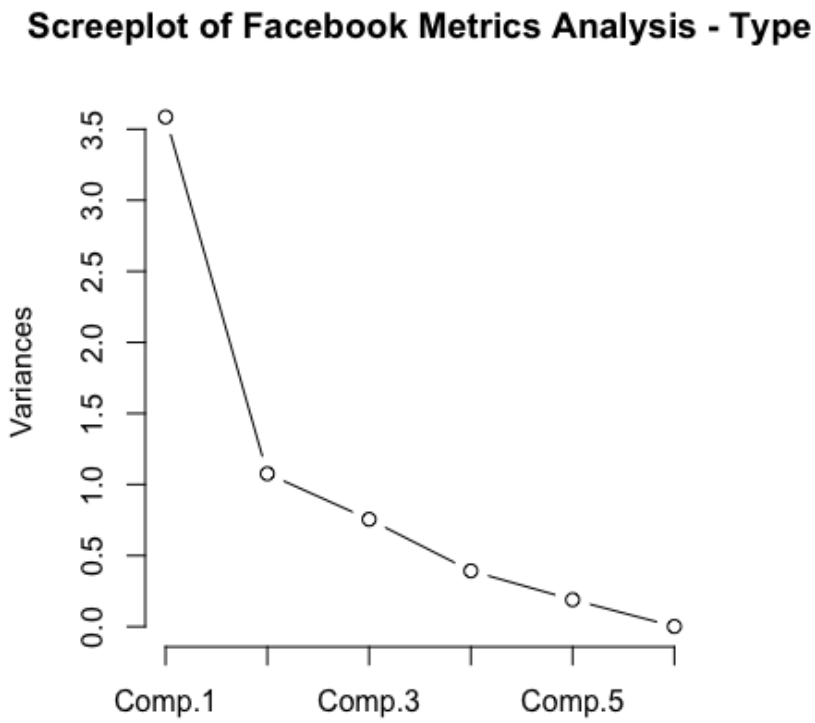
Figure 3 Eigenvalue of PCA related to TYPE

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Eigenvalue	3.58573799	1.07611551	0.75499966	0.39250486	0.18928698	0.00135501

-Scree Plot

According to Figure 4 of Scree plot, it can also be inferred that the first and second components should be kept because the eigenvalues of them are greatly higher than others.

Figure 4 Scree plot of PCA related to TYPE



-Equilibrium Contribution and Mardia's Criterion

When conducting equilibrium and Mardia's Criterion, they suggest that 6 and 3 variables should be kept respectively.

Interpretation of loadings

The table below indicates the loadings of each variable in the principal components. Lifetime engaged users has the highest loading in PC1 with a score of -0.516 whereas lifetime post

consumers has the second largest score of 0.498. It may seems reasonable to conclude that lifetime engaged users and lifetime post consumers contribute most to the variance. However, page total likes comprises the largest in the second component which may also be considered to be influential.

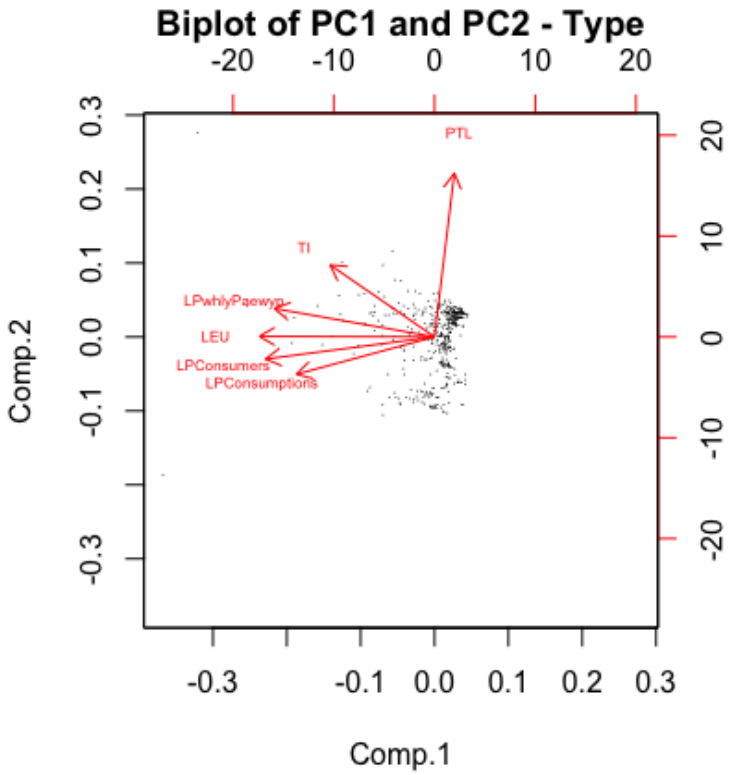
Figure 5 Loadings of PCA related to TYPE

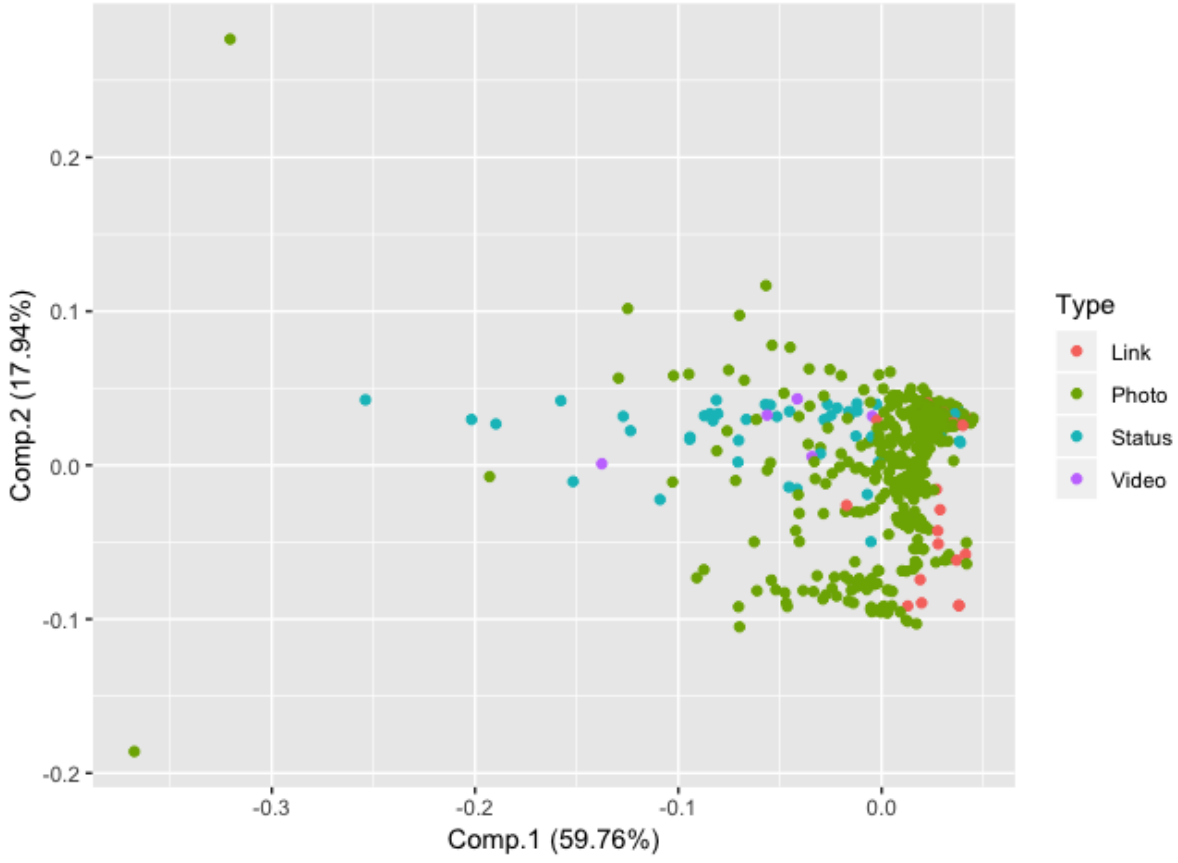
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
PTL		0.879	0.446		0.143	
LEU	-0.516			-0.171	0.411	-0.731
LPConsumers	-0.498	-0.119	0.188	-0.29	0.436	0.654
LPConsumptions	-0.407	-0.201	0.396	0.782	-0.159	
LPwhlyPaewyp	-0.472	0.152		-0.395	-0.771	
TI	-0.307	0.386	-0.777	0.34		0.193

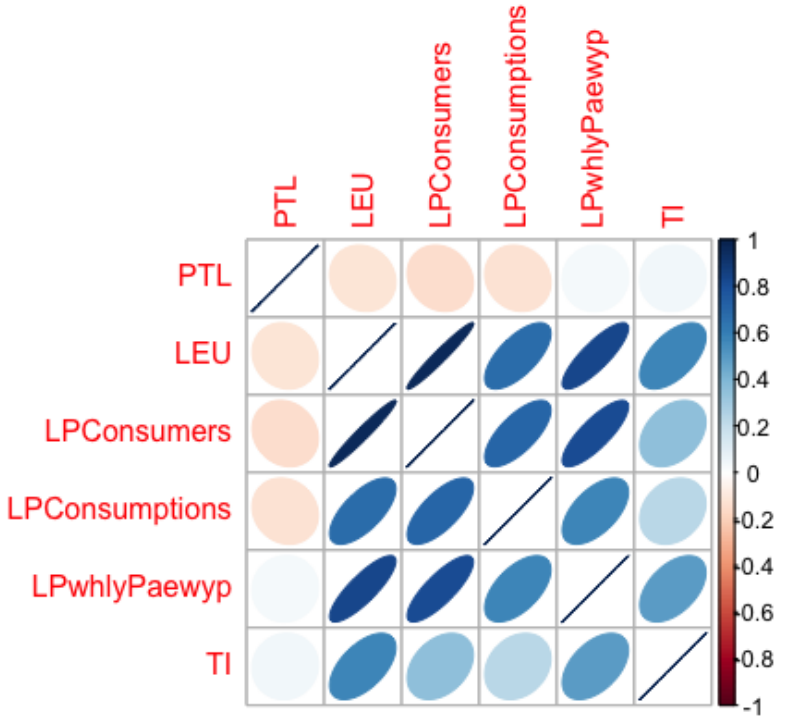
Interpretation of plots

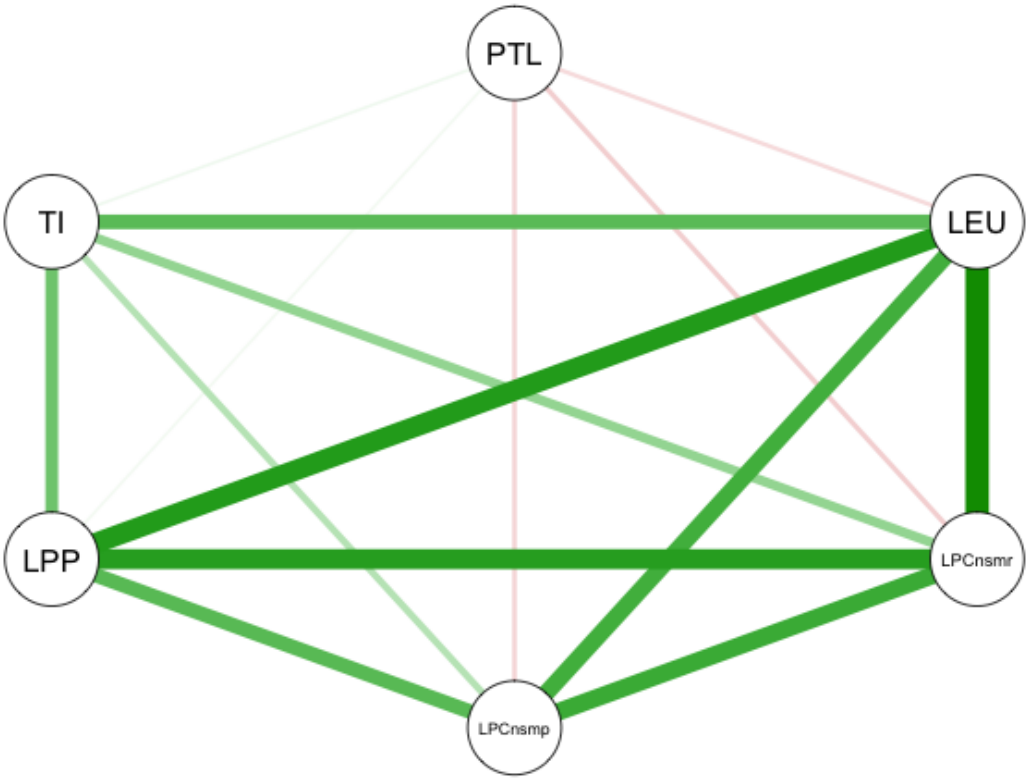
From the plots of the PCA related to type, we can find the similar conclusion in the interpretation of loadings. The biplot clearly shows how they correlated and how the directions of the components agree with each other. The variables pointing in the same direction means they are all essentially aligning on the first component expect for page total likes.

Figure 6 Plots of PCA related to TYPE


Biplot of PC1 and PC2 - Type
PTL
TI
LPwhlyPaewyn
LEU
LPConsumers
LPConsumptions
Comp.2
Comp.1


Comp.2 (17.94%)
Comp.1 (59.76%)
Type
Link
Photo
Status
Video


PTL
LEU
LPConsumers
LPConsumptions
LPwhlyPaewyp
TI
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

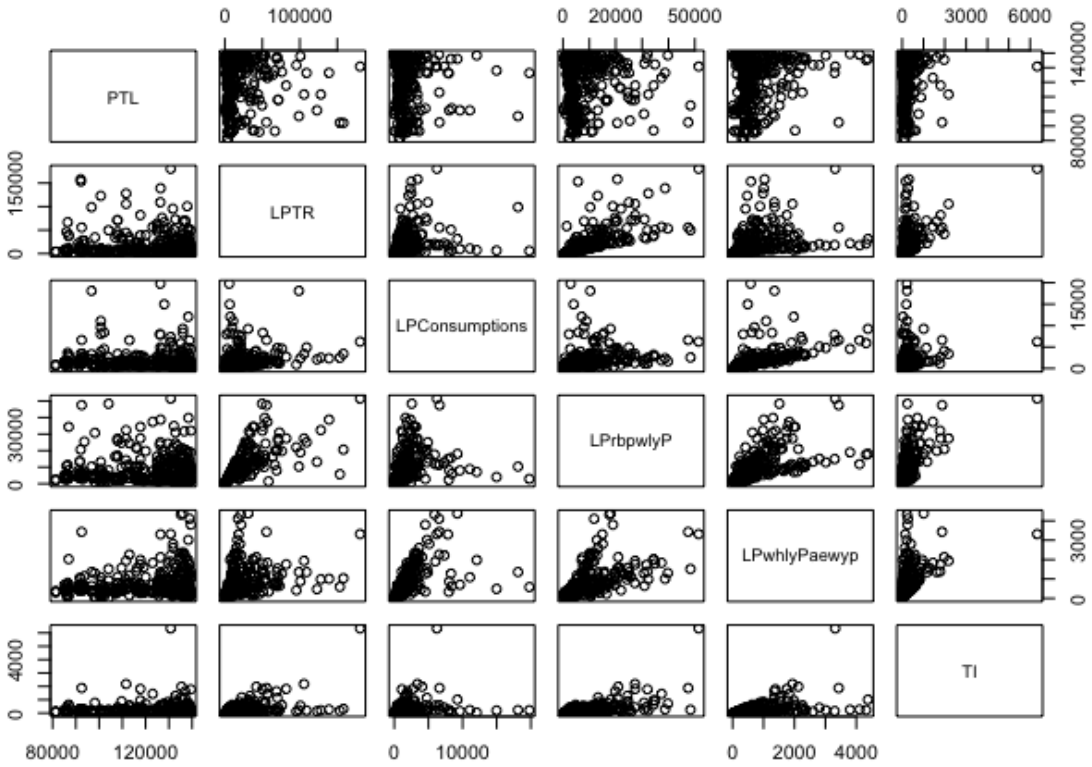

PTL
TI
LEU
LPP
LPCnsmr
LPCnsmp

Category as Response Variable

Multinomial Logistic Regression and Exploratory Analysis

The output of Anova of multinomial logistic regression indicates that page total likes, lifetime engaged users, lifetime post consumers, lifetime post consumptions, lifetime post impressions by people who have liked your page, lifetime post reach by people who like your page are the six covariables that are significant and should be contained in the new dataset to conduct PCA. The scatter plot of every two variables of this new dataset are shown in Figure 7, from which we can clearly see the collinearity may exist between every two of lifetime engaged users, lifetime post impressions by people who have liked your page, lifetime post reach by people who like your page. Furthermore, multinomial logistic regression was fitted again with the first and second principal component as the covariates to see the output, which undoubtfully shows the strong significance of principal component 1 and 2 in the model.

Figure 7 Scatter plot of dataset related to CATEGORY



Summary of PCA output

Figure 8 displays the summary of output of PCA. The first principal component has the largest standard deviation of about 1.73 and comprises 50.09% of the variance in the data cloud whilst the second and third principal component has standard deviation of 1.03 and 0.95 and proportion of 17.63% and 14.90%. And these three approximately contains 82.62% of the data cloud.

Figure 8 Output of PCA related to CATEGORY

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	1.7336675	1.0284424	0.9455783	0.6984533	0.62375133	0.40704092
Proportion of Variance	0.5009338	0.1762823	0.1490197	0.08130617	0.06484429	0.02761372
Cumulative Proportion	0.5009338	0.6772161	0.8262358	0.90754199	0.97238628	1

Number of components to retain

-Kaiser's Criterion

In light of the criterion of Kaiser, the component with eigenvalues that are higher than 1 should be retained when using a correlation matrix. We can see from Figure 3 that the first two components have eigenvalues of 3.01 and 1.06 that should be retained.

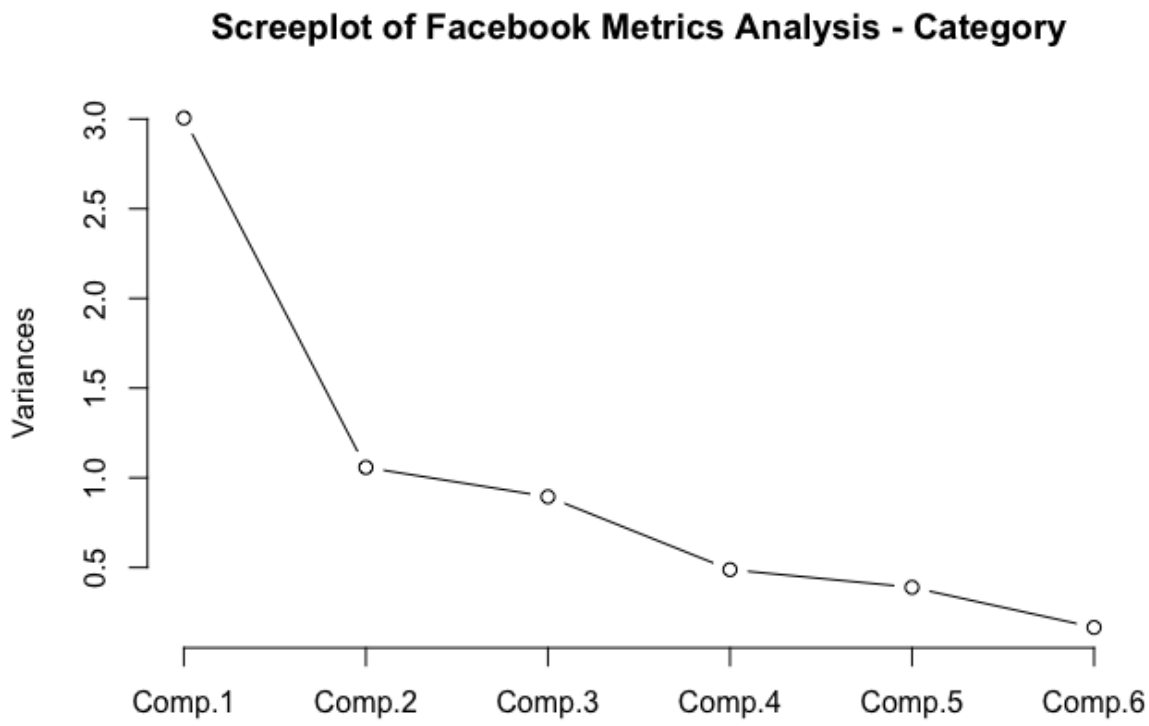
Figure 9 Eigenvalue of PCA related to CATEGORY

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Eigenvalue	3.0056029	1.0576937	0.8941184	0.487837	0.3890657	0.1656823

-Scree Plot

According to Figure 10 of Scree plot, it can also be inferred that the first and second components should be kept because the eigenvalues of them are greatly higher than others.

Figure 10 Scree plot of PCA related to CATEGORY



-Equilibrium Contribution and Mardia's Criterion

When conducting equilibrium and Mardia's Criterion, they suggest that 6 and 4 variables should be kept respectively.

Interpretation of loadings

The table below indicates the loadings of each variable in the principal components. Lifetime page total reach has the highest loading in PC1 with a score of -0.515 whereas lifetime people who have liked your page are engaged with your page of -0.463. It may seems reasonable to conclude that lifetime page total reach and lifetime people who have liked your page are engaged with your page contribute most to the variance. However, page total likes comprises the largest in the second component which may also be considered to be influential.

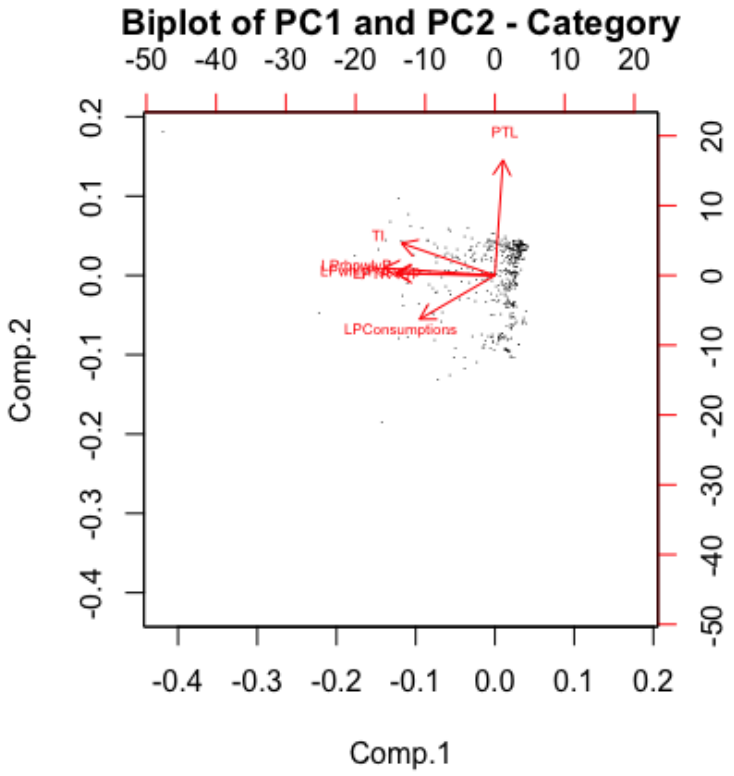
Figure 11 Loadings of PCA related to TYPE

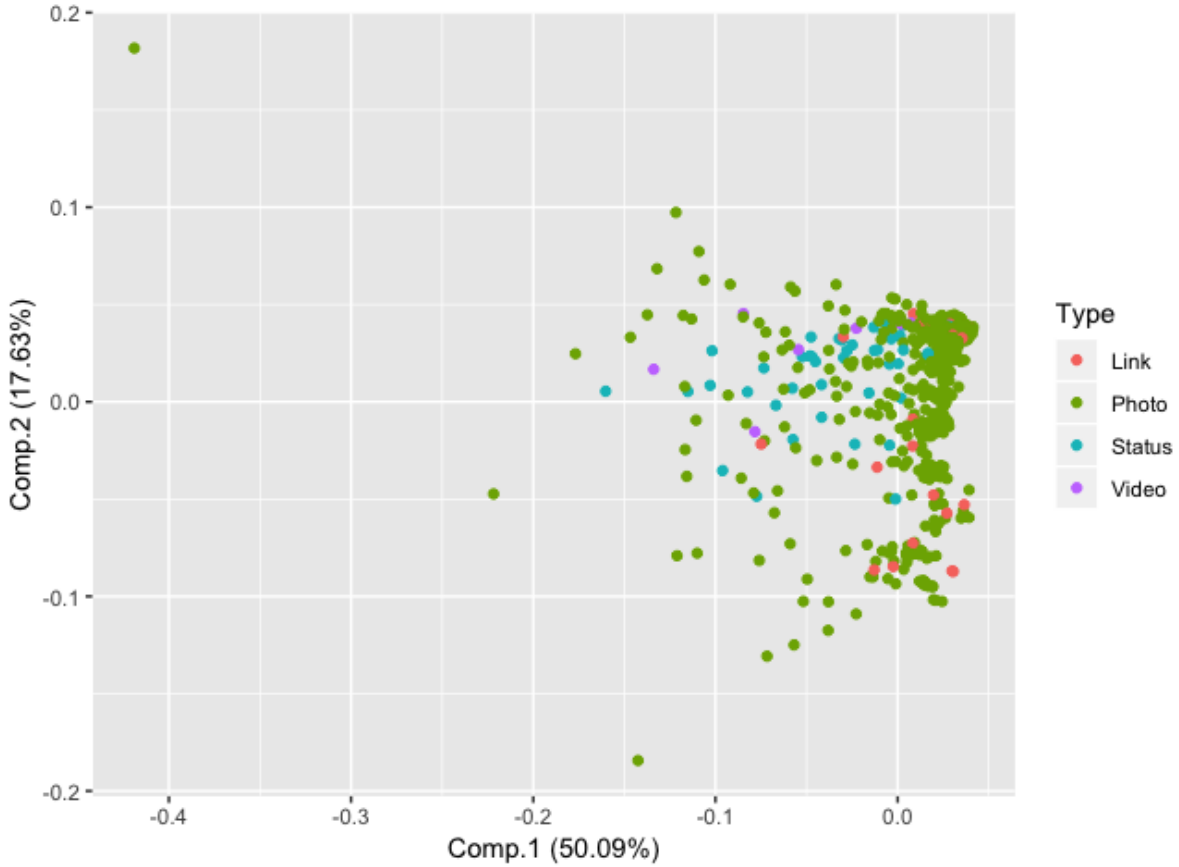
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
PTL		0.903	0.335	0.252		
LPTR	-0.457		-0.396	0.619	0.143	-0.481
LPConsumptions	-0.35	-0.343	0.656	0.232	0.476	0.223
LPrbpwlyP	-0.515		-0.221	0.12	-0.419	0.702
LPwhlyPaewyp	-0.463		0.422	-0.315	-0.533	-0.472
TI	-0.434	0.252	-0.271	-0.622	0.537	

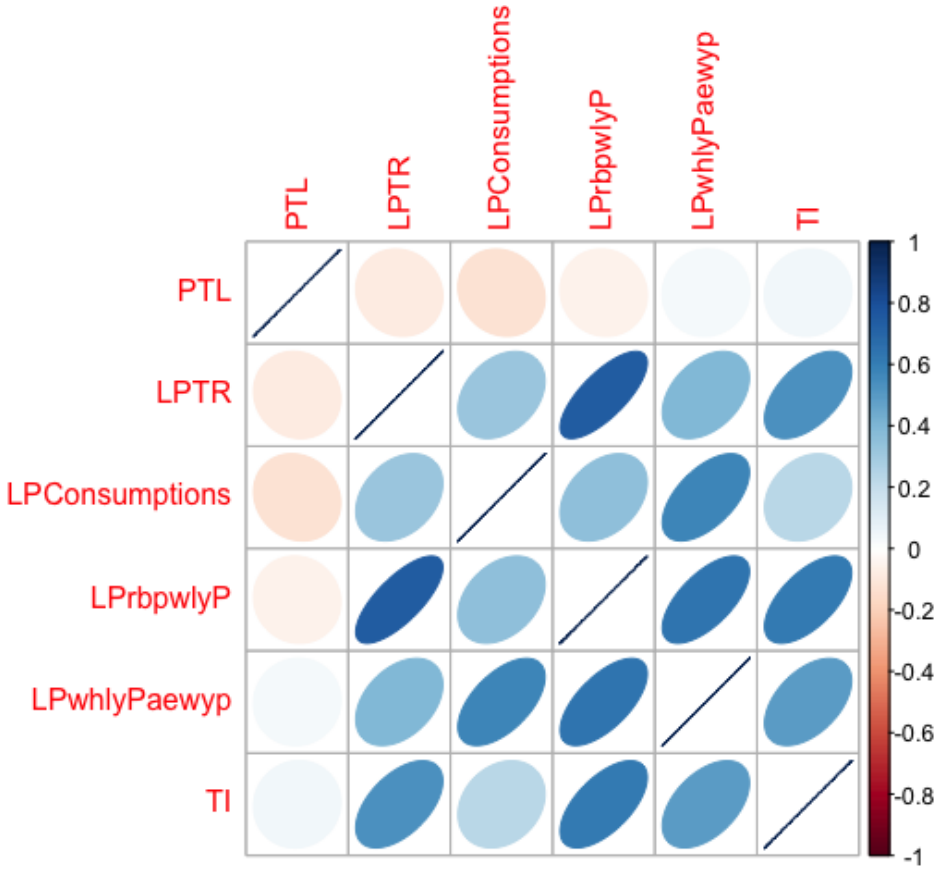
Interpretation of plots

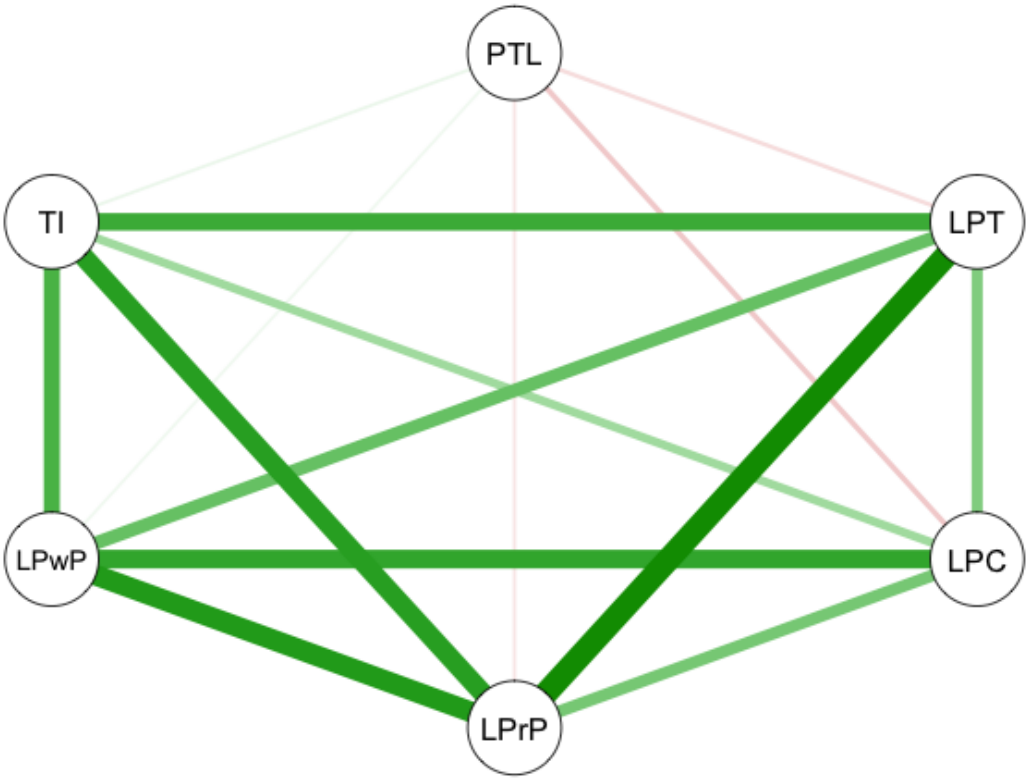
From the plots of the PCA related to category, we can find the similar conclusion in the interpretation of loadings. The biplot clearly shows how they correlated and how the directions of the components agree with each other. The variables pointing in the same direction means they are all essentially aligning on the first component expect for page total likes.

Figure 12 Plots of PCA related to CATEGORY




0.2
0.1
0.0
-0.1
-0.2
Comp.2 (17.63%)
-0.4
-0.3
-0.2
-0.1
0.0
Comp.1 (50.09%)
Type
Link
Photo
Status
Video


PTL
LPTR
LPConsumptions
LPrbpwlyP
LPwhlyPaewyp
TI
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1



Discussion

From the results of PCA related to type, we can see the a negative relationship between lifetime engaged users, lifetime post consumers, lifetime post consumptions, lifetime people who have liked your page and engaged with your post, total interactions and the first principal component while page total likes has a positive relationship with the second principal component. Similarly, the results of PCA related to category shows a positive relationship between lifetime engaged users, lifetime post consumers, lifetime post impressions by people who have liked your page, lifetime post reach by people who like your page and the first principal component while page total likes has a negative relationship with the second principal component.

References

Lariscy, R. W., Avery, E. J., Sweetser, K. D., & Howes, P. (2009) 'Monitoring public opinion in cyberspace: how corporate public relations is facing the challenge', *Public Relations Journal*, 3(4), pp. 1-17.

Moro, S., Rita, P. and Vala, B. (2016) 'Predicting social media performance metrics and evaluation of the impact on brand building: a data mining approach', *Journal of Business Research*, 69(9), pp. 3341-4451. *Science Direct* [Online]. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0148296316000813?via%3Dihub/ (Accessed: 2 February 2019).

R Core Team (2018) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/ (Accessed: 22 February 2019).