# SDA Project Report

# GROUP 7


**Group Members:**

**Jaswanth K - S2018001068**
**Ejurothu Pavan Sai santhosh - S2018001053**
**Sathyanarayanan R - S20180010154**
**Darshan G- S20180010046**

## Abstract:

The dataset given to us was Facebook performance metrics of a renowned cosmetic brand's Facebook page.We did a lot of Exploratory Data Analysis(EDA) including Box Plots,Pie Charts , Bar Charts, etc. to visualise the dataset more and gather information. Principal component Analysis was done to know how much reduction can be achieved. Regression was also done and R-Squared score was calculated and found out to be 0.6

## Introduction:

Exploratory Data Analysis (EDA) refers to the process of performing initial investigations on data so as to discover patterns,to spot anomalies and to check assumptions with the help of summary statistics and graphical representations.

Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

Regression is a statistical method that attempts to determine the strength and character of the relationship between one dependent variable and other independent variables

The given dataset had 19 features with 500 observations. The features were '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','comment', 'like', 'share', 'Total Interactions'

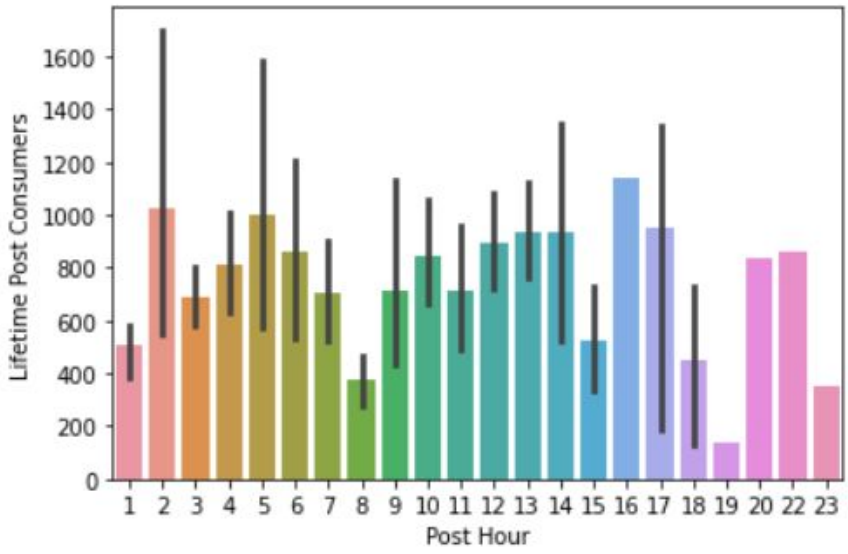
## Methodology:

## i) EDA

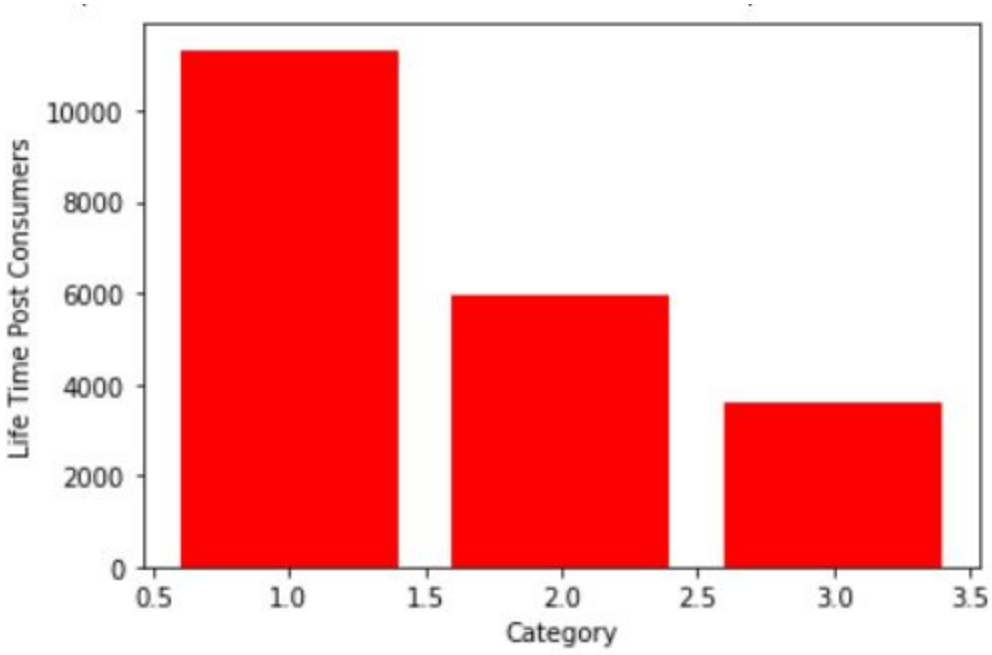
We used histograms, bar graphs, box plots, pie charts, line graphs, scatter plots to plot and analyse the data.

These are few examples of plots we used:

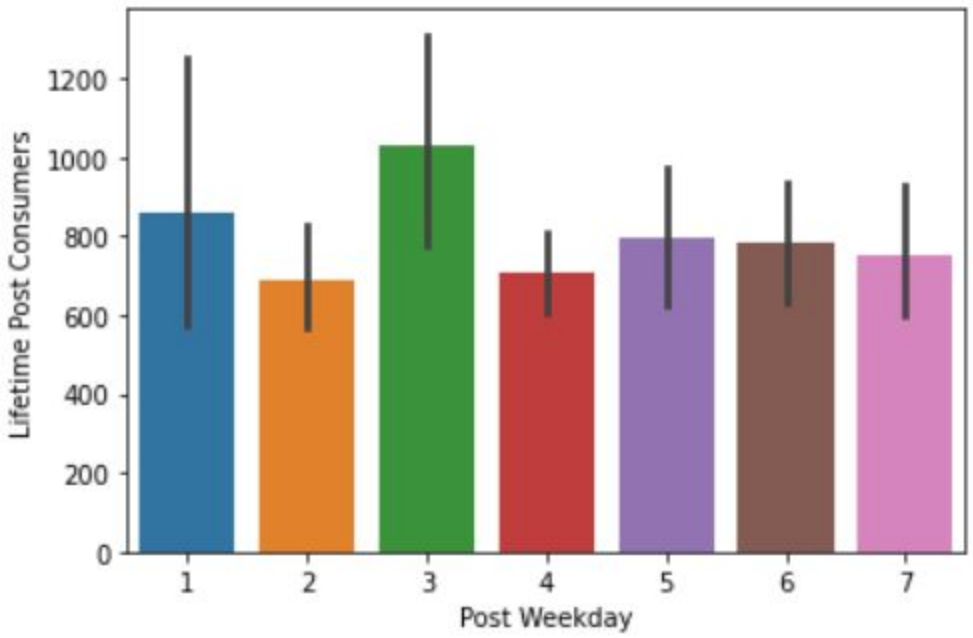
1) Influence of post-hour on lifetime post consumers



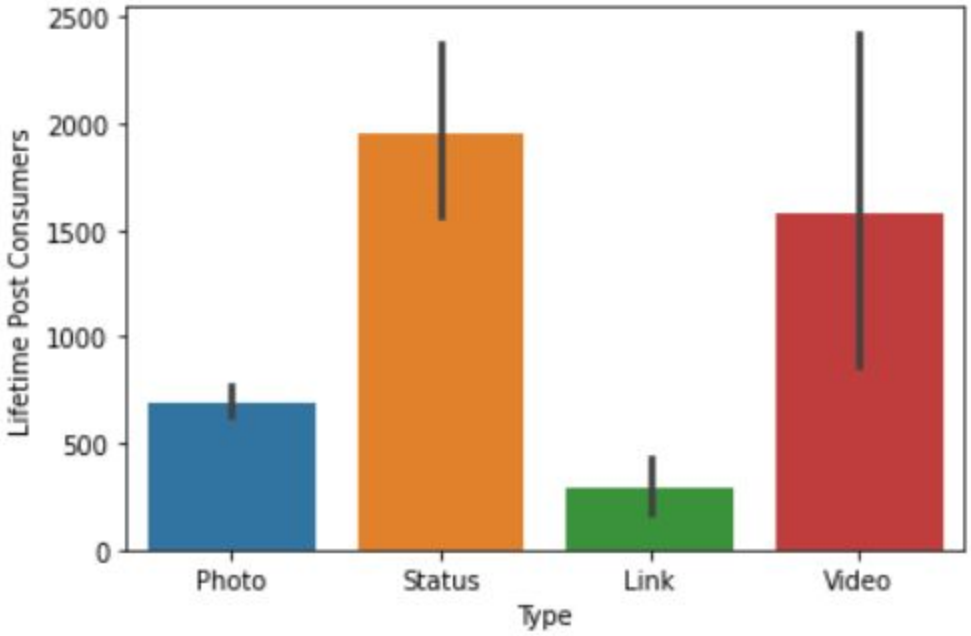

2)Influence of category on lifetime post consumers.



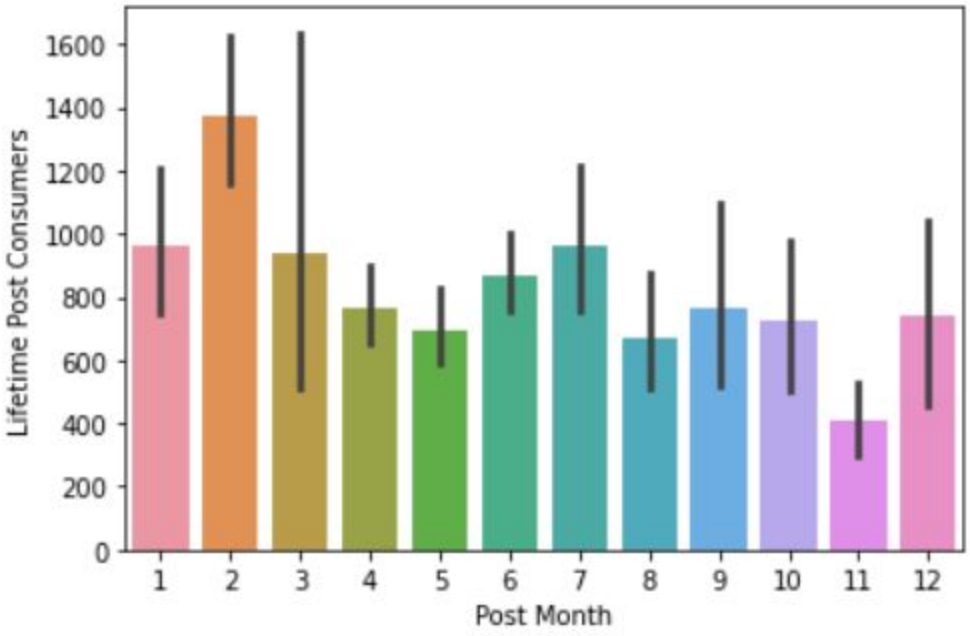
3)Influence of weekdays on consumers on posts



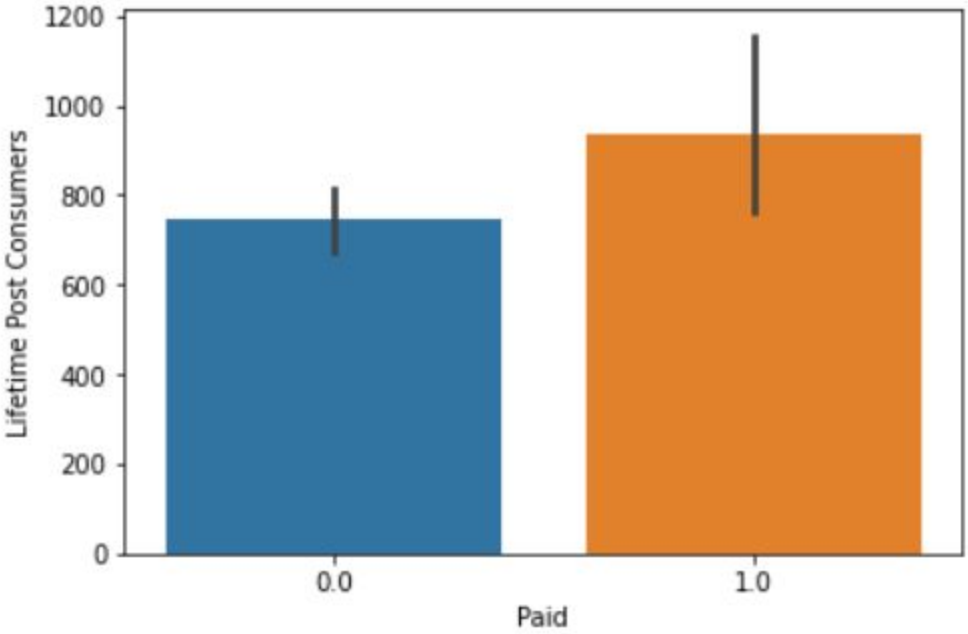

4)Influence of type of posts on consumers .



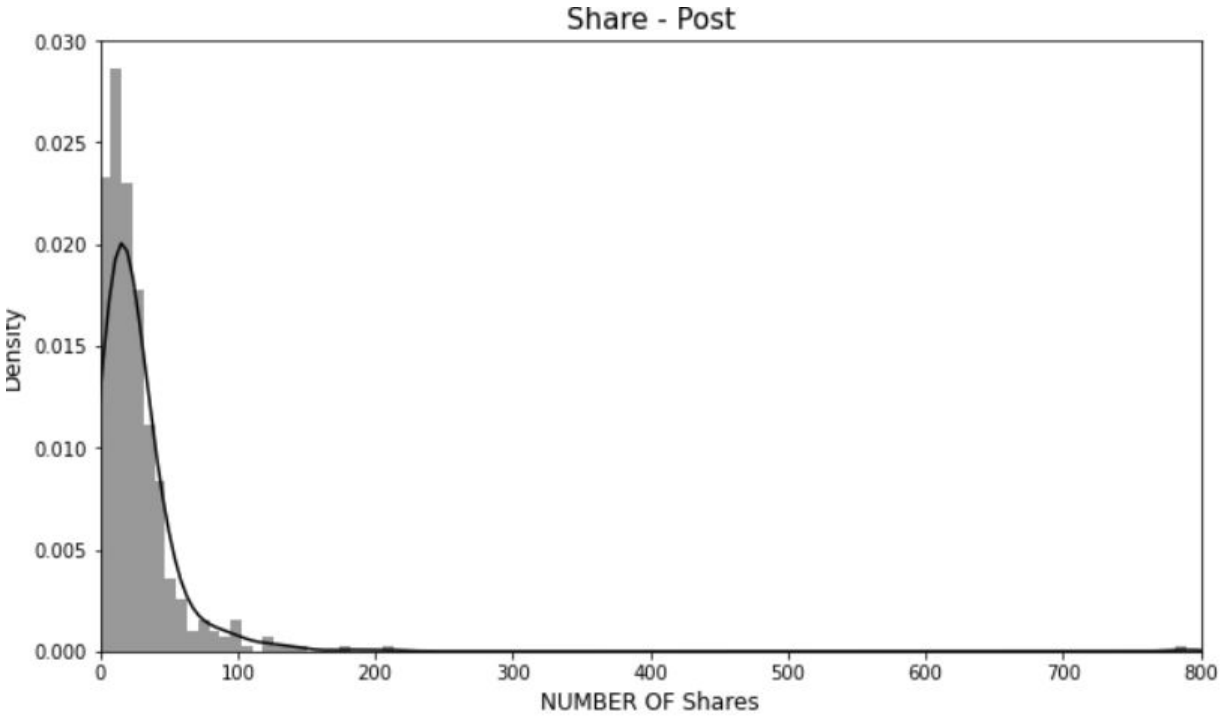
5)Influence of month on life time post consumers .



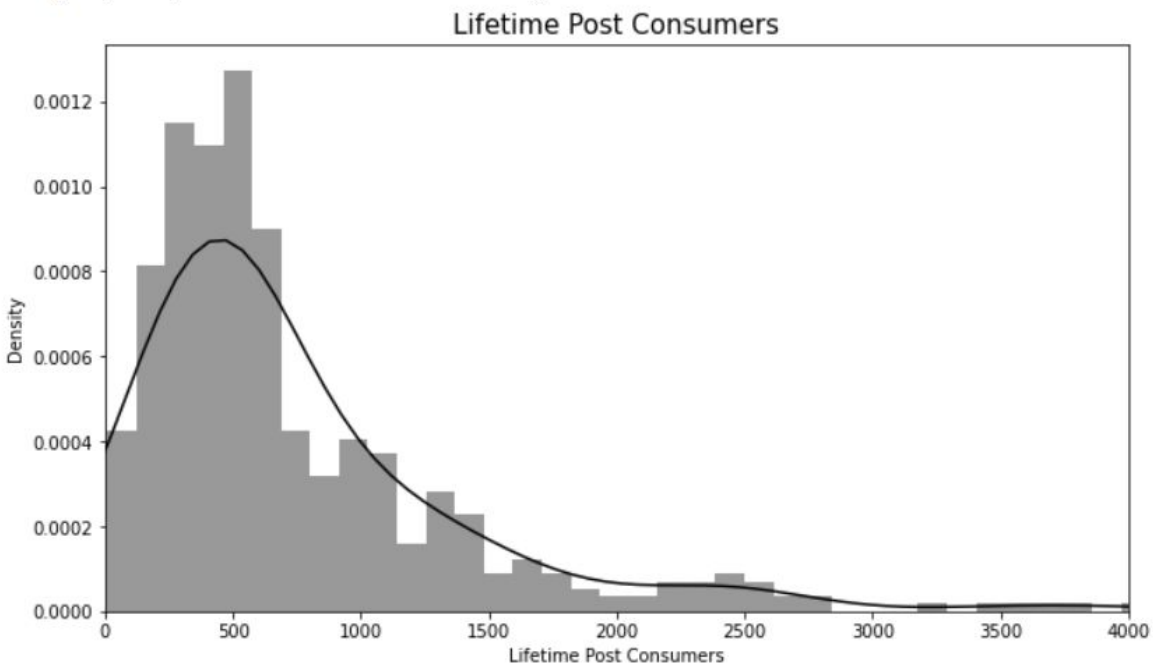

6)Influence of paid on lifetime post consumers .



7)

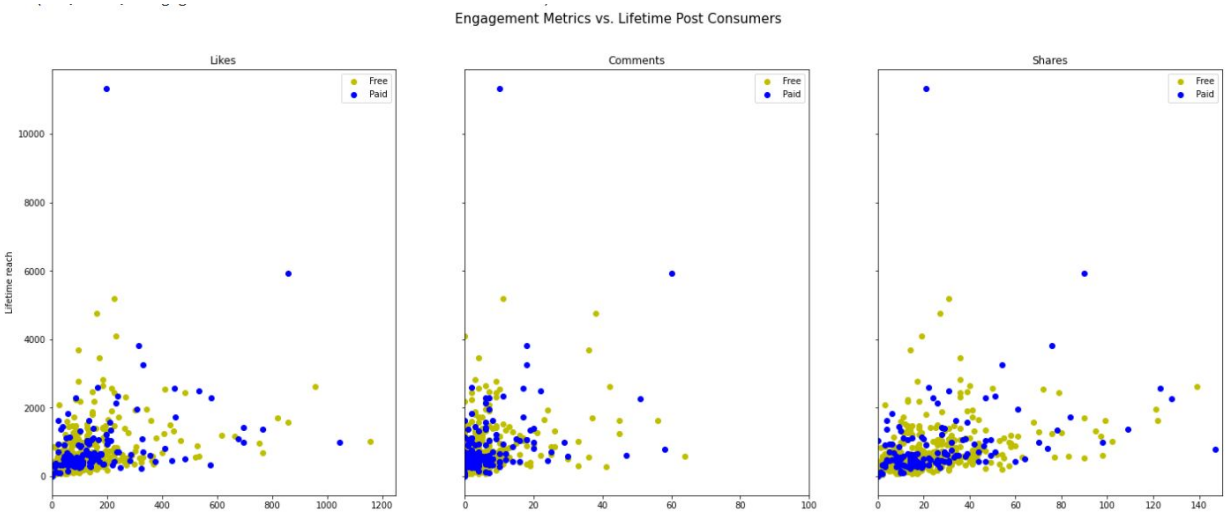

Share - Post
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
500
600
700
800
NUMBER OF Shares

8)


Lifetime Post Consumers
Density
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
500
1000
1500
2000
2500
3000
3500
4000
Lifetime Post Consumers

The distribution is left skewed with most posts around 500 users , with the maximum being around 12000.

9)




## ii) Regression:

Preprocessing for regression:

1. Removing null rows:
   5 rows that had missing information are removed.
   Number of rows = 495
2. Removing outliers:
   10% of the outliers are removed.
3. oneHot encoding for Categorical variables
   Variables like Weekday, Month, Hour, Type, Category are all categorical variables and are separated using oneHot encoding using Pandas library.

<u>Model 1: Lifetime engaged Users</u>

x contains the variables that are available at the time of posting.

```
Index(['Page total likes', 'Paid', 'Video', 'Status', 'Photo', 'Category_1',
       'Category_2', 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday',
       'Saturday', 'Sunday', 'hour_17', 'hour_1', 'hour_2', 'hour_3', 'hour_4',
       'hour_5', 'hour_6', 'hour_7', 'hour_8', 'hour_9', 'hour_10', 'hour_11',
       'hour_12', 'hour_13', 'hour_14', 'hour_15', 'hour_16', 'hour_18',
       'hour_19', 'hour_20', 'hour_22', 'hour_23', 'Month_1', 'Month_2',
       'Month_3', 'Month_4', 'Month_5', 'Month_6', 'Month_7', 'Month_8',
       'Month_9', 'Month_10', 'Month_11', 'Month_12'],
      dtype='object')
```

y contains the target variable ‘Lifetime Engaged Users’

Top 20 best features from x are taken using the k_best function, according to the importance of each.

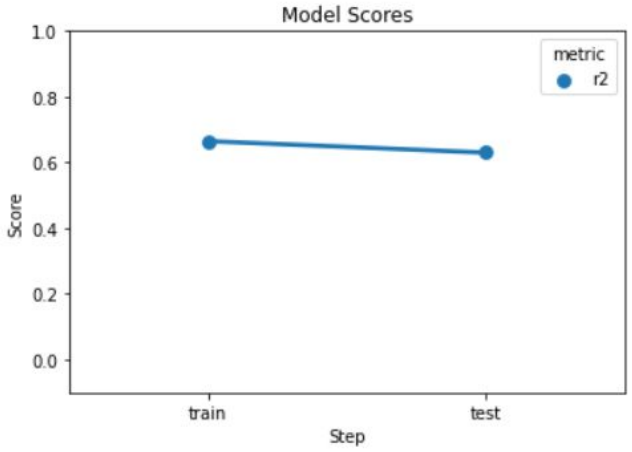
| | Variable | Importance |
|---|---|---|
| **0** | Page total likes | -0.005960 |
| **1** | Paid | 2.743453 |
| **2** | Video | 675.110241 |
| **3** | Status | 1339.185403 |
| **4** | Photo | 0.000000 |

The dataset is dividen into 90-10 ratio for train and test set. The dataset divide is chosen to be 10% trying to optimize the R-squared value.

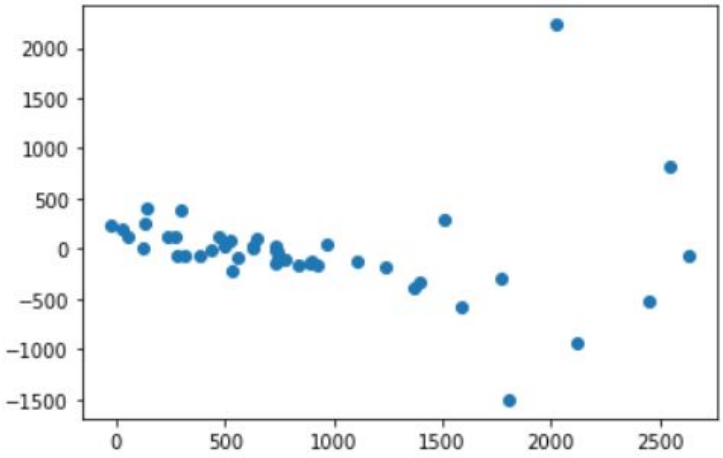
The linear regression gives a result of R-squared value 0.66 for the train and 0.63 for test

```
Train data R-2 score: 0.6646216263013041

Test data R-2 score: 0.6304036765060872
```




As we can see below, the error increases as the number of engaged users increases. This behaviour is against our assumptions of linear regression.




Final Model:
Total Engaged Users =
3045 + (-0.034) * (Page Total Likes) + 76 * (Paid) + 1115 * (Video) + 1745 * (Status) + 332 (Photo) ….

## Model 2: Features important for Page Likes:

Here we try to find what is more important for a page to grow, impressions, engagement, or Reach.

x contains the following variables.

```
Index(['Paid', 'Video', 'Status', 'Photo', 'Total Interactions',
       'Lifetime Post Total Reach', 'Lifetime Post Total Impressions',
       'Lifetime Engaged Users', 'Category_1', 'Category_2', 'Monday',
       'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday',
       'hour_17', 'hour_1', 'hour_2', 'hour_3', 'hour_4', 'hour_5', 'hour_6',
       'hour_7', 'hour_8', 'hour_9', 'hour_10', 'hour_11', 'hour_12',
       'hour_13', 'hour_14', 'hour_15', 'hour_16', 'hour_18', 'hour_19',
       'hour_20', 'hour_22', 'hour_23', 'Month_1', 'Month_2', 'Month_3',
       'Month_4', 'Month_5', 'Month_6', 'Month_7', 'Month_8', 'Month_9',
       'Month_10', 'Month_11', 'Month_12'],
      dtype='object')
```

Y contains the target variable ‘page total likes’
The dataset is divided into 90-10 ratio for train and test set.

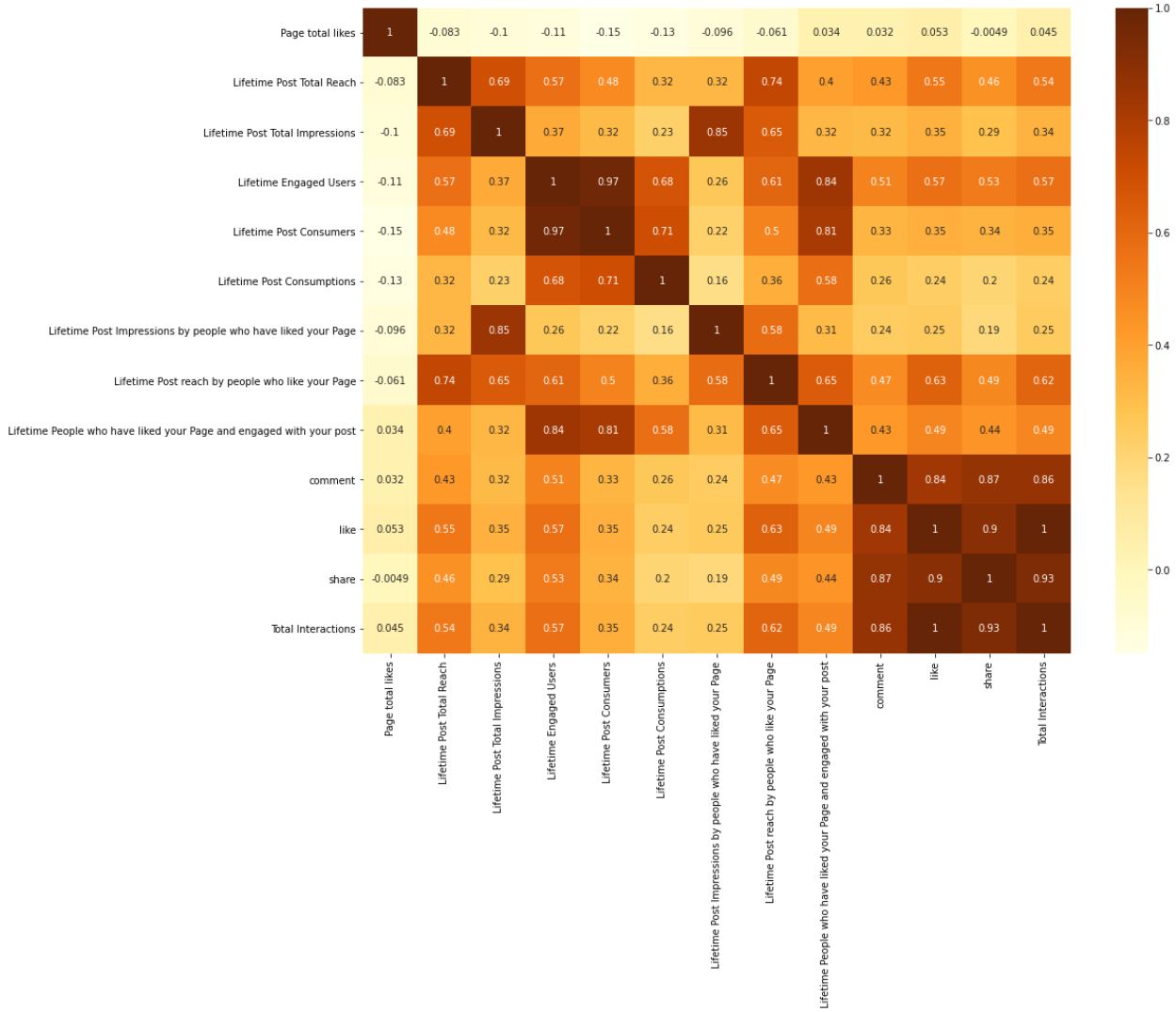
The model was not able to give much information about which parameter is responsible for page likes.
But we can relatively conclude that,

Post Interaction > Post Reach > Post Impressions > Post Engagement

Having post reach higher than engagement makes sense as higher reach would mean chances of more new users finding/liking the page. While, engagement could be the existing community interacting.

## iii) PCA:

The below plot shows correlation of features with each other




Only 13 features were considered omitting categorical features

Few observations:

i) Comments ,like , share and total impressions are completely correlated.
ii) Lifetime post total impressions in general and Lifetime post total impressions by users who have liked the page are correlated
iii) Lifetime engaged users and Lifetime post consumers are correlated

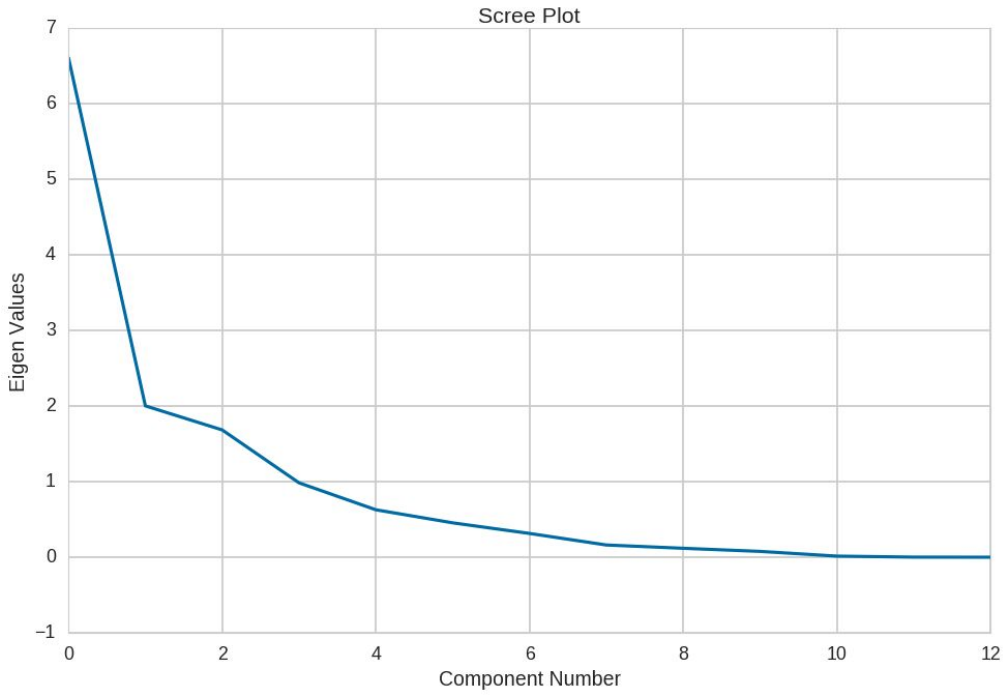
So the data in general has lots of correlated features

So PCA was applied and Dataset was also normalized

Covariance matrix was calculated and Eigen Values and Eigen vectors were found out

| Eigen Values | Eigen Vectors |
|---|---|
| 6.60 | 0.03,0.21,0.07,0.95,0.06,0.16,0.08,0.05,-0.08,-0.08,-0.01,0.00,0.00, |
| 2.00 | -0.29,-0.07,-0.22,-0.03,0.75,0.15,0.11,-0.11,0.11,0.03,0.50,0.01,-0.00, |
| 1.68 | -0.24,-0.17,-0.55,0.05,0.00,0.21,0.29,0.01,0.33,0.13,-0.60,-0.01,0.00, |
| 0.98 | -0.33,-0.23,0.26,0.01,0.02,-0.11,0.30,0.20,-0.28,0.06,-0.06,-0.73,0.05, |
| 0.63 | -0.28,-0.37,0.29,0.02,0.00,-0.11,0.39,0.09,-0.28,-0.09,-0.08,0.65,-0.04, |
| 0.45 | -0.21,-0.36,0.28,-0.01,-0.12,0.72,-0.44,0.06,0.10,-0.06,0.01,-0.01,-0.00, |
| 0.31 | -0.19,-0.17,-0.55,0.09,-0.54,-0.00,0.02,0.15,-0.19,-0.05,0.52,0.01,-0.00, |
| 0.16 | -0.32,-0.09,-0.23,0.09,0.20,-0.33,-0.58,-0.18,-0.36,-0.31,-0.29,-0.00,-0.00, |
| 0.12 | -0.30,-0.23,0.22,0.22,-0.19,-0.44,-0.12,-0.29,0.61,0.18,0.15,-0.01,0.00, |
| 0.08 | -0.30,0.35,0.07,-0.09,-0.21,0.24,0.17,-0.74,-0.27,0.17,-0.00,-0.02,-0.04, |
| 0.01 | -0.33,0.35,0.04,-0.04,0.00,-0.04,-0.17,0.38,-0.02,0.40,-0.03,0.09,-0.65, |
| 0.00 | -0.30,0.38,0.09,-0.14,-0.10,0.03,0.17,0.13,0.30,-0.76,0.02,-0.05,-0.09, |
| 0.00 | -0.33,0.36,0.05,-0.05,-0.02,-0.02,-0.12,0.30,0.01,0.26,-0.02,0.16,0.75, |

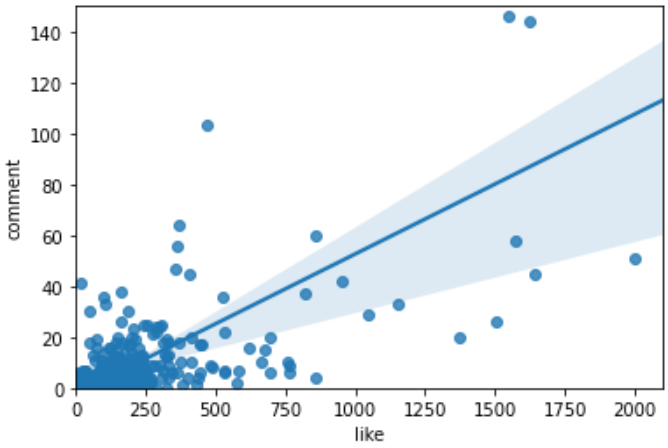
With the help of Scree Plot it can be shown that the given data can be reduced to 7 features with 97.17% of data being retained

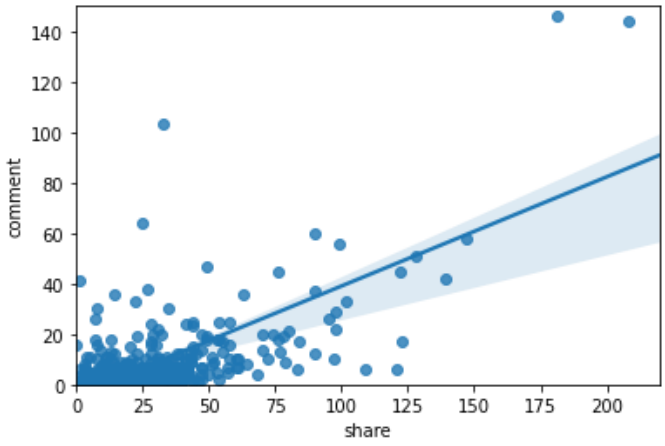


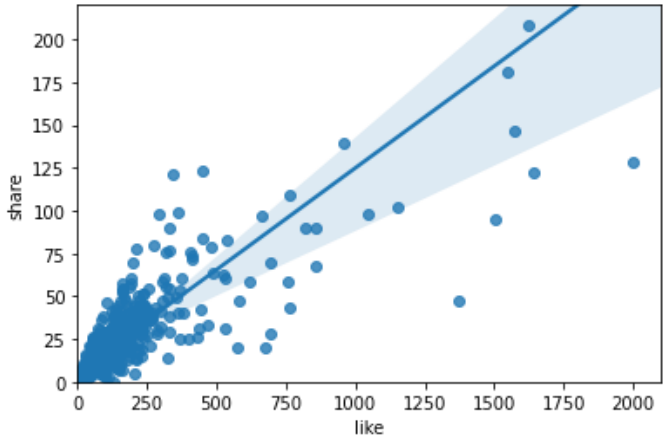
# Results:

## EDA:

### Relationship between like-share-comment:



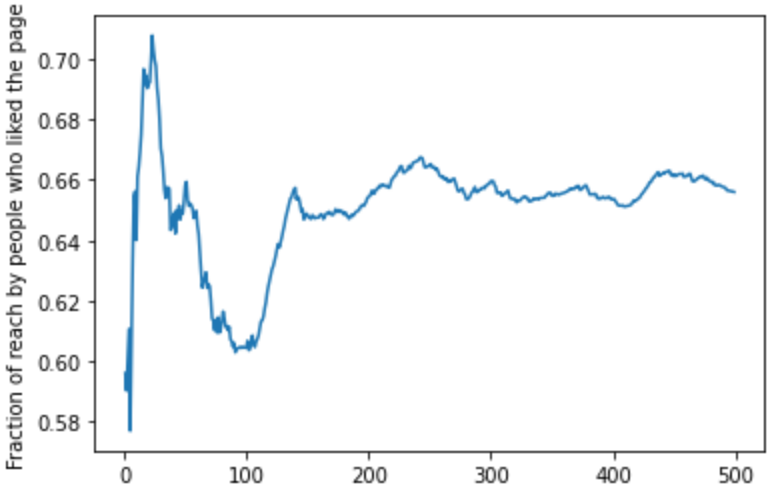



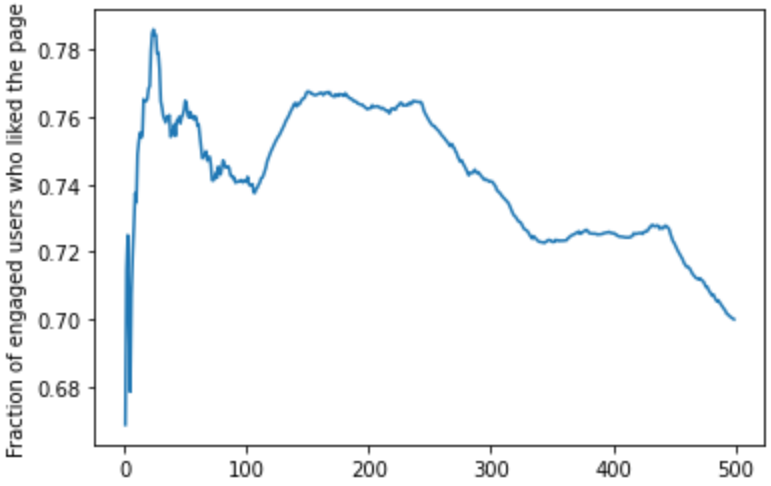

We can clearly see that **comment- share** and **comment- like** plots are more scattered than **like- share**. Like-share is more correlated with each other compared to comment-share and comment- like.

## Fraction of 'Reach', 'Impression' and 'Engagement' coming from people who liked the page:

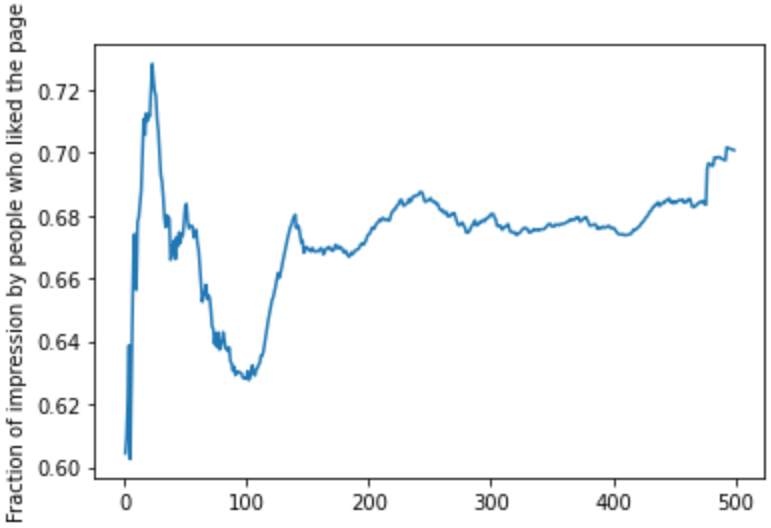
The fraction is calculated by: data['Lifetime Post reach by people who like your Page']/data['Lifetime Post Total Reach']




On average 65% of reach are from people who liked the page.

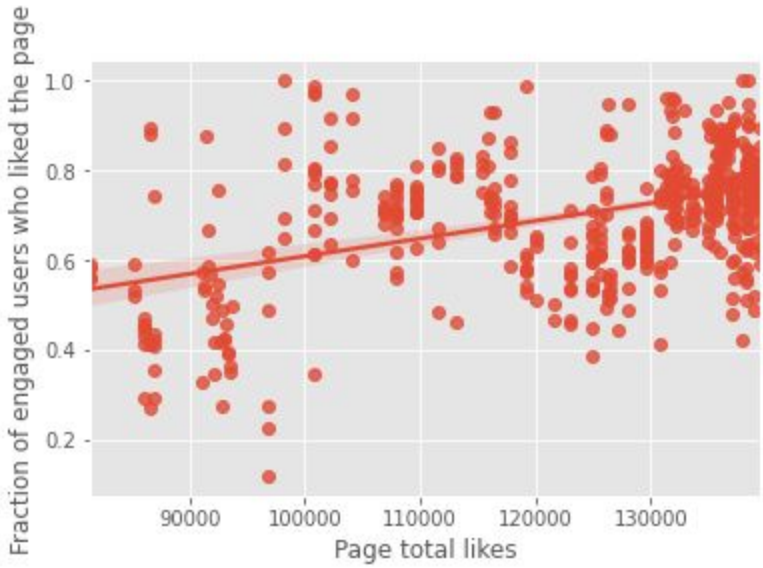



On average 70% of engaged users are who liked the page.

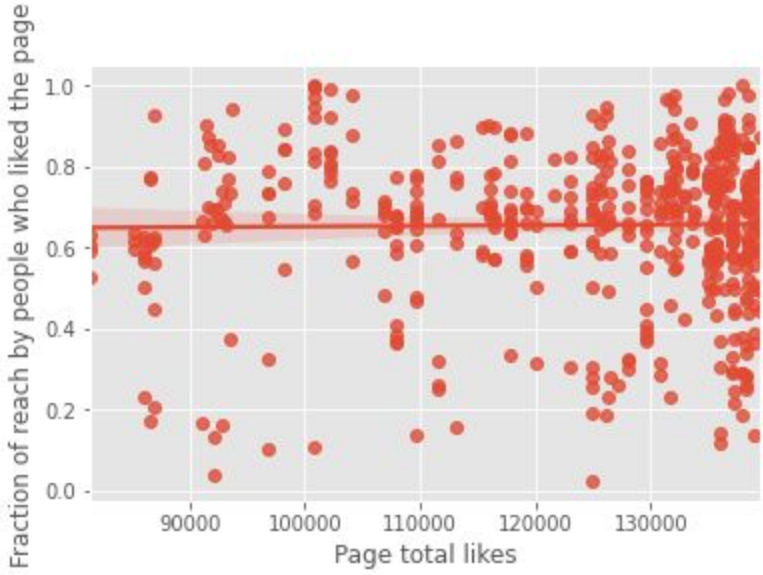



On average 70% of impressions are from people who liked the page.

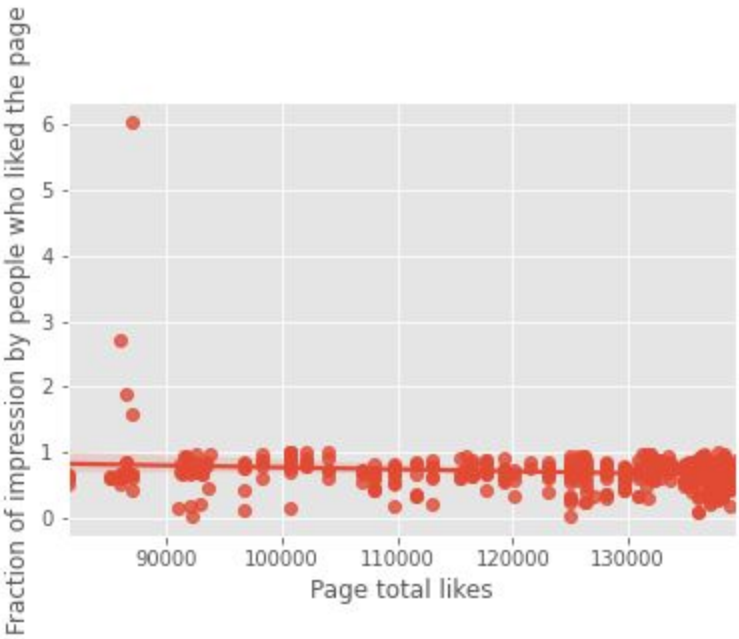
**How does this fraction change for different pages?**



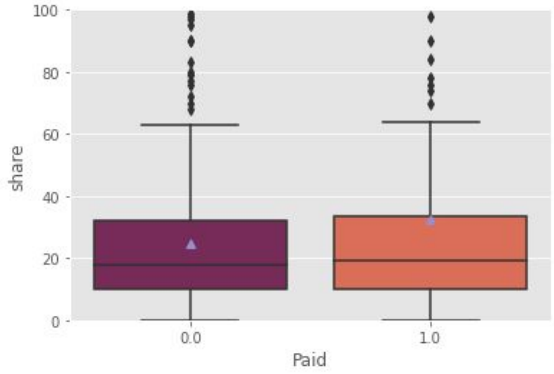

Fraction of engaged users who liked the page **increases with increasing page-likes**.
We can see that pages with likes **more than 130000 have around 80%** of engaged users from people who liked the page whereas the fraction is only **around 55% for pages having likes less than 90000**.

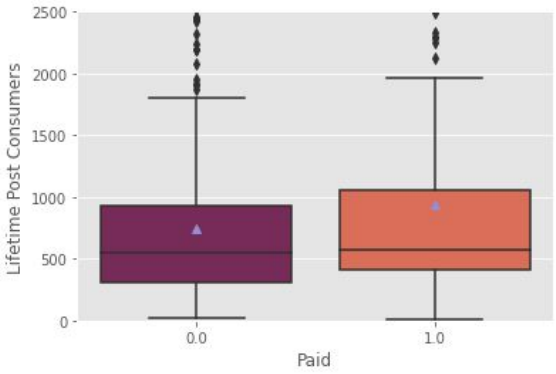


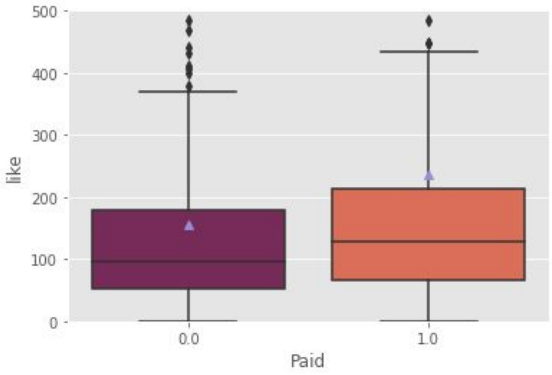

Fraction of reach from users who liked the page **does not change much** with increasing page-likes.

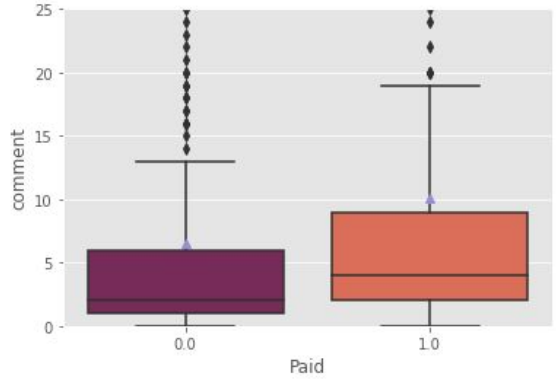

Fraction of impression by people who liked the page
6
5
4
3
2
1
0
90000
100000
110000
120000
130000
Page total likes

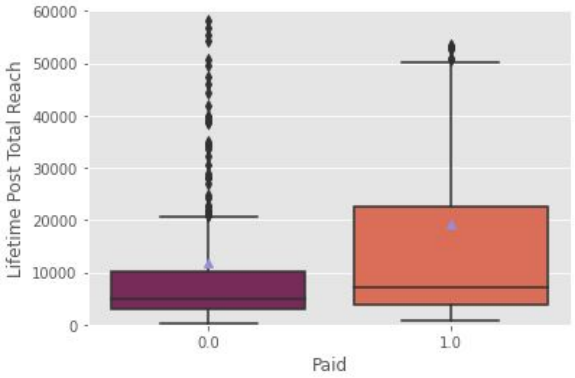
## Paid vs Non- paid: Which attributes of post are most and least influenced by ‘Paid’?

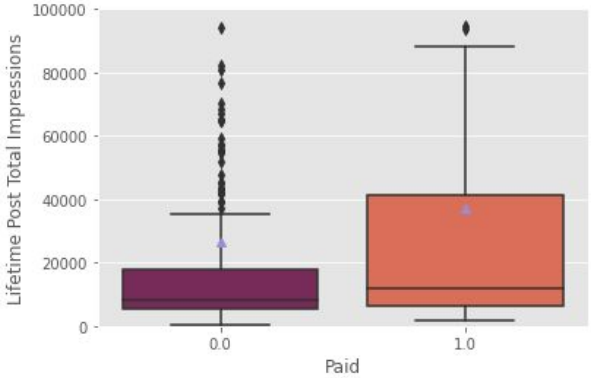


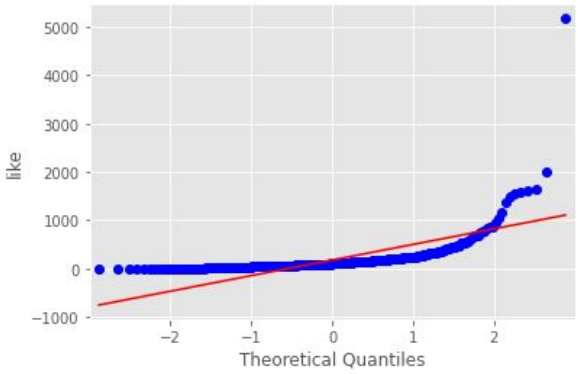



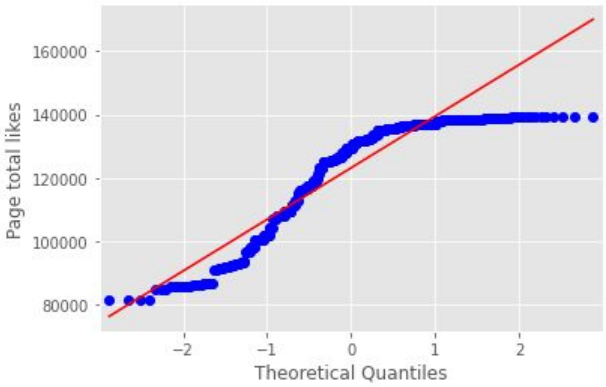



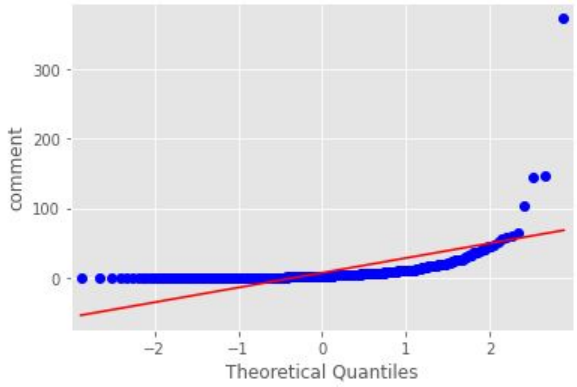



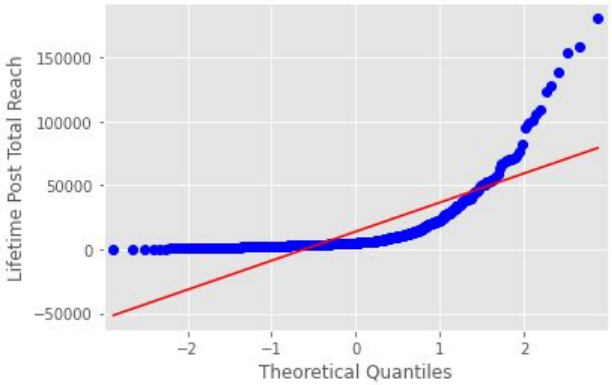

We can see that **share**, **Lifetime post consumers** are **least influenced** by Paid and **Total reach** and **Total Impressions** are **most influenced** by Paid.

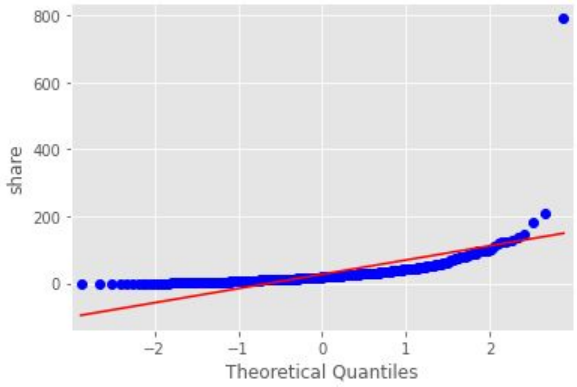
## QQ plots:



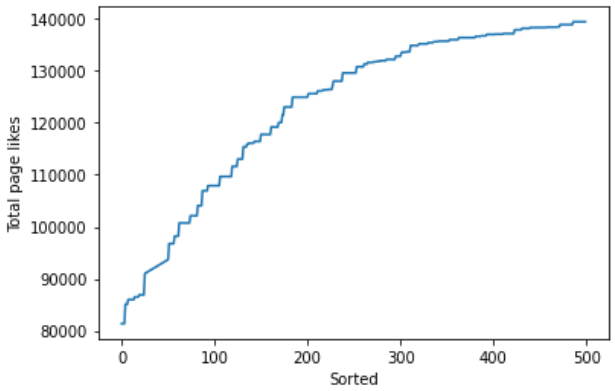



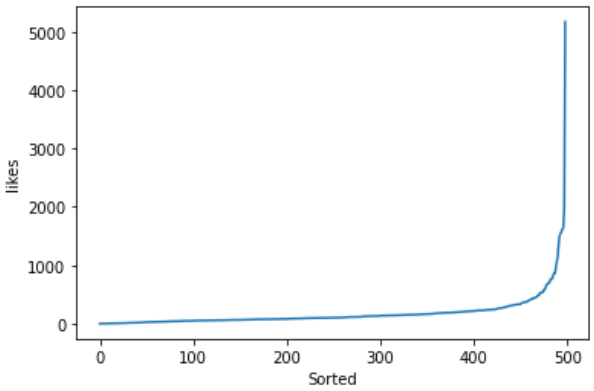



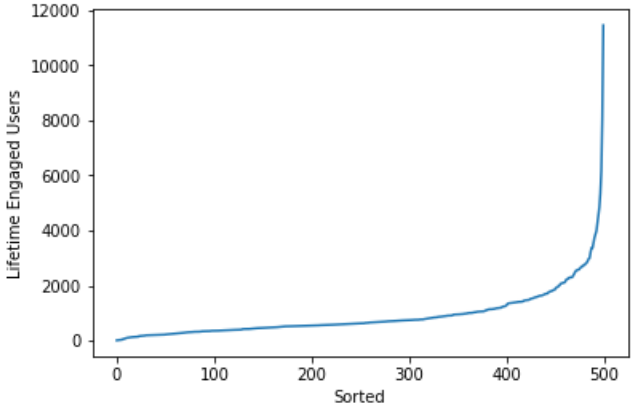

None of the variables are normally distributed.

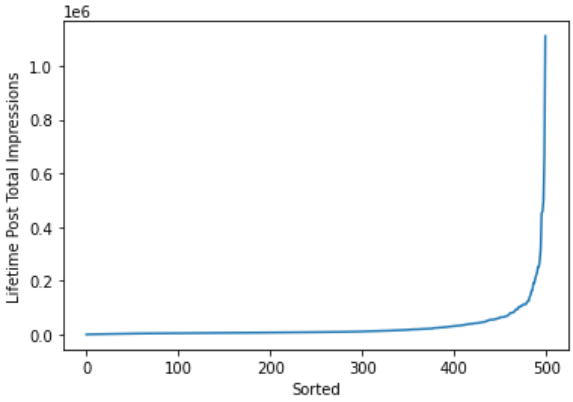
## Values of various attributes were sorted and plotted for 500 Posts:

Sorted low to high,



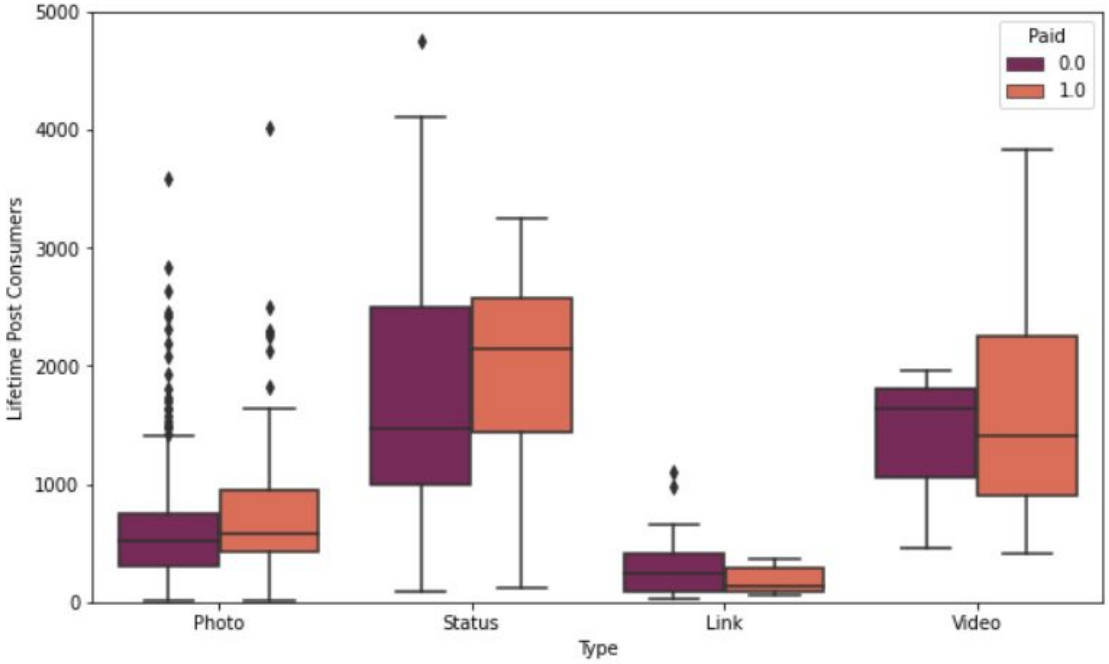



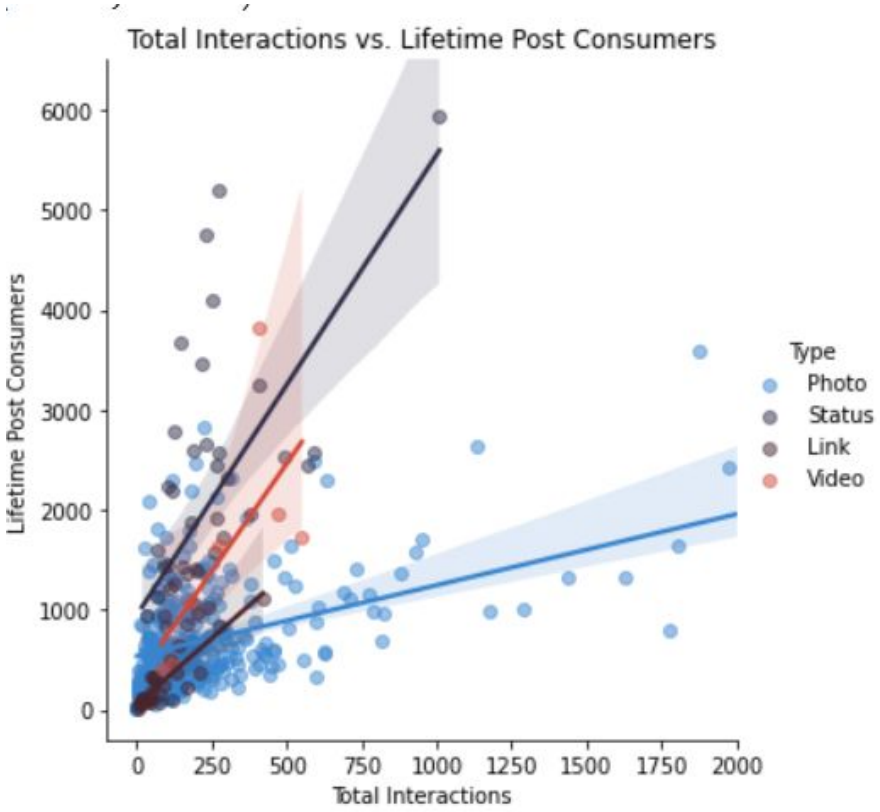




We can see that the slope of Total page likes decreases with up the order.
For every other variable/attribute slope increases steeply at higher ranks.

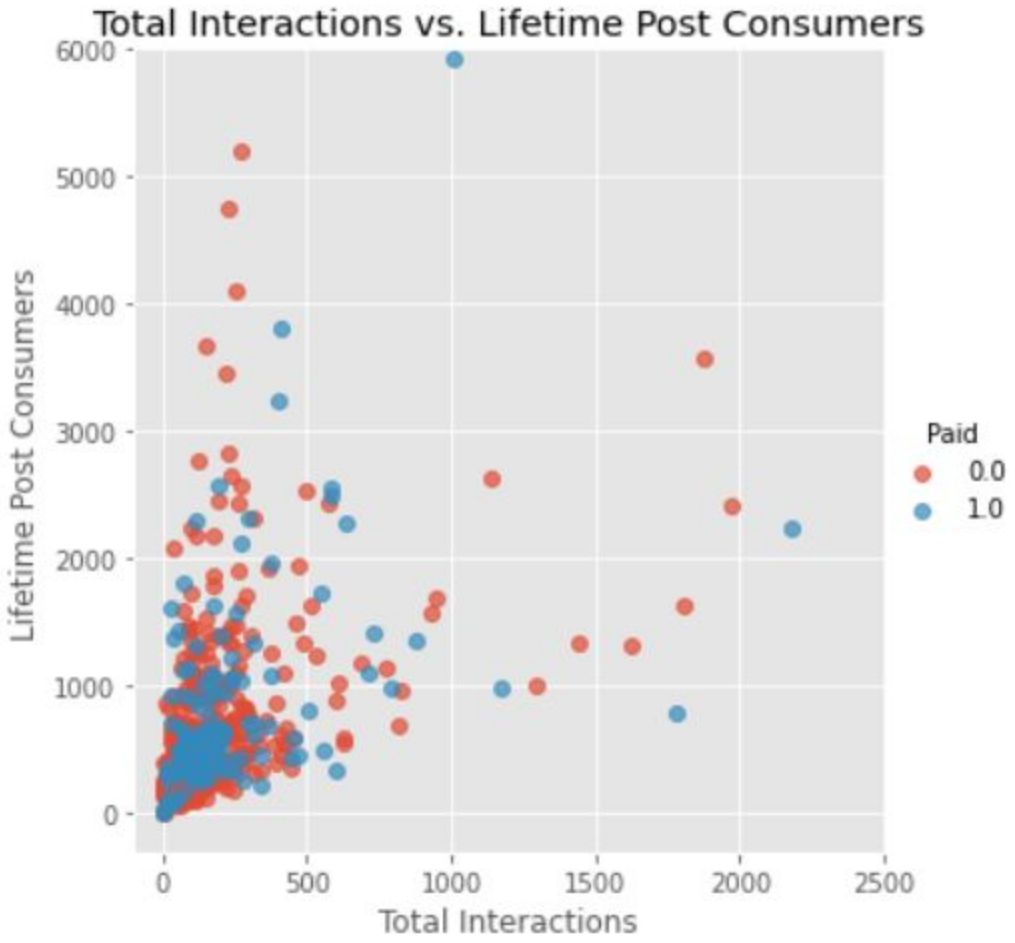
10)Boxplot of lifetime post consumers on the basis of type group by paid or not paid .



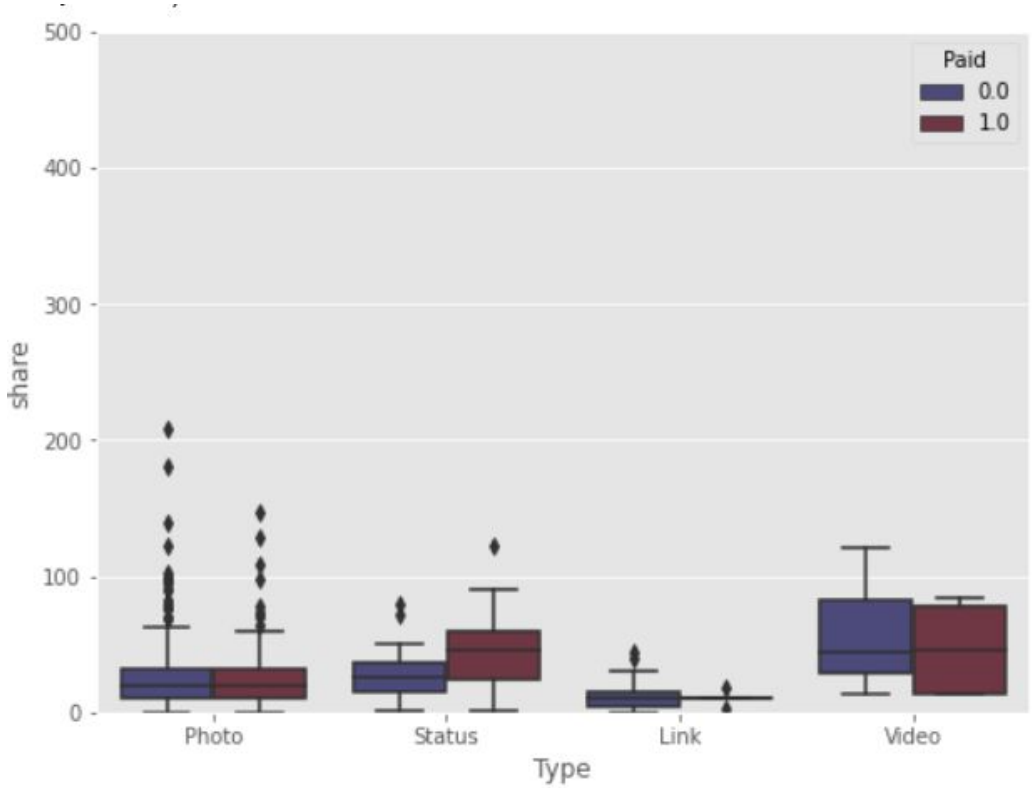

11)



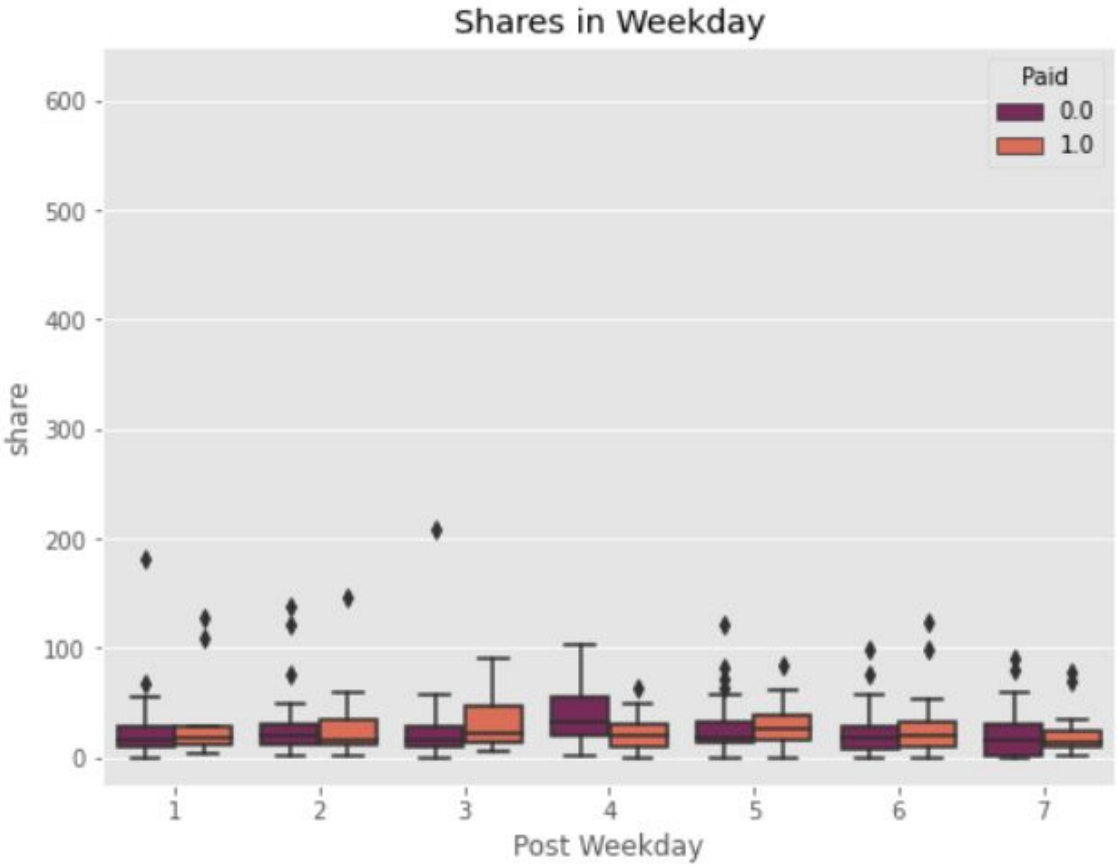
12)



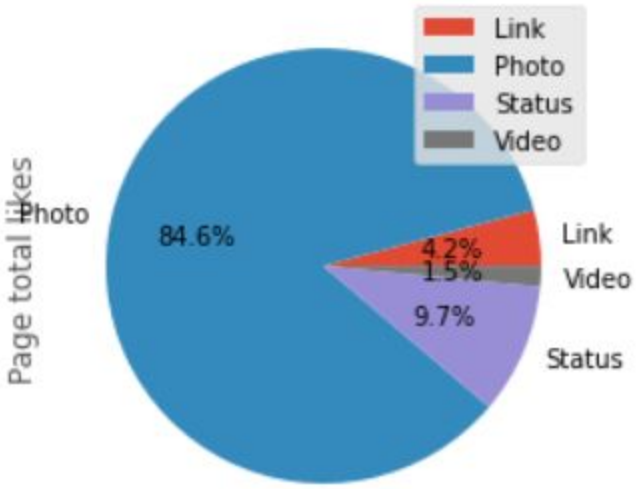

13)Shares on the basis of type


Paid
0.0
1.0
500
400
300
200
100
0
share
Photo
Status
Link
Video
Type

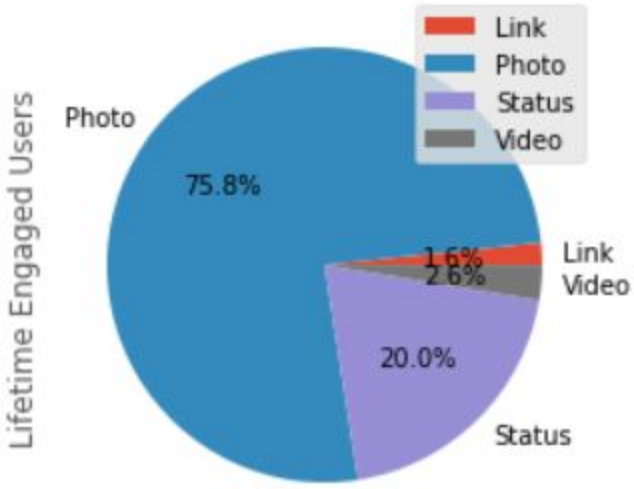
14)



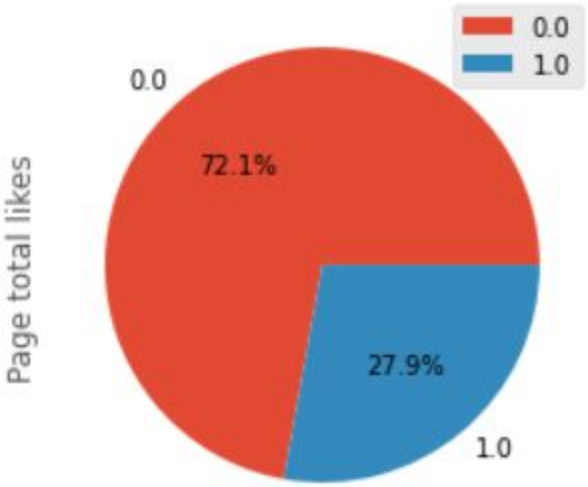

15)observation of which type has more page total likes



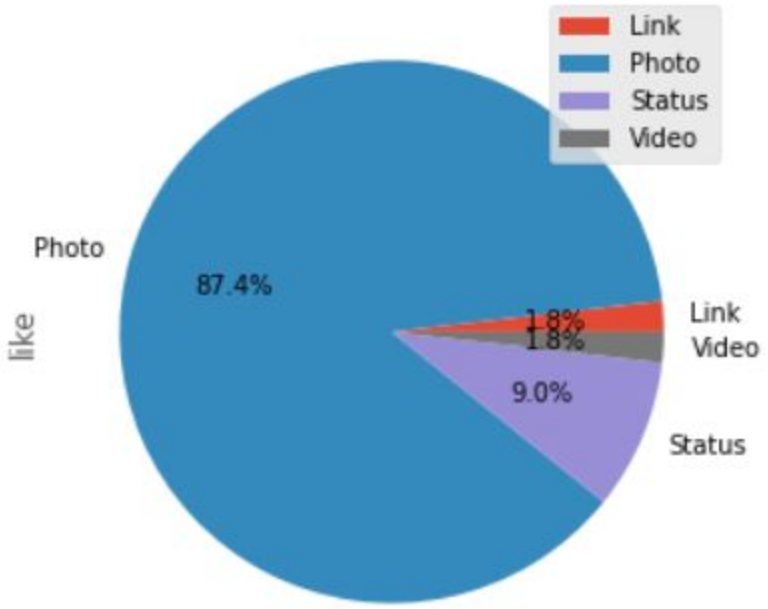

16)Pie chart of influence of type on lifetime engaged Users

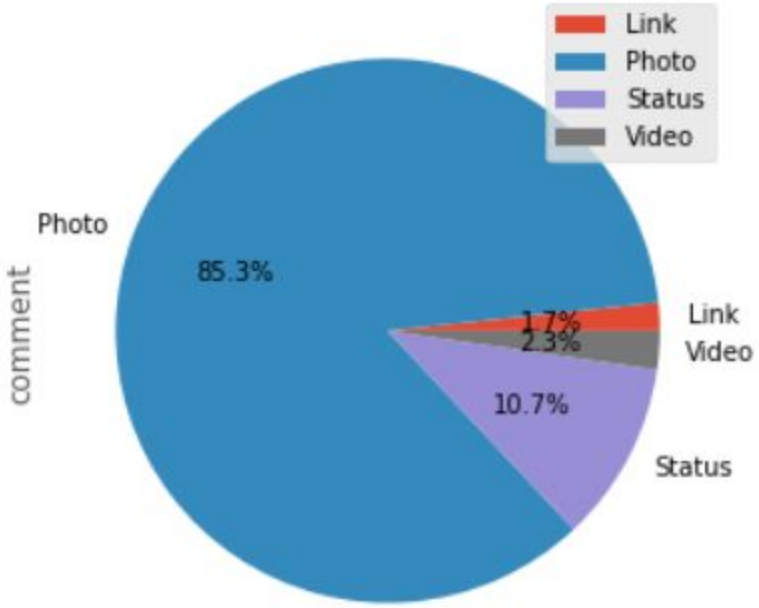


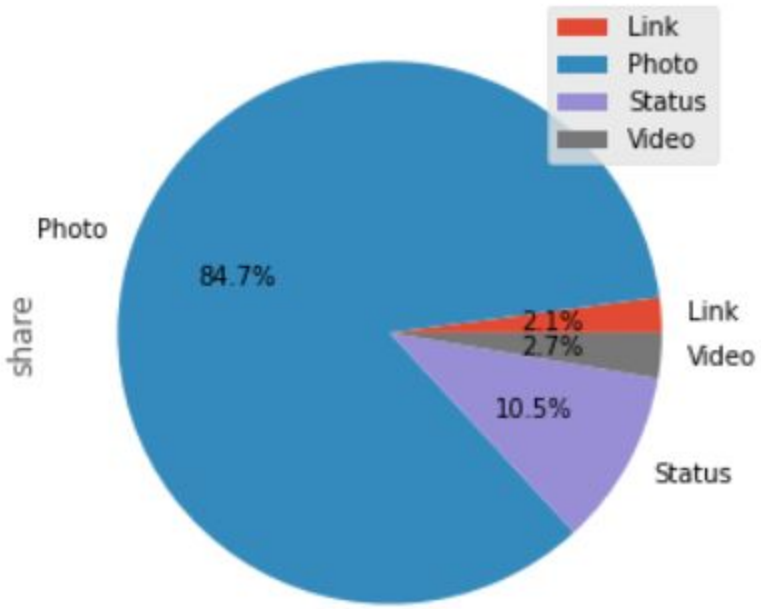

17)Pie chart of page total likes on the basis of paid or not paid.



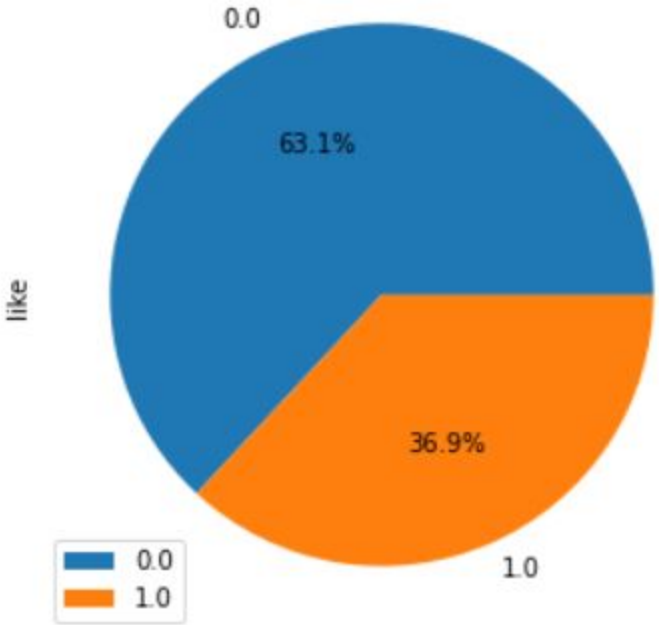

18)Pie chart of likes,shares,comments on the basis of type(photo,link,status,video)

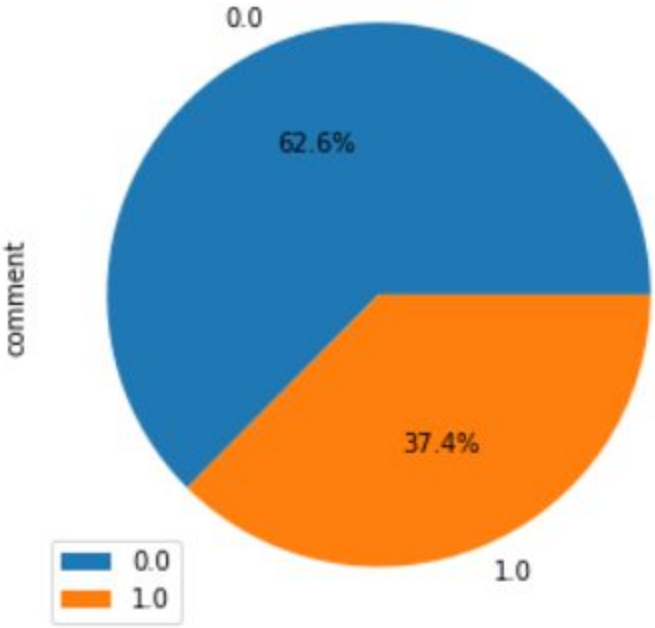

Link
Photo
Status
Video
Photo
87.4%
like
1.8%
1.8%
Link
Video
9.0%
Status

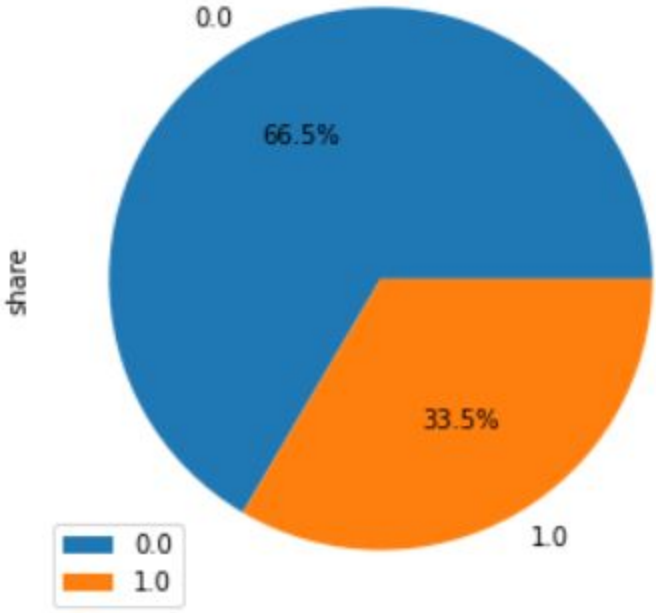

Link
Photo
Status
Video
Photo
85.3%
comment
1.7%
2.3%
Link
Video
10.7%
Status




19)Pie chart of likes,shares,comments on the basis of Paid or not paid.

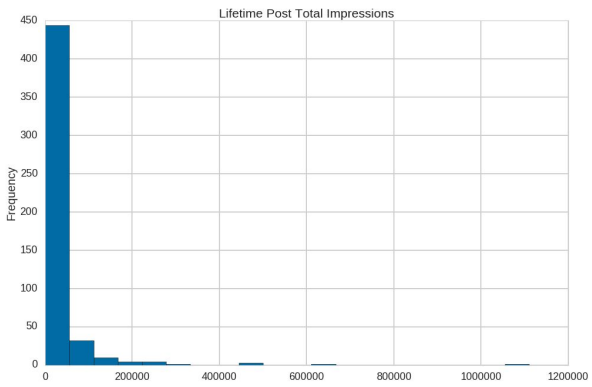


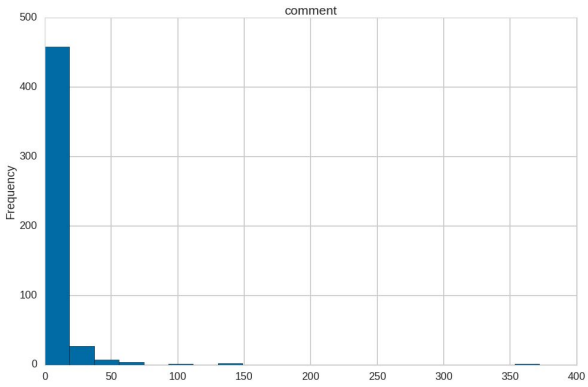

0.0
62.6%
comment
37.4%
1.0
0.0
1.0


0.0
66.5%
share
33.5%
1.0
0.0
1.0

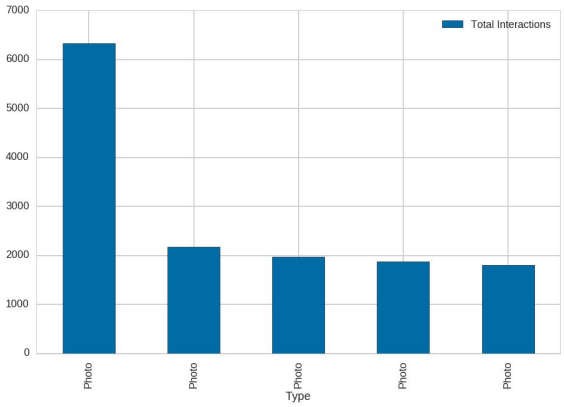
20)

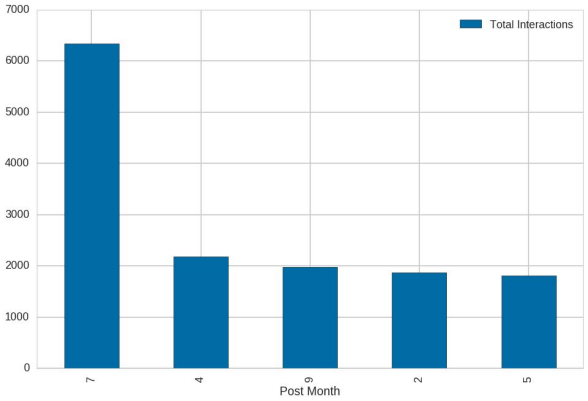
A histogram plot on Lifetime Post Time impressions , comments

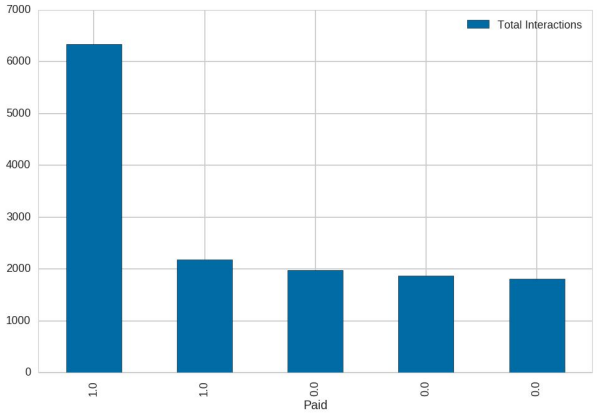


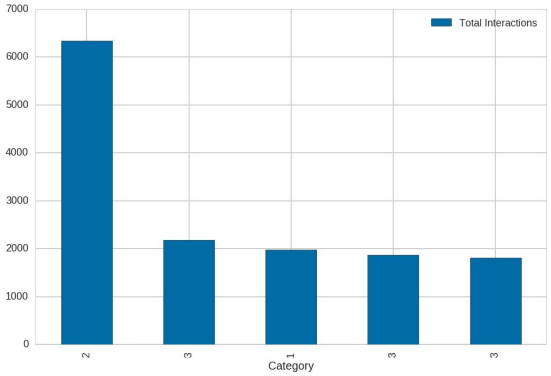




Bar plot made on top 5 posts sorted based on Total Interactions

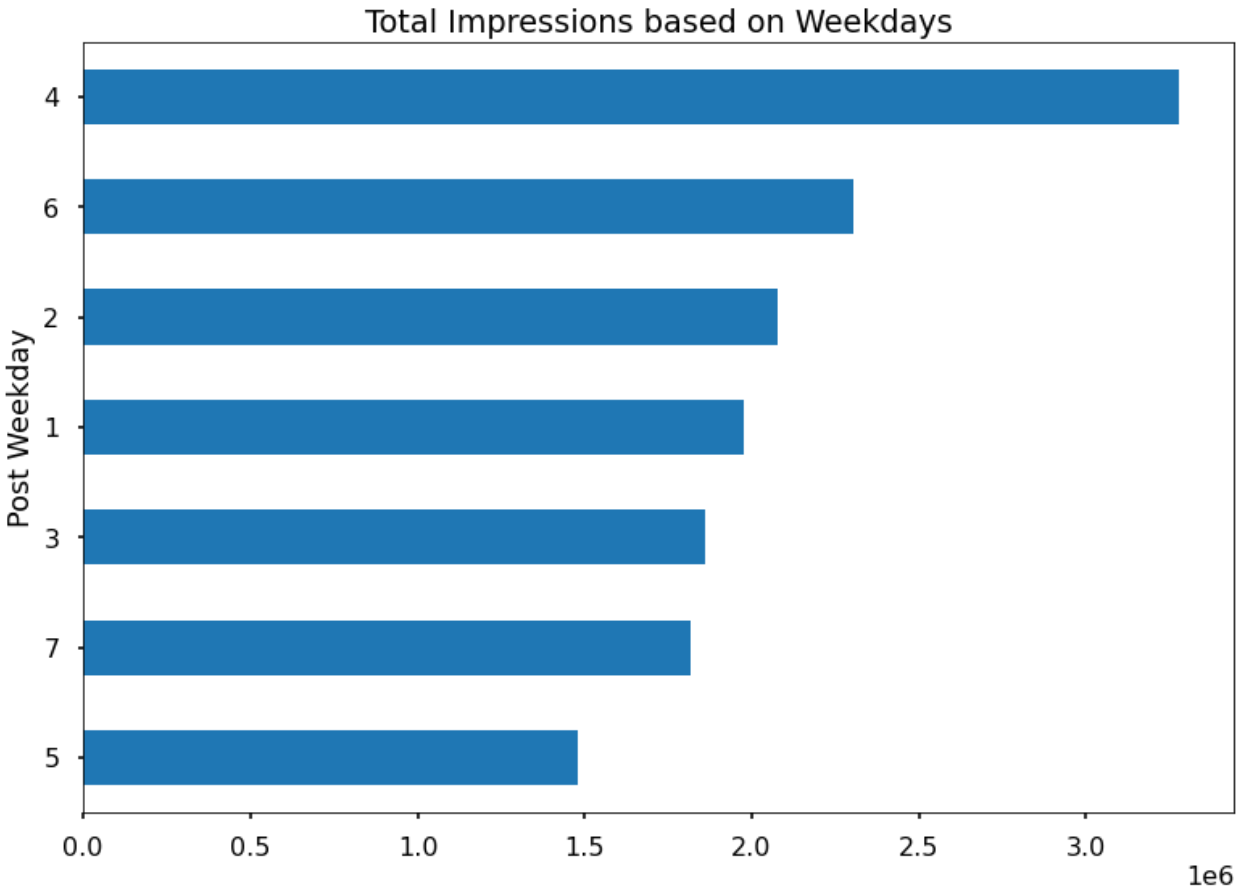


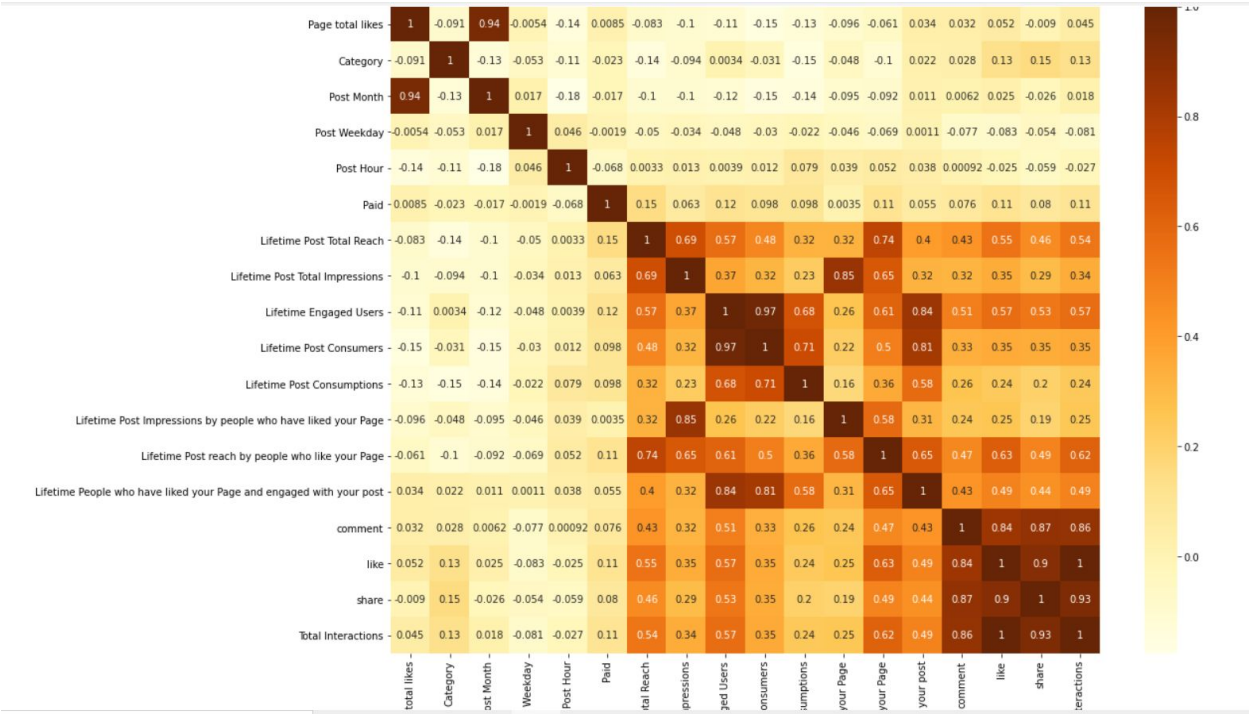




Horizontal Barplot for Total Impressions based on Weekdays

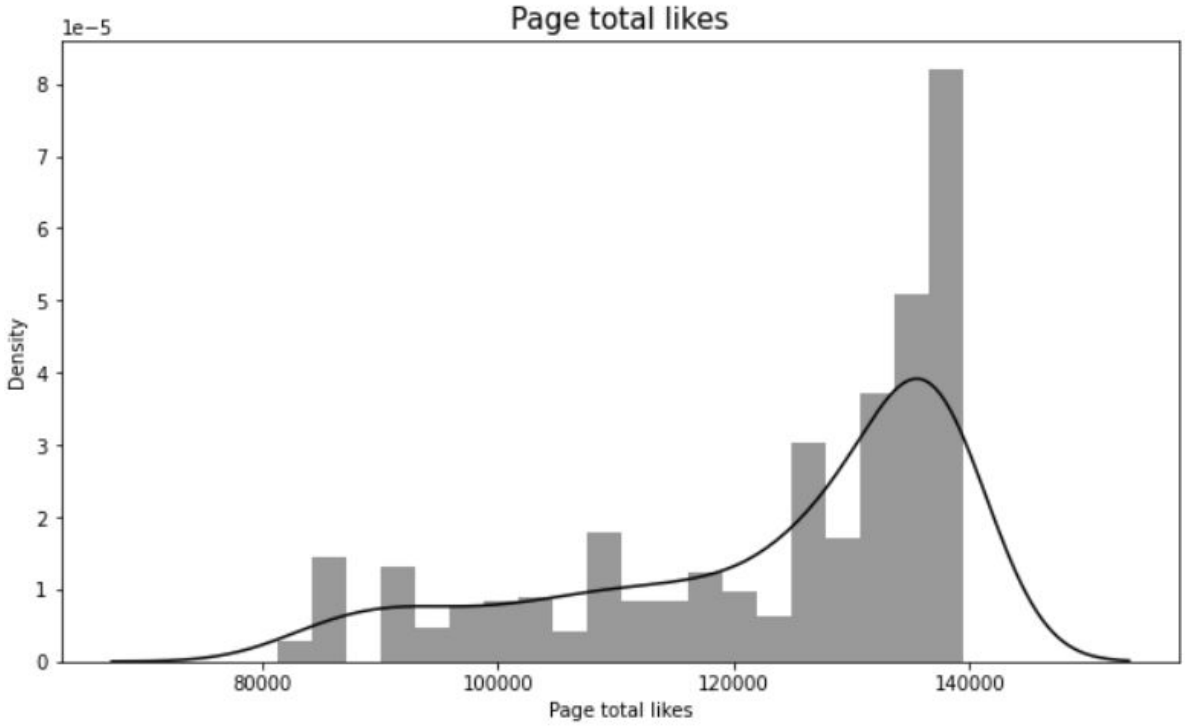



A histogram plot on Lifetime Post Time impressions , comments gave enough evidence on the presence of an outlier.
Bar plot made on top posts based on Total Interactions showed that a Paid Photo Post from category 2 in July had gone viral receiving immense response from the Users. If it has got negative feedback , the page should refrain from posting such content. If it was a positive one ,such content could give them more reach so they should post similar content in their page frequently

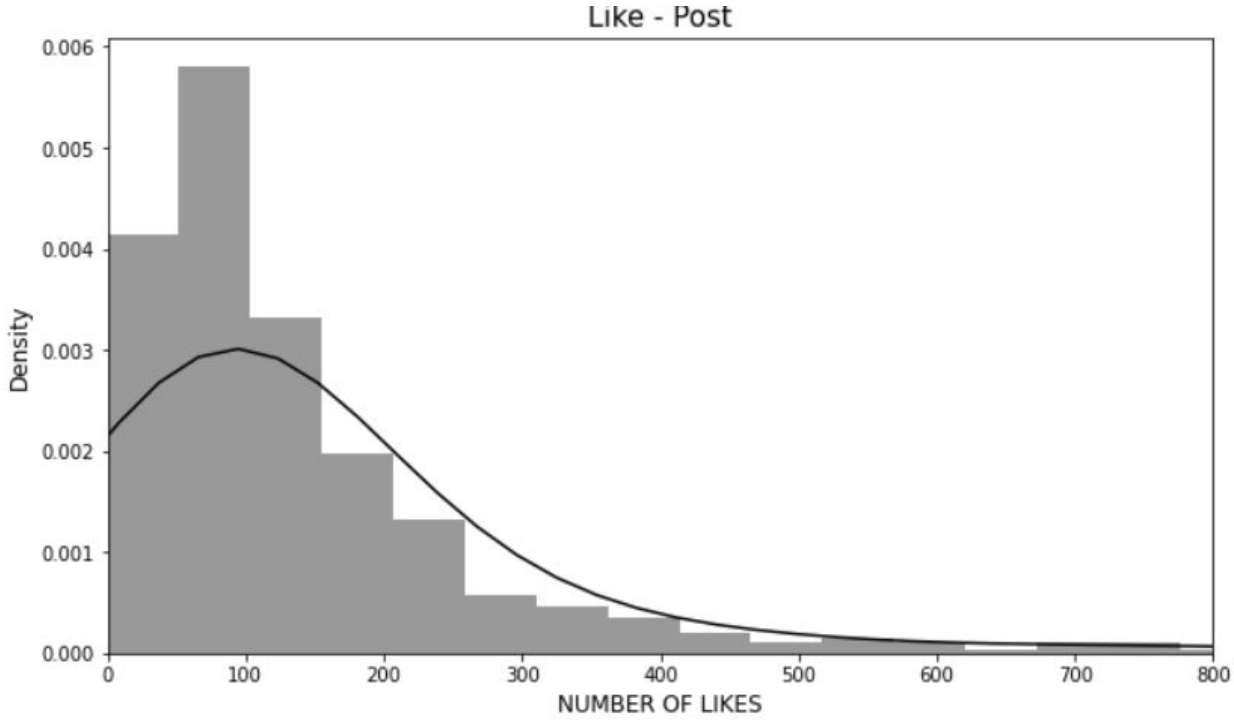
21)


Page total likes 1 -0.091 0.94 -0.0054 -0.14 0.0085 -0.083 -0.1 -0.11 -0.15 -0.13 -0.096 -0.061 0.034 0.032 0.052 -0.009 0.045
Category -0.091 1 -0.13 -0.053 -0.11 -0.023 -0.14 -0.094 0.0034 -0.031 -0.15 -0.048 -0.1 0.022 0.028 0.13 0.15 0.13
Post Month 0.94 -0.13 1 0.017 -0.18 -0.017 -0.1 -0.1 -0.12 -0.15 -0.14 -0.095 -0.092 0.011 0.0062 0.025 -0.026 0.018
Post Weekday -0.0054 -0.053 0.017 1 0.046 -0.0019 -0.05 -0.034 -0.048 -0.03 -0.022 -0.046 -0.069 0.0011 -0.077 -0.083 -0.054 -0.081
Post Hour -0.14 -0.11 -0.18 0.046 1 -0.068 0.0033 0.013 0.0039 0.012 0.079 0.039 0.052 0.038 0.00092 -0.025 -0.059 -0.027
Paid 0.0085 -0.023 -0.017 -0.0019 -0.068 1 0.15 0.063 0.12 0.098 0.098 0.0035 0.11 0.055 0.076 0.11 0.08 0.11
Lifetime Post Total Reach -0.083 -0.14 -0.1 -0.05 0.0033 0.15 1 0.69 0.57 0.48 0.32 0.32 0.74 0.4 0.43 0.55 0.46 0.54
Lifetime Post Total Impressions -0.1 -0.094 -0.1 -0.034 0.013 0.063 0.69 1 0.37 0.32 0.23 0.85 0.65 0.32 0.32 0.35 0.29 0.34
Lifetime Engaged Users -0.11 0.0034 -0.12 -0.048 0.0039 0.12 0.57 0.37 1 0.97 0.68 0.26 0.61 0.84 0.51 0.57 0.53 0.57
Lifetime Post Consumers -0.15 -0.031 -0.15 -0.03 0.012 0.098 0.48 0.32 0.97 1 0.71 0.22 0.5 0.81 0.33 0.35 0.35 0.35
Lifetime Post Consumptions -0.13 -0.15 -0.14 -0.022 0.079 0.098 0.32 0.23 0.68 0.71 1 0.16 0.36 0.58 0.26 0.24 0.2 0.24
Lifetime Post Impressions by people who have liked your Page -0.096 -0.048 -0.095 -0.046 0.039 0.0035 0.32 0.85 0.26 0.22 0.16 1 0.58 0.31 0.24 0.25 0.19 0.25
Lifetime Post reach by people who like your Page -0.061 -0.1 -0.092 -0.069 0.052 0.11 0.74 0.65 0.61 0.5 0.36 0.58 1 0.65 0.47 0.63 0.49 0.62
Lifetime People who have liked your Page and engaged with your post 0.034 0.022 0.011 0.0011 0.038 0.055 0.4 0.32 0.84 0.81 0.58 0.31 0.65 1 0.43 0.49 0.44 0.49
comment 0.032 0.028 0.0062 -0.077 0.00092 0.076 0.43 0.32 0.51 0.33 0.26 0.24 0.47 0.43 1 0.84 0.87 0.86
like 0.052 0.13 0.025 -0.083 -0.025 0.11 0.55 0.35 0.57 0.35 0.24 0.25 0.63 0.49 0.84 1 0.9 1
share -0.009 0.15 -0.026 -0.054 -0.059 0.08 0.46 0.29 0.53 0.35 0.2 0.19 0.49 0.44 0.87 0.9 1 0.93
Total Interactions 0.045 0.13 0.018 -0.081 -0.027 0.11 0.54 0.34 0.57 0.35 0.24 0.25 0.62 0.49 0.86 1 0.93 1
Category
Post Month
Post Hour
Paid
comment
like
share
1.0
0.8
0.6
0.4
0.2
0.0

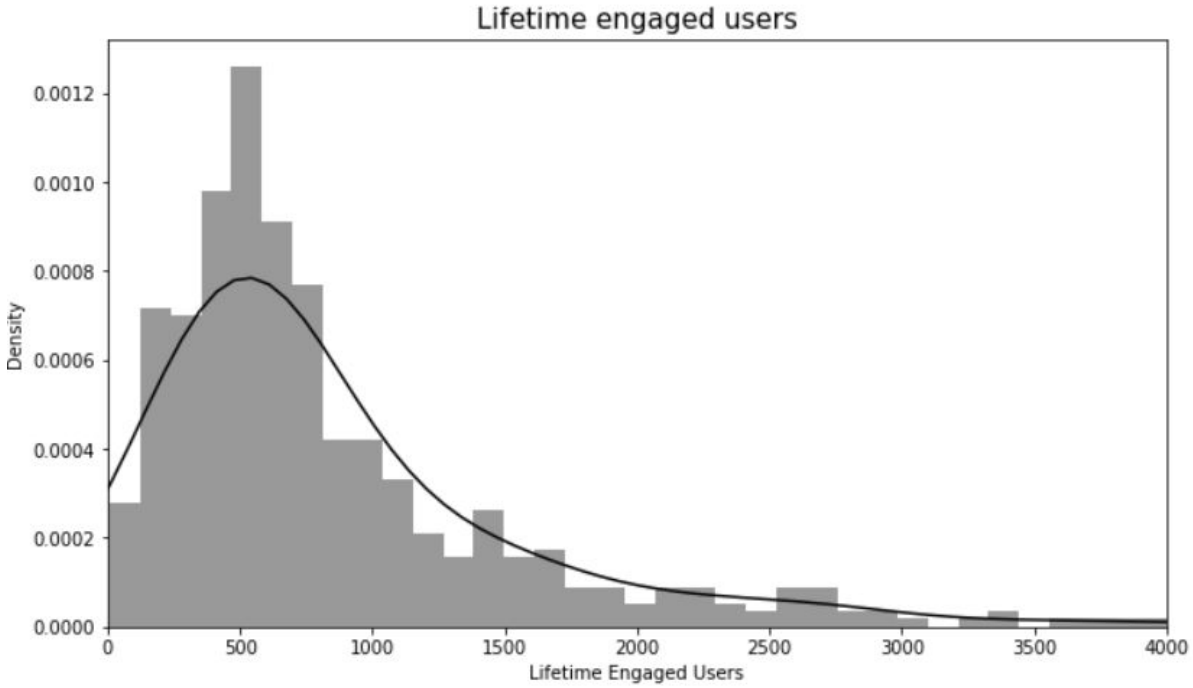
22)


Page total likes
1e−5
8
7
6
5
4
3
2
1
0
Density
80000
100000
120000
140000
Page total likes

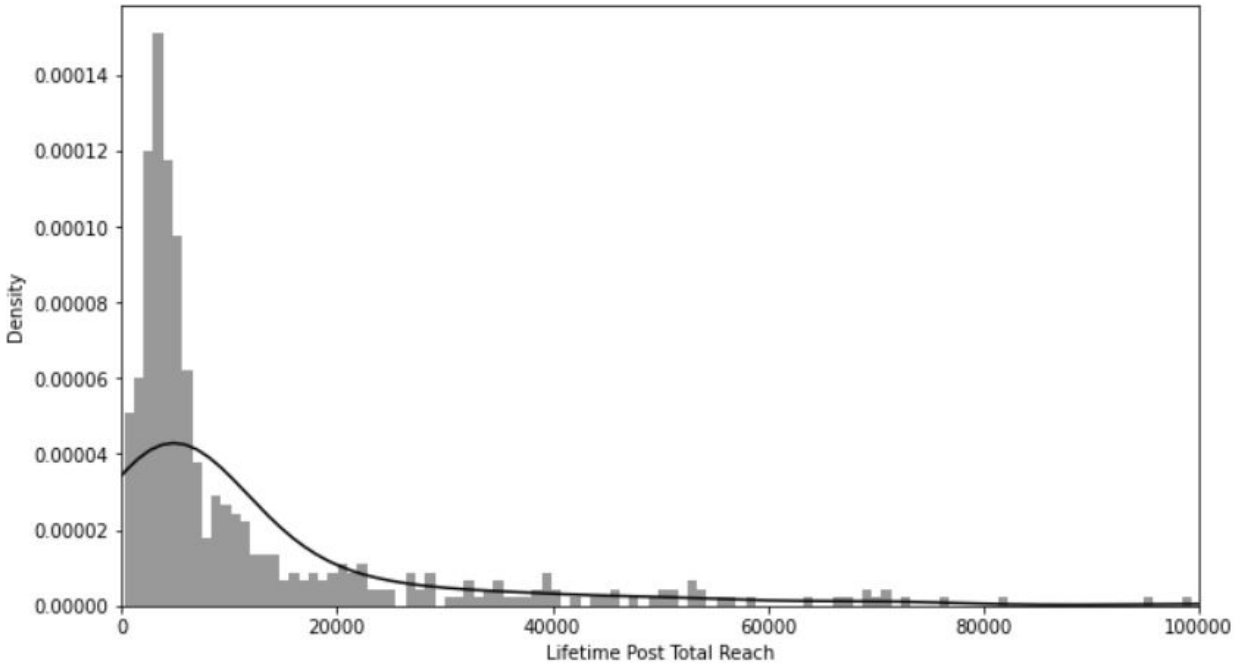
23)Histograms:



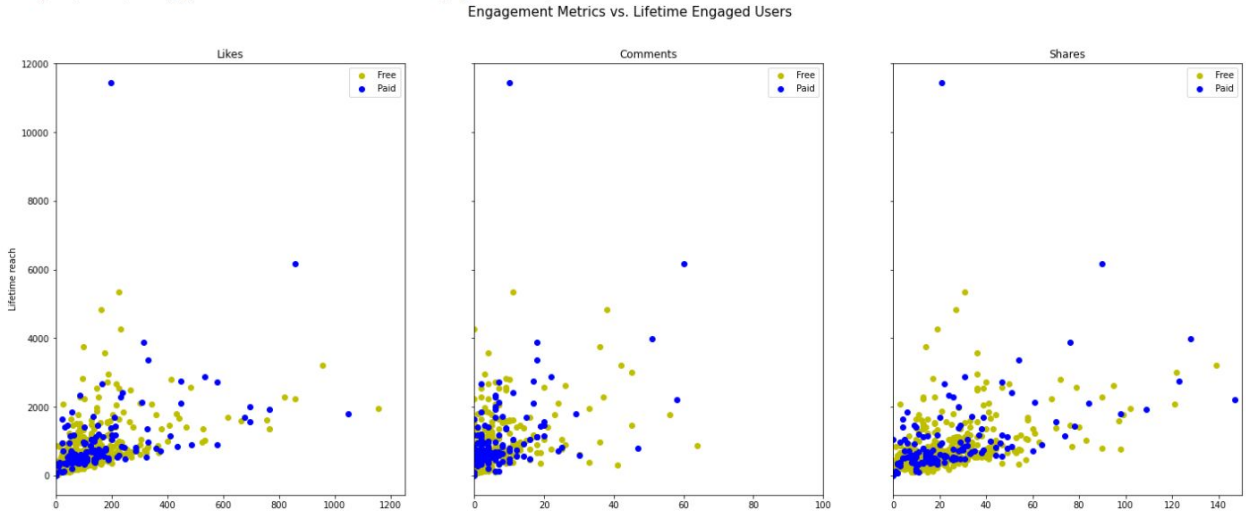

24)



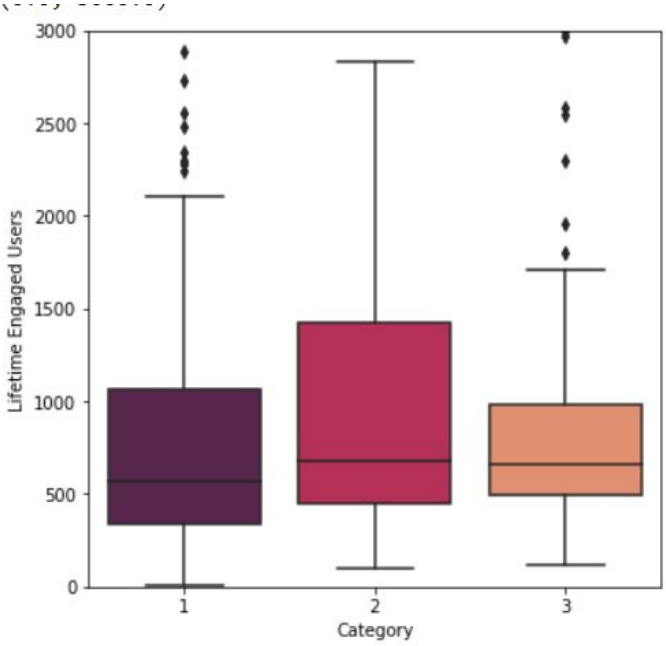
25)

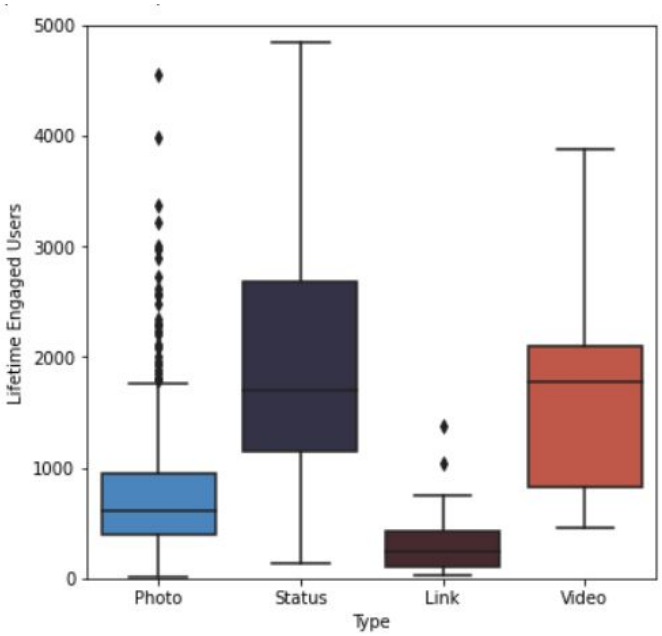

Density
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
20000
40000
60000
80000
100000
Lifetime Post Total Reach

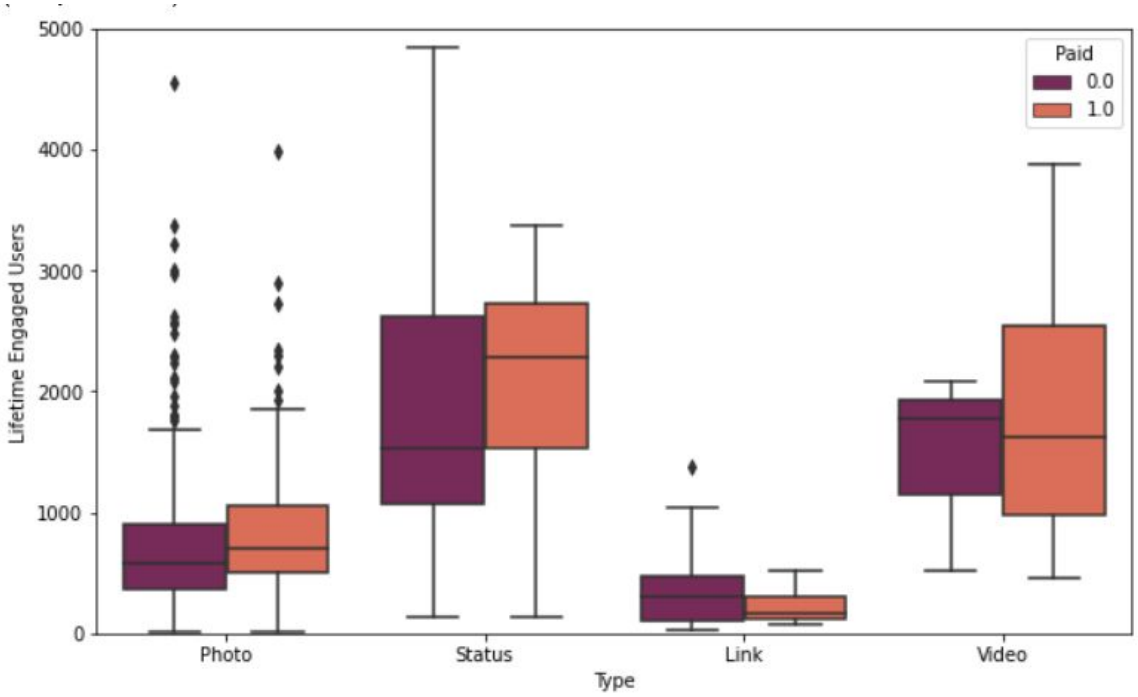
26)


Engagement Metrics vs. Lifetime Engaged Users
Likes
Comments
Shares
Free
Paid
Lifetime reach

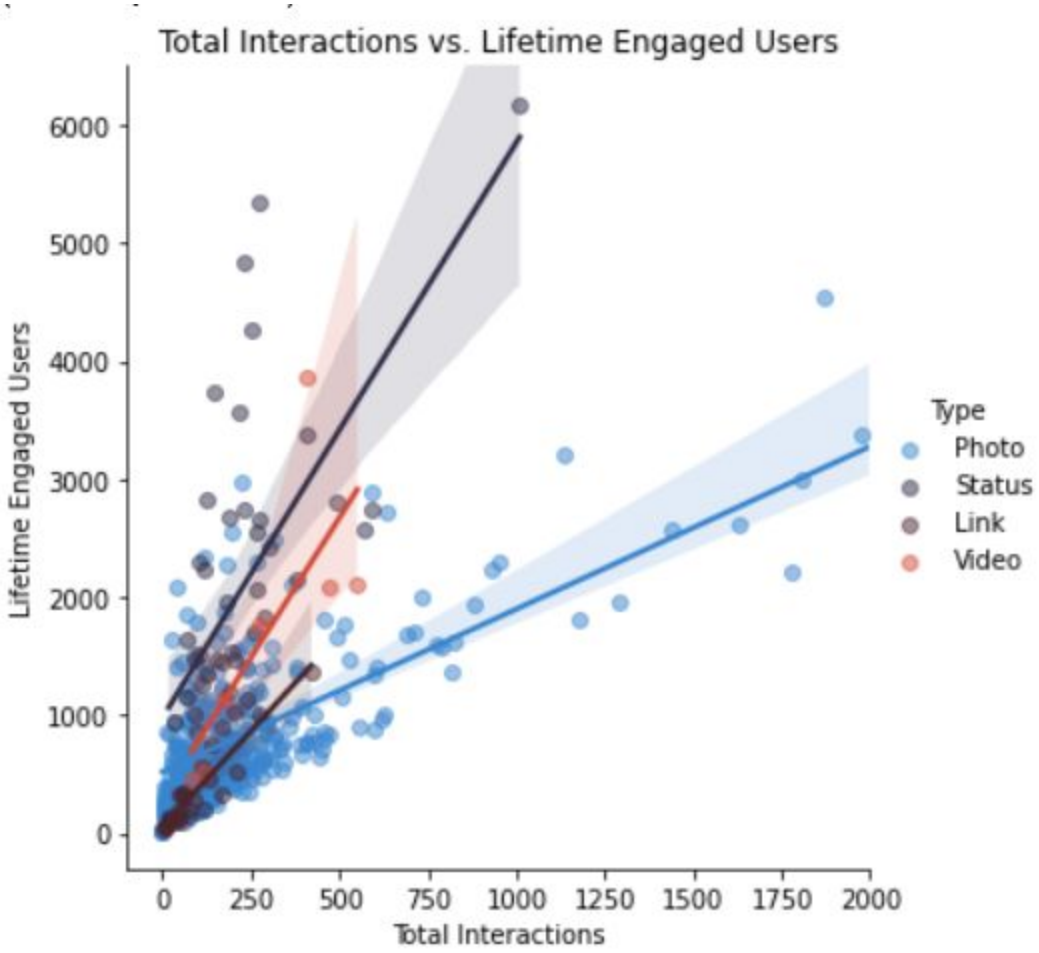
27)Influence of Type and Category on Lifetime engaged users:



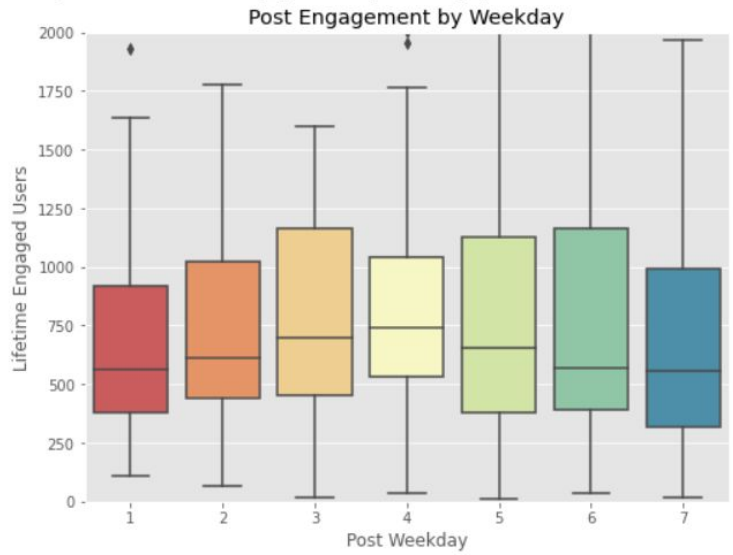



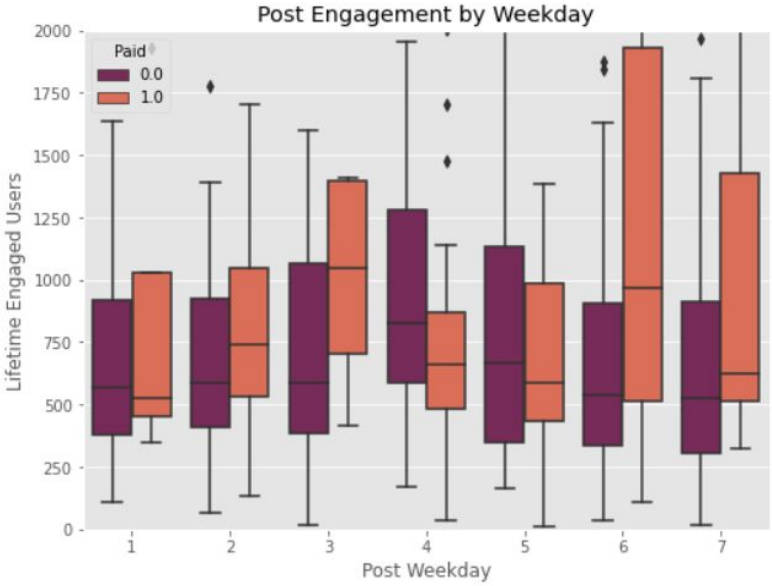
28)

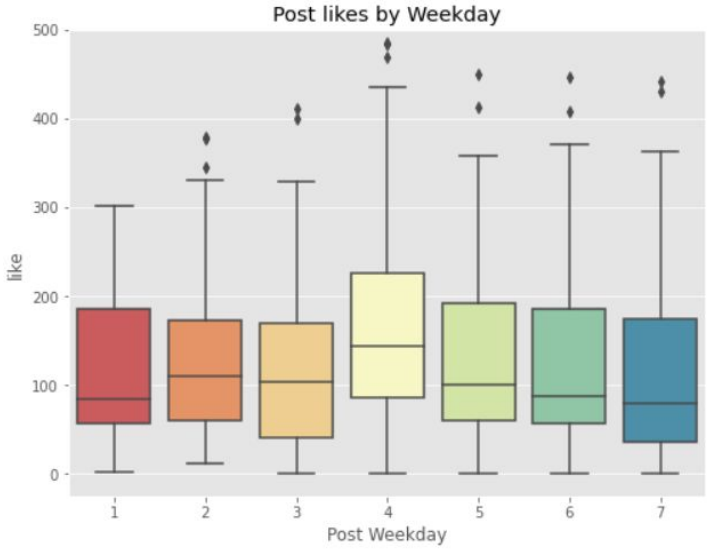


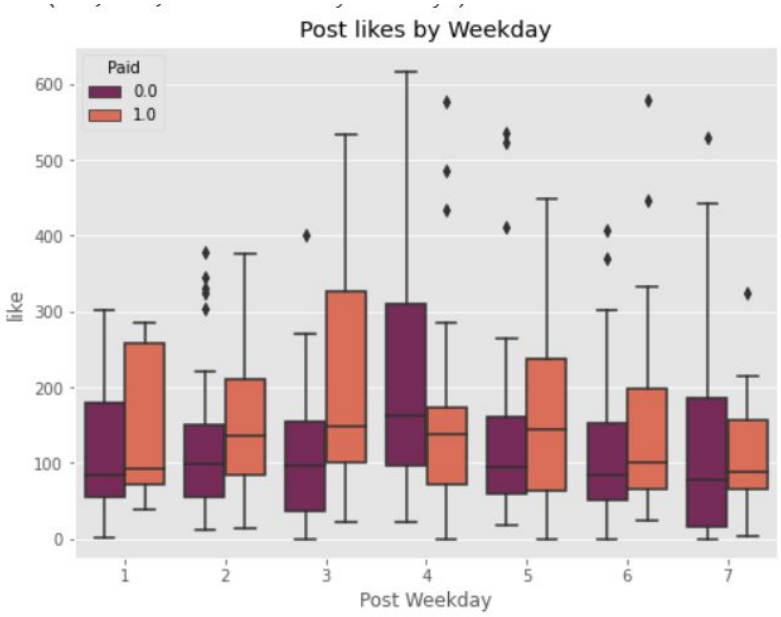

Interactions : Engaged Users ratio is higher for Photo than others

29)We can see how post day influences the lifetime engaged users:

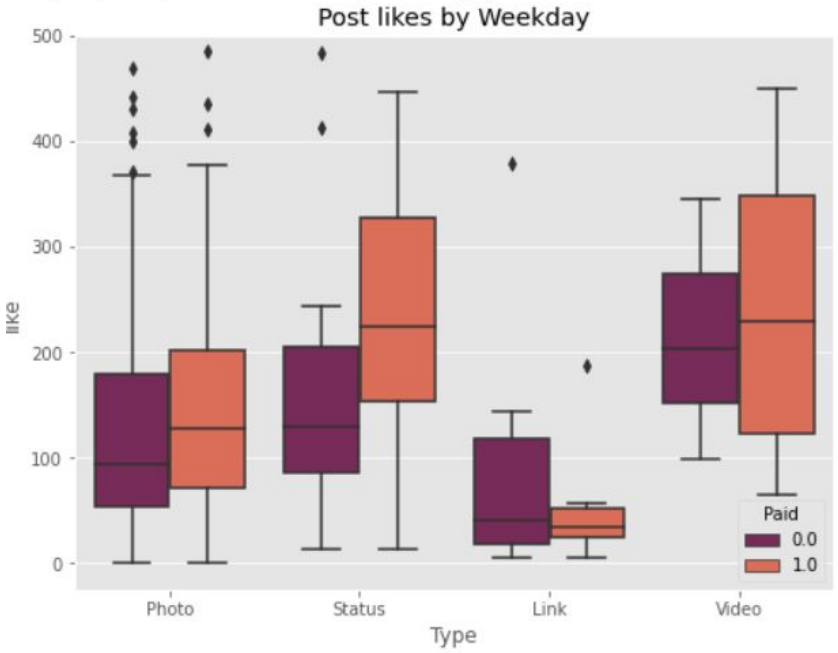


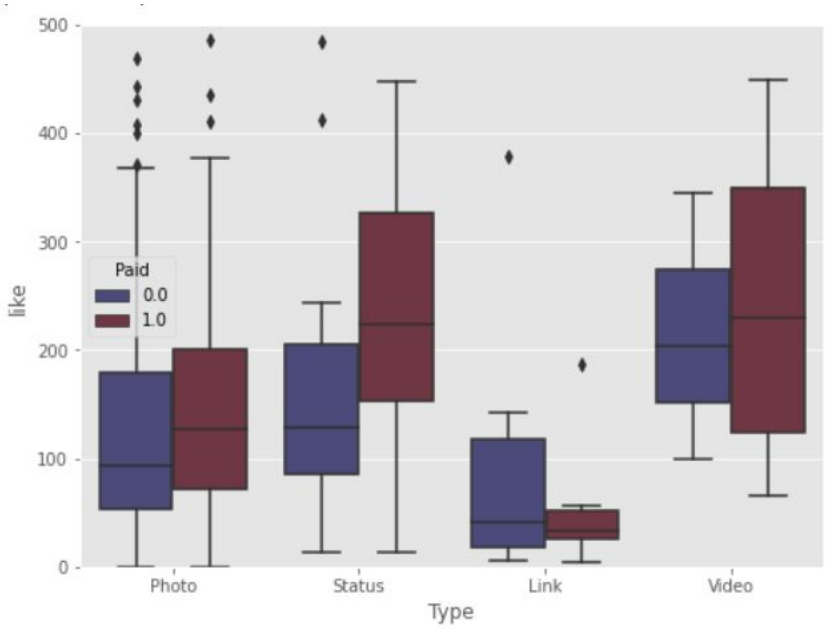



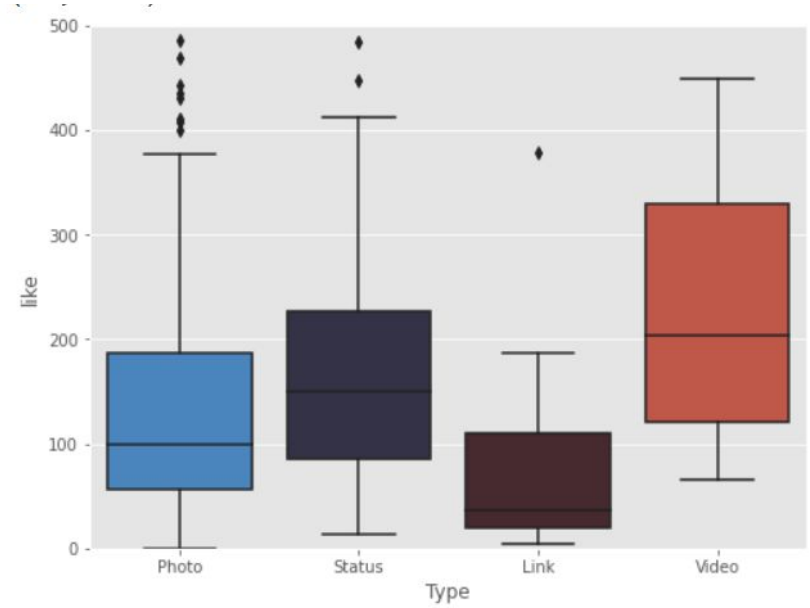



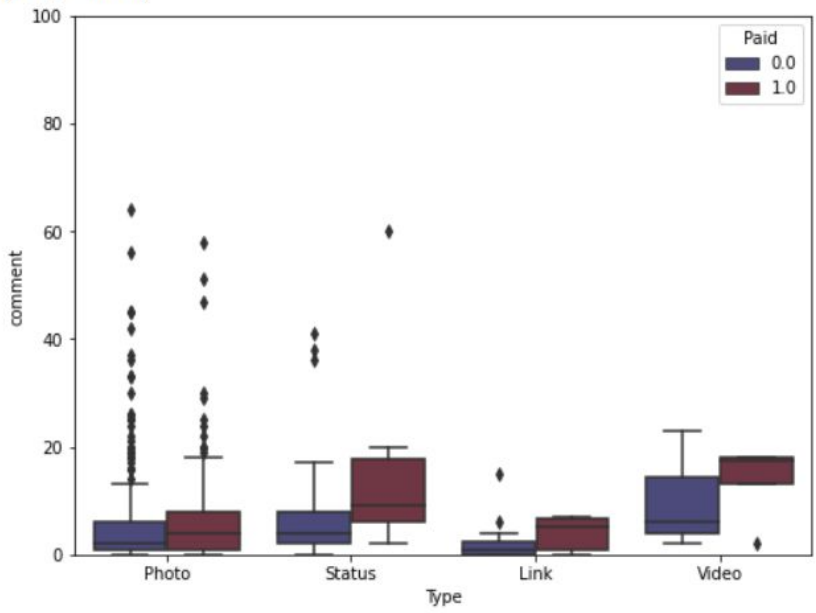

Post likes by Weekday
Paid
0.0
1.0
like
600
500
400
300
200
100
0
1
2
3
4
5
6
7
Post Weekday

30) Influence of post-type on other attributes:

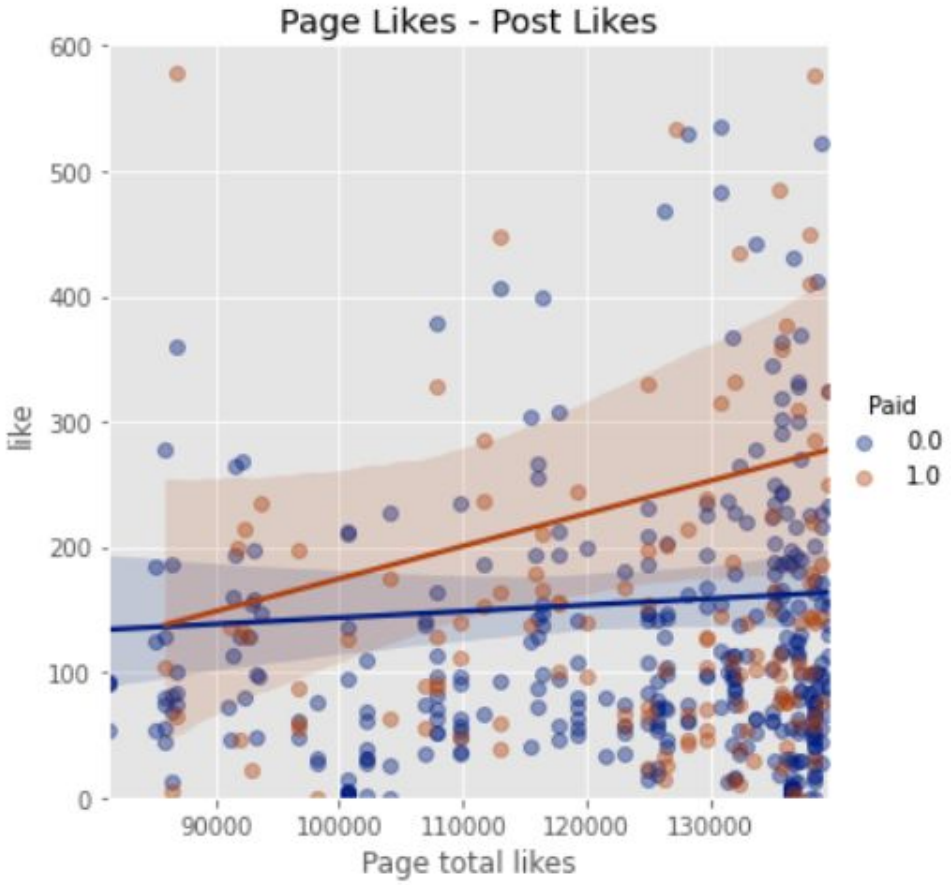



500
400
300
200
100
0
like
Photo
Status
Link
Video
Type


100
80
60
40
20
0
comment
Paid
0.0
1.0
Photo
Status
Link
Video
Type

31) Number of likes paid-posts get increases with total page likes.



## Conclusion:

All the observations made give us a good idea of the dataset. PCA showed that a given dataset can be expressed with 7 principal components which can retain 97% of data. An outlier was also noted from bar plots.

Regression analysis showed that the original 7 attributes could account for “Lifetime Post Engagement” with a R-squared value of 0.6. We also conclude that Post Interaction > Post Reach > Post Impressions > Post Engagement for Page total likes.

Data was thoroughly analysed and many inferences were made from the analysis.

Though we were able to make some valuable conclusions from Exploratory Data Analysis. From PCA and Regression, we find that most of the information in the dataset is given by the lifetime variables, the independent variables had a modest ability to explain the lifetime variables.

# Reference:


1) Moro S, Rita P, & Vala B. (2016), Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach
2) https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce
3) https://en.wikipedia.org/wiki/Exploratory_data_analysis
4) https://en.wikipedia.org/wiki/Principal_component_analysis
5) Lecture Videos of Prof Dr.Mainak Thakur , IIIT Sri City