SDA Project

By,

Pavan Sai Santhosh E
Darshan G
Jaswanth K
Sathyanarayanan R

Dataset Description

1. Number of Observations - 500
2. Features - 19
    - Page total likes
    - Type
    - Category
    - Post Month
    - Post Weekday
    - Post Hour

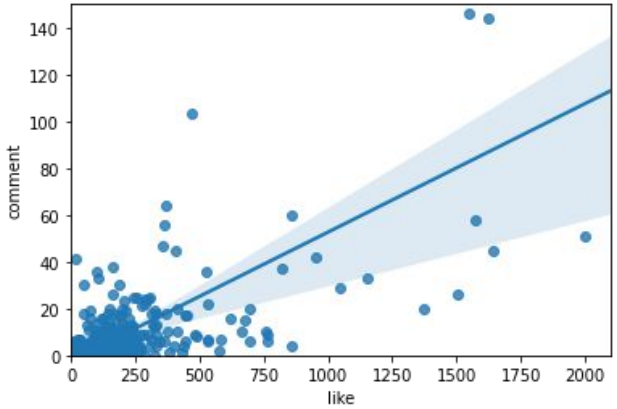
- Paid
- Lifetime Post Total Reach
- Lifetime Post Total Impressions
- Lifetime Engaged Users
- Lifetime Post Consumers
- Lifetime Post Consumptions
- Lifetime Post Impressions by people who have liked your Page
- Lifetime Post reach by people who like your Page
- Lifetime People who have liked your Page and engaged with your post
- Comment
- Like
- Share
- Total Interactions

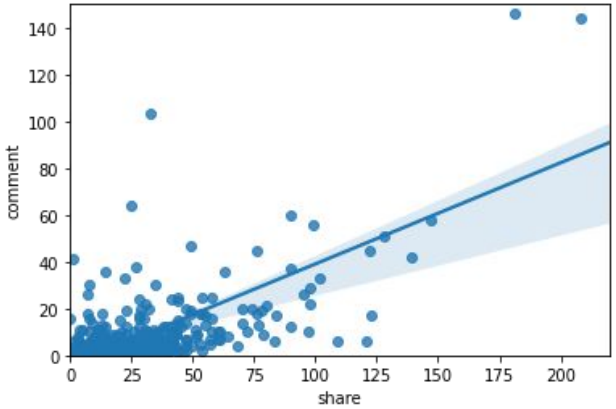
Tasks done:

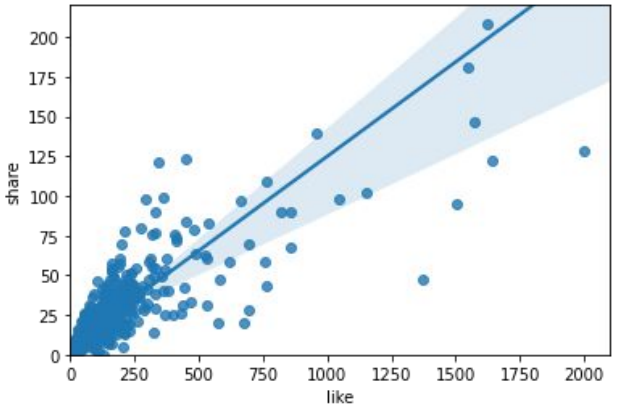
1. Exploratory Data Analysis
2. Regression
3. Principal Component Analysis

Exploratory Data Analysis

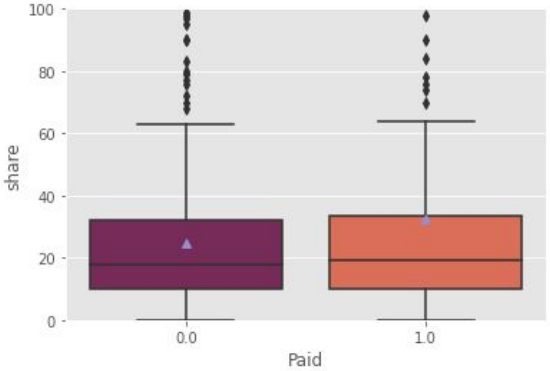
Relationship between Like-Share-Comment:

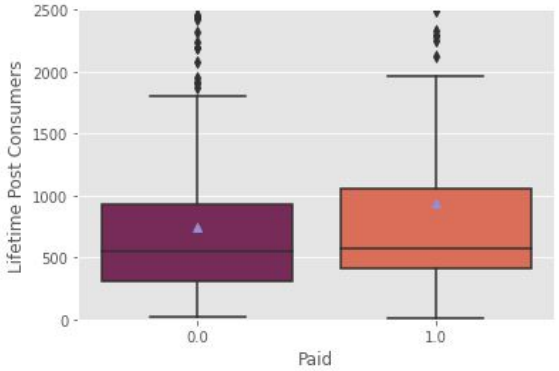


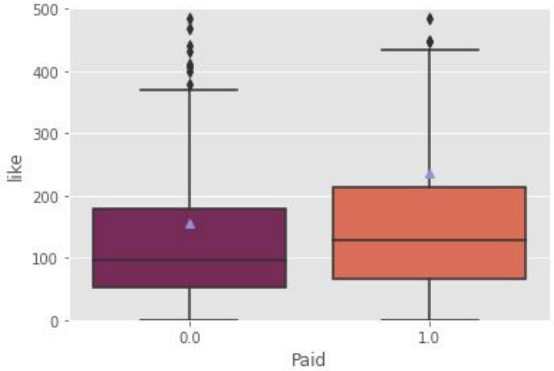


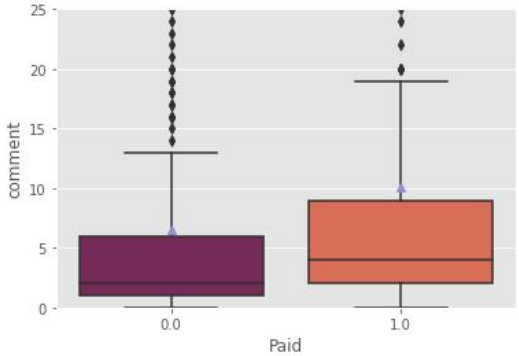


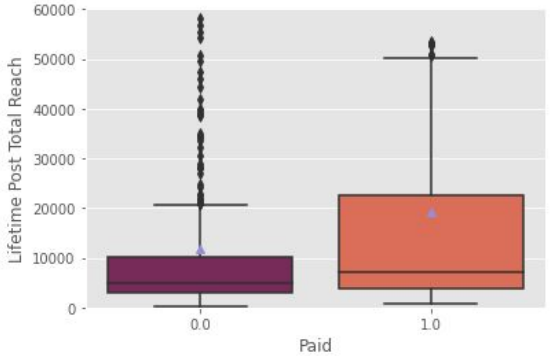
How does ‘Paid’ influence other variables?

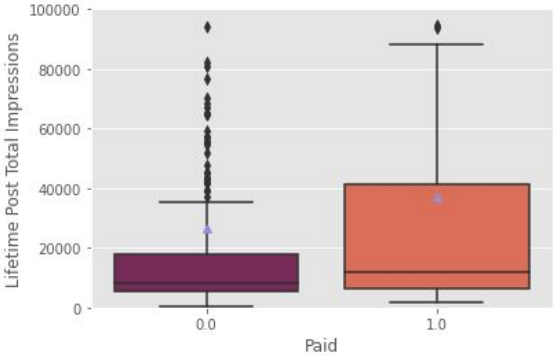




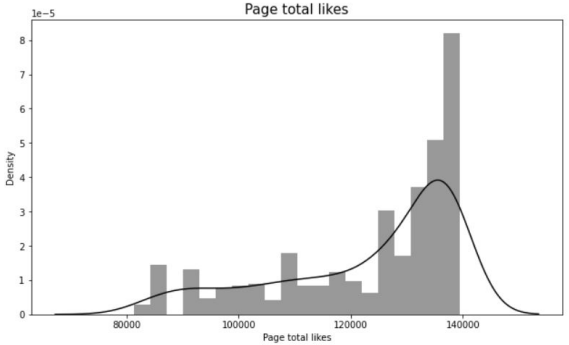


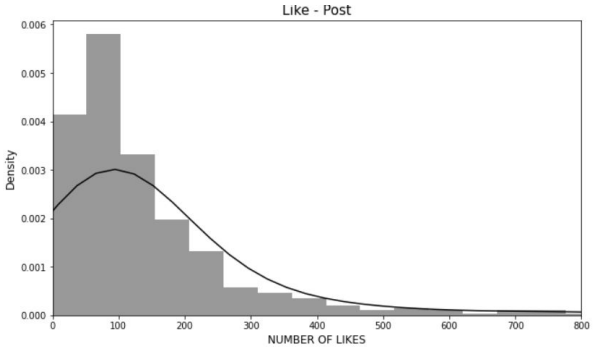


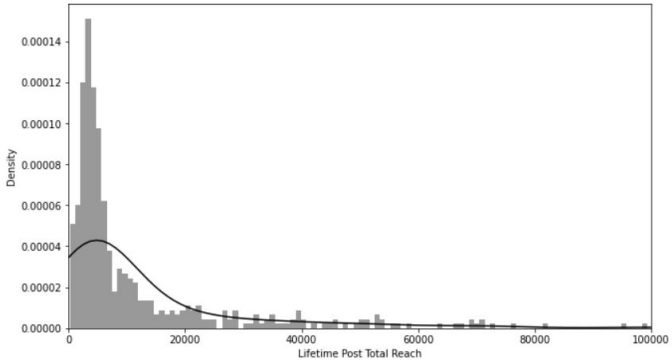


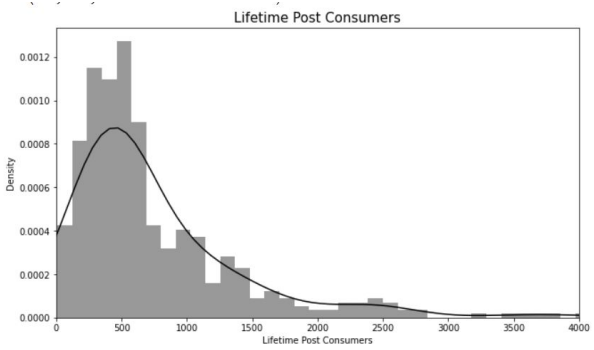


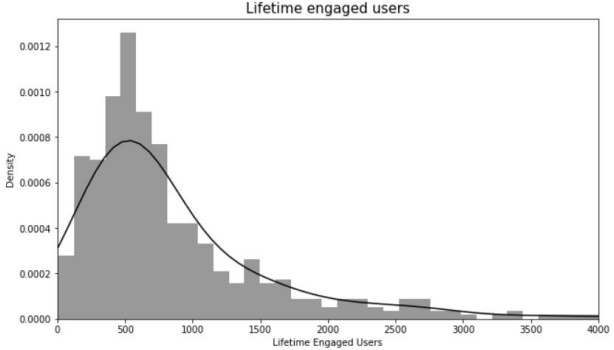
Histograms for various features:

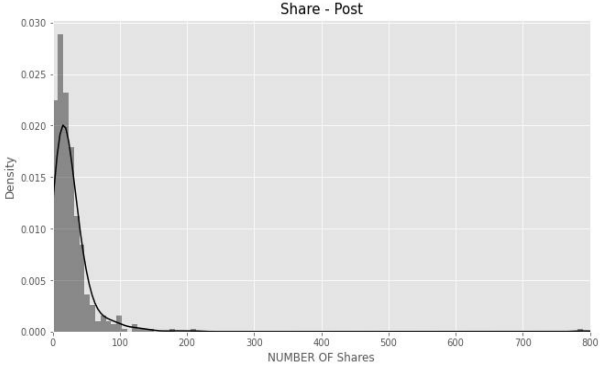




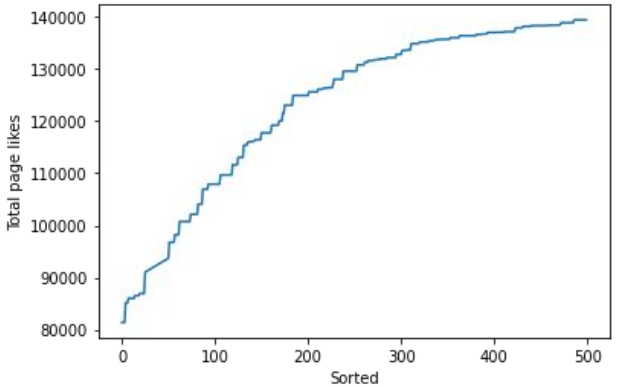


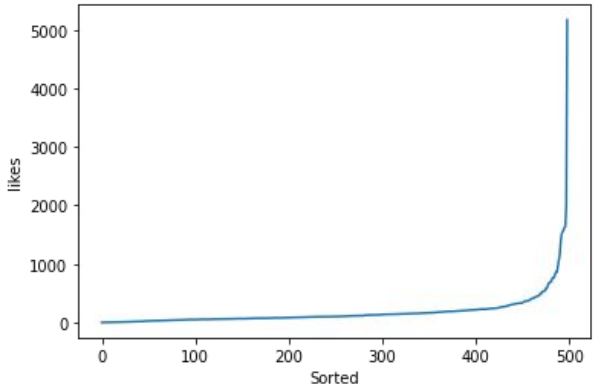


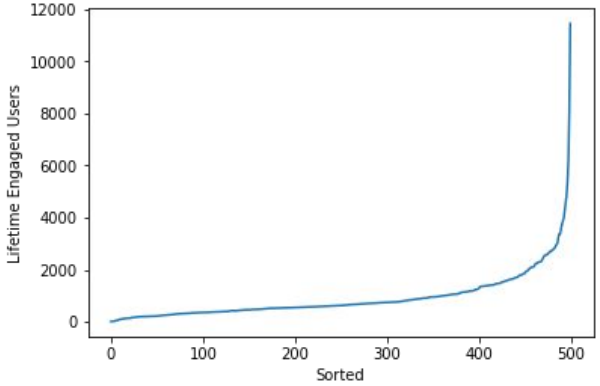


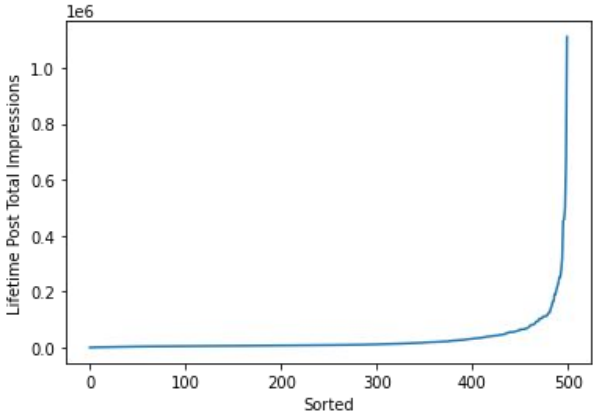


Values of various attributes were sorted and plotted for 500 Posts:





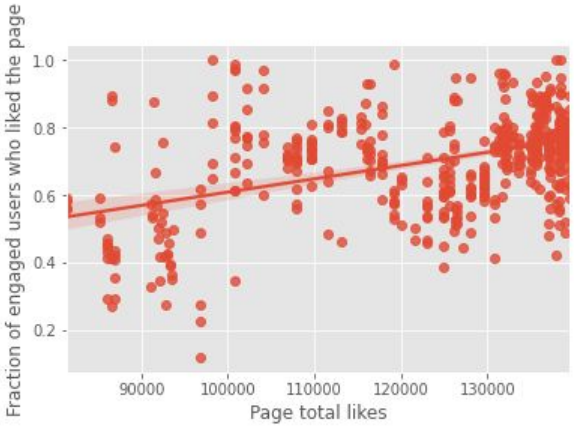


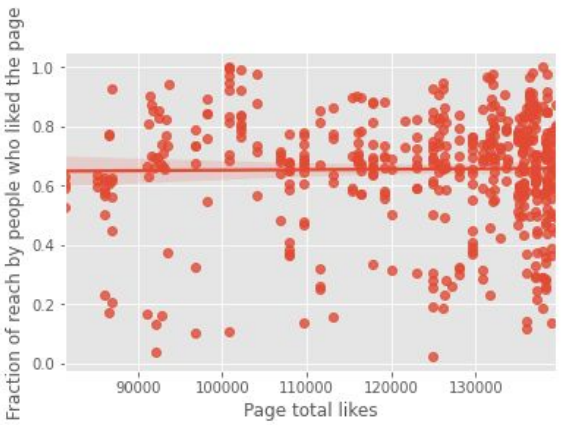


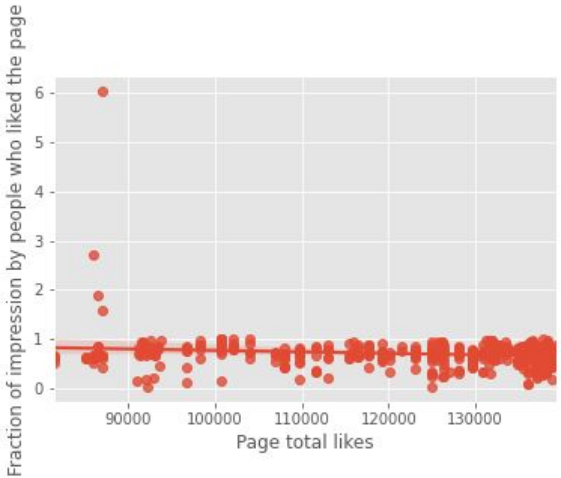
Fraction of 'Reach', 'Impression' and 'Engagement' from people who have liked the page:

Using these features, the fraction is calculated and analysed for ‘Reach’, ‘Impression’ and ‘Engagement’:

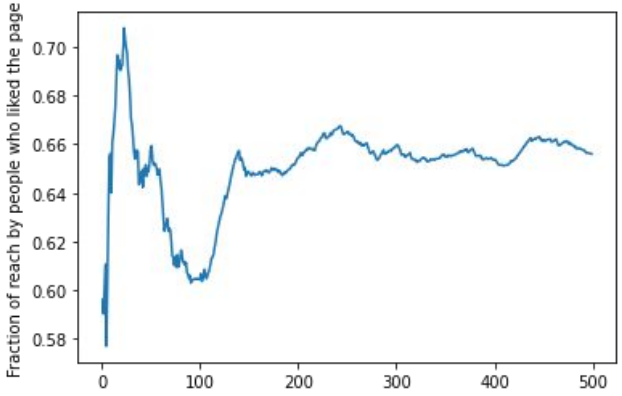
- Lifetime Post Impressions by people who have liked your Page / Total Impressions
- Lifetime Post reach by people who like your Page / Total Reach
- Lifetime People who have liked your Page and engaged with your post / Total Engaged users

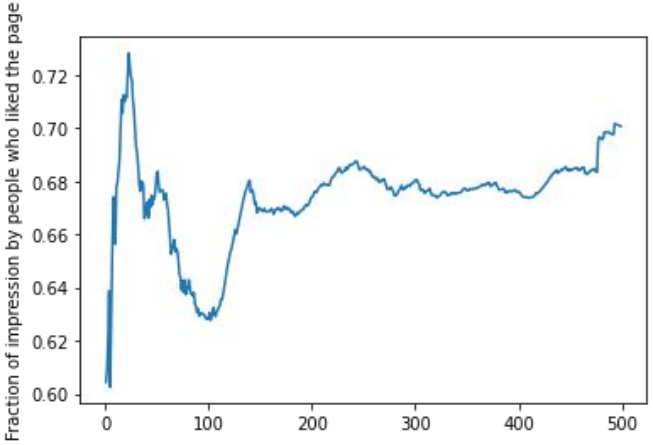


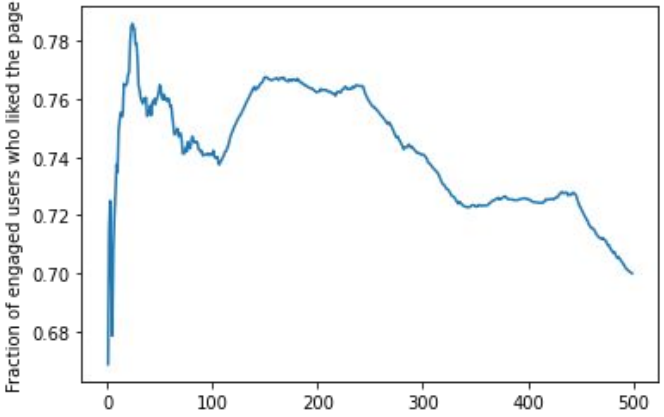




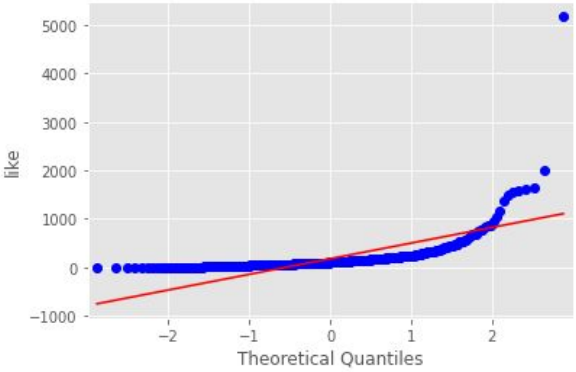
Mean of fractions:

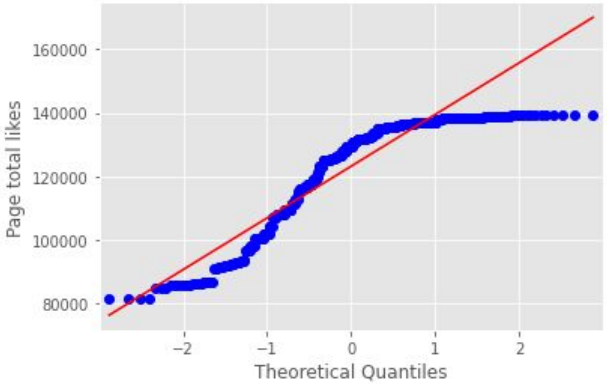


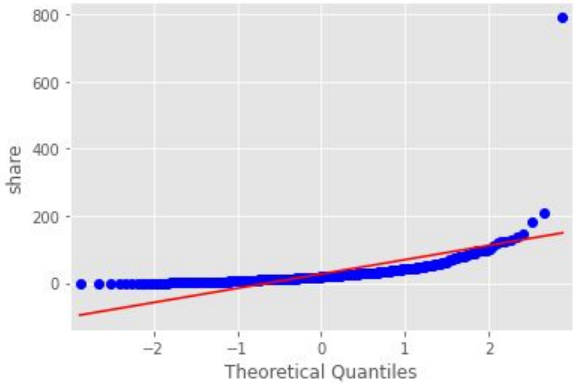


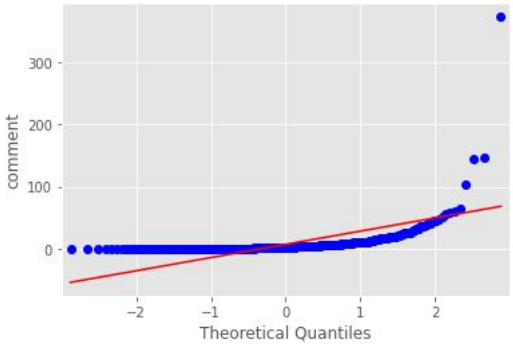


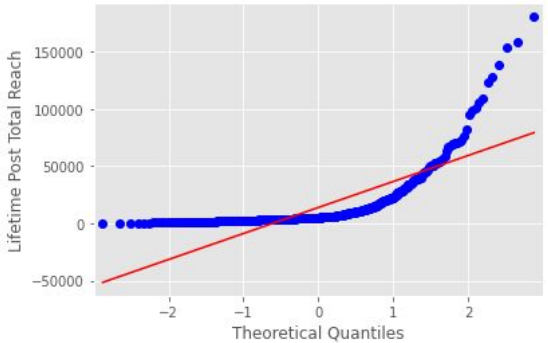
QQ Plots


like
Theoretical Quantiles

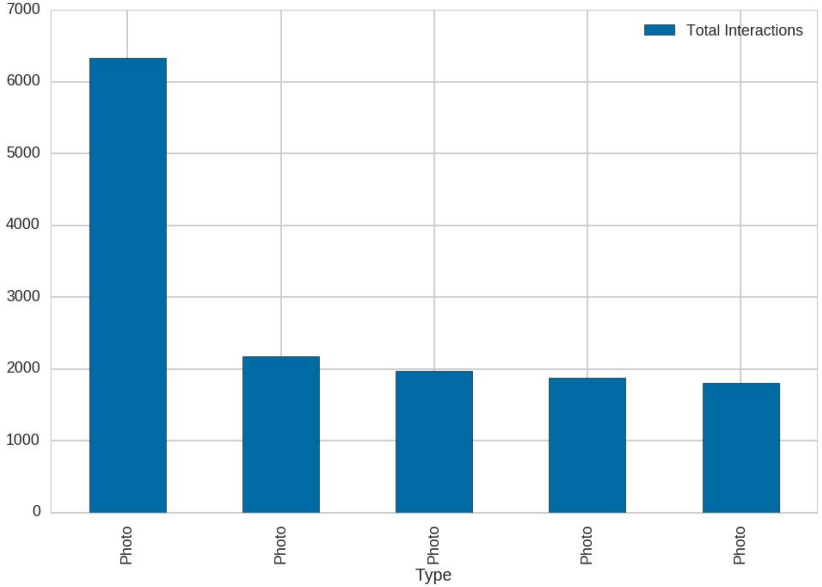

Page total likes
Theoretical Quantiles

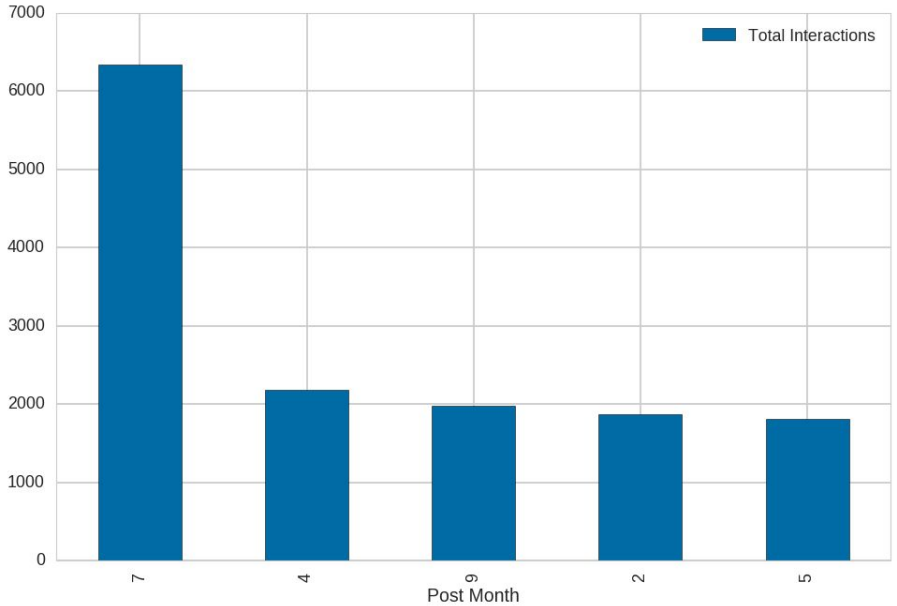

share
Theoretical Quantiles


comment
Theoretical Quantiles

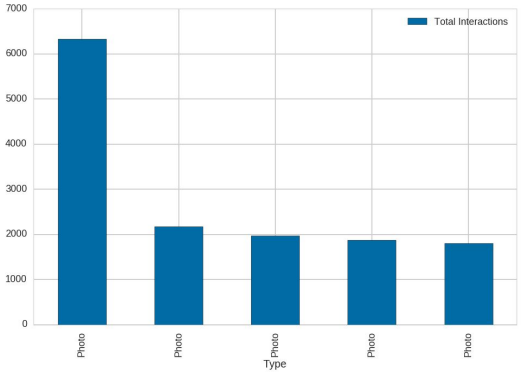

Lifetime Post Total Reach
Theoretical Quantiles

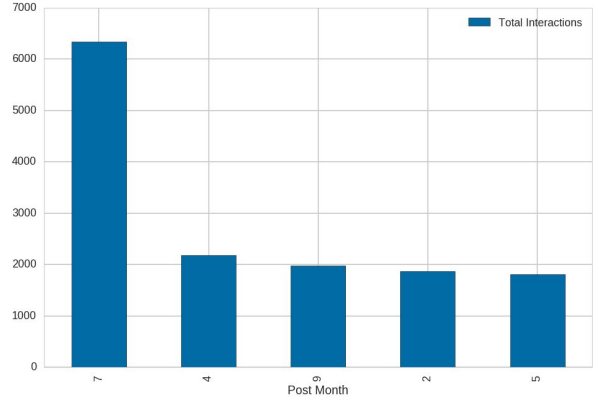
Bar plot made on top 5 posts sorted based on Total Interactions

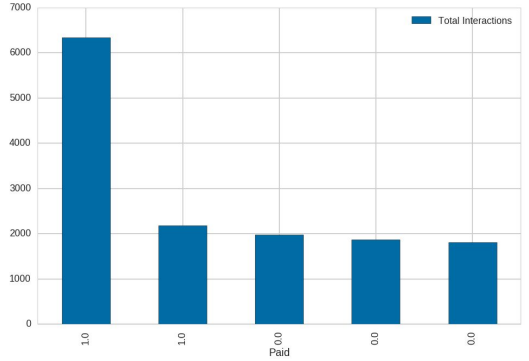


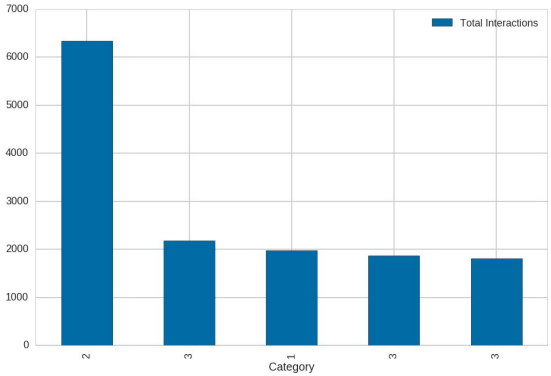


Bar plot made on top 5 posts sorted based on Total Interactions





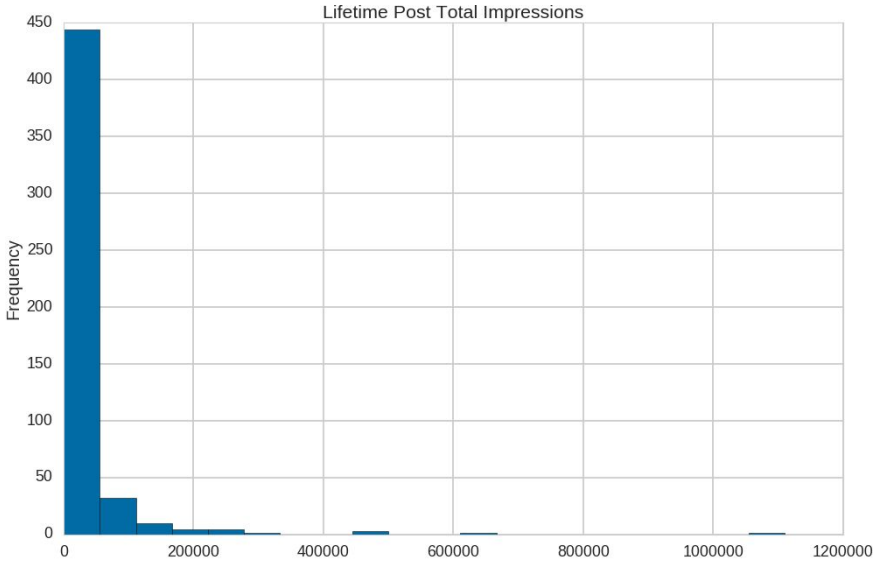


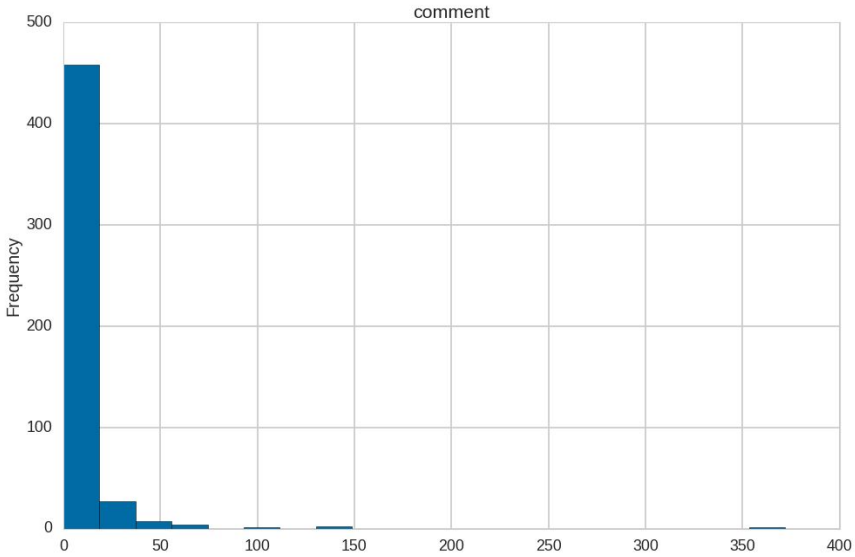


Observation

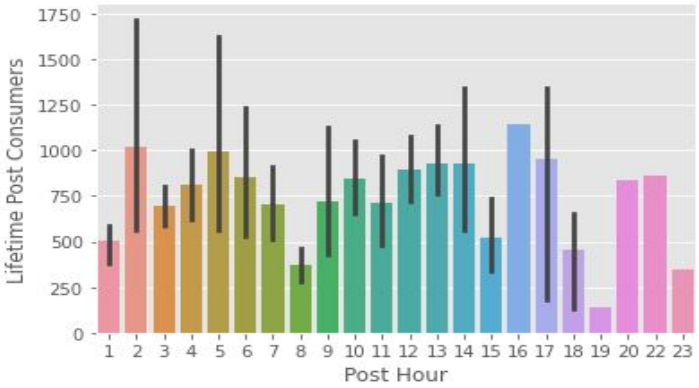
- A histogram plot on Lifetime Post Time impressions , comments
  - Evidence on the presence of an outlier.
- Bar plot made on top posts based on Total Interactions
  - Shows that a Paid Photo Post from category 2 in July had gone viral receiving immense response from the Users.

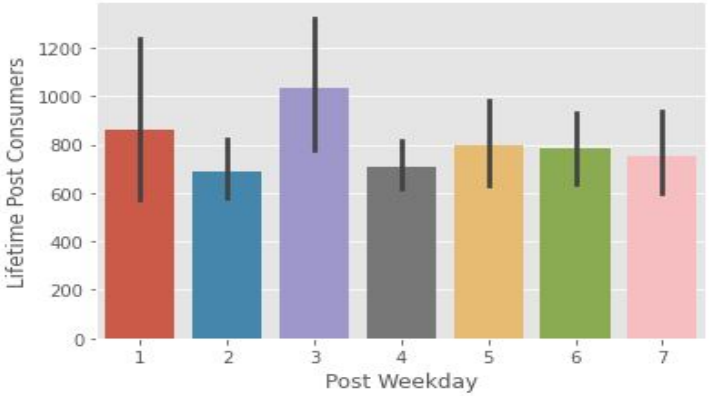
Histograms - Lifetime Post Time impressions and comments

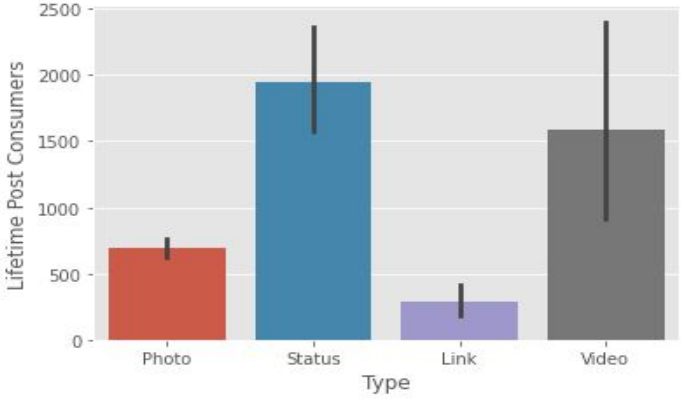


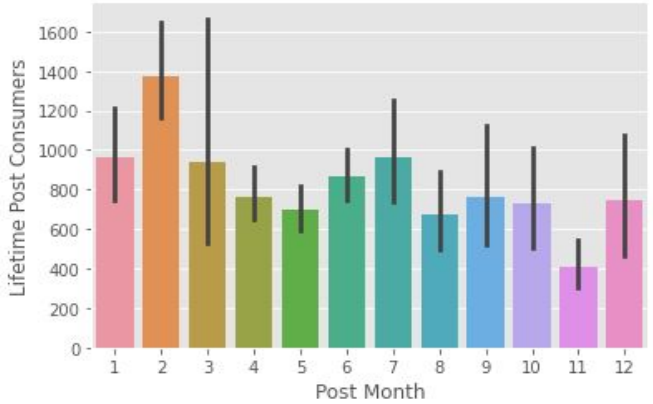


BAR PLOTS ON LIFE TIME POST CONSUMERS

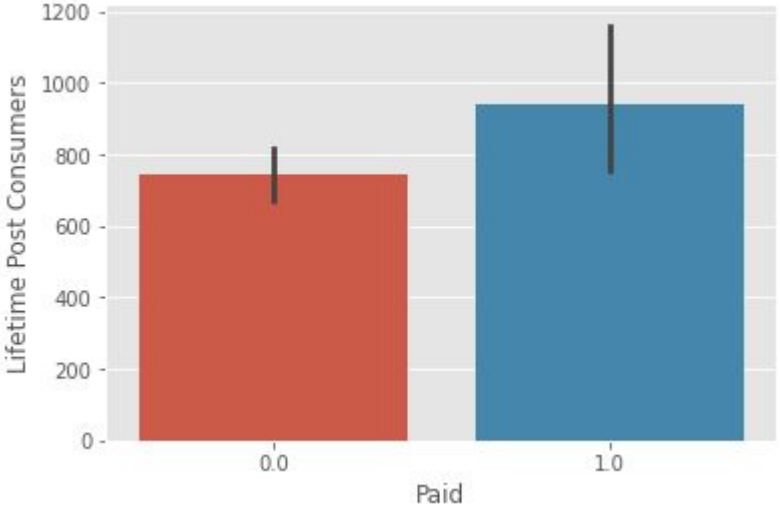


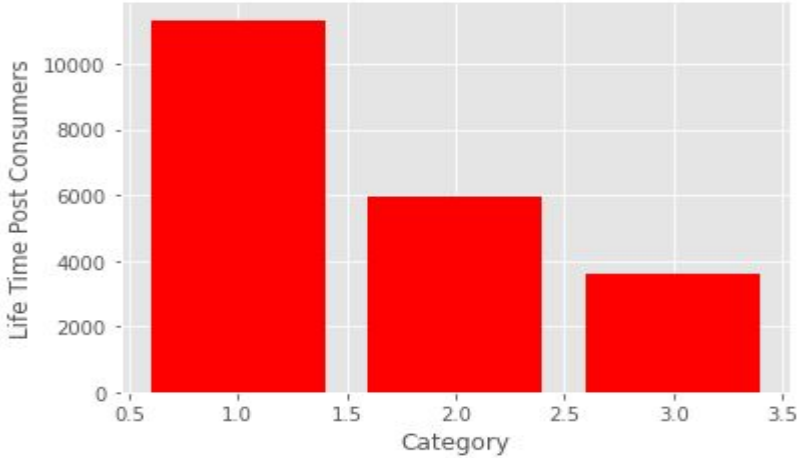




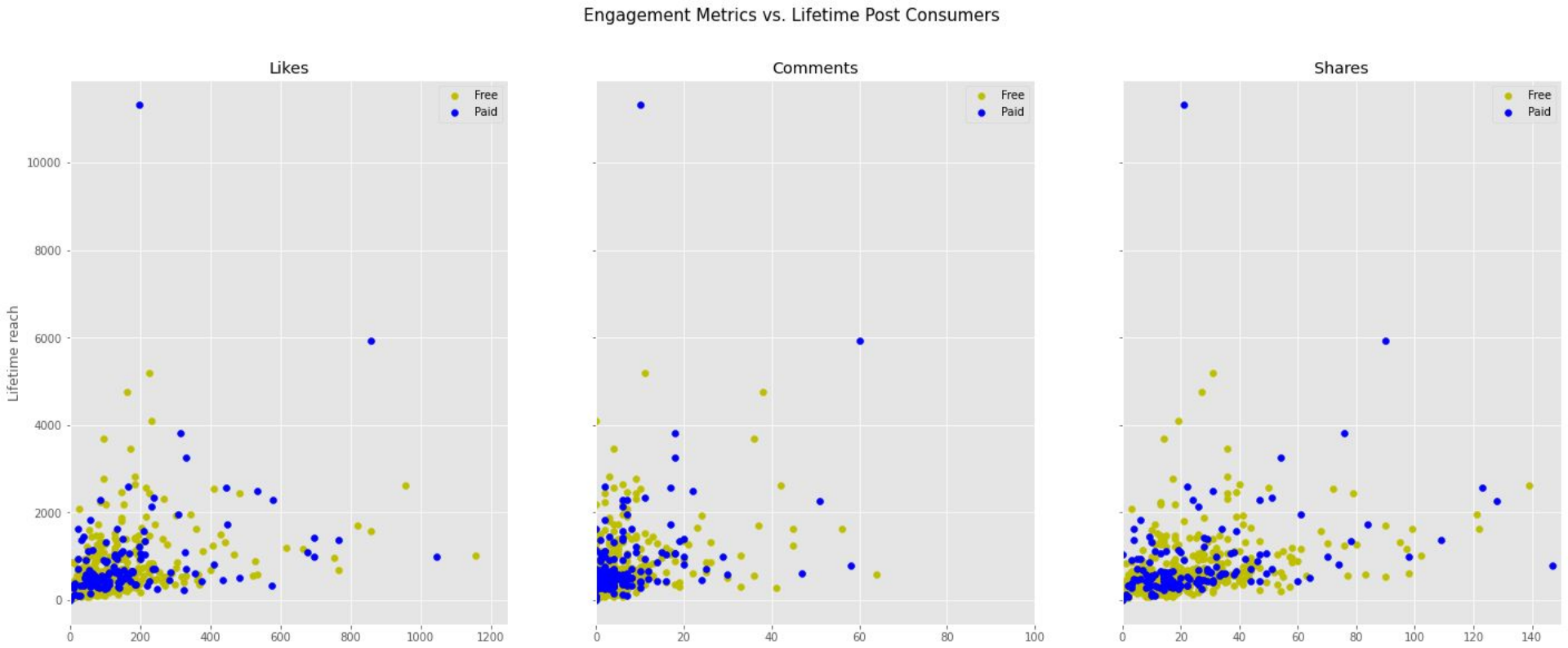


BAR PLOTS ON LIFE TIME POST CONSUMERS

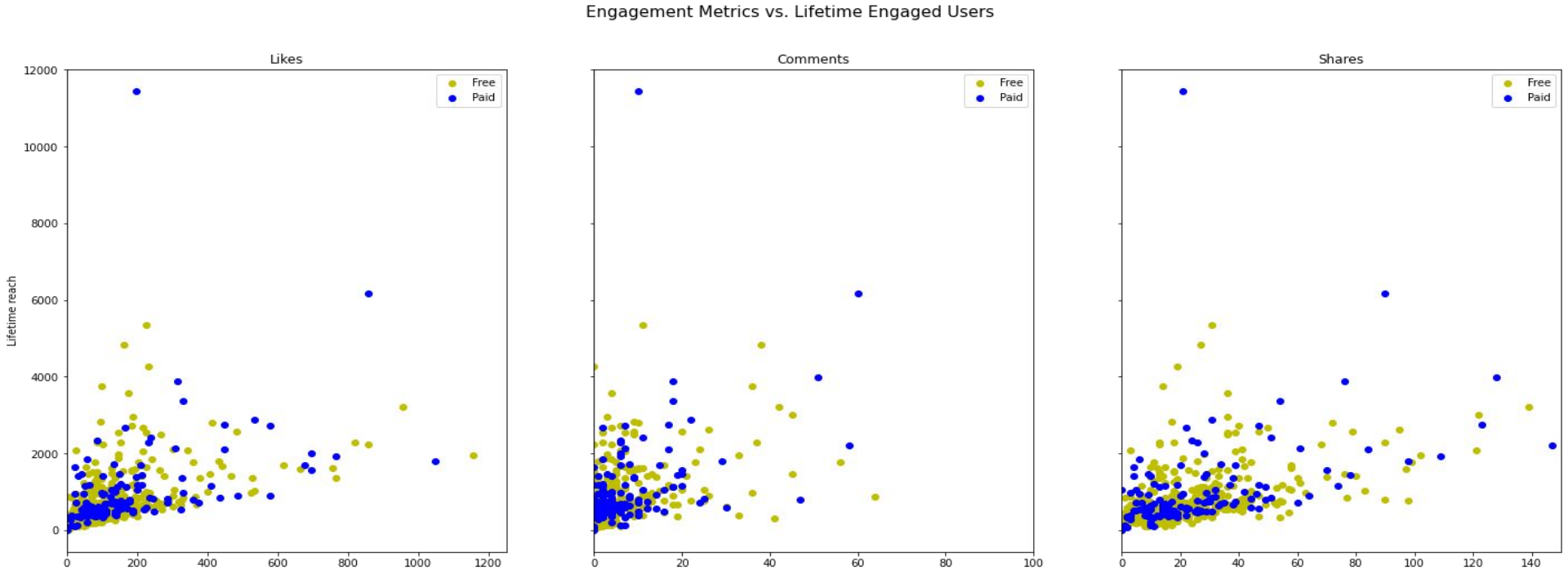


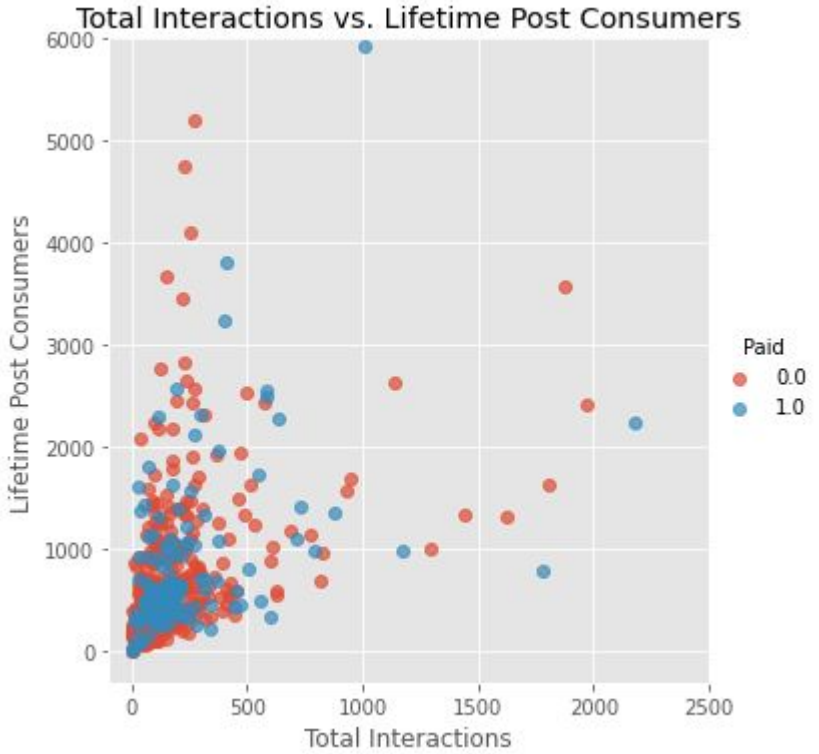


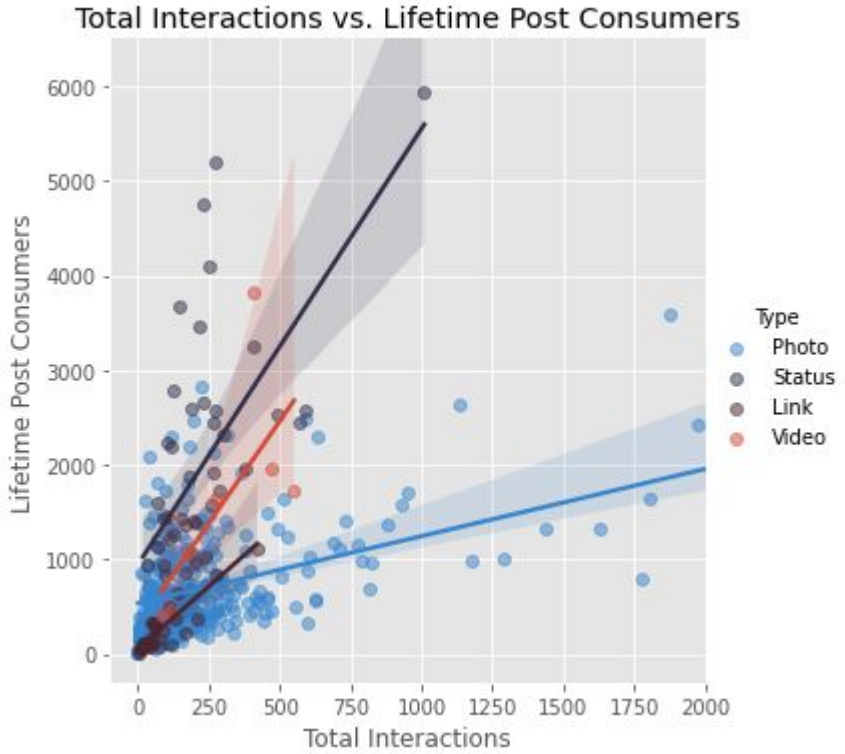
ENGAGEMENT METRICS VS LIFE TIME POST CONSUMERS



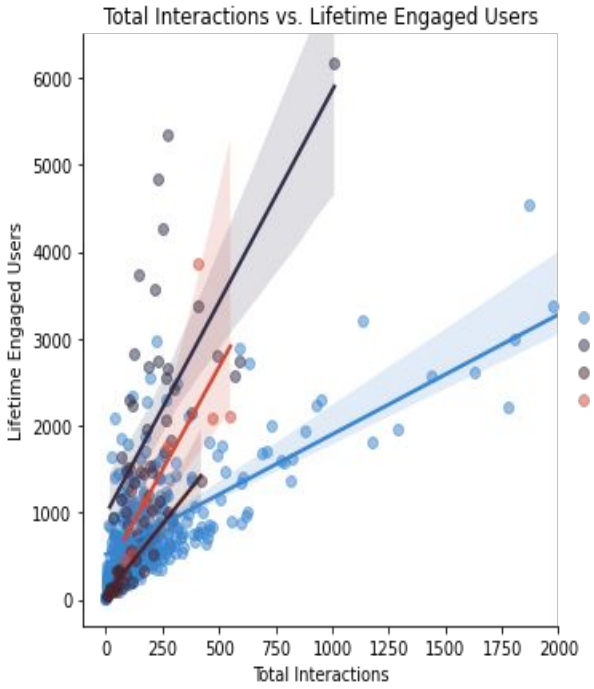
ENGAGEMENT METRICS VS LIFE TIME ENGAGED USERS

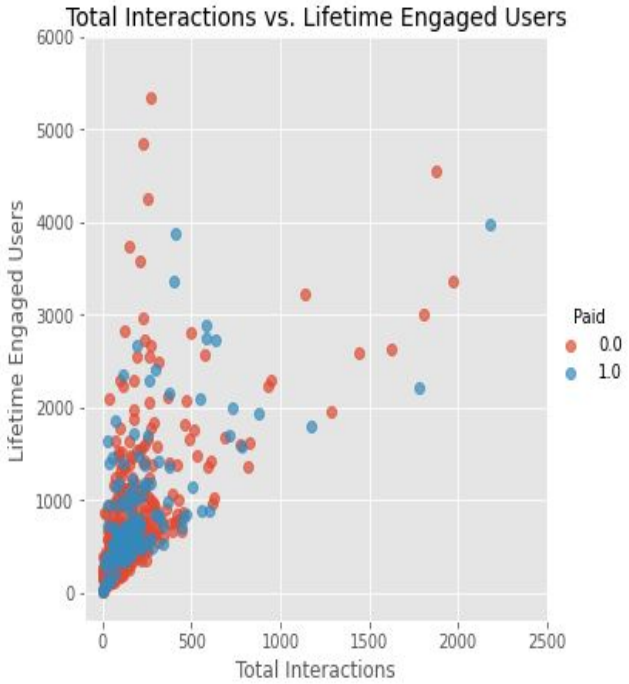


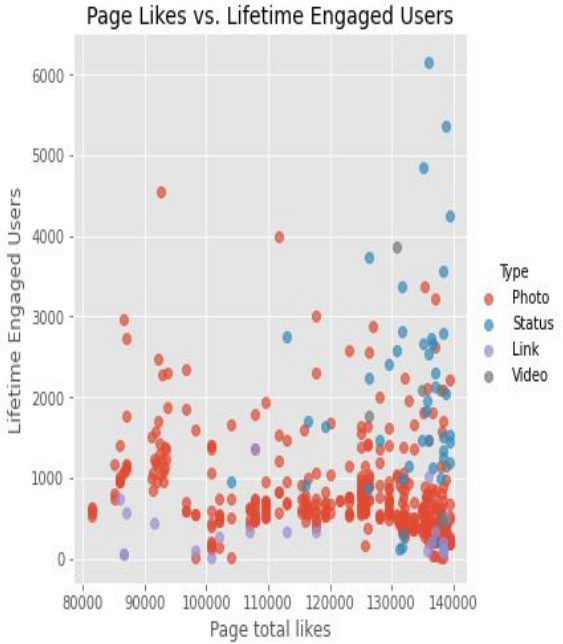




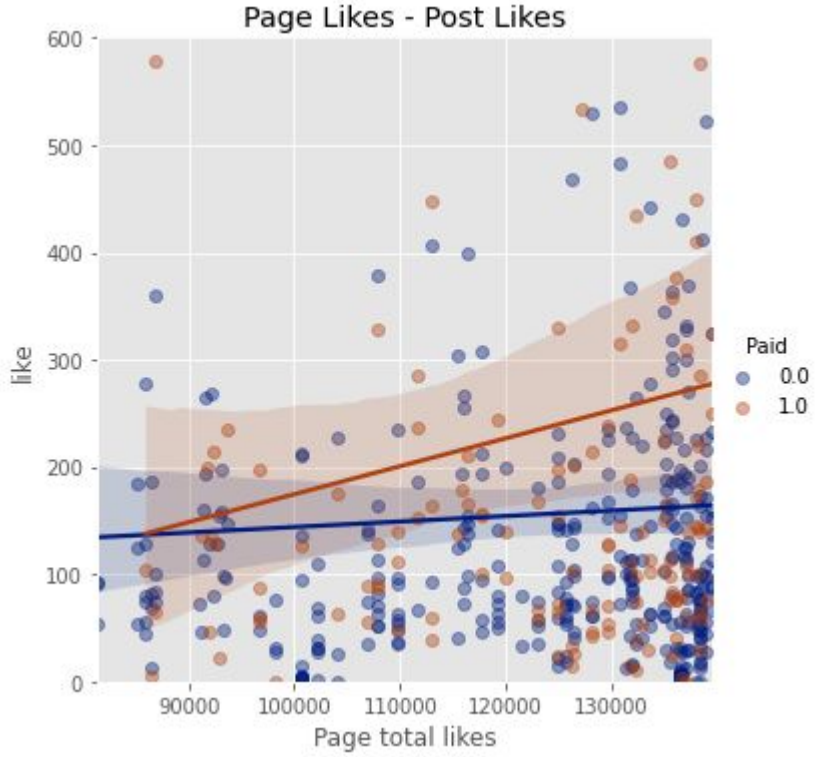
Interactions: consumers ratio is higher for photo than others





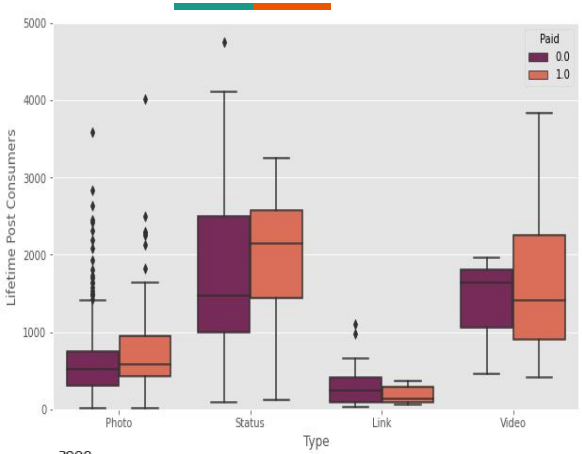


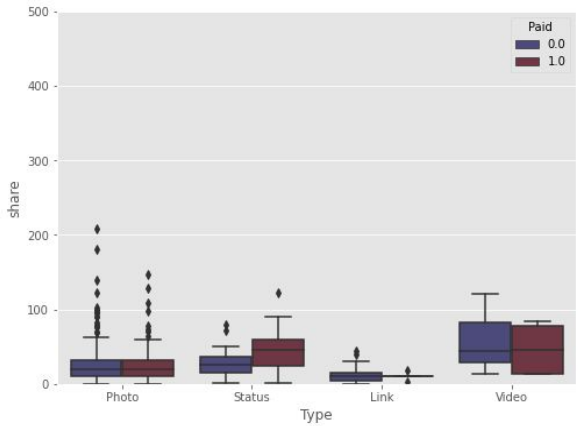
INTERACTIONS:ENGAGED USERS RATIO IS HIGHER FOR USERS THAN OTHERS

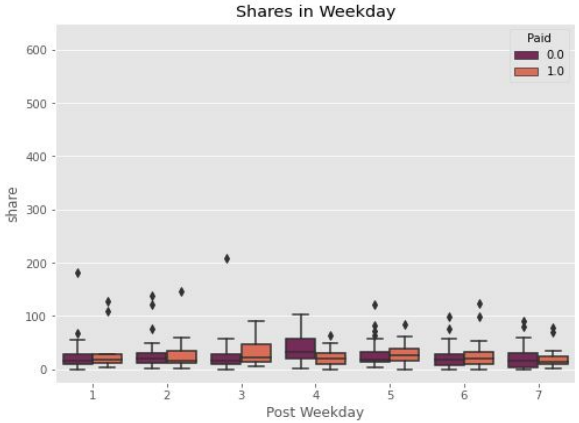


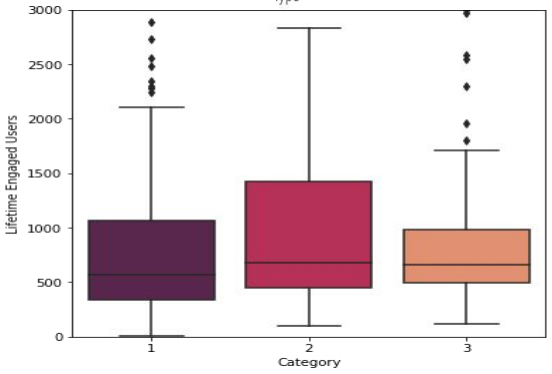
NO INCREASE IN LIKES FOR NON-PAID POSTS, WEAK POSITIVE TREND FOR PAID POSTS

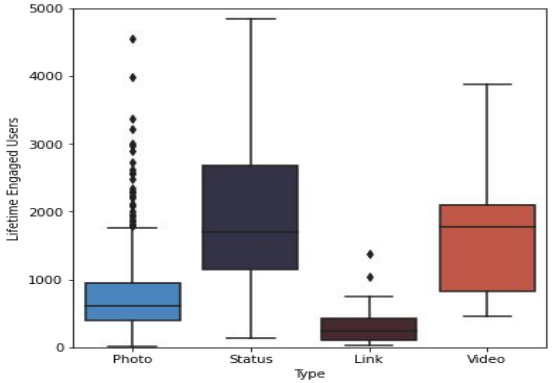
BOX PLOTS

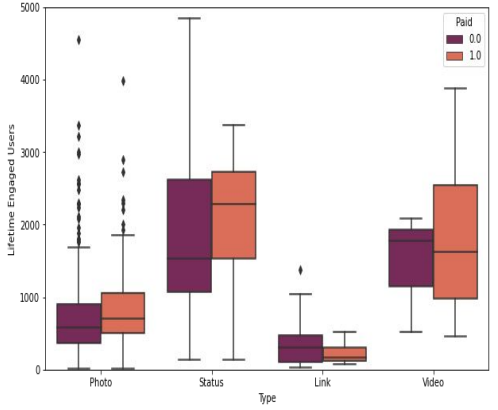

Paid
0.0
1.0
Lifetime Post Consumers
Photo
Status
Link
Video
Type


Paid
0.0
1.0
share
Photo
Status
Link
Video
Type

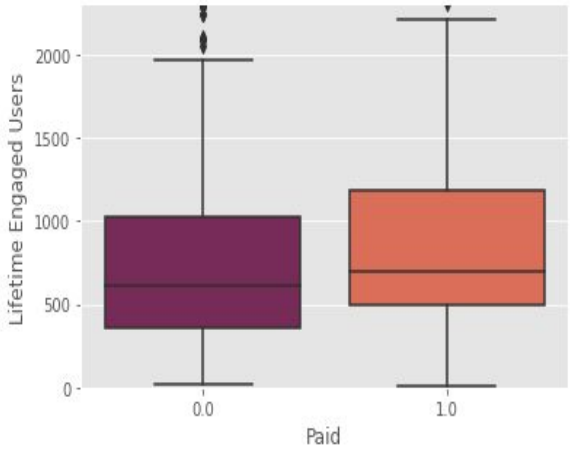

Shares in Weekday
Paid
0.0
1.0
share
Post Weekday

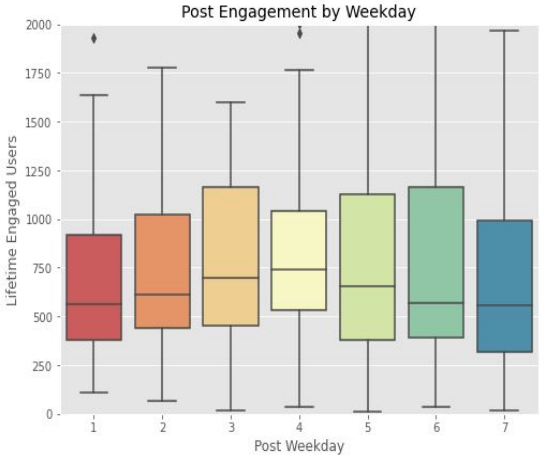

Lifetime Engaged Users
Category

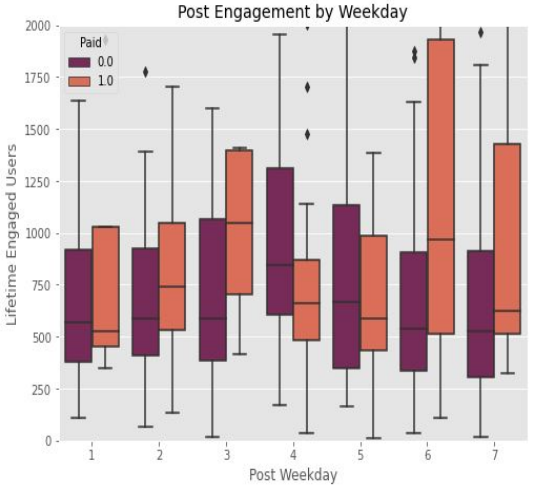

Lifetime Engaged Users
Photo
Status
Link
Video
Type

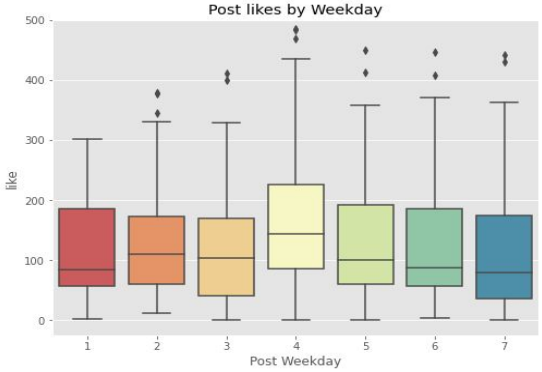

Paid
0.0
1.0
Lifetime Engaged Users
Photo
Status
Link
Video
Type

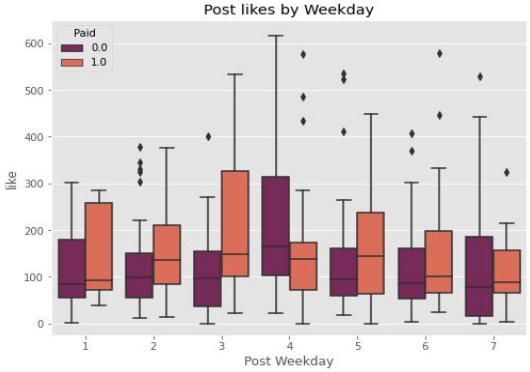
BOX PLOTS

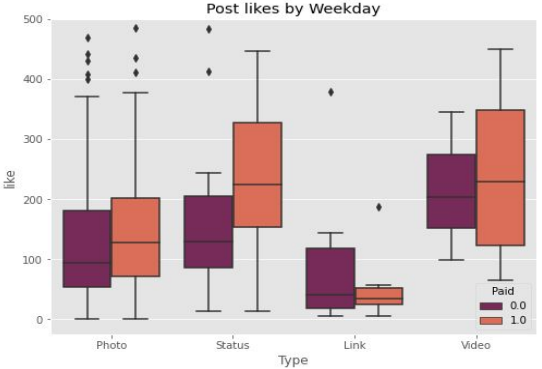

Lifetime Engaged Users
2000
1500
1000
500
0
0.0
1.0
Paid


Post Engagement by Weekday
Lifetime Engaged Users
2000
1750
1500
1250
1000
750
500
250
0
1
2
3
4
5
6
7
Post Weekday

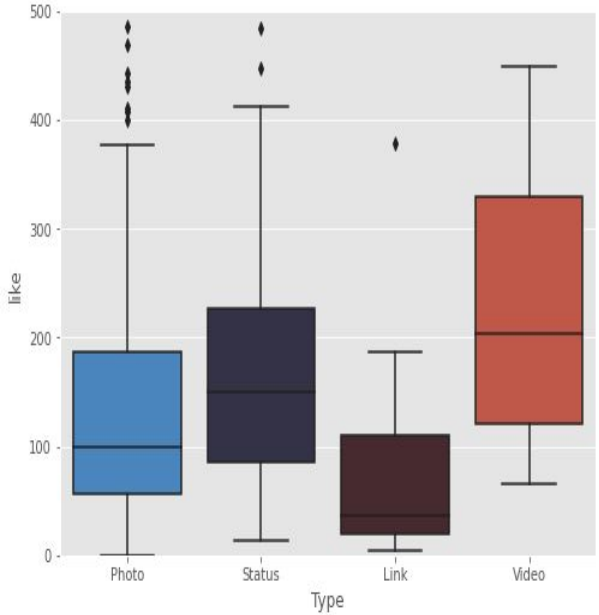

Post Engagement by Weekday
Paid
0.0
1.0
Lifetime Engaged Users
2000
1750
1500
1250
1000
750
500
250
0
1
2
3
4
5
6
7
Post Weekday

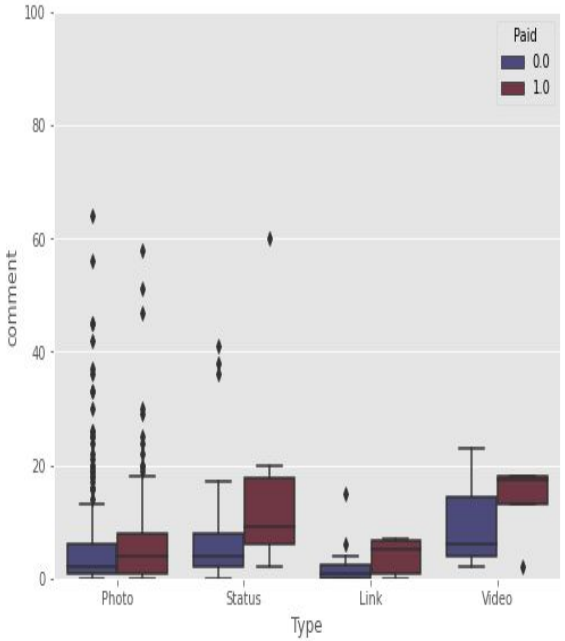

Post likes by Weekday
like
500
400
300
200
100
0
1
2
3
4
5
6
7
Post Weekday

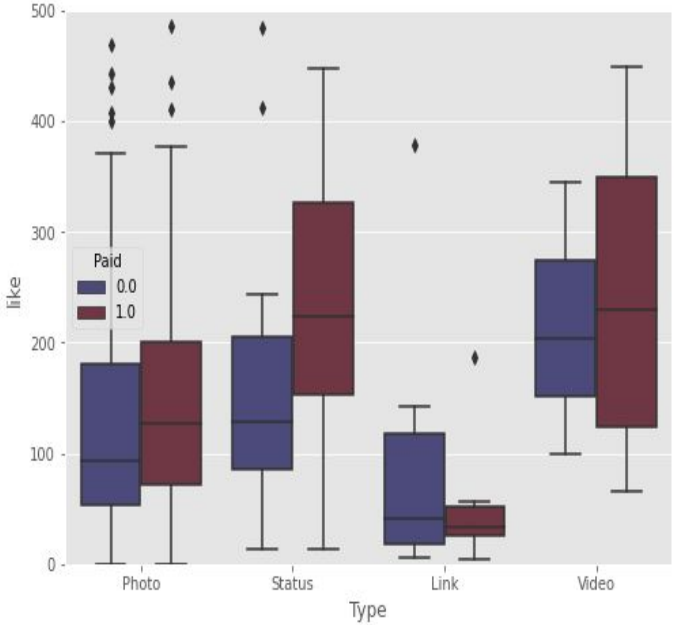

Post likes by Weekday
Paid
0.0
1.0
like
600
500
400
300
200
100
0
1
2
3
4
5
6
7
Post Weekday


Post likes by Weekday
like
500
400
300
200
100
0
Photo
Status
Link
Video
Type
Paid
0.0
1.0

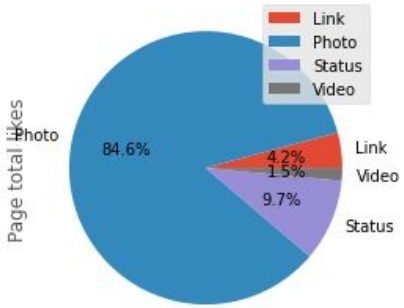
BOX PLOTS

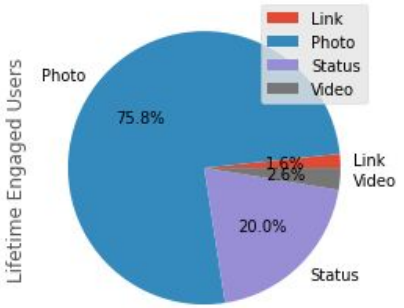

like
500
400
300
200
100
0
Photo
Status
Link
Video
Type

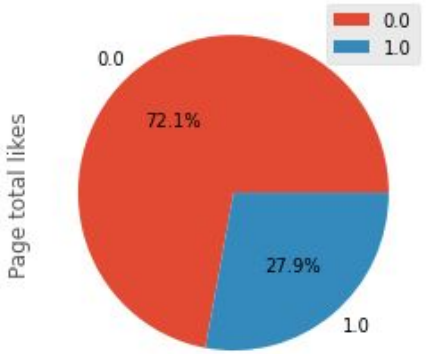

comment
100
80
60
40
20
0
Paid
0.0
1.0
Photo
Status
Link
Video
Type

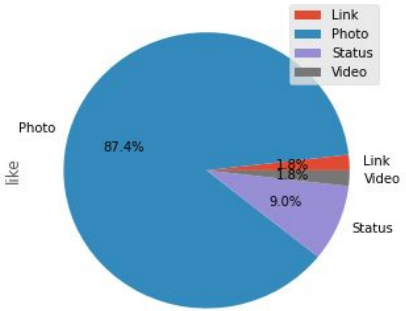

like
500
400
300
200
100
0
Paid
0.0
1.0
Photo
Status
Link
Video
Type

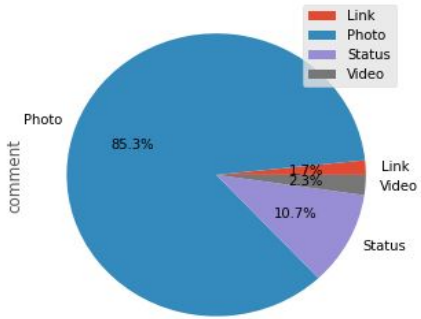
PIE CHARTS

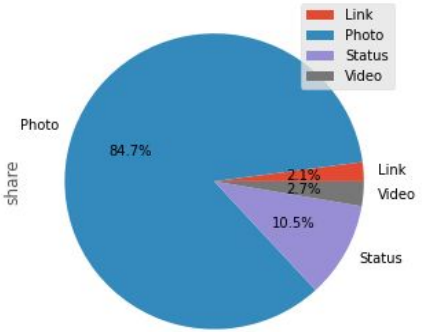

Page total likes
Link
Photo
Status
Video
Photo
84.6%
4.2%
1.5%
9.7%
Link
Video
Status


Lifetime Engaged Users
Link
Photo
Status
Video
Photo
75.8%
1.6%
2.6%
20.0%
Link
Video
Status


0.0
1.0
0.0
72.1%
Page total likes
27.9%
1.0


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


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


share
Link
Photo
Status
Video
Photo
84.7%
2.1%
2.7%
10.5%
Link
Video
Status

OBSERVATIONS

1. Engagement for Status > Video > photo > Link .
2. Interactions : Engaged Users ratio is higher for Photo than others .
3. Interactions : consumers ratio is higher for Photo than others .
4. Paid posts have slightly higher engagement than non-paid posts .
5. Paid posts get higher likes than non-paid posts.
6. No increase in likes for non-paid posts, weak positive trend for paid posts.

Regression

Preprocessing

1. Dropping columns with NaN.
   5 rows were dropped.

2. Remove Outliers.

3. One-Hot encoding for categorical variables like Post month, Post weekday, Post hour, Category.

Model 1

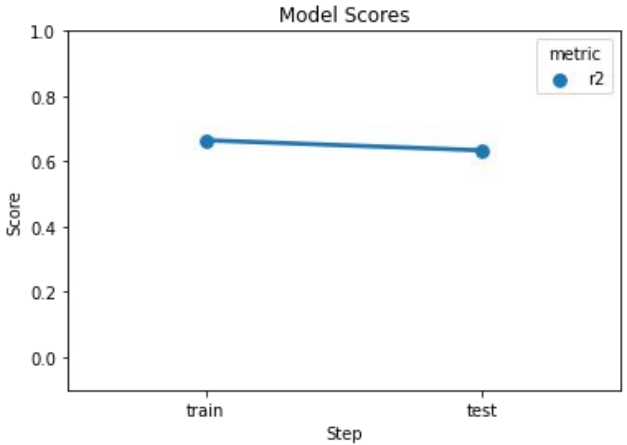
Regressor Variables

X : Page Total Likes,
Type
Category
Post Month
Post Hour
Post Weekday
Paid
Total Interactions

Dependent Variable

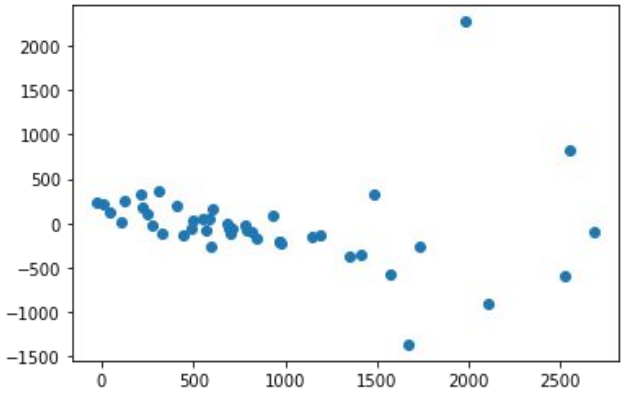
Y : Lifetime
Engaged
Users

R-squared Value



After applying Multiple Linear Regression, we get a **R-squared** value of **0.65**

Prediction vs Residual plot



The error increases as the number of ‘Engaged users increases’

Final Model

Lifetime Engaged Users：

3199
- (0.033 * Page Total Likes)
+ (1115 * Video) + (1745 * Status) + (332 * Photo)
+ (76.5 * Paid)
+ (302 * Category 1) + (103 * Category 2)
+ (130 * Mon) + (345 * Tues) + (200 * Other days)
+ (600 * Month)

Model 2

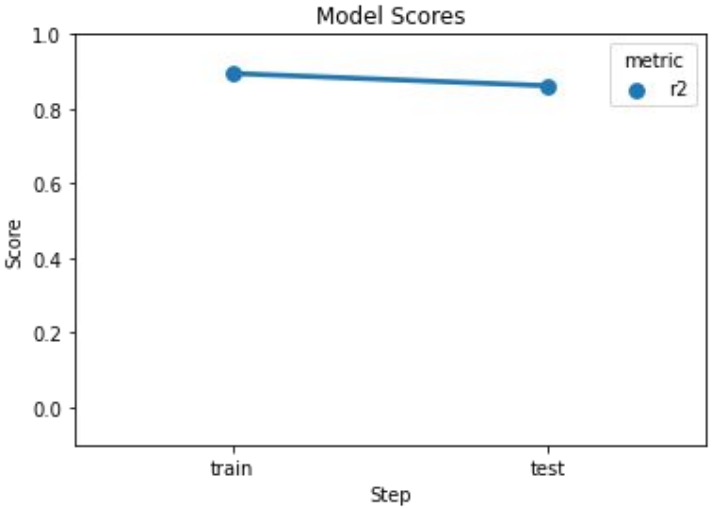
Regressor Variables

X : Total Interactions
Lifetime Engaged Users
Type
Category
Post Month
Post Hour
Post Weekday
Paid

Dependent Variable

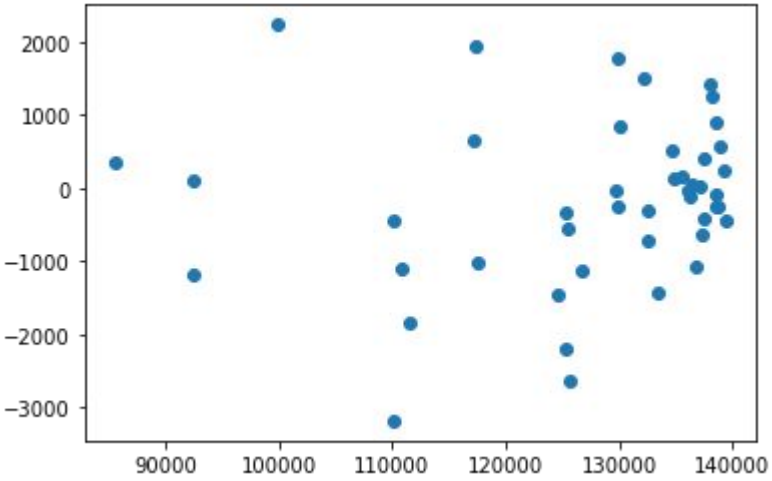
Y : Page
Total
Likes

R-squared Value



After applying Multiple Linear Regression, we get a **R-squared** value of **0.90**

Prediction vs Residual plot



The error decreases as the Page Total Likes increases

Final Model

Page Total Likes ：

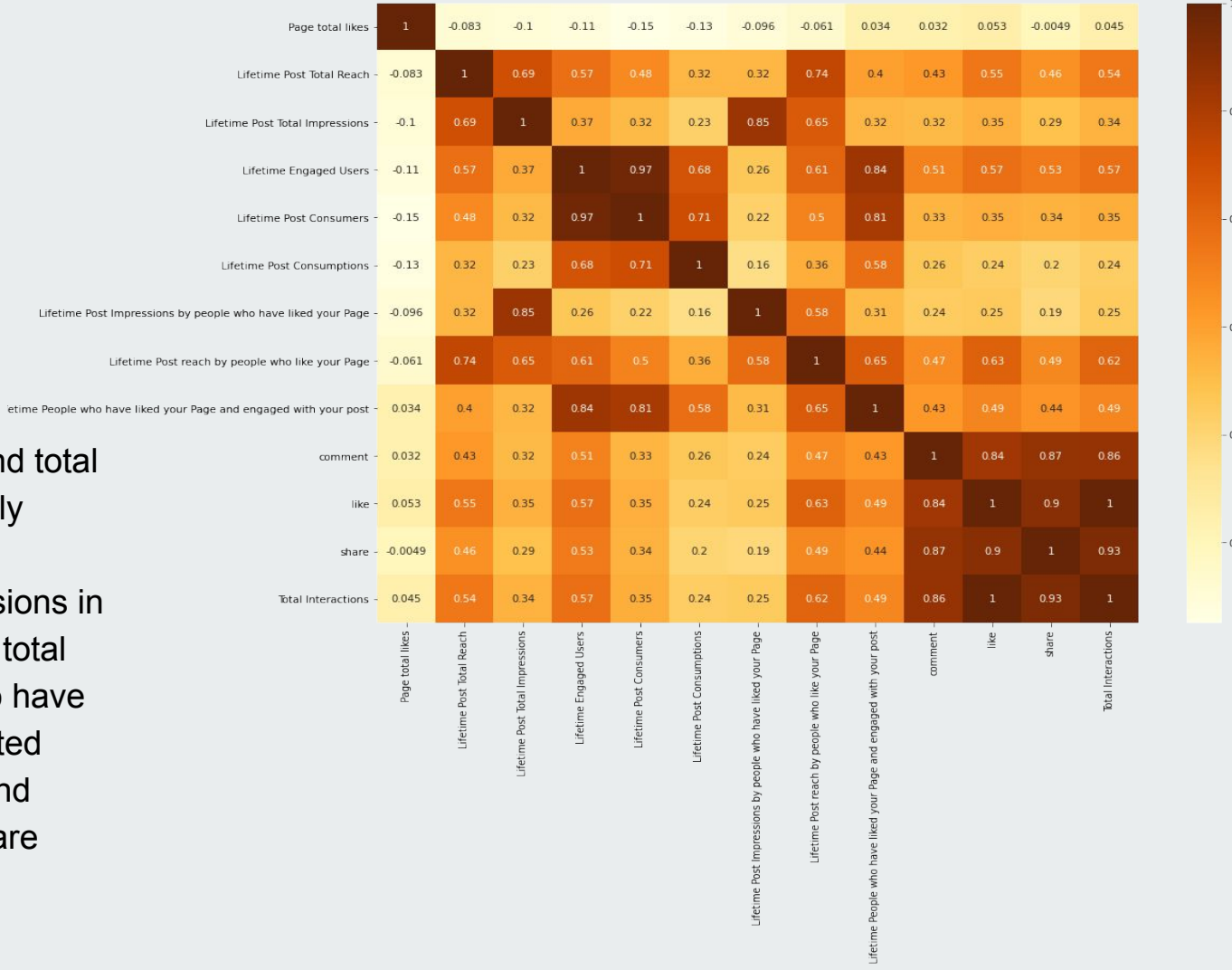
94470
+ (218 * Paid)
+ (308 * Video) + (714 * Status) + (260 * Photo)
+ (2.16 *Total Interactions)
- (0.347 * Lifetime Engaged Users)
+ (430 * Category 1) + (170 * Category 2) + (13300 * Weekday)

Principal Component Analysis

Definition

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.

Correlation of features (Omitting Categorical Data)



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

Method

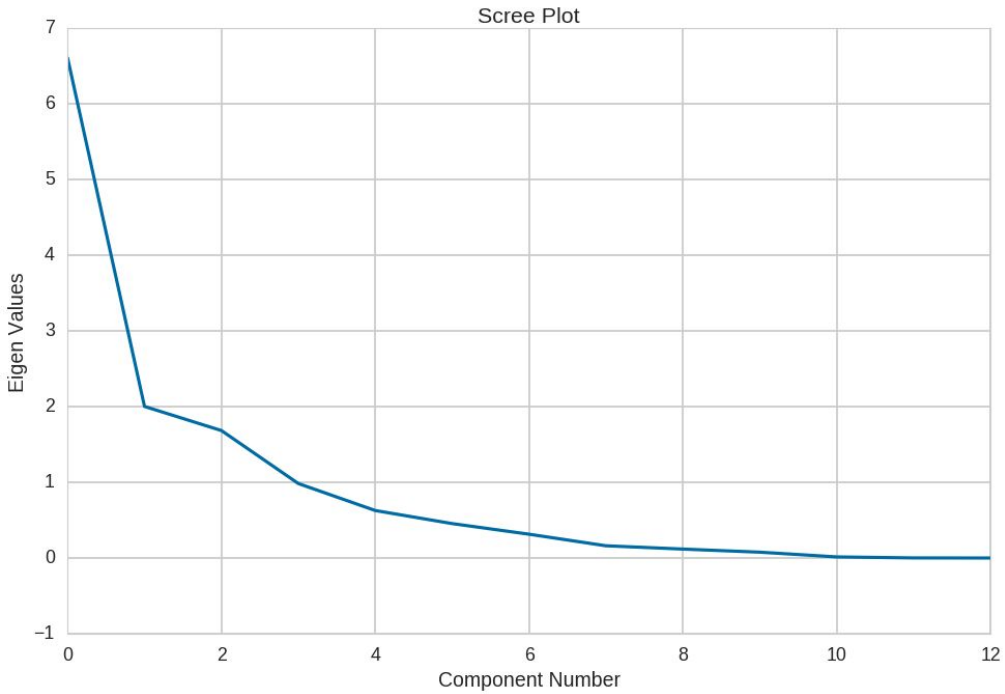
- The data in general has lots of correlated features
- So PCA was applied and Dataset was also normalized
- Covariance matrix was calculated
- Eigen Values and Eigen Vectors were found out

Eigen Values and Vectors

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,

Eigen Values	Eigen Vectors
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,

Scree Plot



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

References

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