1. Data Description

The following dataset is data that relates to the statistics of a specific Facebook page, pertaining particularly to the interactions that its posts receive. The statistics include categories such as number of likes, the type of post, when it was posted (including hour, day, and month), as well as if the post was paid for in order to boost interactions. Below is an extract of the first 3 rows from the data.

Type	PostHour	Paid	comment	like	share	TotalInteractions
{'Photo' }	3	0	4	79	17	100
{'Status'}	10	0	5	130	29	164
{'Photo' }	3	0	0	66	14	80

In this assignment, we will be analyzing the following columns: type, post hour, paid, comment, like, share, and total interactions. We will try and understand correlations such as: if someone likes a post, how likely are they to comment on it? How does the hour of posting something affect how many people will interact with it? If someone likes a post, how likely are they to share it?

2. Descriptive Analysis

After cleaning up the data and deciding upon the seven columns we are interested in analyzing we showcased the first three rows only, since the table has a large amount of data. We then took five of those columns excluding post type and paid, and computed the mean, minimum, and maximum and created a histogram for each one.

The first column post type is represented by either a photo, status, or link.

For the second column post hour, the mean is 7.84, maximum is 23, and minimum is 1. The histogram for this column represents a bimodal distribution since it has two peaks. Most posts were made either during the third hour of the day or the tenth hour of the day.

```
>> mean(T.PostHour)
max(T.PostHour)
min(T.PostHour)
histogram(T.PostHour)

ans =

    7.8400


ans =

    23


ans =

     1
```

Histogram is shown in figure 5.1

For the third column paid, the 0 represented unpaid and 1 represents paid.

For the fourth column comment, the mean is 7.482, maximum is 372, and minimum is 0. The histogram is positively skewed to the right as majority of the posts have very little or no comments. This can also be concluded upon since the mean is in single digits.

```
>> mean(T.comment)
max(T.comment)
min(T.comment)
histogram(T.comment)

ans =

    7.4820


ans =

   372


ans =

     0
```

Histogram is shown in figure 5.2

For the fifth column like, the mean is 177.59, maximum is 5172, and minimum is 0. The histogram is positively skewed to the right as majority of the posts are in the like range of 0-300. This can also be concluded upon since the mean is in between that range.

```
>> mean(T.like)
max(T.like)
min(T.like)
histogram(T.like)

ans =

  177.5900


ans =

        5172


ans =

     0
```

Histogram is shown in figure 5.3

For the sixth column share, the mean is 27.048, maximum is 790, and minimum is 0. The histogram is positively skewed to the right as majority of the posts are in the share range of 0-50. This can also be concluded upon since the mean is in between that range.

```
>> mean(T.share)
max(T.share)
min(T.share)
histogram(T.share)

ans =

   27.0480


ans =

   790


ans =

     0
```

Histogram is shown in figure 5.4

For the seventh column total interactions, the mean is 212.12, maximum is 6334, and minimum is 0. The histogram is positively skewed to the right as majority of the posts that were interacted with are in the range of 0-500. This can also be concluded upon since the mean is in between that range.

```
max(T.TotalInteractions)
min(T.TotalInteractions)
histogram(T.TotalInteractions)

ans =

  212.1200


ans =

        6334


ans =

     0
```

Histogram is shown in figure 5.5

3. Data Cleaning

To draw a clear picture of our data, we had to clean up our dataset and remove some categories and columns that we were not going to be using. This is because our data table was a 500x19 table. By cleaning up, we were able to paint a clear picture of our data and produce a concise data table that only included the data we planned to work on. The following photo is what our data table looked like before cleaning.

500×19 table

PageTotalLikes	Type	Category	PostMonth	PostWeekday	PostHour	Paid	LifetimePostTotalReach	LifetimePostTotalImpressions	LifetimeEngagedUsers	LifetimePostConsumers	LifetimePostConsump
1.3944e+05	{'Photo' }	2	12	4	3	0	2752	5091	178	109	159
1.3944e+05	{'Status'}	2	12	3	10	0	10460	19057	1457	1361	1674
1.3944e+05	{'Photo' }	3	12	3	3	0	2413	4373	177	113	154
1.3944e+05	{'Photo' }	2	12	2	10	1	50128	87991	2211	790	1119
1.3944e+05	{'Photo' }	2	12	2	3	0	7244	13594	671	410	580
1.3944e+05	{'Status'}	2	12	1	9	0	10472	20849	1191	1073	1389
1.3944e+05	{'Photo' }	3	12	1	3	1	11692	19479	481	265	364
1.3944e+05	{'Photo' }	3	12	7	9	1	13720	24137	537	232	305
1.3944e+05	{'Status'}	2	12	7	3	0	11844	22538	1530	1407	1692
1.3944e+05	{'Photo' }	3	12	6	10	0	4694	8668	280	183	250
1.3944e+05	{'Status'}	2	12	5	10	0	21744	42334	4258	4100	4540
1.3944e+05	{'Photo' }	2	12	5	10	0	3112	5590	208	127	145
1.3944e+05	{'Photo' }	2	12	5	10	0	2847	5133	193	115	133

We used the following lines of code to remove the columns we discarded:

```
T = readtable('Facebook.csv')
T.Properties.VariableNames
T = removevars(T,{'PostMonth','PostWeekday','PageTotalLikes','Category','LifetimePostTotalReach'});
T = removevars(T,{'LifetimePostTotalImpressions','LifetimeEngagedUsers','LifetimePostConsumers','LifetimePostConsumptions'});
T = removevars(T,{'LifetimePostImpressionsByPeopleWhoHaveLikedYourPage','LifetimePostReachByPeopleWhoLikeYourPage'});
T = removevars(T,{'LifetimePeopleWhoHaveLikedYourPageAndEngagedWithYourPost'});
T.Properties.VariableNames
vars = {'Paid','share','like'};
T2 = T{:,vars};
T2(isnan(T2)) = 0;
T{:,vars} = T2
writetable(T,'FacebookNew.csv')
```

This is the final output table that resulted from the aforementioned lines of code:

T =

500×7 table

Type	PostHour	Paid	comment	like	share	TotalInteractions
{'Photo' }	3	0	4	79	17	100
{'Status'}	10	0	5	130	29	164
{'Photo' }	3	0	0	66	14	80
{'Photo' }	10	1	58	1572	147	1777
{'Photo' }	3	0	19	325	49	393
{'Status'}	9	0	1	152	33	186
{'Photo' }	3	1	3	249	27	279
{'Photo' }	9	1	0	325	14	339

Data cleaning was an essential part of this project, as we needed to narrow down the columns we were going to work on. We were able to produce a 500x7 table, which was obtained from the initial and crowded 500x19 table.

4. Correlation Analysis

Based on the columns we decided we were going to work on, we came up with the following questions to answer with correlations:

1) If someone likes a post, how likely are they to share it?
2) If someone likes a post, how likely are they to comment on it?
3) If the page pays to advertise their post, will it receive more total interactions than a non-paid post?

4.1. Likes & Shares

To answer the question of "if someone likes a post, how likely are they to share it?", we decided to code a scatter plot of the columns "likes" and "shares". This plot shows us the distribution of likes and shares, and how closely correlated they are. The following screenshot is the code used for the scatterplot, as well as the correlation coefficient.

```
plot(T.like, T.share, 'Linestyle', 'none','Marker','.')
corrcoef(T.like, T.share)
```

The output of the correlation coefficient was the following screenshot, and the output of the scatterplot is shown in figure 5.1.

```
corrcoef(T.like, T.share)

ans =

    1.0000    0.9042
    0.9042    1.0000
```

The output shows us a correlation of 0.9042.

4.2. Likes & Comments

The question of "if someone likes a post, how likely are they to comment on it?" was answered in a similar fashion to the last one. We also drew a scatterplot and determined a correlation coefficient between columns "likes" and "comments". The following screenshot is the code used.

```
plot(T.like, T.comment, 'Linestyle', 'none','Marker','.')
corrcoef(T.like, T.comment)
```

The output of the correlation coefficient function is below, and the output of the scatterplot is shown in figure 5.2.

```
>> plot(T.like, T.comment, 'Linestyle', 'none','Marker','.')
corrcoef(T.like, T.comment)

ans =

    1.0000    0.8379
    0.8379    1.0000
```

The correlation coefficient between likes and comments is 0.8379.

4.3. Paying & Total Interactions

To determine if paying for a post to be advertised yields more results in terms of the total interactions a post receives, we decided to use a bar chart to represent it. We also determined a correlation coefficient between the "paid" and "total interactions" columns. Below is a screenshot of the code.

```
[a1,~,c1] = unique(T.Paid);
A1 = accumarray(c1(:),T.TotalInteractions(:));
bar(a1(:),A1(:));
corrcoef(T.Paid, T.TotalInteractions)
```

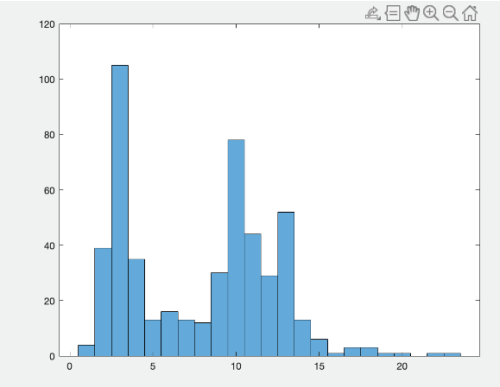
This is the output produced from the above line of code. The bar chart is represented in figure 5.3.

```
>> [a1,~,c1] = unique(T.Paid);
A1 = accumarray(c1(:),T.TotalInteractions(:));
bar(a1(:),A1(:));
corrcoef(T.Paid, T.TotalInteractions)

ans =

    1.0000    0.1080
    0.1080    1.0000
```

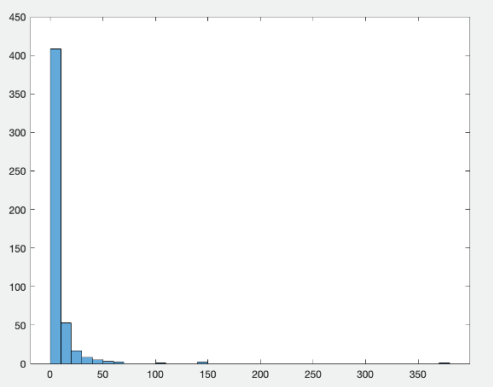
The correlation coefficient comes out as 0.1080.

5. Plots & Charts

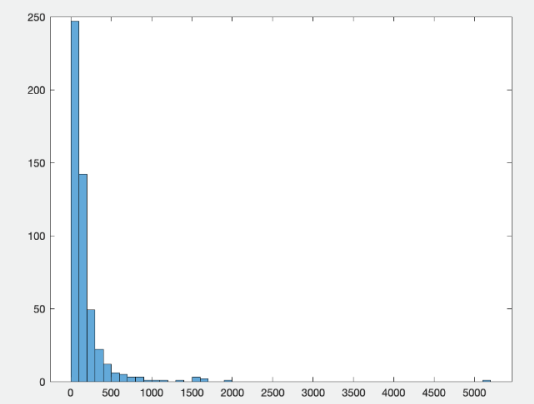
5.1. Post hour histogram



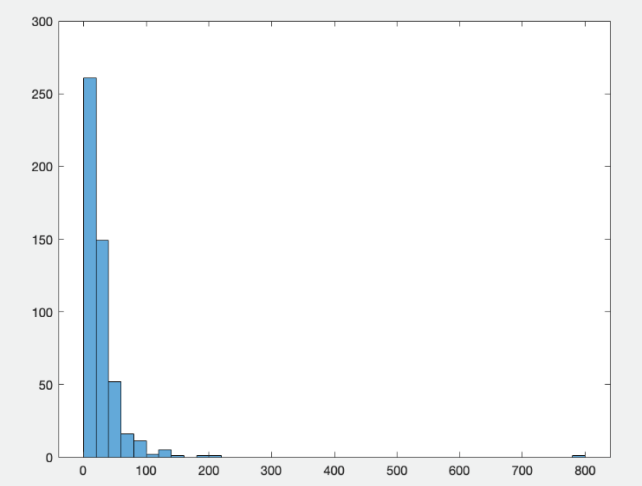
5.2. Comments histogram



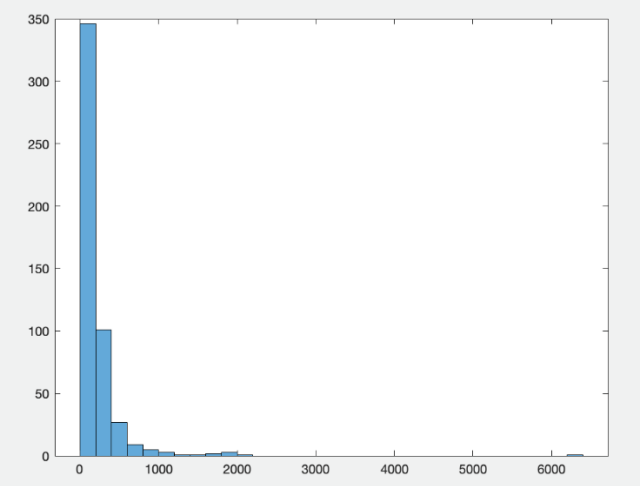
5.3. Likes histogram



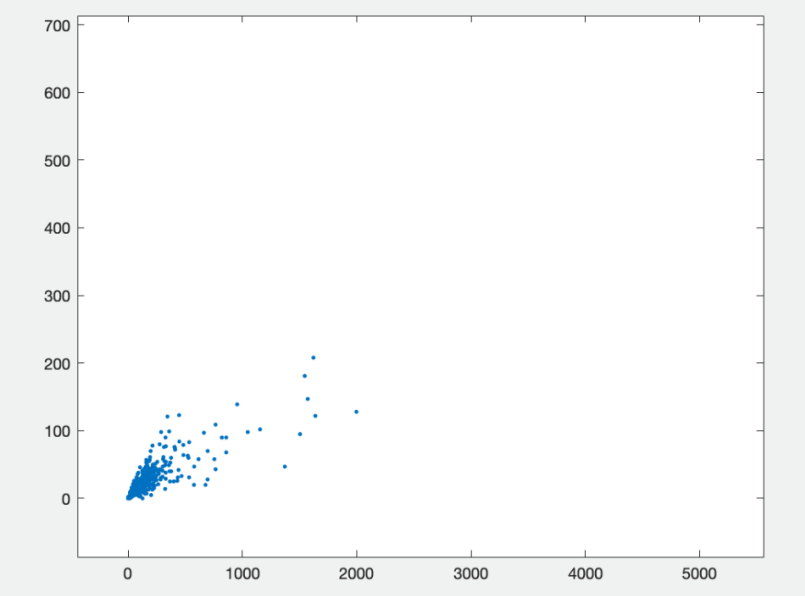
5.4. Shares histogram



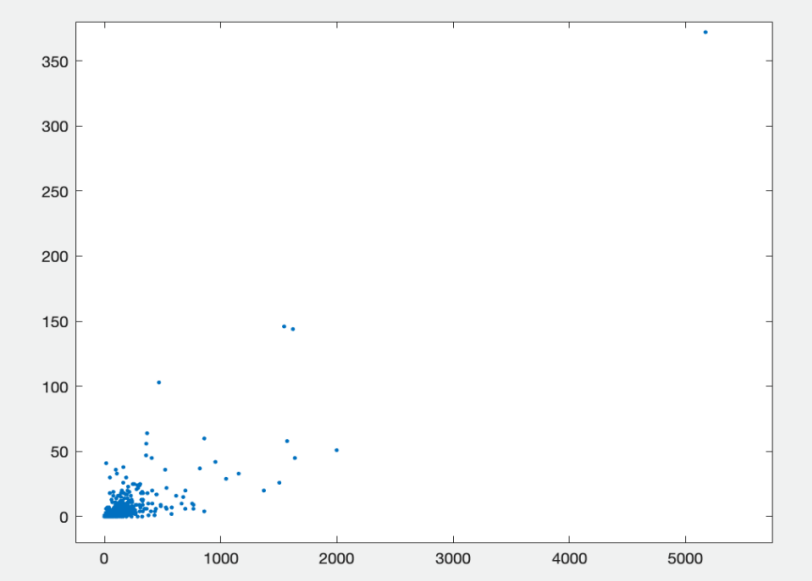
5.5. Total interactions histogram



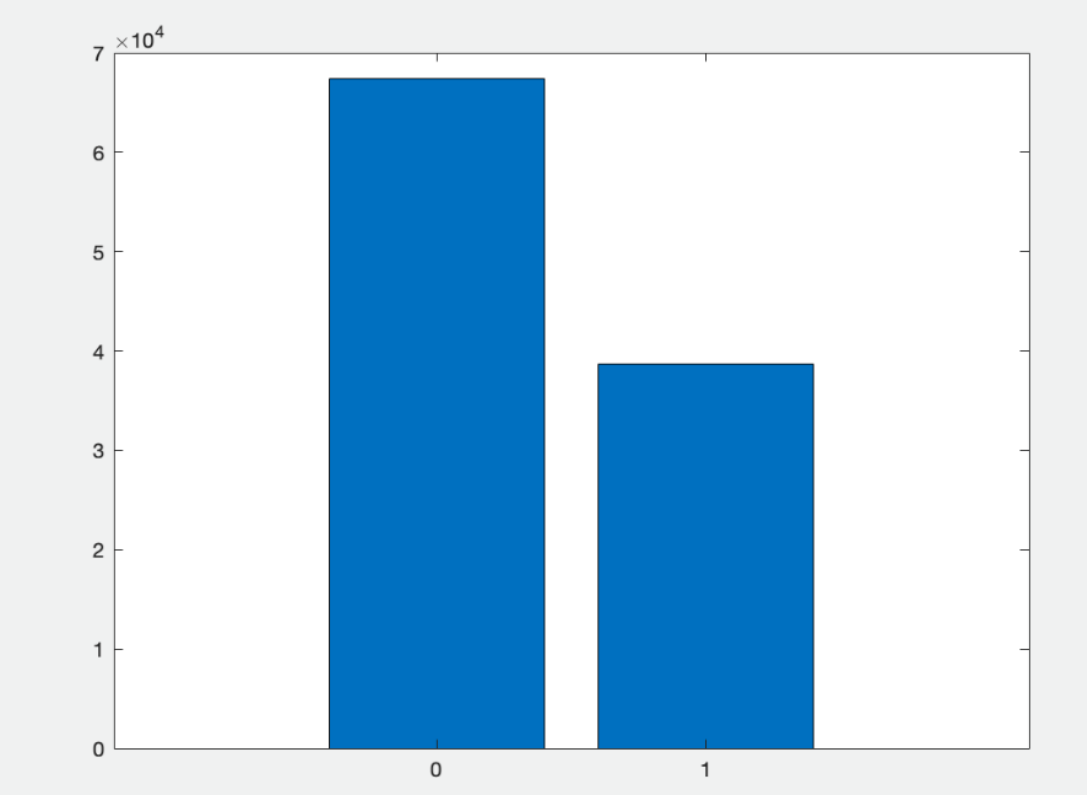
5.6. Likes & Shares scatterplot



5.7. Likes & Comments scatterplot



5.8. Paid & Total Interactions bar chart



6. Inferential Analysis

Using our data, we were able to draw inferences based on the correlation coefficients from our outputted code, as well as using the various plots and charts that were plotted.

For our first question, “if someone likes a post, how likely are they to share it?”, we found that the correlation between data in the “likes” and “shares” category had a correlation of 0.9042. This is a very strong correlation, essentially meaning that the likelihood that every like has a corresponding share is 90.42%. This strong correlation leads us to draw the inference that the more likes a post has, the more shares it will likely have. Looking at the scatterplot, we were able to see that while there were some outliers, they were not extreme outliers in any direction and generally followed the trend of more likes & more shares despite not falling directly on the line of best fit.

As for our second question, “if someone likes a post, how likely are they to comment on it?”, we found the correlation between “likes” and “comments” was 0.8379. While this is not as strong of a correlation as the previous question, it is still considered to be a significant positive correlation. Based on this, we can infer that there is a strong likelihood that the more likes a post receives, the more comments it will receive; but it is not as strong as the previous question. The scatterplot for this question also has few outliers, and the outliers also generally follow the inference that we drew.

The final question of “if the post is paid for, is it more likely to have more total interactions?” yielded an interesting answer. When looking at the bar chart, the total interactions for unpaid interactions vs paid interactions was significant, but in the opposite way we expected. It appears that the unpaid posts had significantly *more* interactions than the posts that had been paid for. This was confirmed by our correlation coefficient, which came out to be 0.1080. Oddly enough, there was a difference of around 3000 total interactions between the paid for posts and unpaid posts. Therefore, the inference we drew from this bar graph and correlation coefficient was that paying for a post did not end up yielding significant results in terms of total interactions for the post.