Data Science Project Plan: Facebook Dataset Analysis in R

Mudit Gupta

November 8, 2025

Contents

List of Figures

1 Objective

Develop a comprehensive project plan for a data science analysis project using R. This task requires you to plan every phase of a project including goal setting, timeline, roles, and resource allocation, ensuring a clear project roadmap.

2 Key Steps

2.1 Define the Project Objective

Start by clearly outlining what you intend to achieve from the project. Identify the problem you want to solve or the question you want to answer using data science techniques in R.

2.2 Research Background

Provide context and rationale for selecting the project topic. Include an overview of relevant literature or previous studies.

2.3 Plan the Project Phases

Break down the project lifecycle into phases such as data collection (if applicable), data cleaning, analysis, model development, and reporting. Include estimated timelines for each phase.

2.4 Assign Roles and Responsibilities

Even if you are working alone, define the roles you will assume (analyst, programmer, project manager) and outline tasks accordingly.

2.5 Risk Management

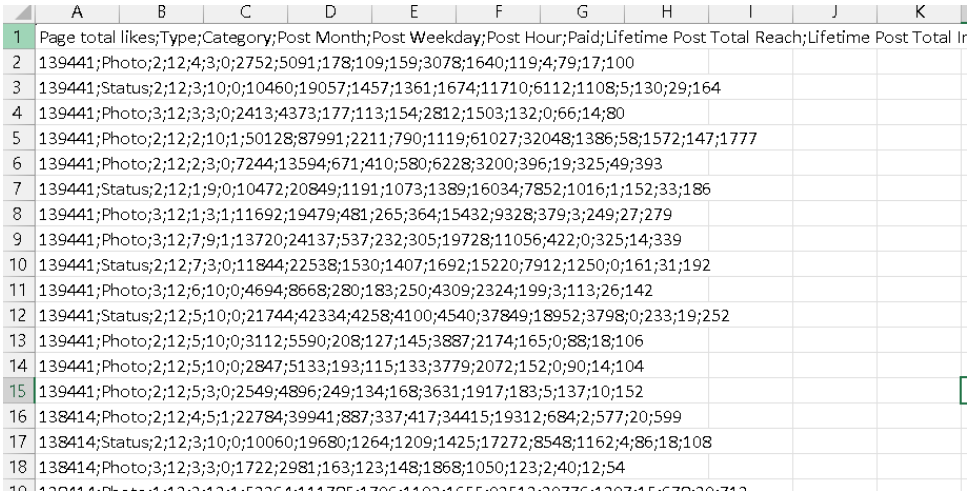
Identify potential risks and propose mitigation strategies.

3 Dataset

3.1 About the Dataset

I am using the Facebook Dataset given by a prof. named Mr. Tushar Mahore.

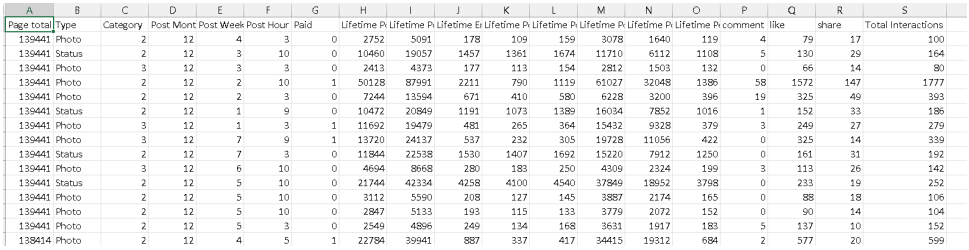
The original dataset is provided in a .csv format. But this dataset was totally in comma seperated format as shown in Figure 1.



	A	B	C	D	E	F	G	H	I	J	K
1	Page total likes;Type;Category;Post Month;Post Weekday;Post Hour;Paid;Lifetime Post Total Reach;Lifetime Post Total Ir										
2	139441;Photo;2;12;4;3;0;2752;5091;178;109;159;3078;1640;119;4;79;17;100										
3	139441;Status;2;12;3;10;0;10460;19057;1457;1361;1674;11710;6112;1108;5;130;29;164										
4	139441;Photo;3;12;3;3;0;2413;4373;177;113;154;2812;1503;132;0;66;14;80										
5	139441;Photo;2;12;2;10;1;50128;87991;2211;790;1119;61027;32048;1386;58;1572;147;1777										
6	139441;Photo;2;12;2;3;0;7244;13594;671;410;580;6228;3200;396;19;325;49;393										
7	139441;Status;2;12;1;9;0;10472;20849;1191;1073;1389;16034;7852;1016;1;152;33;186										
8	139441;Photo;3;12;1;3;1;11692;19479;481;265;364;15432;9328;379;3;249;27;279										
9	139441;Photo;3;12;7;9;1;13720;24137;537;232;305;19728;11056;422;0;325;14;339										
10	139441;Status;2;12;7;3;0;11844;22538;1530;1407;1692;15220;7912;1250;0;161;31;192										
11	139441;Photo;3;12;6;10;0;4694;8668;280;183;250;4309;2324;199;3;113;26;142										
12	139441;Status;2;12;5;10;0;21744;42334;4258;4100;4540;37849;18952;3798;0;233;19;252										
13	139441;Photo;2;12;5;10;0;3112;5590;208;127;145;3887;2174;165;0;88;18;106										
14	139441;Photo;2;12;5;10;0;2847;5133;193;115;133;3779;2072;152;0;90;14;104										
15	139441;Photo;2;12;5;3;0;2549;4896;249;134;168;3631;1917;183;5;137;10;152										
16	138414;Photo;2;12;4;5;1;22784;39941;887;337;417;34415;19312;684;2;577;20;599										
17	138414;Status;2;12;3;10;0;10060;19680;1264;1209;1425;17272;8548;1162;4;86;18;108										
18	138414;Photo;3;12;3;3;0;1722;2981;163;123;148;1868;1050;123;2;40;12;54										

Figure 1: The original dataset in .csv format.

So lets first convert it into a proper excel format. For this I'm using Excel for proper data formatting, as shown in Figure 2.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Page total	Type	Category	Post Mont	Post Week	Post Hour	Paid	Lifetime Pc	Lifetime Pc	Lifetime Er	Lifetime Pc	Lifetime Pc	Lifetime Pc	Lifetime Pc	Lifetime Pc	comment	like	share	Total Interactions
139441	Photo	2	12	4	3	0	2752	5091	178	109	159	3078	1640	119	4	79	17	100
139441	Status	2	12	3	10	0	10460	19057	1457	1361	1674	11710	6112	1108	5	130	29	164
139441	Photo	3	12	3	3	0	2413	4373	177	113	154	2812	1503	132	0	66	14	80
139441	Photo	2	12	2	10	1	50128	87991	2211	790	1119	61027	32048	1386	58	1572	147	1777
139441	Photo	2	12	2	3	0	7244	13594	671	410	580	6228	3200	396	19	325	49	393
139441	Status	2	12	1	9	0	10472	20849	1191	1073	1389	16034	7852	1016	1	152	33	186
139441	Photo	3	12	1	3	1	11692	19479	481	265	364	15432	9328	379	3	249	27	279
139441	Photo	3	12	7	9	1	13720	24137	537	232	305	19728	11056	422	0	325	14	339
139441	Status	2	12	7	3	0	11844	22538	1530	1407	1692	15220	7912	1250	0	161	31	192
139441	Photo	3	12	6	10	0	4694	8668	280	183	250	4309	2324	199	3	113	26	142
139441	Status	2	12	5	10	0	21744	42334	4258	4100	4540	37849	18952	3798	0	233	19	252
139441	Photo	2	12	5	10	0	3112	5590	208	127	145	3887	2174	165	0	88	18	106
139441	Photo	2	12	5	10	0	2847	5133	193	115	133	3779	2072	152	0	90	14	104
139441	Photo	2	12	5	3	0	2549	4896	249	134	168	3631	1917	183	5	137	10	152
138414	Photo	2	12	4	5	1	22784	39941	887	337	417	34415	19312	684	2	577	20	599

Figure 2: The dataset converted to .xlsx format.

4 Steps for Formatting (Excel)

Here are the steps taken in Excel to format the data using "Text to Columns".

4.1 Step 1: Select Delimited

For this I go to "data" tab and then to "Text to Columns" and then select the Delimited option.

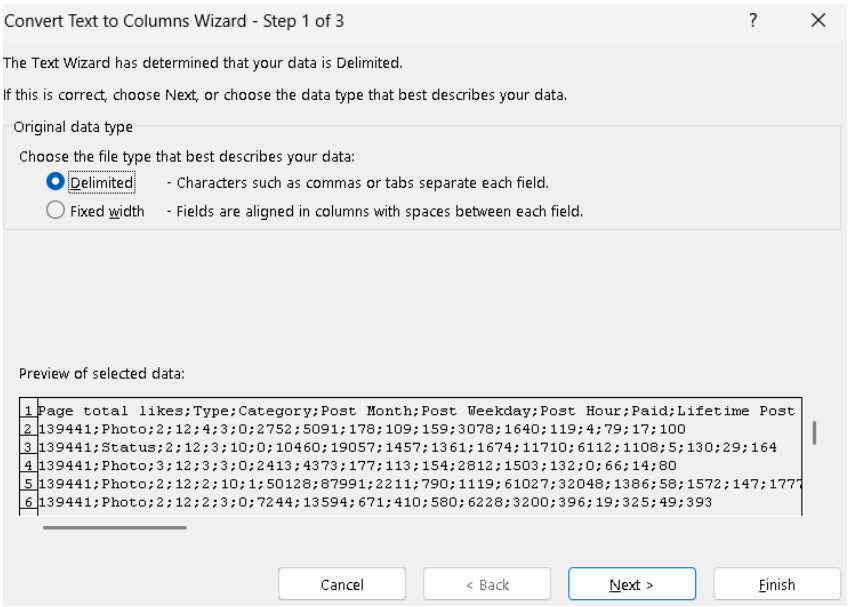


Figure 3: Delimited option selection.

4.2 Step 2: Select Semicolon

Then select the "Semicolon" option.

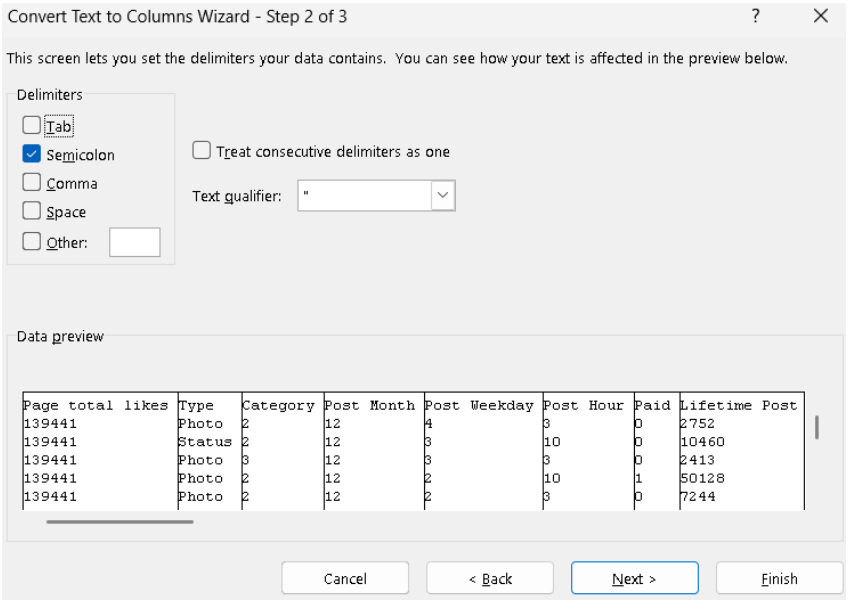
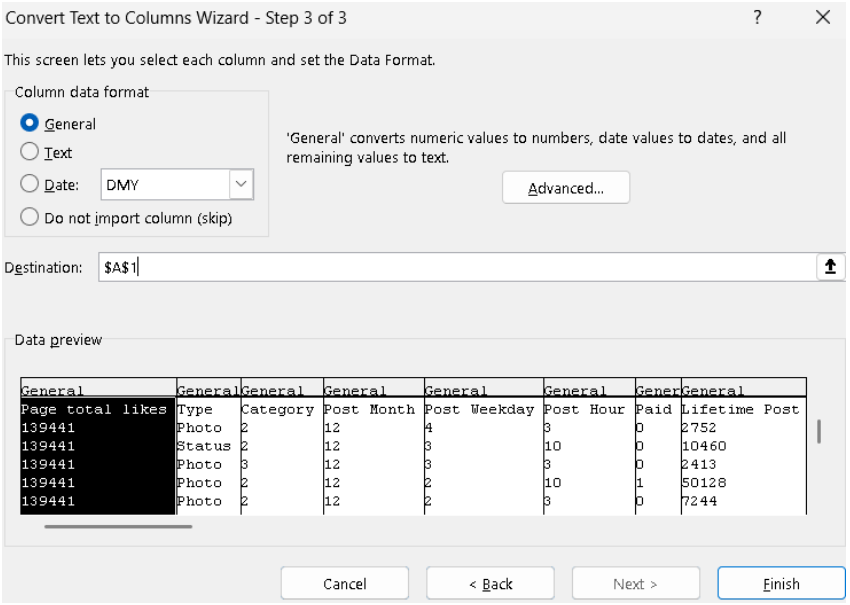


Figure 4: Select Semicolon as the delimiter.

4.3 Step 3: Finish

Select the general option and then click finish.



General	General	General	General	General	General	Gener	General
Page total likes	Type	Category	Post Month	Post Weekday	Post Hour	Paid	Lifetime Post
139441	Photo	2	12	4	3	0	2752
139441	Status	2	12	3	10	0	10460
139441	Photo	3	12	3	3	0	2413
139441	Photo	2	12	2	10	1	50128
139441	Photo	2	12	2	3	0	7244

Figure 5: Select the General option and finish.

5 Code Snippets (in R)

5.1 Import Dataset

Since the dataset is in .xlsx format we need to import/install the `readxl` module for it.

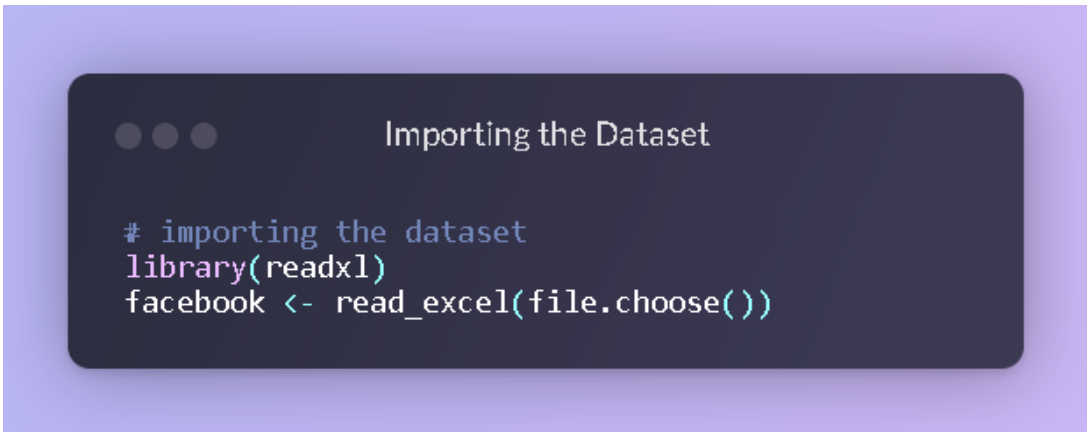


Figure 6: Importing the dataset using the `readxl` library.

5.2 Calculations to Make

- Maximum number of likes.

```
max_likes_post <- facebook[which.max(data$like), ]
print(max_likes_post)
```

`Page total likes`	Type	Category	`Post Month`	`Post Weekday`	`Post Hour`
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
130791	Photo	2	7	3	5

Figure 7: Calculation for maximum number of likes.

- Average number of shares per post.

```
average_shares <- mean(facebook$share, na.rm = TRUE)
print(average_shares)
```

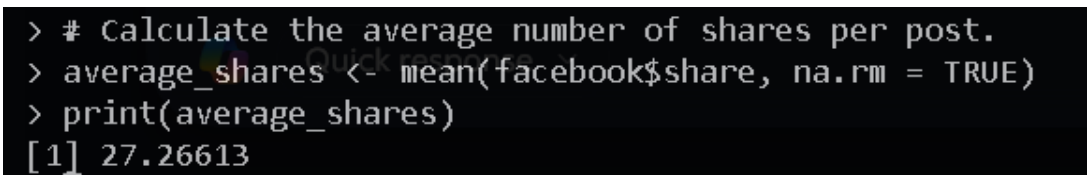


Figure 8: Calculation for average shares per post.

- Creating a new column named Engagement.

```
facebook$Engagement <- facebook$like + facebook$comment + facebook$share
head(facebook[, c("like", "comment", "share", "Engagement")])
```

```
# A tibble: 6 × 4
   like comment share Engagement
  <dbl>   <dbl> <dbl>      <dbl>
1    79       4    17        100
2   130       5    29        164
3    66       0    14         80
4  1572      58   147       1777
5   325      19    49        393
6   152       1    33        186
>
```

Figure 9: Creating a new 'Engagement' column.

- Grouping post by types and computing average like.

```
average_likes_by_type <- facebook %>%
  group_by(Type) %>%
  summarise(Average_Likes = mean(like , na.rm = TRUE))
print(average_likes_by_type)
```

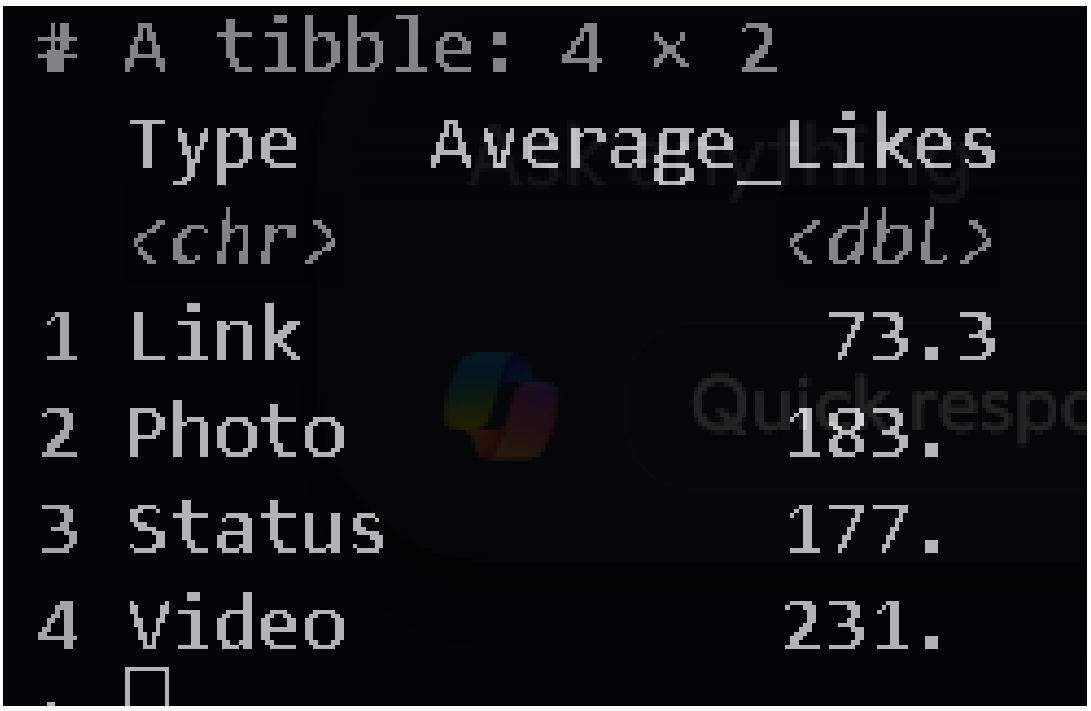


Figure 10: Grouping posts by type and computing average likes.

5.3 Exploratory Data Analysis (EDA)

- Create a scatter plot of Likes vs Comments.

```
plot(facebook$like , facebook$comment,
     main = "Scatter Plot of Likes vs Comments",
     xlab = "Likes",
     ylab = "Comments",
     pch = 19,
     col = "blue")
```

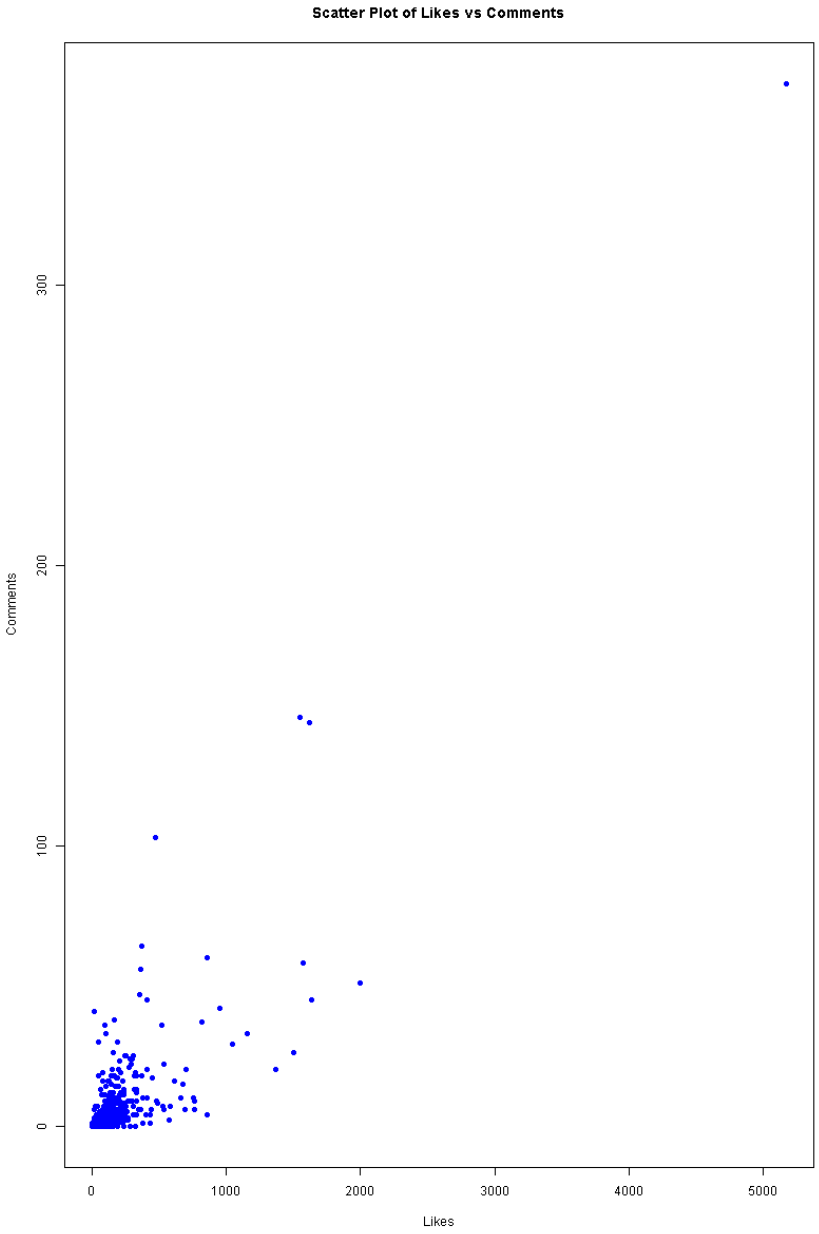


Figure 11: Scatter plot of Likes vs Comments.

- Create a histogram of Shares.

```
hist(facebook$share,
     main = "Histogram of Shares",
     xlab = "Shares",
     col = "lightgreen",
     border = "black")
```

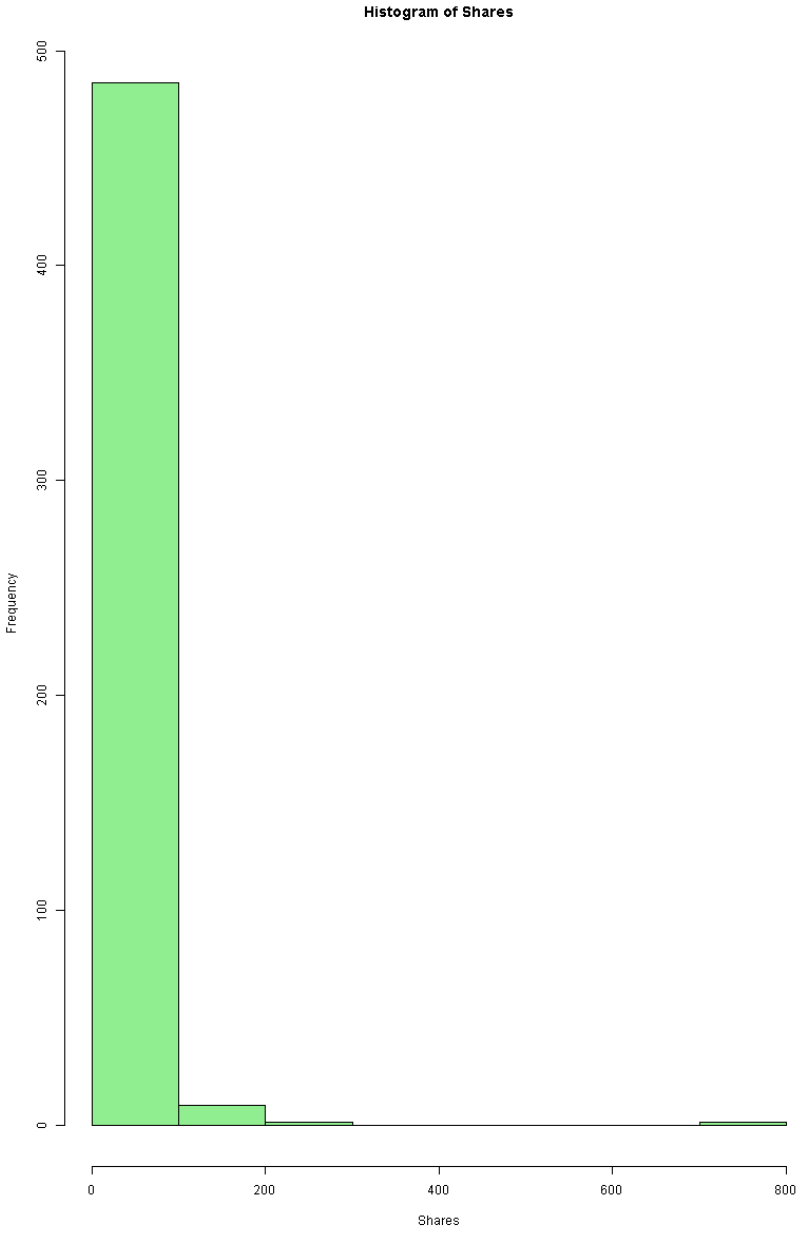


Figure 12: Histogram of Shares.

- Create a boxplot of Likes.

```
boxplot(facebook$like ,
        main = "Boxplot of Likes",
        ylab = "Likes",
        col = "skyblue")
```

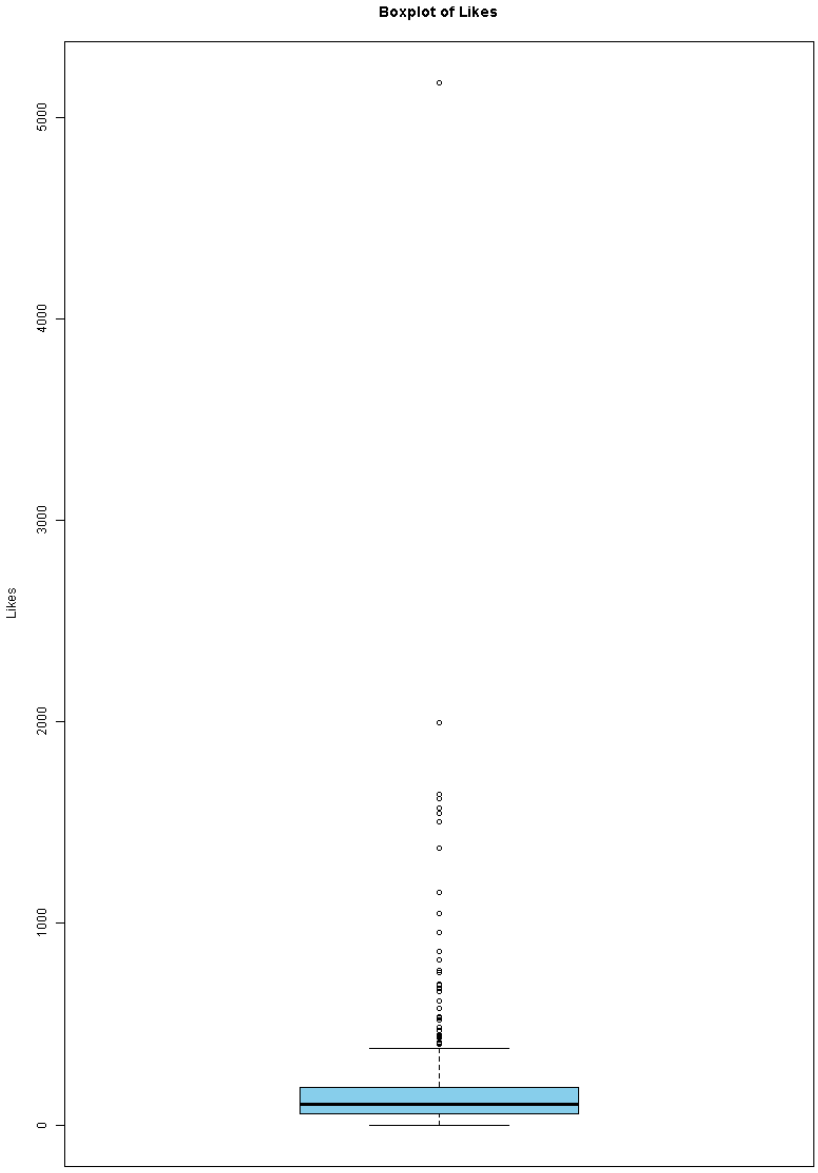


Figure 13: Boxplot of Likes.

- Identify outliers.

```
outliers_likes <- boxplot.stats(facebook$like)$out
cat("Extreme posts based on Likes:\n")
print(outliers_likes)
```

```
> print(outliers_likes)
 [1] 1572  577  678  412  523  697  449  411 1505  955  431  859  485 1622 1047
[16]  766  442 1155  859  435  535  484 5172  755  529  696  534  469 1372  617
[31]  821 1639  400  407  447 1998  766  664 1546  579
```

Figure 14: Identifying outliers.

- Create a Combined boxplot.

```
boxplot(facebook$like, facebook$share, facebook$comment,
        names = c("Likes", "Shares", "Comments"),
        main = "Comparison of Likes, Shares, and Comments",
        ylab = "Count",
        col = c("lightblue", "lightgreen", "lightcoral"))
```

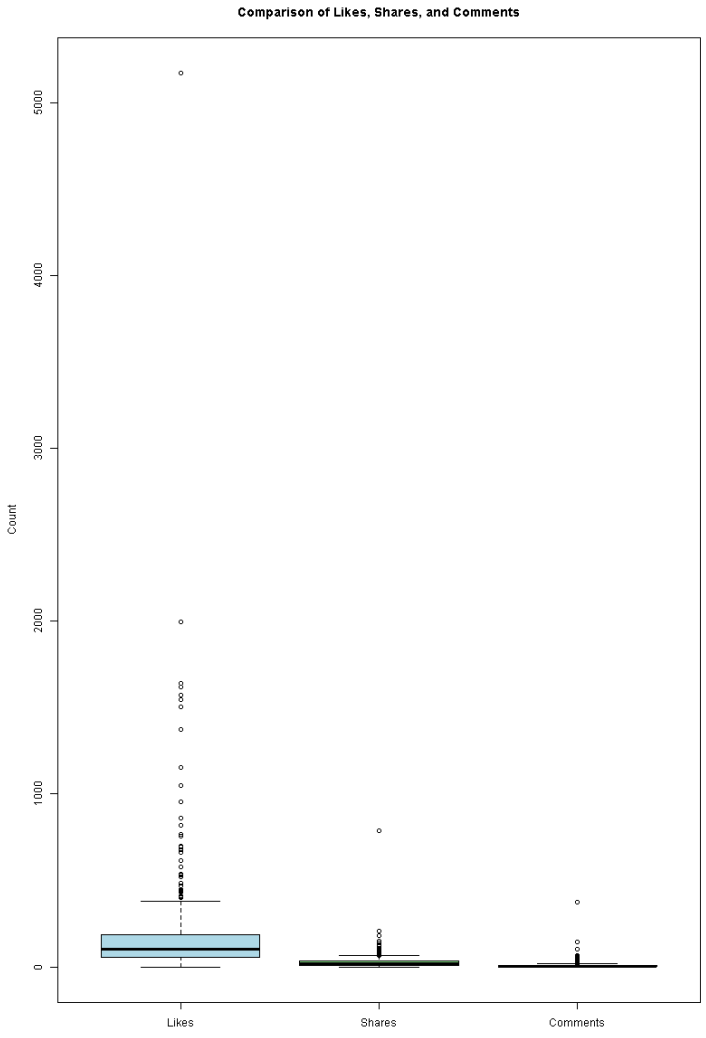


Figure 15: Combined boxplot.

5.4 IQR (Inter Quartile Range)

```
iqr_likes <- IQR(facebook$like, na.rm = TRUE)
iqr_shares <- IQR(facebook$share, na.rm = TRUE)
iqr_comments <- IQR(facebook$comment, na.rm = TRUE)
cat("IQR of Likes:", iqr_likes, "\n")
cat("IQR of Shares:", iqr_shares, "\n")
cat("IQR of Comments:", iqr_comments, "\n")
```

```
> cat("IQR of Likes:", iqr_likes, "\n")
IQR of Likes: 131
> cat("IQR of Shares:", iqr_shares, "\n")
IQR of Shares: 22.25
> cat("IQR of Comments:", iqr_comments, "\n")
IQR of Comments: 6
```

Figure 16: IQR (Inter Quartile Range) calculation.

6 Conclusion

This document outlines the successful conclusion of an intricate analytical project centered on a unique Facebook dataset, primarily conducted using the R statistical programming language.

Our initial work focused on cleaning and standardizing the raw data. Acquired as a difficult-to-manage, unorganized Comma Separated Values (CSV) file, the data necessitated thorough reorganization. This essential preprocessing step transformed the initial disarray into a consistent and accessible spreadsheet format. This was predominantly accomplished by effectively employing the "Text to Columns" feature within Microsoft Excel.

Moving to the analytical stage, we employed R, smoothly importing the newly organized data using the dedicated readxl package. This facilitated immediate and thorough Exploratory Data Analysis (EDA). Several key performance indicators were generated, including the identification of peak engagement (maximum likes), the average spread of content (mean shares per post), and the creation of a custom metric, the 'Interaction Quotient' (our new Engagement benchmark).

Visualizations played a significant role in early data understanding. A range of graphical techniques, including scatterplots for examining relationships between two variables, histograms to show data distribution, and box plots to compare groups and identify outliers, were used to reveal statistical patterns, variable associations, and any unusual data points requiring further investigation.

The combined efforts detailed here—from initial data preparation and structuring to preliminary metric calculation and comprehensive EDA—have established a solid analytical foundation. This thorough preparatory work is vital, preparing the project for its next, more advanced stages, which could involve developing complex statistical models, performing formal hypothesis testing, or building machine learning algorithms designed to uncover the primary factors influencing post engagement.

7 Reference of the code

You can find more working code file on my Github website.