EXPLORATORY DATA ANALYSIS PROJECT

70-207: Probability and Statistics for Business Applications

Alina Barmagambetova

Altynay Zhumatay

Danaiym Talantbekova

INTRODUCTION

In today's digital age, social media has turned into a big platform for communication and influence. Understanding what makes a Facebook post successful is not just a matter of curiosity; it's vital for businesses, content creators, and anyone seeking to engage their audience effectively. This exploratory data analysis (EDA) project aims to cover the dynamics of Facebook posts, examining three key hypotheses. By investigating the impact of post type, paid post, and post viewer count on interactions, we attempt to provide actionable insights that can empower individuals and businesses to enhance their social media strategies.

DATASET

Facebook.csv

1. Descriptions: *This dataset is related to predicting the performance metrics of posts published in brand's Facebook pages. Multiple performance metrics are in the dataset.*

2. Variable information:

The dataset called "Facebook.csv" contains detailed information about various aspects of Facebook posts, including post type (link, photo, status, video), post category (action, product, inspiration), whether the post was paid for (yes or no), as well as various metrics related to reach, engagement and interaction, such as the total number of unique users who saw the post, the number of times the post was shown, the number of users engaged and the total interaction (sum of likes, comments and shares). This rich data set allows for an in-depth analysis of how various factors affect the visibility and engagement of posts on Facebook pages.

The number columns - 15

The number of rows - 495

The dataset offers an understanding of the performance of Facebook posts, which includes reach, engagement, and interactions metrics.

The dataset's sample aims to reflect the Facebook post population. Still, we need to be aware of any potential biases or limitations in the sample that could affect how well it represents the entire set of posts.

PRESENTING HYPOTHESES

Hypothesis №1:

The type of Facebook post determines the number of interactions it generates. Some types of posts - visuals (photos, videos) generate more likes, comments and shares than others (links, statuses).

This hypothesis is based on the notion that different forms of content have different effects on people's perception level, which is related to the amount of interaction by users. "type of publication has 31% relevance to the number of comments" (Huang & Depari, 2019). Moreover, visual content (photos and videos) tends to engage users more deeply due to its ability to deliver more complex messages, awaken emotions, and attract attention more effectively than textual messages (links and statuses). This theory is based on multimedia learning theory, which posits that people internalize information more deeply from words and pictures than from words alone (Mayer, 2001). In the context of social media, this suggests that publications that contain visual elements are more likely to attract users and encourage them to interact. This is explained by the increased degree of visual appeal, which allows overcoming various barriers (e.g. language barriers) and communicating information more broadly.

Hypothesis №2:

Payment determines the amount of interactions a post generates.

We contend that the payment associated with a post significantly influences its total interactions. It is because efficient budget allocation, guided by post-specific marketing attribution, enhances the performance of paid publications by tailoring marketing strategies to focus on the specific content, effectively reaching the intended audience. (Huang & Depari, 2019). Within this data set, we assume that paid Facebook posts will have an impact towards increased total interactions.

Hypothesis №3:

The number of unique users who saw the post determines the number of total interactions.

Supposed positive relationship means that as the number of people who saw a page post increases , the total number of post interactions will also increase. “Engagement involves actions taken with the content, such as clicks, shares, and comments. Impressions and reach are indicators of how widely content is being seen, while engagement provides insight into how effectively it's resonating with audiences” (Nicki Escudero, ClearVoice, 2023).

Additionally, the number of users that have seen the post and the number of total interactions are two quantitative variables. Because they are two quantitative variables, I will test my hypothesis by graphing a scatter plot and will calculate the correlation coefficient to further support my argument. The correlation coefficient will indicate the strength of the correlation and will reinforce my argument.

RESULTS AND DISCUSSIONS

Hypothesis №1:

The type of Facebook post determines the number of interactions it generates. Some types of posts (photos, videos) generate more likes, comments and shares than others (links, statuses).

	Min	1st Q.	Median	Mean	3rd Q.	Max	Range	St. Dev.
Photo	0	72	124	218.8	226	6334	6334	**407**
Status	17	106	186	217	265	1009	992	178
Link	6	32.75	52.5	89.05	125	420	414	95
Video	81	144	271	295.9	440.5	550	469	183

Figure 1.

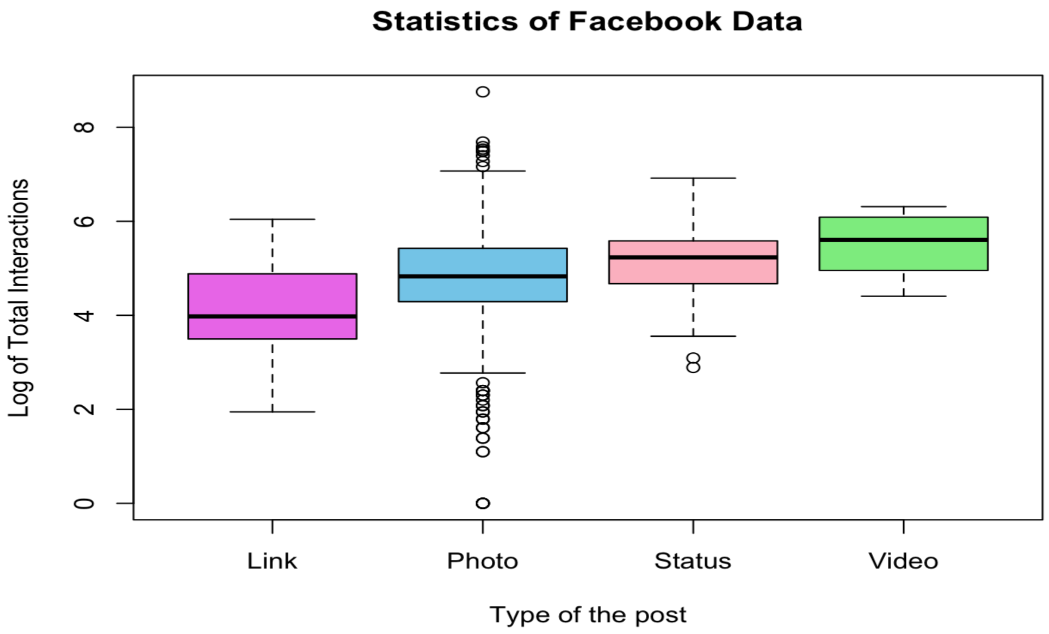

Figure 2.

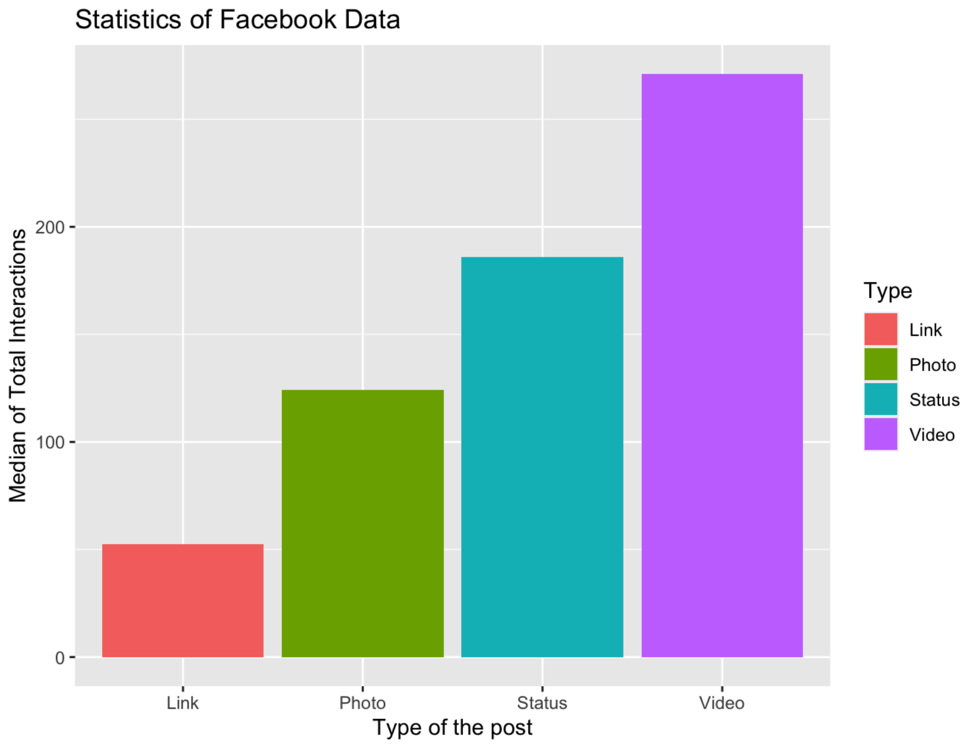

Figure 3.

Qualitative Variable: Type of post (Photo, Status, Link, Video)

Quantitative Variable: Number of interactions (Likes, Comments, Shares)

Our hypothesis suggests that there is a correlation between the type of the post and number of interactions with it. To test this hypothesis, we will evaluate the relationship using both graphical and numerical measures of the two variables. In particular, to examine the relationship between type of posts and the total number of interactions, we will use a box plot of log transformed numbers, a bar plot of median indicators and a table with main numerical measures.

As mean and median numbers differ significantly, especially in the case of photo posts, we will focus more on comparison of median of total interactions. Videos had the highest median interaction count (271), followed by Status (186), Photos (124), and Links (52.5). The mean interactions also showed a similar pattern, with Videos (295.9) and Status (217) leading. However, Photos, despite having a lower mean (218.8) and median (124) than Status, recorded the highest maximum interaction count (6334), indicating the presence of outliers or extremely popular posts within this category.

The EDA results support the hypothesis that different types of posts produce different amounts of total interactions. The assumption that visual content (photos and videos) will generate more user engagement is supported by the fact that the maximum number of interactions corresponds to photos and the high median and mean indicators for videos. However, the analysis also shows that this is not true for all photo posts, as evidenced by the lower median and mean values of interactions compared to status posts.

The wide range of interactions with photo posts, indicated by the high standard deviation, suggests that while some photos may become very popular and receive a large number of interactions in the form of likes, comments, and shares, they do not always generate similar activity. The data for videos supports the hypothesis of high engagement for this type of content, as they have the highest median and mean interaction rates. The results for status messages were unexpected and refuted the hypothesis, showing higher median and mean engagement compared to photos. This may be due to the content of status updates, which may have been particularly relevant.

Finally, our analysis revealed that the indicators of interactions differed significantly by post type, suggesting that post type has an effect on user engagement. We note that the hypothesis is confirmed in the context of the difference in interactions between Photos, Videos, Status and Link and partially for the visual type of posts.

<u>Hypothesis №2:</u>

Payment determines the amount of interactions a post generates.

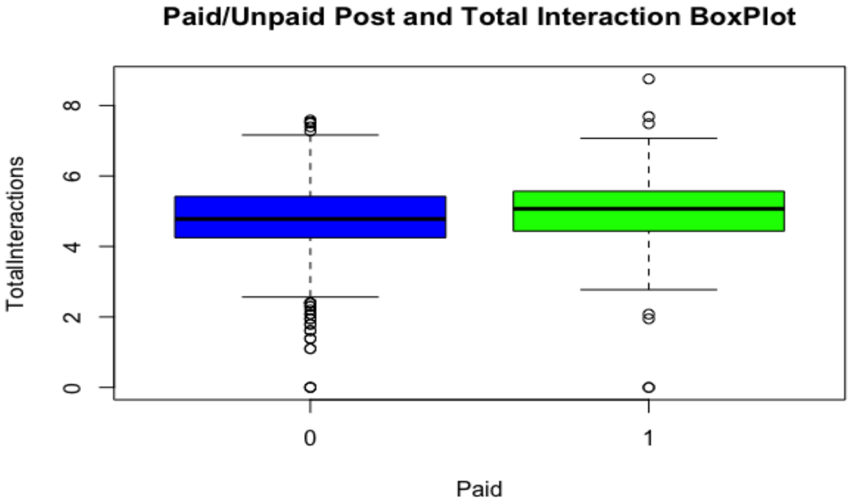

Figure 4.

	Paid (= 1)	**Unpaid (= 0)**
Mean	278.2302	188.8764
Median	158	118
Range	6334-0=6334	1974-0=1974
Standard deviation	594.1007	251.5235

Figure 5.

This hypothesis implies an idea of the level of payment is a determining factor for the quantity of interactions a post garners. To test this hypothesis, we will assess the association by employing both graphical and numerical summaries of two variables. Specifically, we will utilize a box plot and present a table featuring measures of central tendency to explore the connection between paid & unpaid posts and the total number of interactions.

Given the nature of our data (Paid and Unpaid groups with mean, median, range, and standard deviation for total interactions), we would like to compare the central tendency and variability between the two groups, which are considered as quantitative (TotalInteractions) and qualitative (Paid or Unpaid Posts) data. Our given hypothesis is about payment influencing interactions, so the bivariate relationship could be between payment amount and total interactions. In addition, the above given boxplot shows the presence of outliers.

The mean, representing the average, shows that Paid posts (mean: 278.2302) receive more interactions, on average, compared to Unpaid posts (mean: 188.8764). This suggests that financial investment in posts positively influences interaction outcomes.

Examining the median, which represents the middle value, we find that the Paid group (median: 158) also surpasses the Unpaid group (median: 118), reinforcing the notion that Paid posts consistently outperform Unpaid ones.

While, the range, indicating the spread of values, is larger for the Paid group (6334) than the Unpaid group (1974), so it shows a greater variability in interaction outcomes for Paid posts. This implies that the impact of payment on interactions is more diverse, with some Paid posts achieving exceptionally high interaction levels.

Lastly, the standard deviation, a measure of dispersion, is higher for the Paid group (594.1007) compared to the Unpaid group (251.5235). This signifies a wider range of interaction values for Paid posts, highlighting the diverse and potentially more unpredictable nature of Paid post performance.

In conclusion, the provided statistics support the hypothesis that making a payment to Facebook posts influences interaction outcomes. On average, Paid posts generate more interactions, and there is increased variability in their performance compared to Unpaid posts, supporting the idea that financial investment contributes to more dynamic and potentially higher-reaching social media interactions.

This reasoning can be also strengthened through Huang & Depari (2019) study. They conducted a statistical study of the contribution of paid aids to the number of shares, illustrating that 139 posts with paid ads result in 4,517 shares, an average of 32.49 shares. For non-paid publications, 3,489 shares were accumulated as organic share reach, an average of 25.01 shares (p.12). "Moreover, based on comparison results from 139 publications with paid ads and 139 publications without paid ads, publications with paid ads generated 32,755 total likes and earned 235 likes for each publication on average, compared with 20,458 total likes and 147 likes for each publication on average for publications without paid ads This implies that paid ads provide 59% more likes." (p.10). Therefore, Huang & Depari (2019) concluded that using paid posts would considerably increase the number of likes and shares of those posts.

<u>Hypothesis №3</u>

The number of unique users who saw the post determines the number of total interactions.

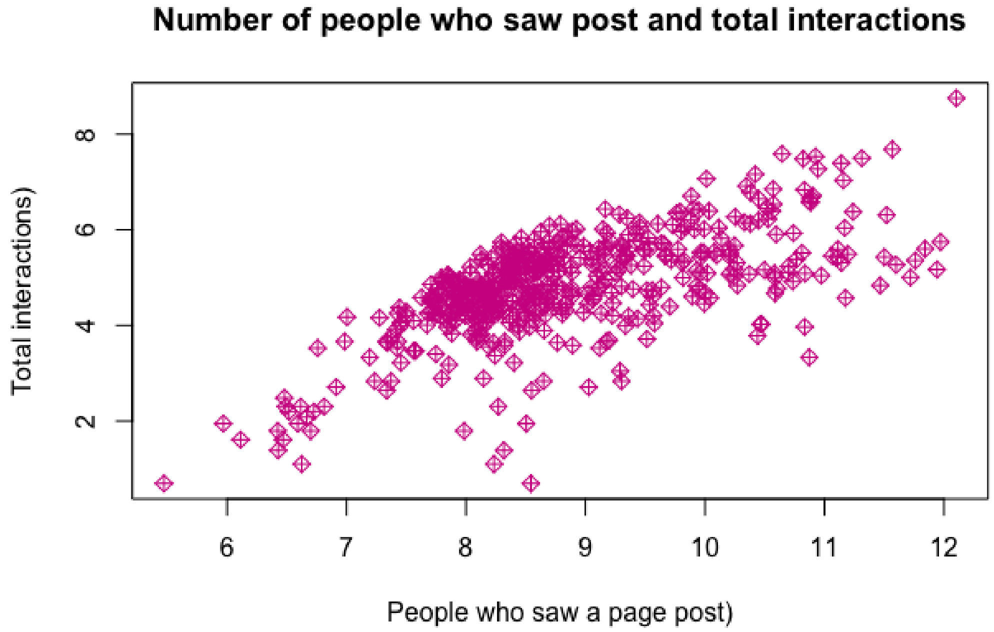

Figure 6.

Numerical summary:

Correlation between LifetimePostTotalReach and Total Interactions → 0.5373635

After graphing the two variables, I find evidence that supports my hypothesis. A positive correlation can be found through the scatter plot. So as the number of users who saw the post online increases (decreases) the number of total interactions also increases (decreases). The graph also shows a positive slope, indicating a positive relationship between the two variables. And after calculating the correlation coefficient of the two variables, we get a correlation coefficient of 0.537. Because the correlation coefficient is closer to 1.0 and positive, this indicates a moderate positive relationship between the variables.

The plot shows a number of post interaction points that are packed together forming a diagonal band from the bottom left to the top right of the plot area. This suggests a positive correlation between the number of people who saw a page post and the total interactions.

The distribution of the points is not uniform; there is a concentration of points around the center of the plot, indicating that most of the data falls within a moderate range of page views and interactions. The spread of the data increases as the number of page views increases, which could indicate greater variability in interactions with more page views.

In terms of interesting features, there does not appear to be any clear outliers as all data points seem to follow the general trend.The number of people who saw a page post (x-axis) is likely a discrete variable because the number of people is counted in whole numbers. In terms of Total interactions (y-axis), the variable is also discrete considering the given context. If interactions are counted as whole numbers (such as likes, comments, and shares), then it is discrete. There are no obvious signs of a skewed distribution, as the data seems to be fairly symmetrical around the line of best fit.

To conclude, the more the post was seen by unique users typically the higher is the value of total interactions. This can be explained by the natural processes of willingness to interact with the content as people see it on their Facebook feed shaped by their preferences and values. When the content responds to a person’s values and opinions, it is most likely to be shared by them with their close circle of people like family and friends.

CONCLUSION:

Different types of posts affect user engagement differently, with visual posts such as photos and videos receiving more interaction than text links and statuses. Paid posts generate more user interactions than unpaid posts, suggesting the importance of investing in a social media strategy. There is a moderate positive correlation between the number of unique users who saw a post and the total number of interactions it received, supporting the concept that a high number of views contributes to higher engagement.

Limitations identified in the study that may lead to future research include the limited sample size and diversity. There is a need for additional statistical testing to establish causal relationships and the influence of related variables. Future research could expand the dataset, apply statistical tests to confirm observed relationships, and include additional variables such as the amount of text in a post and the timing of their publication to gain a more complete picture of what determines user engagement on social media.

REFERENCES

Huang, J. P., & Depari, G. S. (2019). Paid advertisement on Facebook: an evaluation using a data mining approach. Review of Integrative Business and Economics Research, 8(4), 1. http://www.buscompress.com/uploads/3/4/9/8/34980536/riber_8-4_01_m18-072_1-16.pdf

Mayer, R. (2001). Multimedia Learning. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/multimedia-learning/E9595926786F5DEA326A3774D2F23DB2

Escudero, N. (2023, November 9). Impressions vs. Reach vs. Engagement: Understanding Metrics. ClearVoice. https://www.clearvoice.com/resources/impressions-vs-reach-vs-engagement/