Spark Project

Lab Scenario:

You are working as a Big Data expert for a huge cosmetic brand. This brand has a Facebook page with a significant following. As a business page on Facebook, the page moderators and administrators have access to the statistics of the page and the posts made on it with regards to audience outreach. You are provided with access to the dataset that has been created through these statistics along with a set of tasks that you have to perform to assist your organization's marketing team with its strategy planning.

Dataset:

https://drive.google.com/file/d/1pMaeM19Ib9PyaVAlKZQPvW1D1v8VWPG7/view?usp=sharing

Dataset Description:

The dataset filename is **dataset_Facebook_cos**, and the column descriptions are as follows:

- **Page total likes:** The total number of subscriptions/follows on the page feed at the time of the specific row post (photo, status, video or link)
- **Type:** The category of the post (photo, status, video or link)
- **Category:** 1, 2, 3
- **Post Month:** The month in which the post was made
- **Post Weekday:** The day of the week the post was made
- **Post Hour:** The hour (on a 24-hour clock) in which the post was made
- **Paid:** 1 if the post was promoted through paid means and 0 if the post was organically promoted
- **Lifetime Post Total Reach:** The number of unique people who saw your post
- **Lifetime Post Total Impressions:** The number of times your post appeared in people's feeds
- **Lifetime Engaged Users:** The number of unique people who have interacted with the post (liked, shared, or commented)
- **Lifetime Post Consumers:** The number of unique people who clicked anywhere on the post (it shows interest)

- **Lifetime Post Consumptions:** The number of times the post was clicked on (multiple clicks are possible from the same user)
- **Lifetime Post Impressions by people who have liked your page:** The number of times the post appeared in the feed of the users who have liked the page
- **Lifetime Post Reach by people who like your page:** The number of unique people, who are subscribed to (liked) your page, who saw the post appear in their feed
- **Lifetime People who have liked your page and engaged with your post:** The number of unique people who have subscribed to (liked) your page and interacted with the post by liking, commenting or sharing
- **Comment:** The number of comments on the post
- **Like:** The number of likes on the post
- **Share:** The number of times the post has been shared
- **Total Interactions:** The sum of comment, like, and share columns

Tasks To Be Performed:

Find out or solve the following:

1. The total number of posts made.
2. The percentage of the growth or decline of the page, in terms of likes (subscriptions on the page), from the first post to the latest post.

Hint: The first record of the dataset represents the latest post, and the last record of the dataset represents the first post.

3. Which month, on average, has the highest number of post interactions?
4. Which day of the week, on average, has the highest number of post interactions?
5. Which hour of the day, on average, has the highest number of post interactions?

Hint: You can use numbers present in the dataset to define the months, weekdays, and hours in your answer documentation. You don't have to be concerned with naming (e.g., use '12' instead of 'December').

6. Determine if paid (promoted) posts have a higher correlation with a large number of post shares when compared to the post shares of organic

(non-promoted) posts. This is to determine the commercial viability of investing in paid posts for promoting cosmetic products. Answer with either a Yes or a No, and provide the methodology of how you reached your conclusion.

7. Which post type (photo, video, status, or link) is the most attractive to people who have subscribed to your page (people who have liked the page)?
8. Which hour of the day is ideal for posting photographic content? Arrange the hours of the day according to the order of the Lifetime Post Impressions column?
9. Create an additional column with the name Likes-to-Comment Ratio, with the column values having the equation:

   likes to comment ratio = like / comment

Hint: Make sure the ratio is in a decimal format, and correct it to 2 decimal places.

10. Arrange post categories (1, 2, 3) in descending order of the reach that they can accumulate on average.
11. Determine the standard deviation of the average post reach for each of the day hours. This is to determine if the time of the day is an ideal criterion to identify when to create posts.
12. Is there any correlation between the number of post consumptions and the total interactions on the post?
13. Determine the two best days in a week to create posts, when people are extremely active on social media, based on the data that you have.

Hint: Question 13 can have a subjective answer. You are free to choose your own approach to determine the best days to post in a week. Make sure to validate your claims with the relevant code and explanation of your approach.

