**Booklet: Exploratory Data Analysis on Instagram Post Engagement**

**📖 Table of Contents**

1. Introduction
2. Business Objective
3. Dataset Overview
4. Data Preparation
5. Tools and Libraries Used
6. Exploratory Data Analysis
   * Likes vs. Comments
   * Followers vs. Comments
   * Hashtag Usage Analysis
7. Key Findings
8. Recommendations
9. Conclusion
10. Next Steps
11. Appendices

**📖 1. Introduction**

In the era of digital marketing and personal branding, **Instagram has emerged as a critical platform for audience engagement**. Marketers, influencers, and businesses seek to optimize their posts for maximum interaction, particularly focusing on **likes, comments, shares, and follows**.

This booklet presents an **Exploratory Data Analysis (EDA)** to investigate the factors influencing post engagement on Instagram — with a particular emphasis on **the number of comments** a post receives.

**📖 2. Business Objective**

The primary question this project seeks to address is:

**What factors most influence post engagement on Instagram, especially the number of comments received?**

This insight can help content creators and marketers to **refine their content strategy, improve engagement rates, and better connect with their audiences**.

**📖 3. Dataset Overview**

The dataset used contains various metrics related to Instagram posts. Each record represents one post and includes the following fields:

* Impressions
* From Home
* From Hashtags
* From Explore
* From Other
* Saves
* Comments
* Shares
* Likes
* Profile Visits
* Follows
* Caption
* Hashtags
* Followers at Post Time

The dataset was organized in CSV format and imported into **RStudio** for analysis.

**📖 4. Data Preparation**

**📌 Data Cleaning Steps:**

1. **Checked for duplicate posts** based on likes, comments, and follows — none found.
2. **Removed negative values** from likes, comments, and followers.
3. **Removed rows with missing values (NULL)** in the likes or comments fields.
4. **Extracted hashtags** from the captions using string manipulation.

This ensured a clean and reliable dataset for meaningful analysis.

**📖 5. Tools and Libraries Used**

The following tools and libraries were used for data cleaning, analysis, and visualization:

* **RStudio**
* ggplot2 (data visualization)
* dplyr (data manipulation)
* stringr (text processing)

**📖 6. Exploratory Data Analysis**

**📖 6.1 Likes vs. Comments**

A **scatter plot** was created with:

* **X-axis:** Number of Likes
* **Y-axis:** Number of Comments

**Correlation Coefficient:** +0.12 (weak positive)

📌 **Insight:**  
Posts with more likes generally receive more comments, but the relationship is weak.

**📖 6.2 Followers vs. Comments**

A **scatter plot** was created with:

* **X-axis:** Followers at post time
* **Y-axis:** Number of Comments

**Correlation Coefficient:** -0.06 (negligible negative)

📌 **Insight:**  
Follower count does not guarantee higher engagement in the form of comments.

**📖 6.3 Hashtag Usage Analysis**

* Extracted hashtags from post captions.
* Counted occurrences of each hashtag.
* Visualized the frequency using a **bar plot**.

**Most frequent hashtags:**  
#ai, #sunset, #fitness, #adventure, #travel

📌 **Insight:**  
A few hashtags dominate in frequency while many others are rarely used.

**📖 7. Key Findings**

| **Insight** | **Description** |
| --- | --- |
| Weak positive relationship between likes and comments | Likes slightly influence comment count. |
| No relationship between followers and comments | Follower count alone is not a reliable predictor of engagement. |
| Certain hashtags are consistently popular | High-performing hashtags dominate overall usage. |

**📖 8. Recommendations**

1. Focus content strategy around the **top 20-30 high-performing hashtags**.
2. Avoid using too many irrelevant or rarely-used hashtags.
3. Group hashtags by themes (e.g., travel, fitness, AI) for targeted audience engagement.
4. Focus more on **content quality, caption appeal, and timing** rather than follower count.
5. Explore additional variables like **post type, caption length, and posting time**.
6. Examine high-like but low-comment posts for qualitative insights.
7. Incorporate **regression analysis and predictive modeling** for deeper insights.

**📖 9. Conclusion**

This EDA revealed that **Instagram post engagement is driven more by content factors such as effective hashtags, likes, and post timing, than by follower count alone**.

Key takeaways:

* Likes have a weak but positive correlation with comments.
* Follower count has no meaningful relationship with comments.
* Popular hashtags play a vital role in engagement.

The study highlights the importance of **strategic content creation and optimal hashtag use** for audience engagement.

**📖 10. Next Steps**

* Conduct **predictive modeling** to identify key engagement predictors.
* Include other engagement metrics like **saves and shares**.
* Analyze post captions for **sentiment and length analysis**.
* Compare engagement across different **media types**.
* Study the effect of **posting time** on engagement levels.

**📖 11. Appendices**

**📌 SQL Queries for Data Cleaning:**

sql

CopyEdit

SELECT \*

FROM instagram\_posts

WHERE likes < 0

OR comments < 0

OR followers\_at\_post\_time < 0;

SELECT \*

FROM instagram\_posts

WHERE likes IS NULL

OR comments IS NULL;

SELECT \*

FROM instagram\_posts

WHERE likes IS NOT NULL

AND comments IS NOT NULL;

**📌 R Code for Key Visualizations:**

r

CopyEdit

# Likes vs Comments Scatter Plot

ggplot(instagrampost, aes(x = likes, y = comments)) +

geom\_point(color = "blue", size = 3) +

ggtitle("Scatter Plot: Likes vs Comments") +

xlab("Number of Likes") +

ylab("Number of Comments")

# Followers vs Comments Scatter Plot

ggplot(instagrampost, aes(x = followers\_at\_post\_time, y = comments)) +

geom\_point(color = "black", size = 3) +

ggtitle("Scatter Plot: Followers vs Comments") +

xlab("Number of Followers at Post Time") +

ylab("Number of Comments")

# Hashtag Extraction and Count

hashtags <- str\_extract\_all(instagram\_data$caption, "#\\w+")

hashtag\_table <- data.frame(hashtag = unlist(hashtags))

hashtag\_counts <- hashtag\_table %>%

count(hashtag) %>%

arrange(desc(n))

print(hashtag\_counts)