Project Report on

SOCIAL MEDIA ENGAGEMENT ANALYSIS

In partial fulfillment for the award

Of

Professional Certification in Data Analysis and Visualization

Year 2024-2025

Submitted By

FATHIMATH SAFNA P

Tools & Technologies Used:

Python, R Programming, Tableau

Duration:

15/07/2025 – 30/07/2025

Submission Date:

31/07/2025

G - TEC CENTRE OF EXCELLENCE PERINTHALMANNA

ABSTRACT

This project, titled *Social Media Engagement Analysis*, explores how users interact with social media content across different platforms. It focuses on key performance indicators such as Likes, Shares, Comments, Impressions, and Engagement Rate. As brands increasingly depend on digital platforms to connect with audiences, understanding these metrics is essential for strategic decision-making and targeted outreach.

Python was used for data cleaning and exploratory data analysis (EDA) to uncover trends and patterns in user behaviour . R programming validated these insights through statistical hypothesis testing (t-tests, ANOVA, chi-square, and z-tests), ensuring findings were statistically significant. Tableau was used to build interactive dashboards that visually present results in an accessible format for business stakeholders.

Findings suggest significant differences in engagement based on platform, post type, timing, and audience demographics. For instance, Instagram outperformed Facebook in terms of engagement rate, and posts shared in the afternoon had the highest interaction levels. These insights are highly actionable for social media managers and marketers seeking to improve campaign performance.

By combining Python, R, and Tableau, the project offers a scalable, replicable, and stakeholder-friendly framework for monitoring and improving social media performance using real-world data.

INTRODUCTION

Social media has transformed how brands communicate with audiences. With billions of users active across platforms like Facebook and Instagram, measuring performance through Likes, Comments, Shares, and other metrics has become essential. However, the volume of data often makes it difficult to uncover actionable insights.

This project, titled "Social Media Engagement Analysis," aims to explore engagement trends across multiple platforms and content types. It uses survey-like post-level metrics to determine what drives higher engagement. Python was used for cleaning and exploratory analysis, R for statistical testing, and Tableau for visual storytelling. This multi-tool approach ensures both depth and clarity in the results, enabling non-technical stakeholders like content strategists to make informed decisions.

his multi-tool approach ensures both depth and clarity in the results, enabling non-technical stakeholders like content strategists to make informed decisions.

Additionally, the project addresses the growing need for data-driven marketing, where understanding audience behaviour through reliable analytics tools has become critical. With algorithmic changes and shifting user expectations, brands that leverage actionable insights from data can better align their content strategies with user preferences and platform-specific best practices.

LITERATURE REVIEW

Importance of Engagement Metrics

Engagement metrics—such as likes, comments, shares, and saves—have become fundamental indicators of content effectiveness in the digital marketing landscape. According to Chaffey (2021), these interactions reflect user involvement and interest, making them more meaningful than passive metrics like impressions or reach. High engagement rates signal stronger audience connection, increased brand recall, and improved algorithmic visibility across platforms. Marketers now prioritize engagement as a critical benchmark for campaign performance, using it to fine-tune both content and targeting strategies.

Impact of Platform Differences

Each social media platform exhibits distinct user behaviour patterns and content performance dynamics. Sprout Social (2023) found that visual-first platforms such as Instagram and TikTok generally outperform text-based platforms in engagement rates, particularly when video content is used. Facebook, while still effective for broader reach, shows declining organic engagement for static content. Additionally, user intent varies across platforms—LinkedIn users seek professional insights, whereas Instagram users favor aesthetics and entertainment—making it essential for brands to tailor content type and tone accordingly.

Post Timing and User Behavior

Timing plays a pivotal role in how users interact with content. Hootsuite (2022) highlights that audience engagement tends to peak during mid-week, particularly between 12 PM and 3 PM, when users are most active during workday breaks. However, optimal times vary by industry and platform. B2C brands see better engagement on weekends, whereas B2B content performs better on weekdays. Seasonal changes, global events, and even time zone targeting also influence post visibility and interaction rates, demanding a data-driven approach to scheduling.

Analytics in Social Media

The rapid evolution of social media has driven the adoption of advanced analytics for insight generation. Python and R are now standard tools in digital marketing analytics, enabling brands to clean, analyze, and visualize vast datasets. Exploratory Data Analysis (EDA) with Python helps identify performance trends, while R is widely used for hypothesis testing and statistical validation. Tableau complements these tools by turning complex results into accessible dashboards for stakeholders. As noted by Forbes (2022), the integration of data science in marketing workflows is no longer optional but essential for competitive advantage.

RESEARCH GAP

Many social media studies rely solely on descriptive metrics without testing statistical significance. Additionally, few analyses combine Python, R, and Tableau to deliver insights in a comprehensive and stakeholder-friendly format. This project addresses that gap by integrating robust exploratory and confirmatory analysis with interactive dashboarding.

DATA COLLECTION & PREPROCESSING

Data Source and Collection Methods

The dataset used in this project "Social Media Engagement Data" and was sourced from **Kaggle**, contains post-level metrics such as Platform, Post Type, Likes, Comments, Shares, Impressions, Audience Gender and Engagement Rate.

Data Quality Assessment and Cleaning Procedures

Initial inspection of the dataset was done using Python with the help of the pandas and numpy libraries. The following steps were taken:

- Missing values: Detected and handled by imputing basic values where appropriate.
- Duplicate entries: Checked for duplicate entries.
- Data types: Ensured each column had the correct type for further processing (e.g., numerical ratings as float, categorical as factors).

Feature Engineering and Selection Techniques

- **Weekday**: Extracted from the post date to understand engagement trends by day of the week. This enabled analysis of time-based posting behavior and audience response.
- **Total Interactions**: A derived metric representing the total sum of Likes, Comments, and Shares. It was used in Tableau visualizations to assess raw post performance.
- **Filtered Subsets**: For statistical testing in R (such as T-tests and ANOVA), subsets of the data were created. Examples include selecting only Facebook and Instagram posts for platform comparison, or filtering by Male and Female for gender-based tests.

Columns selected for analysis included:

Platform, Post Type, Likes, Comments, Shares, Impressions, Reach, Engagement Rate, Audience Gender, High_ Engagement, Weekday, Year, Post Type, Post Time

METHODOLOGY

This project follows a multi-tool analytical approach combining Python, R programming, and Tableau to explore and validate social media engagement patterns. The methodology was designed to provide a full-cycle analysis—from data cleaning and exploration to statistical testing and dashboard visualization.

Tools and Technologies Used

- Python was used for data cleaning, transformation, and exploratory analysis. Key libraries like pandas, matplotlib, seaborn, plotly express enabled efficient data handling, visualization, and the identification of patterns.
- R programming was utilized for statistical hypothesis testing to validate the patterns discovered during EDA. Key packages such as t.test(), aov(), and chisq.test() helped in performing parametric and non-parametric tests.
- Tableau was used for building interactive dashboards and data visualization, presenting the findings to non-technical stakeholders in a clear and intuitive format.

Exploratory Data Analysis – Python

With the data cleaned, exploratory analysis was conducted to identify trends and patterns using matplotlib and seaborn. This step allowed for the visualization of key relationships between variables, which guided the choice of statistical tests in R.

Visualizations:

- **Bar Plots**: Used to compare different groups.
- **Heatmap**: To visualize **patterns, intensity, or relationships** between two variables using color.
- **Boxplot**: To visualize the **distribution and spread** of a continuous variable.

Groupby and Crosstab:

- Used groupby() to aggregate and analyse means, counts, and distributions based on different features.

- The crosstab() function was applied to understand relationships between two categorical variables helping in determining the association between them.

Statistical Testing (R Programming)

The exploratory insights from Python were validated using statistical hypothesis testing in R.

T-test: An independent two-sample t-test was conducted to evaluate whether the mean engagement rate differed significantly between Facebook and Instagram.

Chi-square Test: Chi-square test of independence was used to test for a significant association between the platform and post type

ANOVA: A one-way Analysis of Variance (ANOVA) was conducted to compare the mean engagement rate across all social media platforms present in the dataset**.**

Z-test: A Z-test for difference in proportions was then applied to determine whether the proportion of high-engagement posts differed significantly between Facebook and Instagram.

F-test: An F-test for equality of variances was performed to assess whether the variability **in** engagement rate was significantly different between Facebook and Instagram.

Data Visualization and Dashboarding - Tableau

After completing the data preparation and analysis in Python, **Tableau** was used to create an interactive and user-friendly dashboard. Tableau was selected for its ability to present complex data in a visual format, making it easier for stakeholders to interpret and act on the insights.

- Interactive Filters: Filters such as platform, Year, Gender allowing to explore different subsets of the data.

- Simple Layout and Clear Visuals: The dashboard was designed with a clean layout, using color coding and labels to highlight the key findings, ensuring that users could easily identify trends and patterns.

RESULTS AND ANALYSIS

This section presents the key findings of the project, focusing on Analyse engagement patterns across different social media platforms (like Instagram and Facebook).

Python- Based Results

- Instagram and Facebook were the most used platforms. Instagram had a little more posts, so it was the most common.
- The average engagement rate was about 46.92, meaning most posts had a normal level of interaction—not too high, not too low.
- The most popular types of posts were images and videos, with image posts appearing the most.
- Instagram posts got better engagement than Facebook posts. This means Instagram is generally better for reaching and engaging people.
- Posts shared in the afternoon had higher engagement than those posted in the morning, evening, or night.
- Weekday posts performed slightly better than weekend posts.
- Most posts had a positive tone (sentiment), followed by neutral and negative ones.
- The age group 25–34 engaged the most, followed by people aged 18–24. Younger audiences were more active.
- People from North America and Europe interacted more with the posts than those from other regions.
- On Facebook, male and female audiences were more balanced. On Instagram, there were slightly more female viewers.
- Posts with positive reactions were usually videos or images. These types of posts got more likes and comments.
- The top 5 posts with the highest engagement were either videos or images—and all were posted on Instagram.
- The top 3 countries with the most audience activity were the USA, India, and Brazil.
- A chart also showed that positive posts had the best engagement, followed by neutral, then negative posts.

- LinkedIn favors link posts, Twitter is balanced with more images, Facebook prefers images/videos over links, Instagram uses all types equally, and video posts are consistently popular across all platforms.
- Engagement rates tend to be slightly higher on Mondays and Sundays, with Monday showing the highest median among all weekdays.

R-Based Statistical Results

- Instagram had a significantly higher engagement rate than Facebook (T-test).
- Engagement rate variation was different between platforms (F-test).
- Average engagement rate was different across all platforms (ANOVA).
- No significant difference in high-engagement post proportions between Facebook and Instagram (Z-test).
- Platform and post type were significantly related (Chi-square test).

Tableau- Based Results

- Engagement is strong, with an average rate of 46.92 and nearly 50 million likes.
- Likes are the most common type of interaction, followed by comments and shares.
- In 2024, all activity suddenly dropped, which might be due to a system issue or change.
- People are most interested in topics like "charge," "table," and "rule".
- Posts are shared almost equally across platforms, with Twitter and LinkedIn having the most posts.
- Men, women, and others all have about the same engagement rate.
- Older people (senior adults) give the most likes on social media.
- Comments and shares are much lower than likes for all age groups.
- Posts were made on Facebook at night (2021), Twitter in the morning (2024), and LinkedIn in the afternoon (2022) with different moods (mixed, positive, neutral).
- Likes are the most common type of engagement, especially from older users.
- The people more interact with night.

CONCLUSION

This Social Media Engagement Analysis project successfully identified key factors that influence how users interact with content across different platforms. By leveraging Python for data preparation and exploration, R for rigorous statistical validation, and Tableau for interactive visualization, the project provides a comprehensive and actionable understanding of engagement trends.

The findings reveal that engagement varies significantly by platform, content type, posting time, and audience demographics, with Instagram outperforming Facebook in engagement rates and visual content driving higher interaction. The project also highlights the importance of tailoring strategies based on platform-specific user behavior and timing to maximize reach and impact.

By combining multiple analytical tools, this approach offers a scalable and replicable framework for brands and marketers to monitor, analyse, and optimize social media performance effectively. Continued use of data-driven insights will enable organizations to adapt to evolving user behaviors and platform algorithms, ultimately enhancing their digital marketing success.

FUTURE WORKS

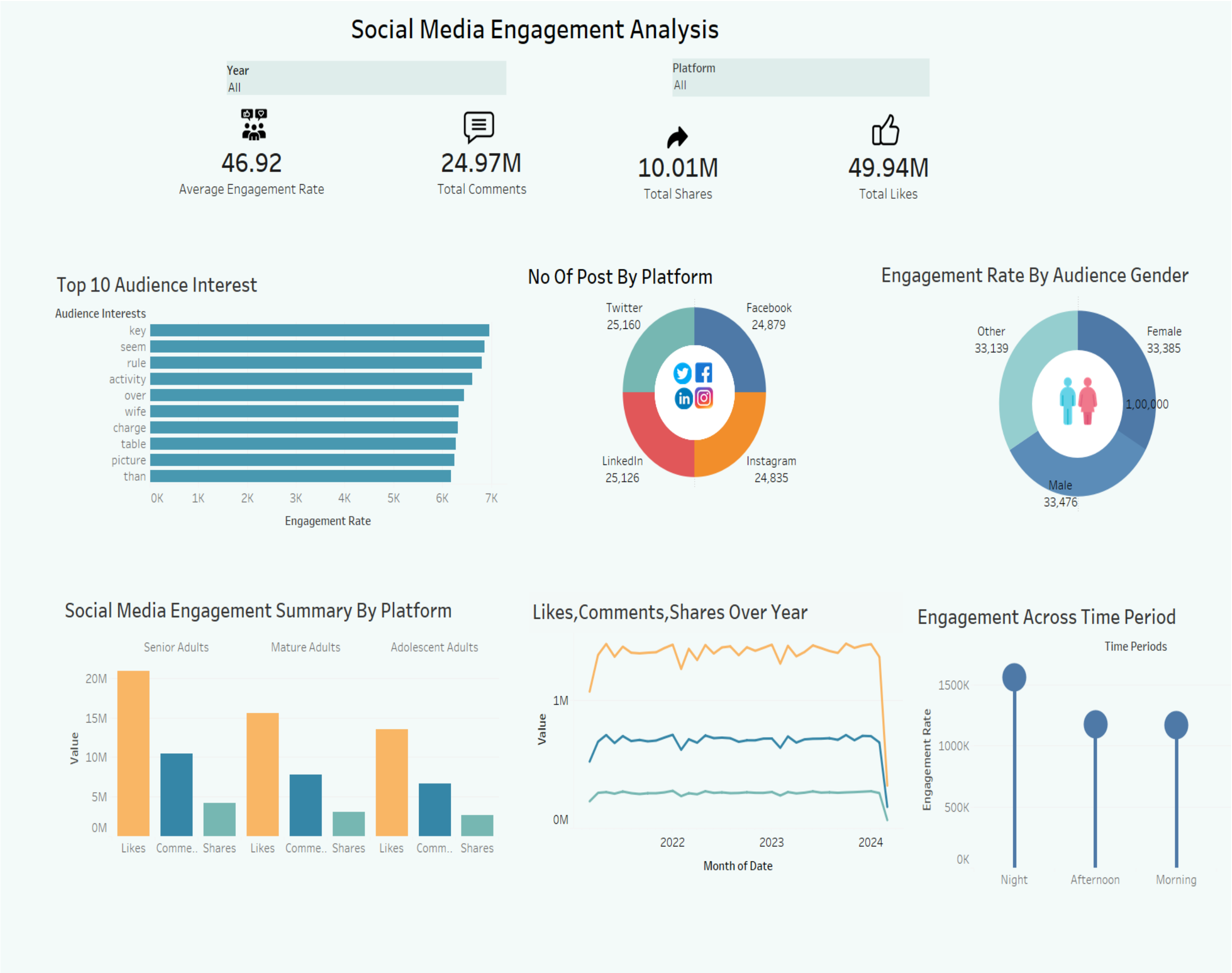
- Use live data from social media for up-to-date analysis.
- Study the feelings and topics in posts to understand what people like.
- Build tools to predict which posts will get more attention.
- Add newer social media platforms like TikTok to the study.
- Look at differences between paid ads and regular posts.
- Check how things like video length or hashtags affect engagement.
- Group users by age, interests, or habits to make better content.

REFERENCES

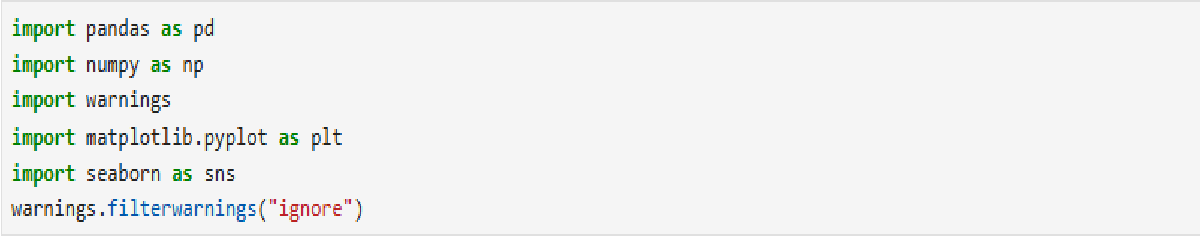
- Chaffey, D. (2021). *Digital marketing: Strategy, implementation and practice* (8th ed.). Pearson.
- Forbes. (2022). How data science is transforming digital marketing. Retrieved from https://www.forbes.com/sites/forbestechcouncil/2022/05/18/how-data-science-is-transforming-digital-marketing
- HubSpot. (2023). The impact of personalization on engagement rates. Retrieved from https://blog.hubspot.com/marketing/personalization-statistics
- Hootsuite. (2022). Best times to post on social media in 2022. Retrieved from https://blog.hootsuite.com/best-time-to-post-on-social-media/
- Sprout Social. (2023). Social media engagement benchmarks by platform. Retrieved from https://sproutsocial.com/insights/social-media-benchmarks/
- Kaggle. (n.d.). Social Media Engagement Report Dataset. Retrieved from https://www.kaggle.com/datasets/username/social-media-engagement-report

Supporting Files

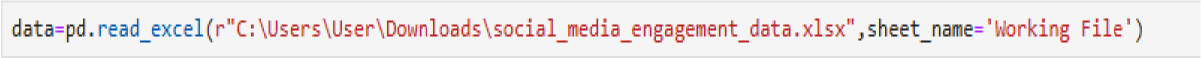
Tableau (Dashboard)



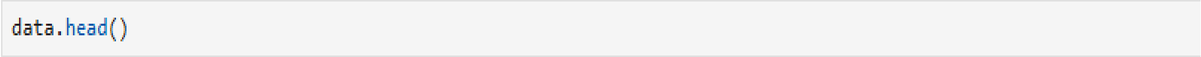
Python



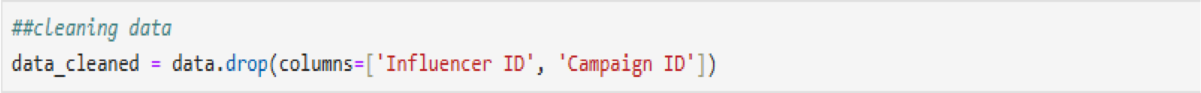
```
import pandas as pd
import numpy as np
import warnings
import matplotlib.pyplot as plt
import seaborn as sns
warnings.filterwarnings("ignore")
```



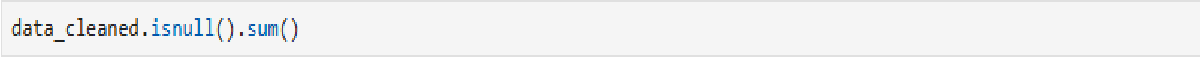
```
data=pd.read_excel(r"C:\Users\User\Downloads\social_media_engagement_data.xlsx",sheet_name='Working File')
```



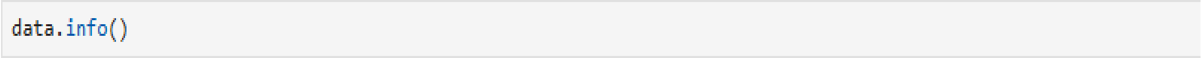
```
data.head()
```



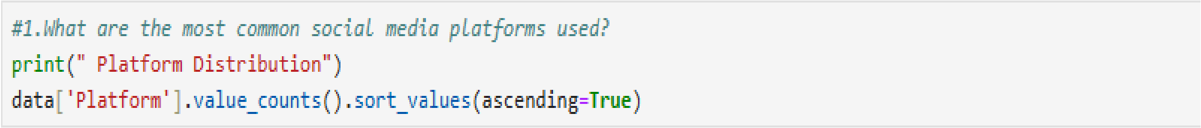
```
##cleaning data
data_cleaned = data.drop(columns=['Influencer ID', 'Campaign ID'])
```



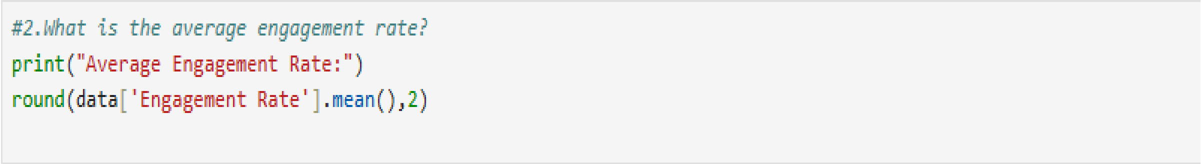
```
data_cleaned.isnull().sum()
```



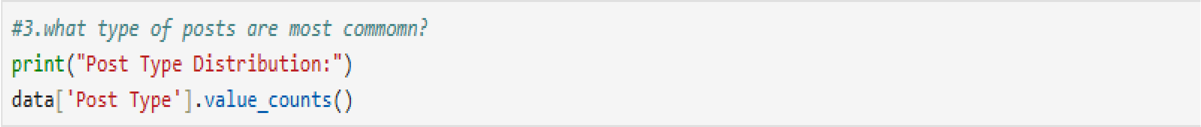
```
data.info()
```



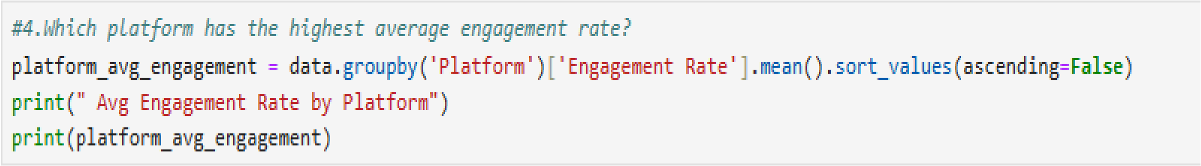
```
#1.What are the most common social media platforms used?
print(" Platform Distribution")
data['Platform'].value_counts().sort_values(ascending=True)
```



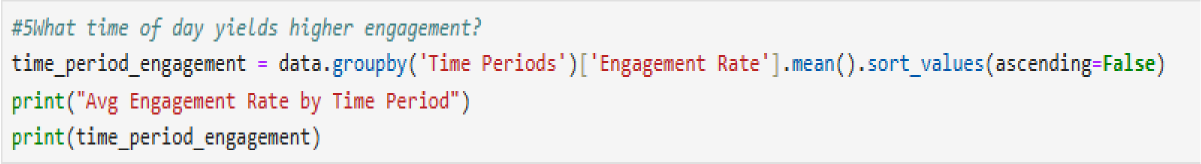
```
#2.What is the average engagement rate?
print("Average Engagement Rate:")
round(data['Engagement Rate'].mean(),2)
```



```
#3.what type of posts are most commomn?
print("Post Type Distribution:")
data['Post Type'].value_counts()
```



```
#4.Which platform has the highest average engagement rate?
platform_avg_engagement = data.groupby('Platform')['Engagement Rate'].mean().sort_values(ascending=False)
print(" Avg Engagement Rate by Platform")
print(platform_avg_engagement)
```



```
#5What time of day yields higher engagement?
time_period_engagement = data.groupby('Time Periods')['Engagement Rate'].mean().sort_values(ascending=False)
print("Avg Engagement Rate by Time Period")
print(time_period_engagement)
```

[ ]:

```
#6.Do weekends perform better than weekdays?
weekday_engagement = data.groupby('Weekday Type')['Engagement Rate'].mean()
print("Weekday vs Weekend Engagement")
print(weekday_engagement)
```

[ ]:

```
#7.What is the sentiment distribution of posts?
print("Sentiment Distribution")
data['Sentiment'].value_counts()
```

[ ]:

```
#8.Which audience age group engages the most?
age_group_engagement = df.groupby('Age Group')['Engagement Rate'].mean().sort_values(ascending=False)
print("Engagement Rate by Age Group")
print(age_group_engagement)
```

[ ]:

```
#9.Which audience continent has highest engagement?
continent_engagement = df.groupby('Audience Continent')['Engagement Rate'].mean().sort_values(ascending=False)
print("Engagement Rate by Continent")
print(continent_engagement)
```

[ ]:

```
#10.Audience Gender vs Platform
pd.crosstab(data['Audience Gender'], data['Platform'])
```

[ ]:

```
#11. Sentiment vs Platform
pd.crosstab(data['Sentiment'], data['Platform'])
```

[ ]:

```
#12
pd.crosstab(data['Platform'], data['Post Type'])
```

[ ]:

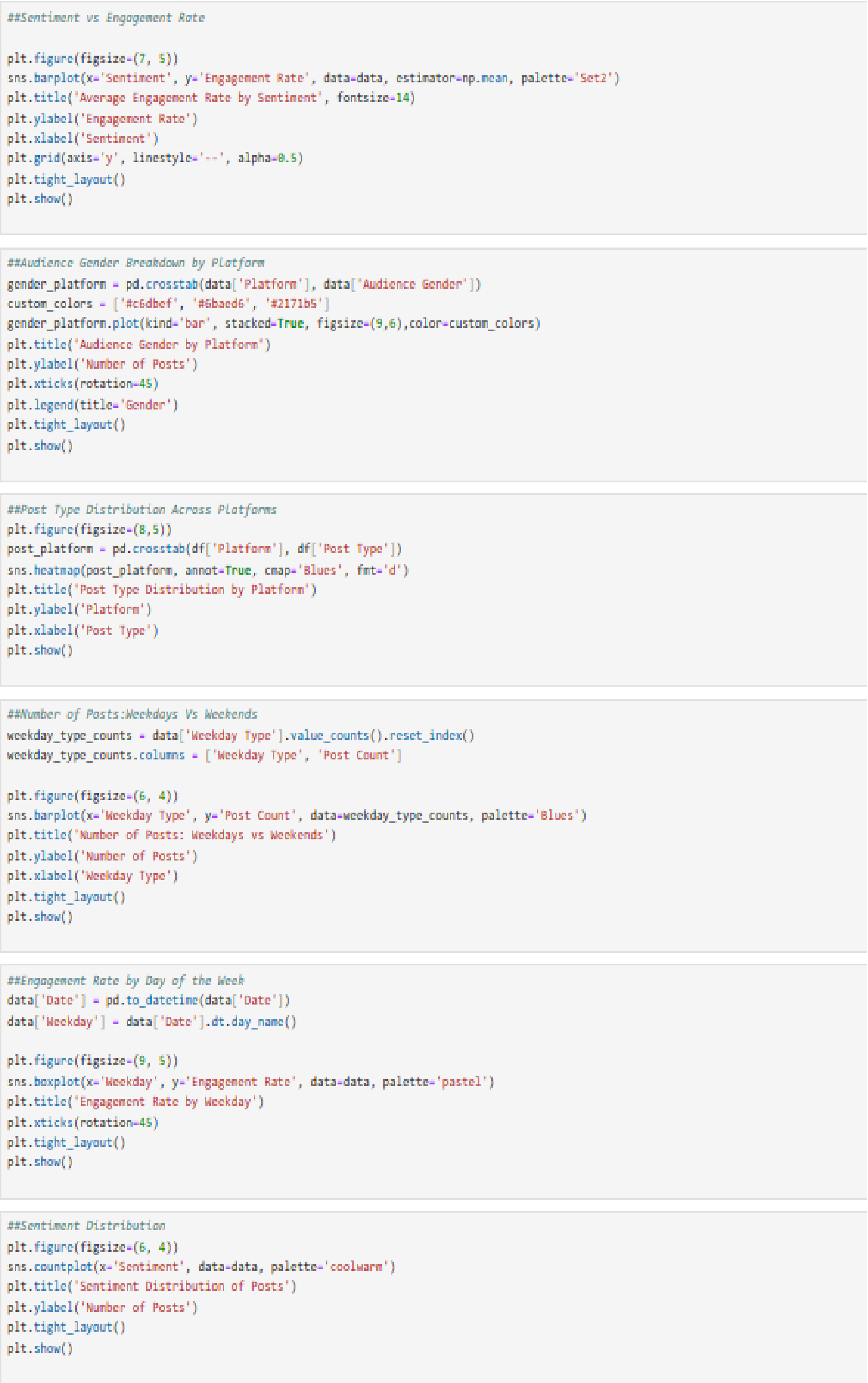
```
#13.Top 5 highest performing posts (Engagement Rate)
data.sort_values('Engagement Rate', ascending=False).head(5)
```

[ ]:

```
#14.Are certain post types more likely to generate positive sentiment?
pd.crosstab(data['Post Type'],data['Sentiment'])
```

[ ]:

```
#15. Which countries have the highest volume of audience activity?
data['Audience Location'].value_counts().head(10)
```



```
##Sentiment vs Engagement Rate

plt.figure(figsize=(7, 5))
sns.barplot(x='Sentiment', y='Engagement Rate', data=data, estimator=np.mean, palette='Set2')
plt.title('Average Engagement Rate by Sentiment', fontsize=14)
plt.ylabel('Engagement Rate')
plt.xlabel('Sentiment')
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()
```

```
##Audience Gender Breakdown by Platform
gender_platform = pd.crosstab(data['Platform'], data['Audience Gender'])
custom_colors = ['#c6dbef', '#6baed6', '#2171b5']
gender_platform.plot(kind='bar', stacked=True, figsize=(9,6),color=custom_colors)
plt.title('Audience Gender by Platform')
plt.ylabel('Number of Posts')
plt.xticks(rotation=45)
plt.legend(title='Gender')
plt.tight_layout()
plt.show()
```

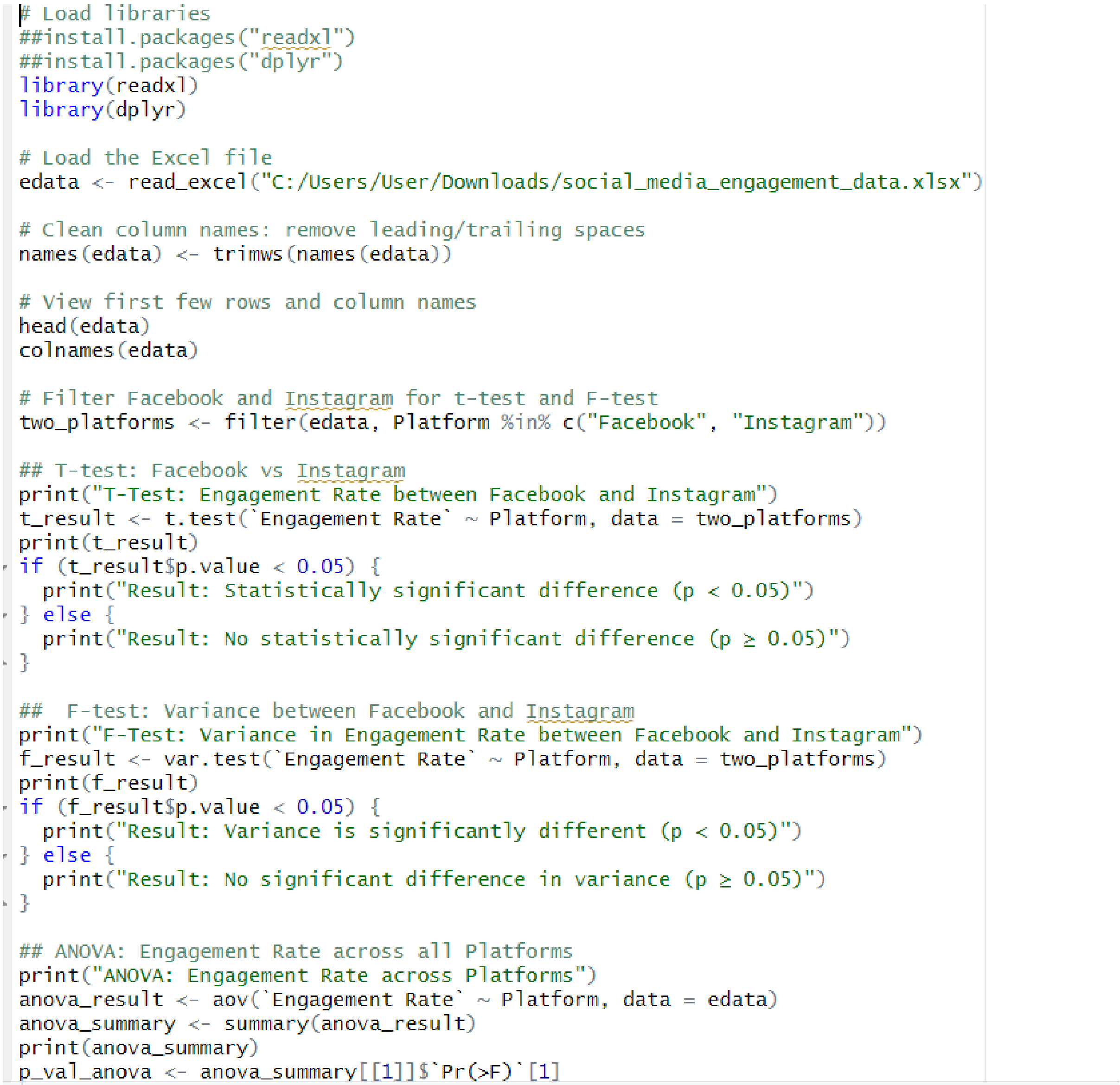
```
##Post Type Distribution Across Platforms
plt.figure(figsize=(8,5))
post_platform = pd.crosstab(df['Platform'], df['Post Type'])
sns.heatmap(post_platform, annot=True, cmap='Blues', fmt='d')
plt.title('Post Type Distribution by Platform')
plt.ylabel('Platform')
plt.xlabel('Post Type')
plt.show()
```

```
##Number of Posts:Weekdays Vs Weekends
weekday_type_counts = data['Weekday Type'].value_counts().reset_index()
weekday_type_counts.columns = ['Weekday Type', 'Post Count']

plt.figure(figsize=(6, 4))
sns.barplot(x='Weekday Type', y='Post Count', data=weekday_type_counts, palette='Blues')
plt.title('Number of Posts: Weekdays vs Weekends')
plt.ylabel('Number of Posts')
plt.xlabel('Weekday Type')
plt.tight_layout()
plt.show()
```

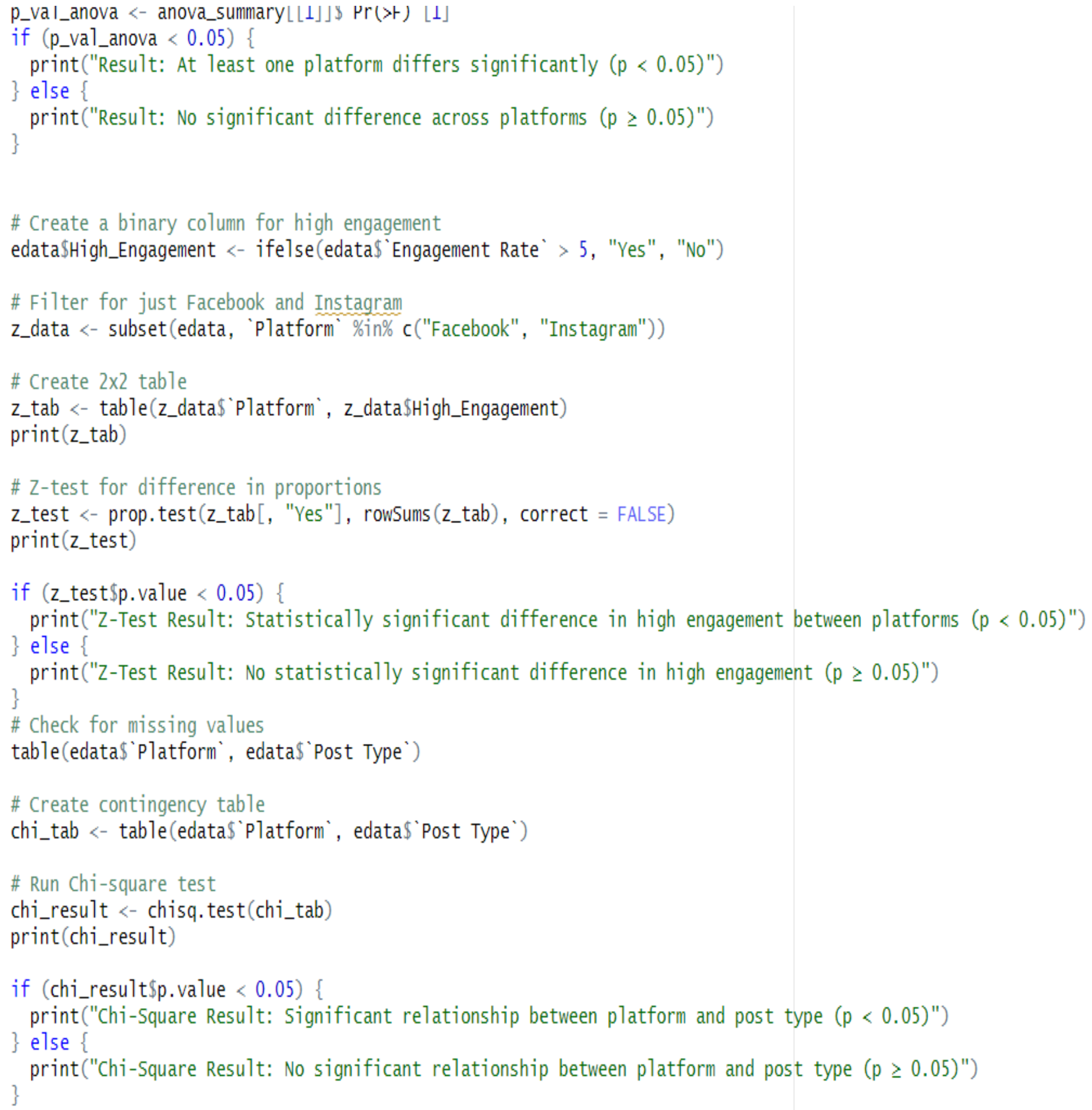
```
##Engagement Rate by Day of the Week
data['Date'] = pd.to_datetime(data['Date'])
data['Weekday'] = data['Date'].dt.day_name()

plt.figure(figsize=(9, 5))
sns.boxplot(x='Weekday', y='Engagement Rate', data=data, palette='pastel')
plt.title('Engagement Rate by Weekday')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

```
##Sentiment Distribution
plt.figure(figsize=(6, 4))
sns.countplot(x='Sentiment', data=data, palette='coolwarm')
plt.title('Sentiment Distribution of Posts')
plt.ylabel('Number of Posts')
plt.tight_layout()
plt.show()
```

R Programming



```
# Load libraries
##install.packages("readxl")
##install.packages("dplyr")
library(readxl)
library(dplyr)

# Load the Excel file
edata <- read_excel("C:/Users/User/Downloads/social_media_engagement_data.xlsx")

# Clean column names: remove leading/trailing spaces
names(edata) <- trimws(names(edata))

# View first few rows and column names
head(edata)
colnames(edata)

# Filter Facebook and Instagram for t-test and F-test
two_platforms <- filter(edata, Platform %in% c("Facebook", "Instagram"))

## T-test: Facebook vs Instagram
print("T-Test: Engagement Rate between Facebook and Instagram")
t_result <- t.test(`Engagement Rate` ~ Platform, data = two_platforms)
print(t_result)
if (t_result$p.value < 0.05) {
  print("Result: Statistically significant difference (p < 0.05)")
} else {
  print("Result: No statistically significant difference (p ≥ 0.05)")
}

##  F-test: Variance between Facebook and Instagram
print("F-Test: Variance in Engagement Rate between Facebook and Instagram")
f_result <- var.test(`Engagement Rate` ~ Platform, data = two_platforms)
print(f_result)
if (f_result$p.value < 0.05) {
  print("Result: Variance is significantly different (p < 0.05)")
} else {
  print("Result: No significant difference in variance (p ≥ 0.05)")
}

## ANOVA: Engagement Rate across all Platforms
print("ANOVA: Engagement Rate across Platforms")
anova_result <- aov(`Engagement Rate` ~ Platform, data = edata)
anova_summary <- summary(anova_result)
print(anova_summary)
p_val_anova <- anova_summary[[1]]$`Pr(>F)`[1]
```



```r
p_val_anova <- anova_summary[[1]]$`Pr(>F)`[1]
if (p_val_anova < 0.05) {
  print("Result: At least one platform differs significantly (p < 0.05)")
} else {
  print("Result: No significant difference across platforms (p ≥ 0.05)")
}


# Create a binary column for high engagement
edata$High_Engagement <- ifelse(edata$`Engagement Rate` > 5, "Yes", "No")

# Filter for just Facebook and Instagram
z_data <- subset(edata, `Platform` %in% c("Facebook", "Instagram"))

# Create 2x2 table
z_tab <- table(z_data$`Platform`, z_data$High_Engagement)
print(z_tab)

# Z-test for difference in proportions
z_test <- prop.test(z_tab[, "Yes"], rowSums(z_tab), correct = FALSE)
print(z_test)

if (z_test$p.value < 0.05) {
  print("Z-Test Result: Statistically significant difference in high engagement between platforms (p < 0.05)")
} else {
  print("Z-Test Result: No statistically significant difference in high engagement (p ≥ 0.05)")
}
# Check for missing values
table(edata$`Platform`, edata$`Post Type`)

# Create contingency table
chi_tab <- table(edata$`Platform`, edata$`Post Type`)

# Run Chi-square test
chi_result <- chisq.test(chi_tab)
print(chi_result)

if (chi_result$p.value < 0.05) {
  print("Chi-Square Result: Significant relationship between platform and post type (p < 0.05)")
} else {
  print("Chi-Square Result: No significant relationship between platform and post type (p ≥ 0.05)")
}
```