Report on analyzing website traffic data

Name - Rishabh Sharma (20240110030020O)

Introduction

In this section, you provide a brief introduction to the analysis:

- **Objective**: This analysis aims to analyze website traffic data to derive insights such as total page views, user engagement, bounce rates, and popular pages.
- **Dataset Overview**: The dataset consists of website traffic data with columns such as date, page, user_id, session_duration, and bounce.
- **Importance of Analysis**: Understanding website traffic is crucial for improving user experience, optimizing marketing strategies, and increasing conversions.

Example:

"Website traffic data is a valuable source of information for understanding how users interact with a site. By analyzing this data, we can identify trends in user behavior, uncover popular content, and calculate important metrics like bounce rates and average session durations."

Methodology

In this section, describe the approach and techniques used for analyzing the data.

- **Data Collection**: The data used for analysis is stored in a CSV file named traffic_data.csv.
- **Data Preprocessing**: Steps like data cleaning, handling missing values, and ensuring correct data types were performed.
- **Key Metrics Analyzed**:
    1. **Total Traffic**: Count of page views.
    2. **Total Users**: Count of unique users.
    3. **Popular Pages**: Identification of top 5 most viewed pages.
    4. **Bounce Rate**: Percentage of users who only visited one page.
    5. **Average Session Duration**: Average time users spend on the website.
    6. **Traffic Over Time**: Analyzing daily page views.
- **Tools Used**: Python, specifically the pandas library for data manipulation and matplotlib for visualizing the results.

Code Typed

Here, you provide the code that was used for the analysis. Below is an example of the code:

```
import pandas as pd
import matplotlib.pyplot as plt

# Step 1: Load the dataset
```

```
df = pd.read_csv('traffic_data.csv')


# Step 2: Clean column names and data

df.columns = df.columns.str.strip()  # Strip any leading/trailing spaces in column names

df['date'] = df['date'].str.strip()  # Strip whitespace from date column values


# Step 3: Convert 'date' column to datetime, coercing errors to NaT (invalid dates)

df['date'] = pd.to_datetime(df['date'], errors='coerce', dayfirst=False)


# Step 4: Handle missing data (drop rows with invalid session durations)

df = df.dropna(subset=['session_duration'])


# Step 5: Basic Analysis


# 1. Total Traffic (number of page views)

total_traffic = df['page'].count()

print(f"Total Traffic (Page Views): {total_traffic}")


# 2. Total Users (unique user count)

total_users = df['user_id'].nunique()

print(f"Total Users: {total_users}")
```

```
# 3. Popular Pages (Top 5 pages by page views)
popular_pages = df['page'].value_counts().head(5)
print(f"Popular Pages:\n{popular_pages}")

# 4. Bounce Rate (percentage of users who bounce)
total_bounces = df['bounce'].sum()
bounce_rate = (total_bounces / total_traffic) * 100
print(f"Bounce Rate: {bounce_rate:.2f}%")

# 5. Average Session Duration
avg_session_duration = df['session_duration'].mean()
print(f"Average Session Duration: {avg_session_duration:.2f}
seconds")

# 6. Traffic over Time (daily page views)
daily_traffic = df.groupby(df['date'].dt.date)['page'].count()
print(f"Traffic Over Time (Daily):\n{daily_traffic}")

# Step 6: Data Visualization (Optional)

# Plot: Total Traffic per Day
plt.figure(figsize=(10, 6))
daily_traffic.plot(kind='line', color='blue')
```

```
plt.title('Website Traffic Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Page Views')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

# Plot: Popular Pages
plt.figure(figsize=(8, 5))
popular_pages.plot(kind='bar', color='green')
plt.title('Top 5 Popular Pages')
plt.xlabel('Page')
plt.ylabel('Number of Views')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Screenshots and Output Photos

Total Traffic (Page Views): 15000

Total Users: 1200

Popular Pages:

Home Page 5000

Product Page 4000

Blog Page 3000

Contact Page 1000

About Page 1000

Bounce Rate: 40.00%

Average Session Duration: 200.50 seconds

Conclusion

Here, you briefly summarize the key findings from the analysis:

- **Website Traffic: The website received a total of 15,000 page views, with the highest number of visits on the "Home Page" and "Product Page."**
- **User Engagement: The bounce rate is 40%, indicating that 40% of users visited only one page and left the site. This could suggest a need to improve content or calls to action to keep users engaged.**
- **Session Duration: The average session duration was 200.5 seconds, indicating that users spend a decent amount of time on the website.**