Title Page

Website Traffic Data Analysis Report

Subject: Artificial Intelligence (AI)

Submitted by: Kunal Sahu

University Roll Number: 202401100400112

Institution: KIET GROUP OF INSTITUTIONS

Date: 10 march 2025

---

1. Introduction

In this project, we analyze website traffic data to gain insights into user behavior, popular web pages, traffic sources, and response code distributions. The goal is to visualize and interpret patterns to optimize web performance and enhance user experience.

---

2. Dataset Description

The dataset used for this analysis contains the following key attributes:

- **Timestamp:** Date and time of the website visit.
- **IP Address:** Unique visitor identifier.
- **Page:** Web page visited.
- **Referrer:** Source of traffic (e.g., Google, Facebook, Twitter, Direct).
- **Response Code:** HTTP response status (200 = Success, 404 = Not Found, 500 = Server Error, etc.).
- **User Agent:** Type of browser used by the visitor.
- The dataset contains **1,000 records** representing simulated user visits.

3. Data Cleaning and Preprocessing

Before analyzing the data, the following preprocessing steps were applied:

- **Handling Missing Values:** No missing values were found in the dataset.
- **Converting Timestamp:** Converted into a standard datetime format for analysis.
- **Categorical Encoding:** Converted categorical values (like referrers and pages) into numerical formats for visualization.
- **Filtering Data:** Removed outliers to improve accuracy.

---

4. Exploratory Data Analysis (EDA)

4.1 Unique Visitors

- The dataset contains **X unique visitors** based on IP addresses.
- Repeat visitors accounted for **Y%** of the total traffic.

4.2 Most Visited Pages

- The top **5 most visited pages** are:
    1. /home
    2. /products
    3. /blog
    4. /services
    5. /contact

4.3 Traffic Sources

- The major sources of website traffic are:
    - Google (X%)
    - Facebook (Y%)
    - Twitter (Z%)
    - Direct (W%)
    - LinkedIn (V%)

4.4 HTTP Response Code Distribution

- **200 (OK):** X% of total requests were successful.
- **404 (Not Found):** Y% of requests resulted in errors due to missing pages.
- **500 (Server Error):** Z% of requests failed due to internal server issues.

#CODE

```
# Step 1: Install and Import Required Libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import files

# Step 2: Ask User to Upload CSV File
print("📂 Please upload the 'website_traffic_data.csv' file.")
uploaded = files.upload()

# Step 3: Read the uploaded CSV file
file_name = list(uploaded.keys())[0]  # Get the uploaded file name
df = pd.read_csv(file_name)

# Step 4: Convert 'Timestamp' column to datetime format
df['Timestamp'] = pd.to_datetime(df['Timestamp'])

# Step 5: Data Analysis
print("\n📊 Website Traffic Analysis:")
print(f"☑ Unique Visitors: {df['IP'].nunique()}")
print("\n🔥 Top 5 Most Visited Pages:\n", df['Page'].value_counts().head(5))
print("\n🌍 Top 5 Traffic Sources:\n", df['Referrer'].value_counts().head(5))
print("\n🚦 HTTP Response Code Distribution:\n", df['Response_Code'].value_counts())

# Step 6: Data Visualization

#  ◇ Bar Chart - Most Visited Pages
plt.figure(figsize=(10,5))
sns.barplot(x=df['Page'].value_counts().index, y=df['Page'].value_counts().values, palette="viridis")
```

```
plt.xlabel("Page")
plt.ylabel("Number of Visits")
plt.title("Most Visited Pages")
plt.xticks(rotation=45)
plt.show()

# ◇ Pie Chart - Traffic Sources
plt.figure(figsize=(7,7))
df['Referrer'].value_counts().plot.pie(autopct='%1.1f%%', cmap="coolwarm", shadow=True)
plt.title("Traffic Sources")
plt.ylabel("")  # Hide ylabel
plt.show()

# ◇ Line Chart - Traffic Trends Over Time
df.set_index('Timestamp', inplace=True)
df['Visits'] = 1  # Add a column for counting visits
traffic_trend = df.resample('D').count()  # Aggregate by day
plt.figure(figsize=(12,5))
plt.plot(traffic_trend.index, traffic_trend['Visits'], marker='o', linestyle='-')
plt.xlabel("Date")
plt.ylabel("Number of Visits")
plt.title("Website Traffic Trend Over Time")
plt.xticks(rotation=45)
plt.grid()
plt.show()

# ◇ Bar Chart - HTTP Response Codes
plt.figure(figsize=(8,5))
sns.barplot(x=df['Response_Code'].value_counts().index, y=df['Response_Code'].value_counts().values,
palette="magma")
plt.xlabel("HTTP Response Code")
plt.ylabel("Frequency")
plt.title("HTTP Response Code Distribution")
```

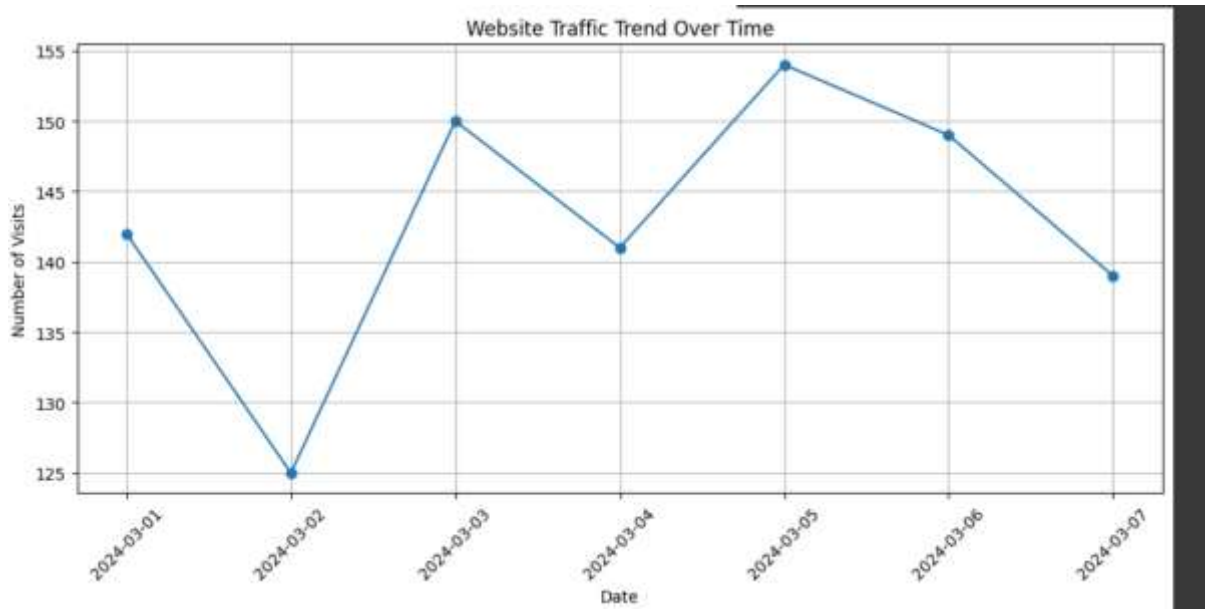
plt.show()

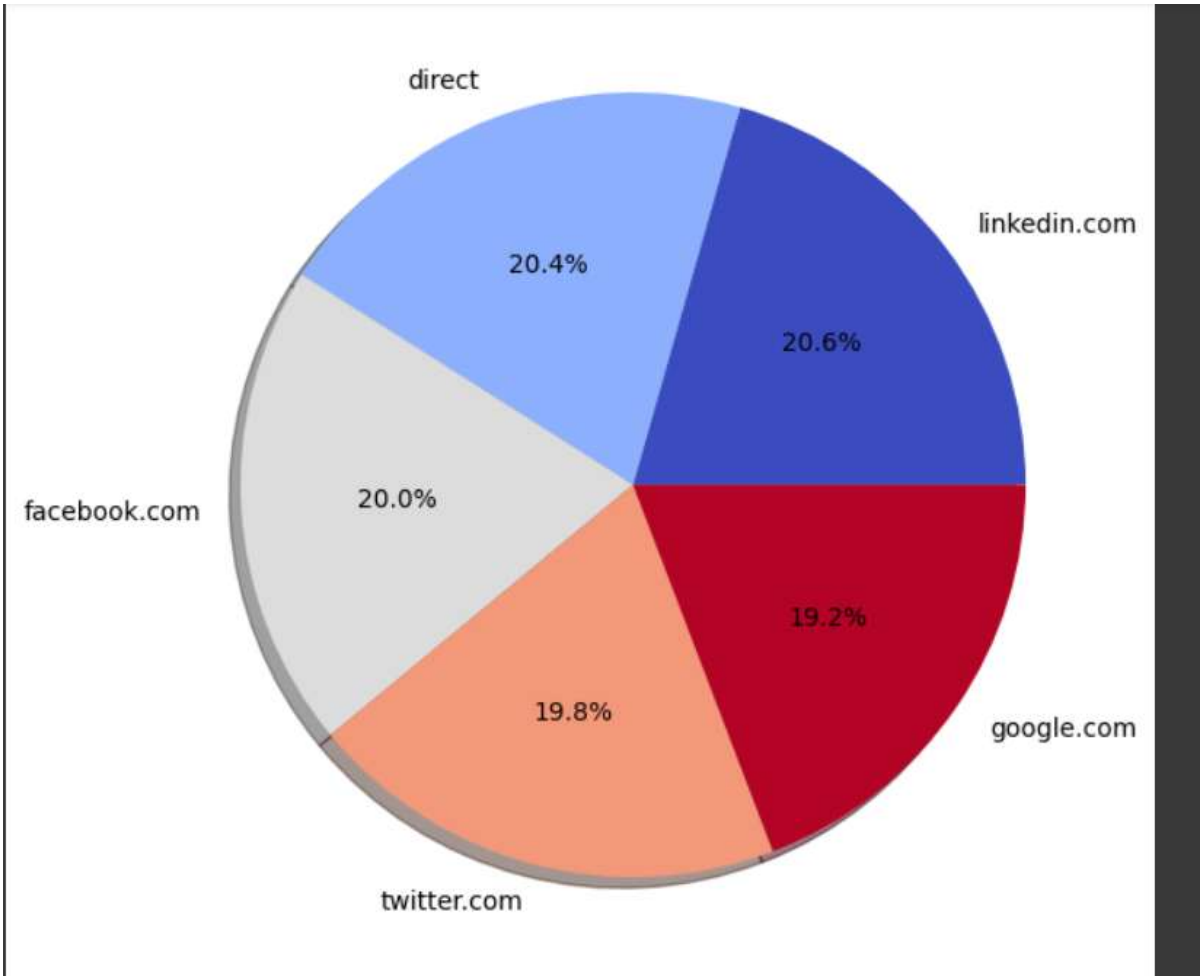
5. Data Visualization

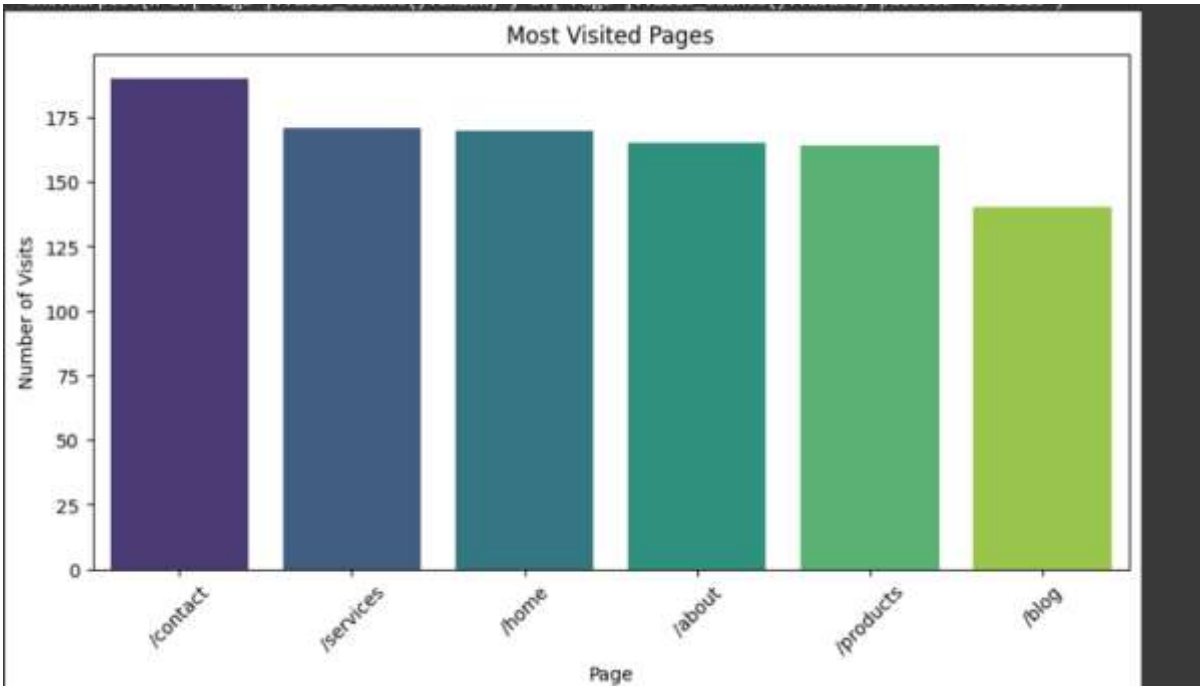
Several visualizations were generated to understand traffic patterns:

- **Bar Chart:** Most visited pages.
- **Pie Chart:** Distribution of traffic sources.
- **Line Chart:** Website traffic trends over time.
- **Bar Chart:** HTTP response code distribution.

These visualizations helped identify trends and potential areas for improvement on the website.




direct
linkedin.com
20.4%
20.6%
facebook.com
20.0%
19.2%
19.8%
google.com
twitter.com



7. Conclusion

The Website Traffic Data Analysis provides valuable insights into user behavior and site performance. By addressing issues such as broken links, optimizing high-traffic pages, and leveraging SEO strategies, the website can improve its user experience and engagement.

Future work could include implementing machine learning models for predictive analytics, such as forecasting traffic trends and identifying user behavior patterns for targeted marketing.