**ANP-D0449**

**DATA ANALYSIS USING PYTHON**

**WEBSITE TRAFFIC ANALYSIS**

**SUBMITTED BY:**

**Dharshana Sri J**

**Jesintha C**

**ABSTRACT:**

**Website Traffic Analysis for News Portals (BBC, The New York Times, etc.)**

In the digital news industry, understanding user engagement is essential for optimizing content, increasing readership, and improving user retention. This project focuses on analyzing website traffic patterns for news portals such as BBC, The New York Times, or Medium to gain insights into user behavior, reading time, peak browsing hours, and content preferences.

The analysis includes key metrics such as session duration, bounce rates, and clickstream behavior to determine how users interact with news articles. Using data filtering, we remove irrelevant traffic (such as users leaving within 5 seconds), ensuring a more accurate analysis of reader engagement. Clickstream analysis helps track user navigation patterns, showing how visitors move between articles, sections, and categories.

Additionally, trend identification techniques are applied to detect peak browsing hours, most-read topics, and seasonal trends in news consumption. The project also utilizes heatmaps and visual analytics to highlight which parts of a webpage receive the most attention (e.g., article headlines, images, or specific sections). By leveraging data-driven approaches, news organizations can enhance reader engagement, optimize advertisement placements, and personalize content recommendations to improve user experience and increase revenue generation.

**PROBLEM STATEMENT:**

News portals face challenges in retaining users and maximizing engagement due to a lack of comprehensive insights into visitor behavior. High bounce rates indicate that many users leave the site quickly without interacting with content, while inconsistent session durations suggest varying levels of interest. Furthermore, inadequate understanding of user preferences leads to ineffective content strategies. Additionally, improper ad placements reduce revenue potential. A structured traffic analysis system is required to identify trends, understand audience behavior, and provide data-driven solutions to enhance engagement and content effectiveness.

* Identify peak user activity times and engagement variations.
* Analyze session duration and bounce rates to evaluate content effectiveness.
* Assess traffic sources (direct, organic, referral, social media) to measure marketing impact.
* Use heatmaps and trend graphs to visualize user navigation behavior.
* Optimize ad placement and personalized content recommendations**.**

**SOLUTION APPROACH:**

1. **Data Collection:** Traffic data is extracted from web analytics tools such as Google Analytics to track user visits, session durations, referral sources, and page interactions.
2. **Data Processing:** The data is cleaned by handling missing values, filtering out bot-generated traffic, and segmenting users based on behavior.
3. **Exploratory Data Analysis (EDA):** High-traffic pages, user demographics, and browsing patterns are analyzed to identify trends and behavioral insights.
4. **Data Visualization:** Heatmaps, graphs, and trend charts are generated to visually interpret engagement patterns, peak activity times, and frequently visited pages.
5. **Optimization Strategies**: Based on insights, content placement is refined, ad positions are optimized for better revenue, and personalized recommendations are implemented to improve user retention.

**IMPLEMENTATION:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

# Load dataset (assuming CSV file format)

data = pd.read\_csv("website\_traffic.csv")

# Data Preprocessing

# Handling missing values

data.fillna({"session\_duration": data["session\_duration"].median(), "device\_type": "unknown"}, inplace=True)

# Filtering out short sessions (less than 5 seconds)

data = data[data["session\_duration"] >= 5]

# Exploratory Data Analysis (EDA)

plt.figure(figsize=(10, 5))

plt.hist(data['session\_duration'], bins=30, edgecolor='black', alpha=0.7)

plt.title("Distribution of Session Durations")

plt.xlabel("Session Duration (seconds)")

plt.ylabel("Frequency")

plt.show()

# Bounce rate calculation

bounce\_rate = (len(data[data["bounce"] == 1]) / len(data)) \* 100

print(f"Bounce Rate: {bounce\_rate:.2f}%")

# Heatmap visualization for engagement (simulated example)

pivot\_table = data.pivot\_table(index="hour", columns="day", values="session\_duration", aggfunc='mean')

plt.figure(figsize=(10, 6))

plt.imshow(pivot\_table, cmap="coolwarm", aspect='auto')

plt.colorbar(label="Avg. Session Duration")

plt.xticks(range(len(pivot\_table.columns)), pivot\_table.columns, rotation=90)

plt.yticks(range(len(pivot\_table.index)), pivot\_table.index)

plt.title("User Engagement Heatmap")

plt.xlabel("Day")

plt.ylabel("Hour")

plt.show()

# Clickstream Analysis (Navigation Flow)

navigation\_flow = data.groupby(["previous\_page", "current\_page"]).size().reset\_index(name='count')

print(navigation\_flow.sort\_values(by='count', ascending=False).head(10))

# Additional Data Visualization

# Bar Chart for Device Usage

plt.figure(figsize=(8, 5))

device\_counts = data["device\_type"].value\_counts()

plt.bar(device\_counts.index, device\_counts.values, color="skyblue", edgecolor="black")

plt.title("Device Usage Distribution")

plt.xlabel("Device Type")

plt.ylabel("Count")

plt.xticks(rotation=45)

plt.show()

# Time-Series Analysis for Peak Browsing Hours

hourly\_sessions = data.groupby("hour")["session\_id"].count()

plt.figure(figsize=(10, 5))

plt.plot(hourly\_sessions.index, hourly\_sessions.values, marker="o", linestyle="-", color="b")

plt.title("Website Traffic by Hour")

plt.xlabel("Hour of the Day")

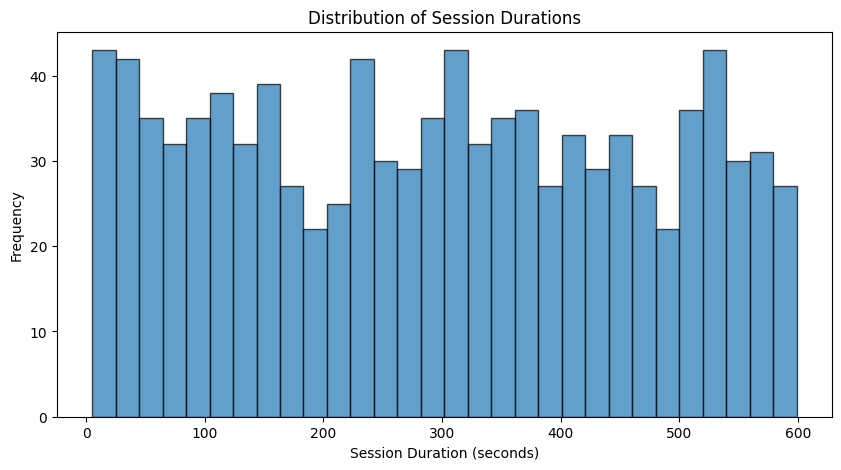
plt.ylabel("Number of Sessions")

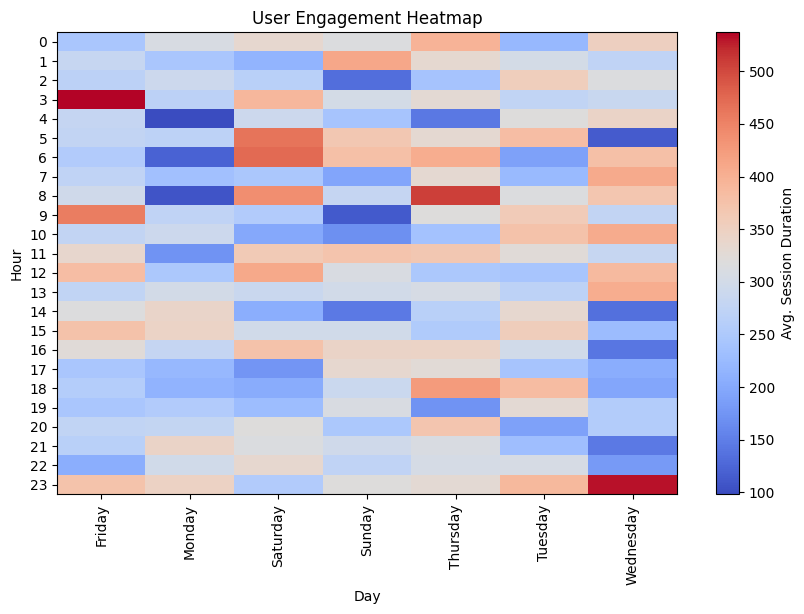
plt.xticks(range(0, 24))

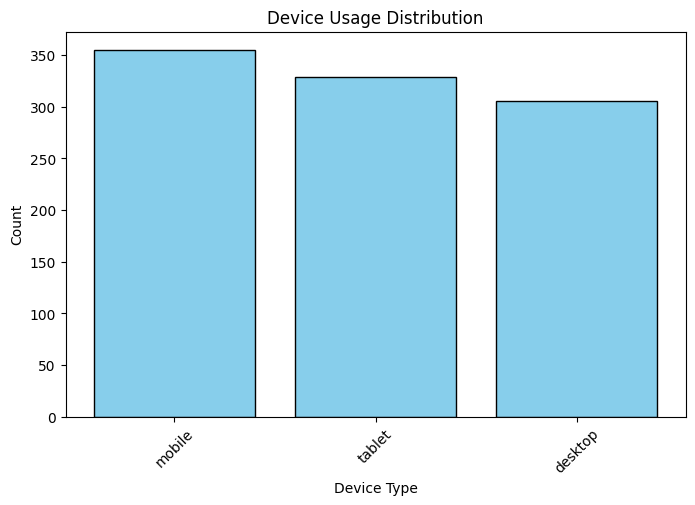
plt.grid()

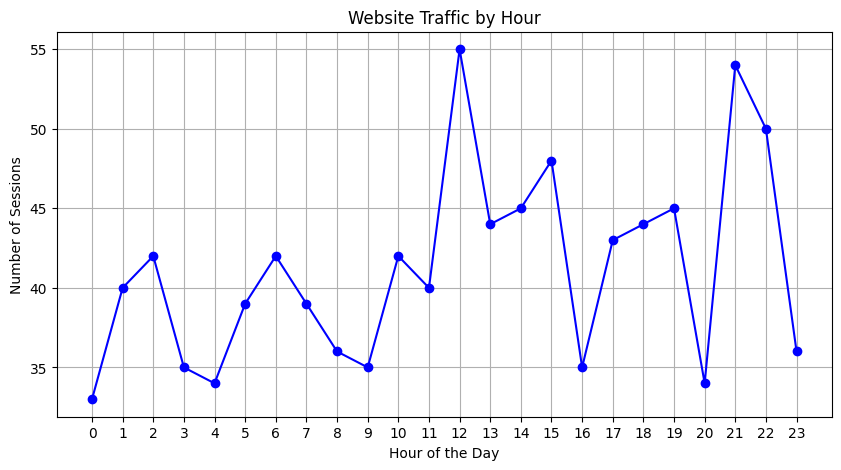
plt.show()

**OUTPUT VISUALIZATION:**









**CONCLUSION:**

The website traffic analysis provides crucial insights into user engagement and browsing behavior on news portals. By examining session durations, bounce rates, clickstream patterns, and device usage, we identify factors affecting user retention. Heatmaps highlight areas of high interaction, while peak browsing hours guide content scheduling and marketing strategies. The findings enable news organizations to optimize content placement, improve user experience, and enhance engagement through data-driven decisions, ultimately leading to better audience retention and increased ad revenue.