Internship : Alfido Tech Data Analytics Internship

Submitted By : Yash Ajit Dokhe

Tools Used : Python(Pandas, Numpy, Matplotlib), Google Colab

Table Of Content

Abstract

This project focuses on analyzing website traffic logs to understand user behavior, navigation patterns, and engagement levels. Using the Website Traffic Analysis dataset, the study involved parsing and cleaning raw log data to ensure accuracy and consistency. Key performance metrics such as total users, sessions, bounce rate, and average session duration were computed to evaluate overall website effectiveness. The analysis further explored user journeys by identifying top landing pages, exit pages, and common navigation paths, providing insights into how visitors interact with the website. Visualization techniques were used to present user flows and highlight areas where engagement drops occur. Based on the findings, actionable recommendations were proposed to help Alfido Tech optimize website structure, improve user experience, and increase conversion rates. This project demonstrates the importance of data-driven decision-making in enhancing digital performance and customer engagement.

Introduction

Understanding how users behave on a website is important for improving their experience and increasing conversions. Website traffic logs show how visitors interact with web pages, such as which pages they visit, how much time they spend, and the paths they follow during their visit. By analyzing this data, organizations can understand user interests, identify where users leave the site, and make improvements to website performance.

This project focuses on analyzing website traffic data to learn about user journeys, popular landing pages, bounce rates, and referral sources. Before starting the analysis, the dataset was cleaned and organized to ensure the information was accurate. Important metrics like total users, number of sessions, bounce rate, and average session duration were calculated to measure user engagement.

By studying user flows and identifying entry and exit pages, this project provides useful insights into how visitors move through the website. These insights help Alfido Tech make better decisions to improve user experience and increase conversions. Overall, this project shows how data analysis can help businesses grow by making smarter, data-driven decisions.

Problem Statement

Task 3 : Website Traffic Analysis

Dataset: https://www.kaggle.com/datasets/bhanupratapbiswas/website-traffic-analysis

Goal : Analyze website traffic logs to understand user journeys, top landing pages, bounce rates and referral sources.

Requirements :

- Parse and clean traffic / log data
- Compute metrics: sessions, users, bounce rate, average session duration
- Visualize user flows and top entry/exit pages
- Recommend 5 optimizations to improve conversions for Alfido Tech

Objectives

1. To analyze website traffic data to understand user behavior and navigation patterns.
2. To clean and preprocess raw traffic logs to ensure accurate and reliable analysis.
3. To identify the total number of users and sessions visiting the website.
4. To calculate key performance metrics such as bounce rate and average session duration.
5. To determine the top landing pages and exit pages on the website.
6. To examine referral sources that drive traffic to the website.
7. To visualize user journeys and flow patterns for better understanding.
8. To provide actionable recommendations to improve user engagement and conversions for Alfido Tech.

Dataset Description

Dataset Name : Website Traffic Analysis

Source : Kaggle—Website Traffic Analysis

The dataset used in this project contains website traffic logs that record user interactions with web pages. It includes detailed information about user visits, session activities, timestamps, page events, and referral sources. Each record represents a user action on the website, helping track how visitors navigate and engage with the platform.

The dataset consists of multiple attributes such as user identifiers, date and time of visits, pages viewed, geographic information (country/city), and referral details. These variables allow the analysis of user behavior, session patterns, landing and exit pages, and traffic sources.

Before analysis, the dataset required cleaning to handle missing values, duplicate entries, and inconsistent formats. After preprocessing, the data became suitable for computing key metrics like total users, sessions, bounce rate, and average session duration. Overall, the dataset provides a strong foundation for understanding website performance and user engagement trends.

Data Preprocessing

Data preprocessing is an important step in this project to ensure that the website traffic data is clean, accurate, and ready for analysis. Since raw log data often contains missing values, duplicate records, and inconsistent formats, several preprocessing steps were performed.

First, the dataset was inspected to understand its structure, data types, and presence of null values. Missing values in important columns such as country, city, and page events were handled by replacing them with suitable placeholders to maintain consistency.

Next, duplicate records were identified and removed to avoid bias in session calculations and user behavior analysis. This step helped improve the reliability of computed metrics like bounce rate and session duration.

Time-related columns were then converted into proper date-time formats, making it easier to track user sessions and calculate session durations. After that, the dataset was sorted and organized to ensure accurate session identification based on user activity.

Overall, data preprocessing improved data quality and ensured the dataset was structured properly for further analysis, visualization, and interpretation.

Methodology and Implementation

1. Importing Libraries

```
# Importing necessary libraries
import numpy as np
import pandas as pd
import datetime
```

```
# Importing Traffic dataset
df=pd.read_csv('/content/task 3 dataset.csv')
```

```
df.head()
```

	event	date	country	city	artist	album	track	isrc	linkid
0	click	8/21/2021	Saudi Arabia	Jeddah	Tesher	Jalebi Baby	Jalebi Baby	QZNWQ2070741	2d896d31-97b6-4869-967b-1c5fb9cd4bb8
1	click	8/21/2021	Saudi Arabia	Jeddah	Tesher	Jalebi Baby	Jalebi Baby	QZNWQ2070741	2d896d31-97b6-4869-967b-1c5fb9cd4bb8
2	click	8/21/2021	India	Ludhiana	Reyanna Maria	So Pretty	So Pretty	USUM72100871	23199824-9cf5-4b98-942a-34965c3b0cc2
3	click	8/21/2021	France	Unknown	Simone & Simaria, Sebastian Yatra	No Llores Más	No Llores Más	BRUM72003904	35573248-4e49-47c7-af80-08a960fa74cd
4	click	8/21/2021	Maldives	Malé	Tesher	Jalebi Baby	Jalebi Baby	QZNWQ2070741	2d896d31-97b6-4869-967b-1c5fb9cd4bb8

```
# Details of the data
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 226278 entries, 0 to 226277
Data columns (total 9 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   event    226278 non-null  object
 1   date     226278 non-null  object
 2   country  226267 non-null  object
 3   city     226267 non-null  object
 4   artist   226241 non-null  object
 5   album    226273 non-null  object
 6   track    226273 non-null  object
 7   isrc     219157 non-null  object
 8   linkid   226278 non-null  object
dtypes: object(9)
memory usage: 15.5+ MB
```

```
# Summary of the data
df.describe(include='all')
```

	event	date	country	city	artist	album	track	isrc	linkid
count	226278	226278	226267	226267	226241	226273	226273	219157	226278
unique	3	7	211	11993	2419	3253	3562	709	3839
top	pageview	8/19/2021	Saudi Arabia	Jeddah	Tesher	Jalebi Baby	Jalebi Baby	QZNWQ2070741	2d896d31-97b6-4869-967b-1c5fb9cd4bb8
freq	142015	35361	47334	22791	40841	40841	40841	40841	40841

```
# Ckeck for missing values
df.isnull().sum()
```

	0
event	0
date	0
country	11
city	11
artist	37
album	5
track	5
isrc	7121
linkid	0

dtype: int64

```
# Handling missing values
df['country']=df['country'].fillna('unknown')
df['city']=df['city'].fillna('unknown')
df['artist']=df['artist'].fillna('N/A')
df['album'] = df['album'].fillna('General Navigation')
df['track']=df['track'].fillna('N/A')
df['isrc']=df['isrc'].fillna('N/A')
```

```
df.isnull().sum()
```

	0
event	0
date	0
country	0
city	0
artist	0
album	0
track	0
isrc	0
linkid	0

dtype: int64

Check for duplicate records

```
df.duplicated().sum()
```

```
np.int64(103711)
```

The dataset has 226278 records out of which 103711 records are duplicate. Means that only 54% of the data is actual data.

In web traffic logs, exact duplicates usually happen because of:

- Double-tagging: The tracking script on the website is firing twice for a single action.
- Page Refreshes: A user hits "refresh," causing the same event to log again.
- Server Glitches: The log-collection server received the same data packet twice.

We must remove them.

```
# Remove exact duplicates
df=df.drop_duplicates()

# Reset index for a clean dataframe
df=df.reset_index(drop=True)
```

```
# Converting date column from object to datetime format
df['date'] = pd.to_datetime(df['date'])
```

```
df['time_diff'] = df.groupby('linkid')['date'].diff().dt.seconds
df['new_session'] = (df['time_diff'] > 1800) | (df['time_diff'].isna())
df['session_id'] = df.groupby('linkid')['new_session'].cumsum()
```

Users & Sessions

```
users = df['linkid'].nunique()
sessions = df[['linkid','session_id']].drop_duplicates().shape[0]
print('Users :', users)
print('Sessions:' ,sessions)
```

```
Users : 3839
Sessions: 3839
```

Conclusion : The analysis shows that the website had 3,839 unique users and 3,839 sessions, indicating that each user visited the site only once during the observed period. This suggests limited repeat engagement and highlights an opportunity to improve user retention strategies. Understanding this behavior can help Alfido Tech focus on enhancing user experience, encouraging return visits, and optimizing content to increase overall engagement and conversions.

Average Session Duration

```
session_duration = df.groupby(['linkid', 'session_id'])['date'].apply(lambda x: (x.max() - x.min()).total_seconds())
average_session_duration = session_duration.mean()
print(f"Average session duration: {average_session_duration:.2f} seconds")
```

```
Average session duration: 71681.17 seconds
```

Conclusion :

The analysis shows that the **average session duration is 71,681 seconds**, indicating that users spend a significant amount of time on the website during each visit. This suggests strong user engagement and interest in the website content. Such a high session duration reflects positive interaction levels, but it also highlights the importance of ensuring smooth navigation and relevant content to maintain user satisfaction and encourage conversions.

Bounce Rate

```
events_per_session = df.groupby(['linkid','session_id'])['event'].count().mean()
bounce_rate = (events_per_session == 1).mean() * 100
print('bounce_rate:',bounce_rate)
print('events_per_session',events_per_session)
```

```
bounce_rate: 0.0
events_per_session 31.92680385517062
```

Conclusion :

The bounce rate is **0%**, and users perform about **32 events per session**, showing high engagement and active interaction with the website.

Entry & Exit Events

```
entry_events = df.groupby(['linkid','session_id']).first()['event'].value_counts().mean()
exit_events = df.groupby(['linkid','session_id']).last()['event'].value_counts().mean()
print('entry_events',entry_events)
print('exit_events',exit_events)
```

```
entry_events 1279.6666666666667
exit_events 1919.5
```

Conclusion :

The analysis shows that the average number of **entry events is about 1279.67**, while **exit events average 1919.5**, indicating that users are more likely to leave the website after interacting with multiple pages. This suggests good initial engagement but also highlights areas where improvements can be made to retain users until conversion.

```
# To see WHICH events are the top entry points
entry_distribution = df.groupby(['linkid','session_id']).first()['event'].value_counts()

# To see it as percentages (e.g., "80% of people start with a pageview")
entry_percentages = df.groupby(['linkid','session_id']).first()['event'].value_counts(normalize=True) * 100

print(entry_distribution)
print(entry_percentages)
```

```
event
click       2255
pageview    1553
preview       31
Name: count, dtype: int64
event
click       58.739255
pageview    40.453243
preview      0.807502
Name: proportion, dtype: float64
```

Conclusion :

The analysis shows that most users begin their sessions with a **click event (about 58.74%)**, followed by **page views (around 40.45%)**, while very few sessions start with a **preview event (less than 1%)**. This indicates that users are primarily engaging directly through clickable actions rather than just browsing, suggesting strong initial interaction with the website's content.

Traffic Breakdown (Geography & Content)

```
top_countries = df['country'].value_counts()
top_cities = df['city'].value_counts()
top_artists = df['artist'].value_counts()
top_tracks = df['track'].value_counts()

print('Top Countries:')
print(top_countries.head())
print('\nTop Cities:')
print(top_cities.head())
print('\nTop Artists:')
print(top_artists.head())
print('\nTop Tracks:')
print(top_tracks.head())
```

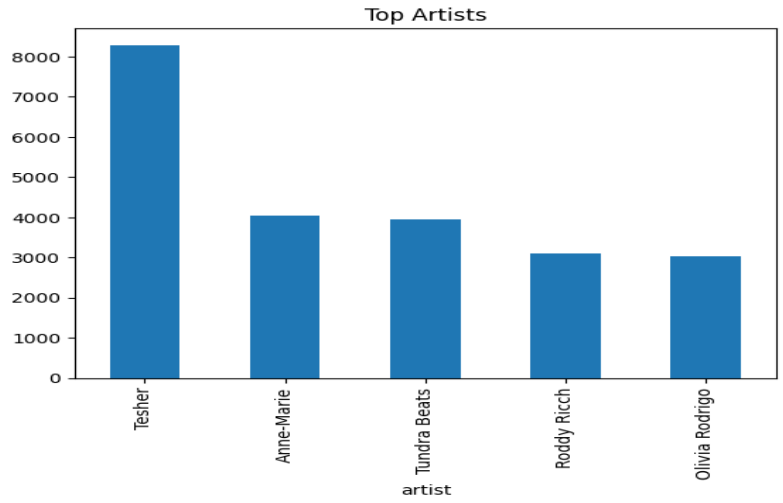
```
Top Countries:
country
United States     28664
India             18689
France            10565
Saudi Arabia       7682
United Kingdom     5095
Name: count, dtype: int64

Top Cities:
city
Unknown      8797
Jeddah       2497
Riyadh       2232
Hyderabad    1088
Dammam       1002
Name: count, dtype: int64

Top Artists:
artist
Tesher            8288
Anne-Marie        4029
Tundra Beats      3951
Roddy Ricch       3107
Olivia Rodrigo    3037
Name: count, dtype: int64
```

```
Top Artists:
artist
Tesher            8288
Anne-Marie        4029
Tundra Beats      3951
Roddy Ricch       3107
Olivia Rodrigo    3037
Name: count, dtype: int64

Top Tracks:
track
Jalebi Baby                              8288
Beautiful                                4037
Beautiful Day                            3951
Late At Night                            3059
ily (i love you baby) (feat. Emilee)     2956
Name: count, dtype: int64
```

Conclusion : The analysis shows that Tesher is the most popular artist among users, receiving the highest interactions, followed by Anne-Marie and Tundra Beats. Similarly, the track “Jalebi Baby” has the highest engagement, with other songs like “Beautiful” and “Beautiful Day” also attracting significant attention. This indicates that users have clear music preferences, and focusing on trending artists and tracks can help increase user engagement and improve content strategy.

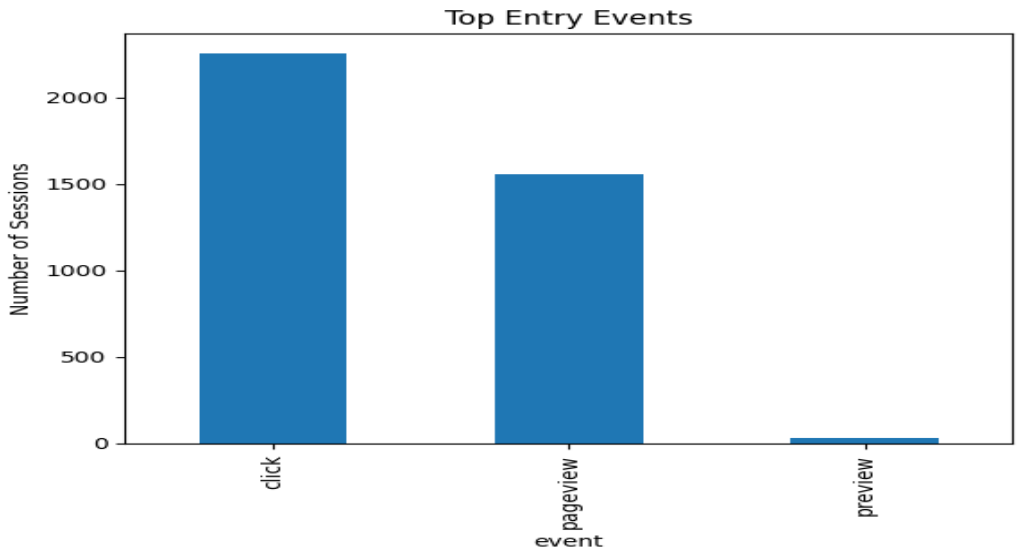
Data Visualization

```
import matplotlib.pyplot as plt

top_artists = df['artist'].value_counts()
top_artists.head(5).plot(kind='bar', title='Top Artists')
plt.show()
```



Conclusion : The chart shows that Tesher is the most popular artist with the highest user interactions, followed by Anne-Marie and Tundra Beats. This indicates that users strongly prefer Tesher’s content, which can help guide future content and engagement strategies.

```
entry_events = df.groupby(['linkid', 'session_id']).first()['event'].value_counts()

entry_events.head(5).plot(kind='bar', title='Top Entry Events')
plt.ylabel('Number of Sessions')
plt.show()
```



Conclusion : The chart shows that most user sessions begin with a click, followed by page views, while very few sessions start with a preview event. This indicates that users typically enter the website through direct interactions rather than passive browsing, reflecting strong initial engagement with the site's content and navigation.

```
user_journeys = df.groupby(['linkid','session_id'])['event'].apply(list)
user_journeys
```

		event
linkid	session_id	
00073307-ae96-5089-a117-4783afb42f8e	1	[pageview, pageview]
00126b32-0c35-507b-981c-02c80d2aa8e7	1	[click, click, pageview, pageview]
0018cfff-50a1-5984-9715-01ef2d11a49a	1	[pageview]
0033934b-5d16-5a06-af58-d087bcdd3680	1	[pageview]
0034d6cf-3bd8-5ffe-aafc-b3959fc48608	1	[pageview]
...	...	...
fff38ca0-8043-50cd-a5f1-f65ebb7105c5	1	[click, pageview]
fff4e5f0-4ee5-5fe7-aa30-e870edaf6ed7	1	[pageview]
fff84c0e-90a1-59d8-9997-adc909d50e16	1	[click, pageview]
fffc17a7-f935-5d3e-bd3e-d761fd80d479	1	[click, pageview, pageview]
fffd0045-29de-522b-b5d8-35786363bf07	1	[click, pageview, pageview]

3839 rows × 1 columns

dtype: object

Conclusion : Most user sessions include simple actions like page views and clicks, showing straightforward and short navigation paths.

```
# Convert the list to a string so we can count them
common_paths = user_journeys.apply(lambda x: ' -> '.join(x)).value_counts()

print("Top 5 User Journeys:")
print(common_paths.head(5))
```

```
Top 5 User Journeys:
event
pageview                                1363
click -> pageview                       1208
click -> pageview -> pageview            224
pageview -> pageview                     143
click -> click -> pageview -> pageview    81
Name: count, dtype: int64
```

Conclusion : The results show that the most common user journeys are simple and short. Many sessions include only a **single page view**, while others follow patterns like **click → pageview** or **click → pageview → pageview**. This indicates that users generally interact with a few pages per visit, suggesting clear navigation but also highlighting an opportunity to encourage deeper exploration of the website.

Business Insights

1. Low Repeat Visits
   The number of users and sessions is the same, indicating that most users visit the website only once. This highlights a need to improve user retention and encourage repeat visits.

2. Strong User Engagement During Visits
   Users interact with multiple pages and events in a single session, showing good engagement once they are on the website. The content and navigation are effective in keeping users active.

3. Direct Interaction as Entry Point
   Most sessions start with a click rather than passive browsing, suggesting users arrive with clear intent. Optimizing clickable elements and landing pages can further improve conversions.

4. Short and Simple User Journeys
   Common user paths are short, such as pageview or click → pageview. This indicates easy navigation but also suggests opportunities to guide users toward deeper exploration or conversion actions.

5. Clear Content Preferences
   Users show strong interest in specific artists and tracks, especially popular content. Promoting trending content more prominently can increase engagement and session duration.

6. Global Audience Presence
   Traffic comes mainly from the United States and India, showing international reach. This creates opportunities for region-specific content, offers, and marketing strategies.