**WEB TRAFFIC Data Analysis**

**Project submitted to the**

# APSSDC

**Bachelor of Technology**

**In**

**Computer Science and Engineering**

**Submitted By**

**CHAKRADHAR PARIMI – 23VV5A0574**

**SHAIK KUDHAB- 23VV5A0578**

**DUGGIVALASA PRAGATHI – 23BCE20173**



**Under the guidance of**

**K. Meenakshi**

**M. Ruthuma**

**JUNE 2025**

# Abstract

Understanding user behavior through website analytics is critical for businesses seeking to enhance digital engagement and optimize conversion strategies. In this project, a structured approach was undertaken to analyze user interaction patterns using a comprehensive web traffic dataset. The dataset included key performance indicators such as session duration, time on page, bounce rate, conversion rate, and traffic source. These metrics were explored to determine how visitors engage with the website and what factors influence successful user conversion.

Data preprocessing was a vital phase, ensuring accuracy and consistency through removal of duplicates, handling of missing values, and standardization of column formats. Post-cleaning, exploratory data analysis (EDA) was conducted using Python libraries—Pandas for data manipulation, NumPy for statistical computation, and Matplotlib for visualization. Over 20 curated queries provided insights into conversion behaviors, high bounce segments, and engagement differences across traffic channels. Visualizations such as histograms, scatter plots, pie charts, boxplots, and line graphs were employed to simplify complex relationships between variables and reveal actionable trends.

Filtering techniques helped isolate sessions of interest, including those with significant time on page but no conversion—highlighting potential UX shortcomings. Grouping and aggregation, such as calculating average conversion rates by traffic source, revealed which channels consistently performed well and which required strategic improvement. The analysis illuminated that organic and direct traffic tend to yield higher engagement, while referral traffic showed variability in user retention and conversion performance.

This project culminated in a set of insights that can directly influence real-world decisions, such as optimizing landing page designs, reallocating marketing budgets, and improving content strategy to retain users. The outcome demonstrates the power of a well-executed data analytics workflow—from data cleaning and transformation to insightful storytelling and recommendation. By combining technical rigor with communicative clarity, this report lays a foundation for smarter digital experiences and data-driven growth strategies.

Introduction

In the age of digital transformation, websites serve as the primary interface between businesses and users. Every interaction—whether a click, a scroll, or a form submission—generates data that holds valuable insights into user intent, interest, and satisfaction. Organizations that can effectively interpret this data gain a strategic advantage in crafting better user experiences, allocating resources wisely, and improving overall conversion performance.

This project focuses on the systematic analysis of website traffic data to uncover patterns in user engagement and identify factors that influence conversions. The dataset comprises detailed attributes such as session duration, time on page, bounce rate, conversion rate, and traffic source, allowing for a multidimensional understanding of visitor behavior.

To process and extract insights from this data, the project uses Python’s powerful analytical ecosystem—leveraging **Pandas** for structured data manipulation, **NumPy** for numerical computation, and **Matplotlib** for generating visual representations. By combining descriptive statistics, logical filtering, and targeted visualizations, the analysis aims to answer key questions: Which traffic sources deliver quality engagement? What is the relationship between session depth and conversion likelihood? Where does user attention drop off?

Beyond technical exploration, the project emphasizes clear communication of findings.

## Problem Statement

## In the digital ecosystem, a business’s success is increasingly determined by how effectively it can understand and respond to user behavior. Every session on a website generates data points—duration, origin, engagement level, conversion outcome—that collectively reflect the site's ability to attract, retain, and convert visitors. However, raw web analytics often lack context and clarity, making it difficult to translate numerical trends into actionable insights.

## Organizations frequently face challenges such as high bounce rates, inconsistent conversion performance across traffic sources, and unclear engagement patterns. Without targeted analysis, it becomes difficult to identify which acquisition channels are genuinely effective, which sections of the site are underperforming, and what user behaviors precede successful conversions. Furthermore, decisions based on intuition rather than data can lead to misaligned marketing strategies, wasted resources, and suboptimal user experiences.

## This project seeks to address these challenges by conducting an in-depth analysis of website traffic data. It aims to filter out noise, discover high-impact patterns, and visualize key metrics like session duration, traffic source composition, time on page, and conversion rates. By leveraging Python’s analytical stack and applying structured querying, aggregation, and filtering techniques, the project identifies crucial behavioral insights that inform smarter decisions in content strategy, traffic acquisition, and user experience design.

## Ultimately, the problem lies not in the availability of web traffic data—but in the ability to extract and interpret meaningful signals from it. This project solves that by transforming raw interaction logs into a well-articulated story about user behavior and its impact on digital performance.

## Project Objectives

# Identify high-performing traffic sources based on conversion rate and session duration

# Examine the relationship between bounce rate, time on page, and conversion likelihood

# Apply data filtering and grouping techniques to isolate meaningful subsets for analysis

# Visualize metrics using Python to uncover trends and support decision-making

# Transform raw web logs into a structured narrative that aids strategy development

# Create a presentation and report that clearly communicate analytical findings

# Strengthen proficiency with data science tools including Pandas, NumPy, and Matplotlib

# Demonstrate the effectiveness of simple analytical workflows in solving real-world problems

# Lay groundwork for future integration of machine learning or dashboard deployment

# System Requirements

**Hardware Requirements**

* Processor: Intel Core i3 or higher (recommended: Core i5/i7 for faster computation)
* RAM: Minimum 4 GB (recommended: 8 GB or higher for handling larger datasets)
* Storage: At least 500 MB free disk space for Python environment and project files
* Display: 1366×768 resolution or higher for viewing Jupyter Notebooks and plots clearly

**Software Requirements:**

# Operating System: Windows 11 (compatible with Windows 10, Linux, or macOS)

# Python Version: Python 3.10 or above (used: Python 3.11)

# Code Editor: Jupyter Notebook (via Anaconda distribution or VS Code extension)

**Libraries required:**

** Pandas:** For data manipulation and querying

 **NumPy:** For statistical and numerical operations

** Matplotlib:** For visualizations and plotting

# Architecture

The architecture of this project is structured around a modular and sequential data analytics pipeline using Python. It begins with the **Data Ingestion Layer**, where raw traffic data is imported into a Jupyter Notebook environment using Pandas. This is followed by the **Preprocessing Layer**, which includes data cleaning operations—such as handling null values, correcting data types, and removing duplicates—to ensure analytical readiness. Next, the **Exploratory Analysis Layer** performs descriptive statistics, filtering, grouping, and aggregation to uncover patterns and summarize user behavior across traffic sources. The **Visualization Layer** leverages Matplotlib to generate clear and informative charts that represent engagement metrics and conversion trends. These insights are further structured into a **Presentation Layer**, comprising narrative interpretations, graphical slides, and documentation that make results accessible to both technical and non-technical stakeholders. This layered architecture ensures a logical flow, scalability for future enhancements, and clarity in both analysis and communication.

## ◆ 1. Library Imports & Dataset Loading

Essential libraries such as Pandas, NumPy, Matplotlib, Seaborn, datetime, and warnings are imported. The APSRTC dataset is loaded from a CSV file into a Pandas DataFrame for processing.

## ◆ 2. Initial Data Exploration

Basic exploration is performed using functions like .head(), .info(), and .describe() to understand the structure, data types, and key variables such as revenue, distance, fuel, and passenger count.

## ◆ 3. Data Cleaning

Missing values are identified and handled. Data types are standardized, duplicates are removed, and categorical values are cleaned to ensure consistency.

## ◆ 4. Feature Engineering & Aggregation

New metrics are created to enhance analysis:

* Revenue per kilo meter
* Fuel efficiency
* Passengers per kilo meter
* Occupancy category (based on thresholds) Data is grouped by route and bus type to enable comparative analysis.





## ◆ 5. Data Visualization

A wide range of plots are generated using Matplotlib and Seaborn to uncover trends and patterns:

* Revenue by bus type
* Occupancy rate distribution
* Passengers by route
* Monthly revenue trends
* Fuel efficiency by bus type
* Occupancy by day of week
* Revenue vs distance
* Depot performance
* Occupancy category distribution
* Correlation heatmap
* Revenue distribution by bus type
* Daily revenue time series with rolling average

◆ **6. Advanced Analytics & Insights** The analysis identifies:

* Top 5 routes by revenue
* Best-performing bus types
* Seasonal trends across months
* Efficiency metrics like average occupancy, revenue per km, and fuel usage Key statistics are summarized to support strategic decision-making.

# vz‘’ Libraries Used

This project utilizes four essential Python libraries—Pandas, NumPy, Matplotlib—each contributing uniquely to the data analysis workflow.

**Pandas**

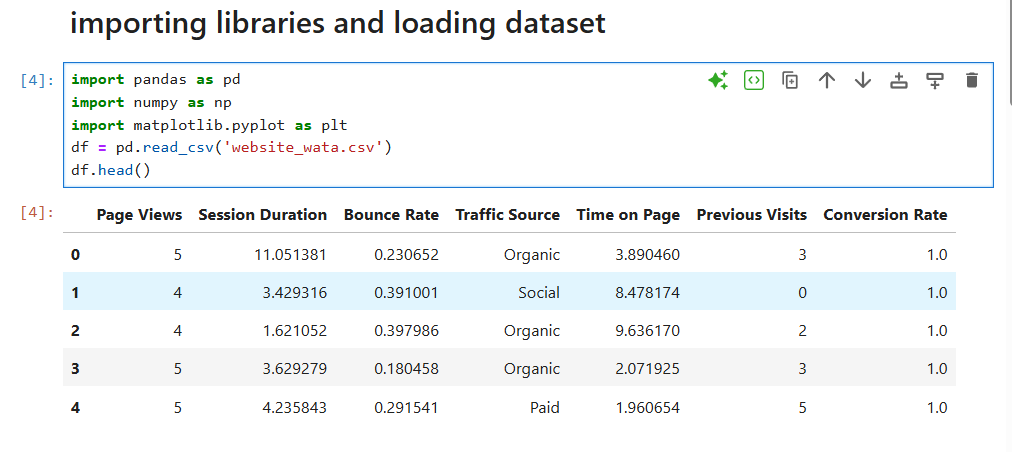
Pandas is a powerful library designed for data manipulation and analysis. It allows users to work seamlessly with tabular data, such as CSV or Excel files, through intuitive data structures like DataFrames . In this project, Pandas was used extensively for importing the dataset, cleaning inconsistent entries, filtering sessions based on conditions, and performing group-wise aggregations. Its expressive syntax made it easy to transform raw web traffic logs into structured insights, enabling efficient exploration of user behavior metrics**.**

**NumPy** serves as the backbone for numerical computations in Python. With its high-performance array operations and a wide range of mathematical functions, it was used in this project to calculate statistical measures such as mean, median, variance, and standard deviation. These metrics played a crucial role in interpreting session durations, bounce rates, and conversion distributions. NumPy’s optimized computations allowed for fast processing of large numeric arrays and supported deeper statistical exploration beyond basic summaries.

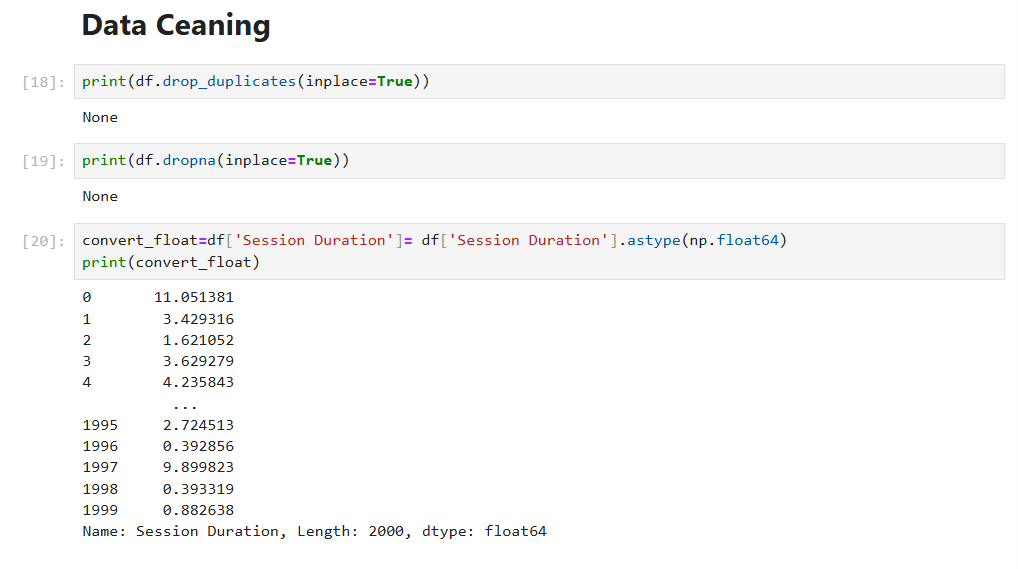
**Matplotlib** is the go-to library for creating static, animated, and interactive visualizations in Python. In this project, it was used to develop various chart types—such as histograms, scatter plots, line graphs, and pie charts—that visually narrated the story behind the user interaction data. Each plot highlighted specific aspects of user engagement, from traffic source distribution to correlations between page duration and conversion likelihood. By integrating visuals into the analysis, Matplotlib helped simplify complex relationships and enhance the communicative power of the report.

ization.

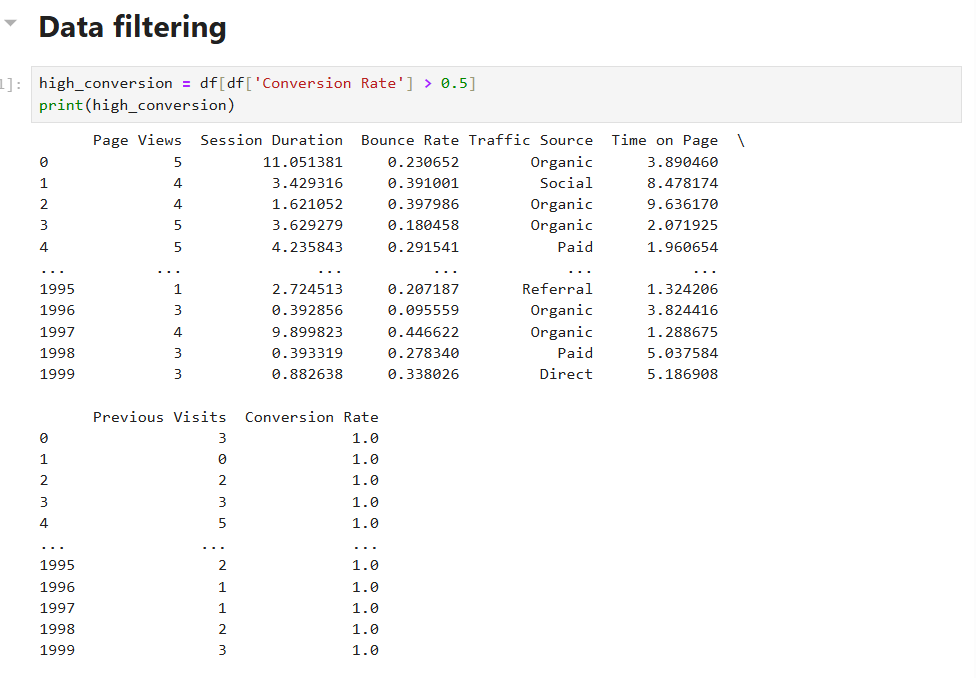
# Project Code



**Fig 1. Import libraries, Frameworks and Dataset**



**Fig 2.1 Data cleaning**



**Fig 3. Data filtering**

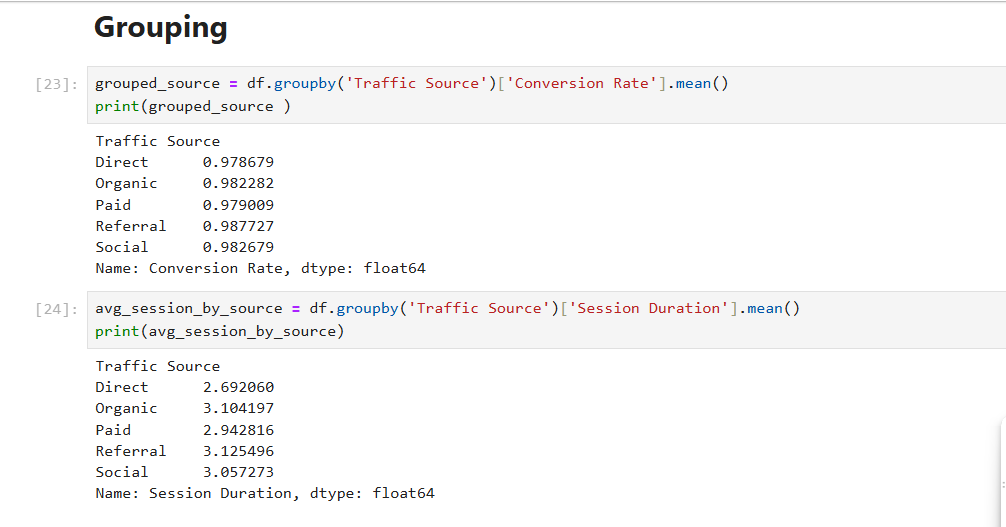
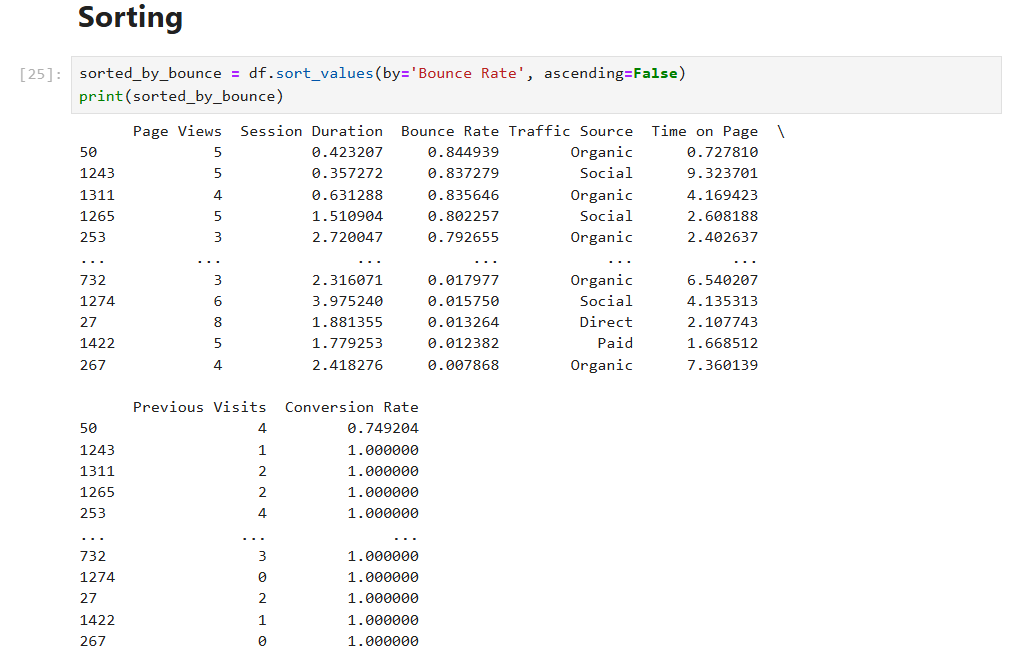
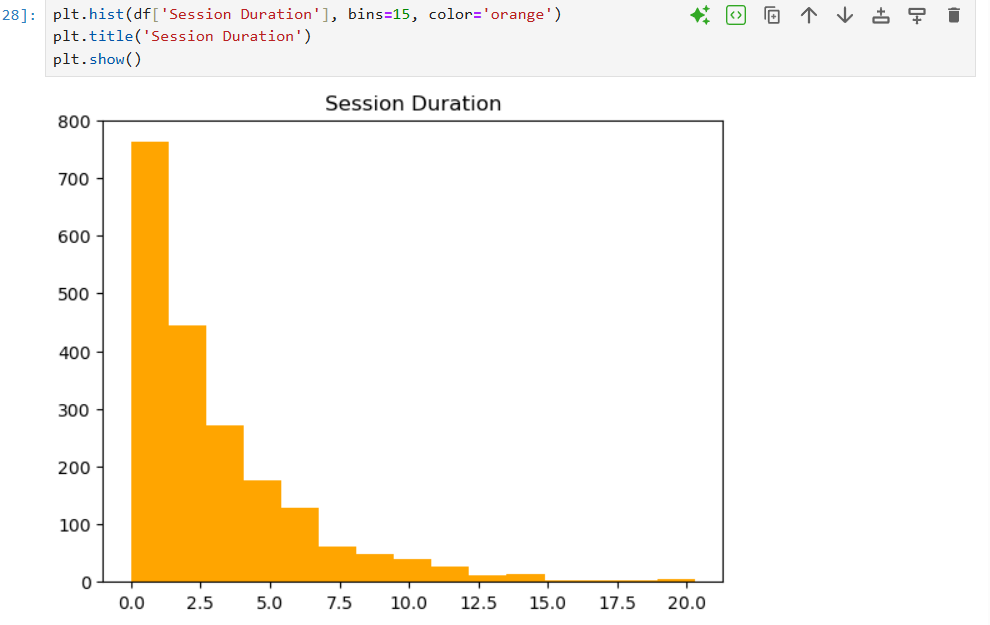
 **Fig 3. Data Grouping** **Fig 4.Data sorting**

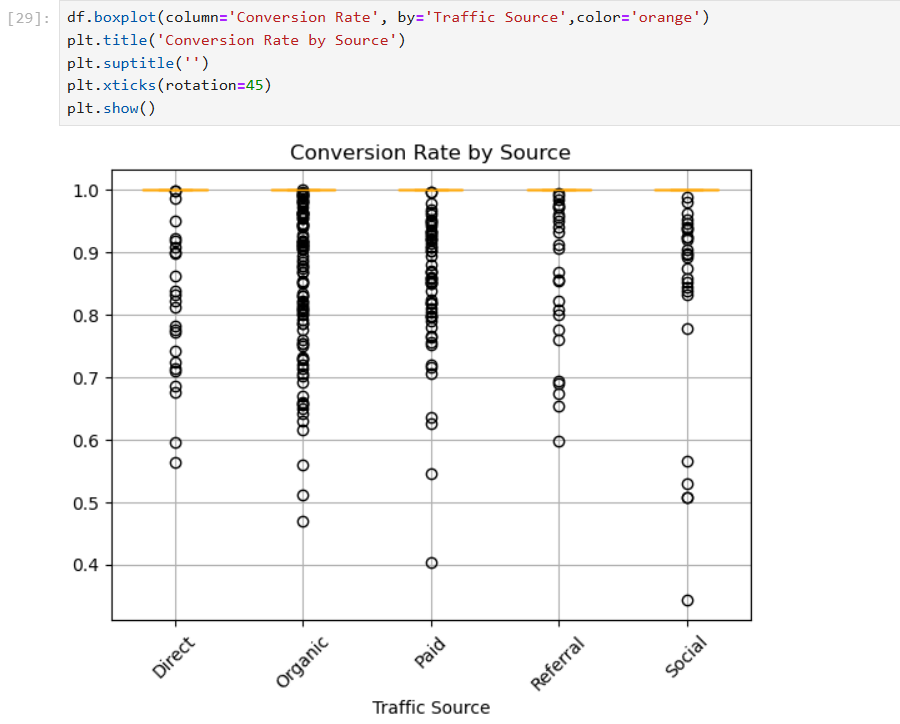


Fig 5. Data aggregation

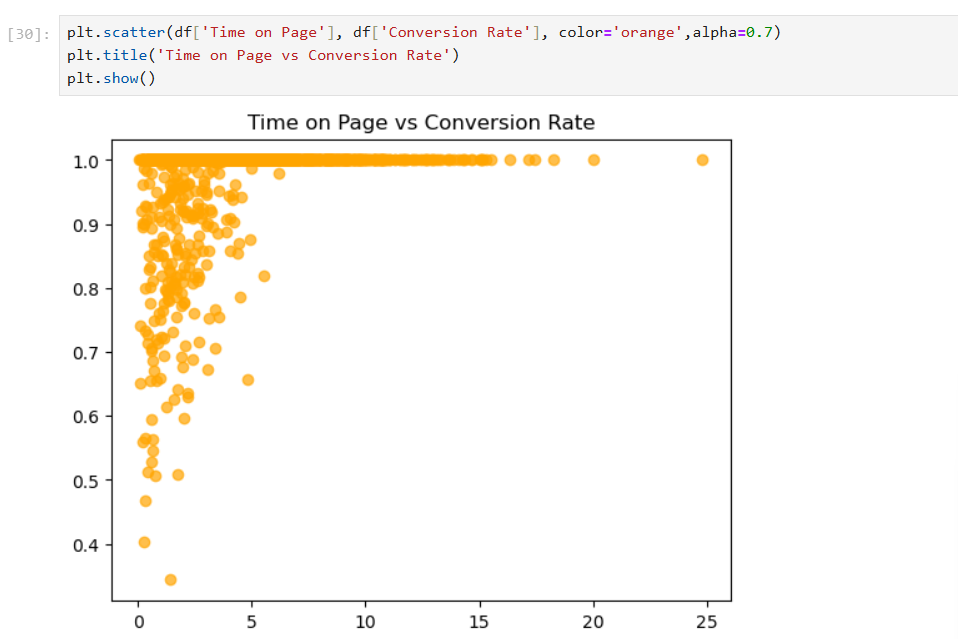
**DATA VISUALIZATION**



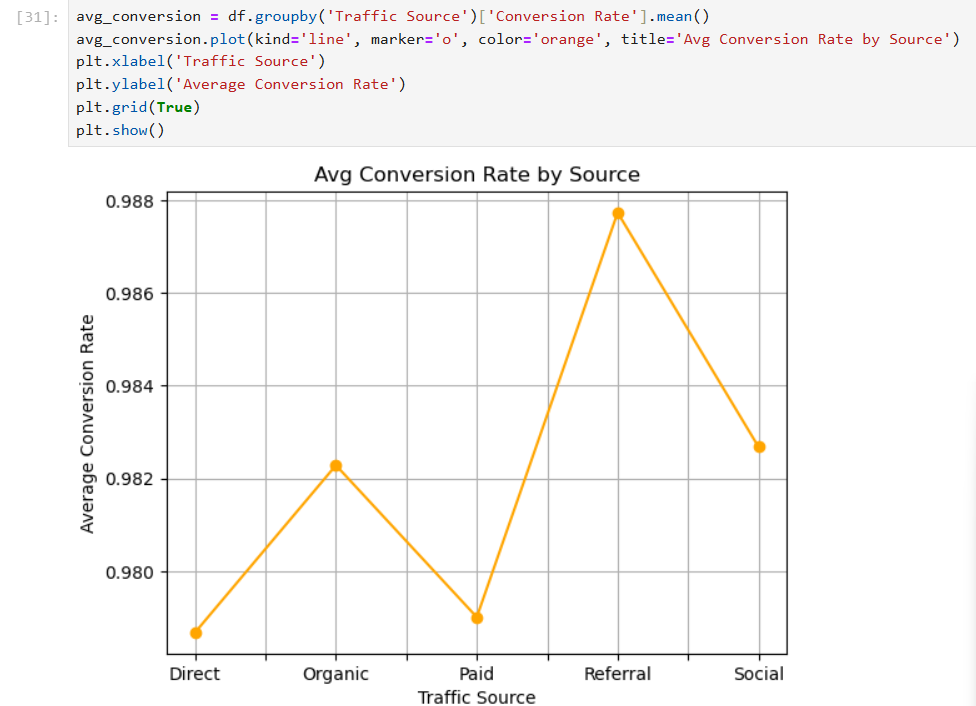
**Fig 5.1 histogram plot**



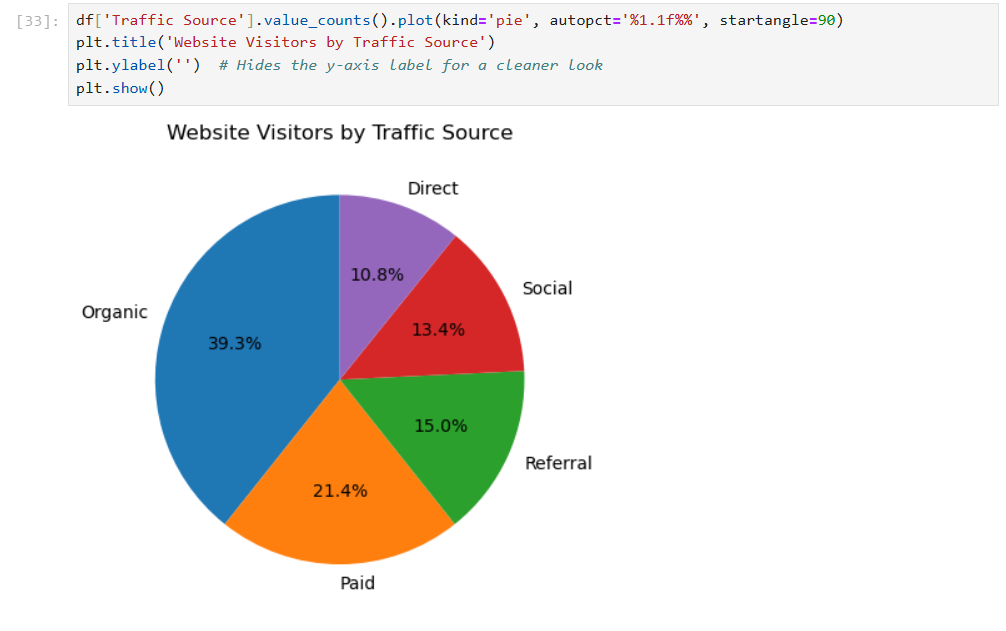
**Fig 4.2 boxplot**



**Fig 5.3 scatterplot**



**Fig 5.4 line plot**



**Fig 5.5 pie chart**

**Advantages of the Project**

* Provides actionable insights into user behavior and conversion trends
* Uses open-source tools, ensuring cost-effective and reproducible analysis
* Simplifies complex data through clear visualizations for easy interpretation
* Helps identify high-performing traffic sources to focus marketing efforts
* Reveals engagement gaps that can guide UX improvements
* Strengthens decision-making with evidence-based recommendations
* Enhances technical proficiency in Python, Pandas, NumPy, and Matplotlib
* Prepares groundwork for future dashboard development or machine learning integration
* Promotes clarity in communication through structured reporting and presentation
* Can be adapted for real-world use by digital teams across industries

**Conclusion**

This project effectively demonstrates the power of structured data analysis in uncovering meaningful insights about user engagement and conversion dynamics. By employing Python tools such as Pandas, NumPy, and Matplotlib, raw website traffic data was transformed into a coherent narrative that highlights key behavioral patterns, traffic source performance, and areas for optimization. Visualizations played a pivotal role in simplifying complex metrics, allowing decision-makers to grasp trends and act on them with confidence. Filtering and aggregation techniques helped segment data for focused exploration, revealing which channels consistently drive value and where attention must be directed to reduce bounce rates and enhance user experience. Overall, this analysis not only meets its technical objectives but also bridges the gap between raw data and strategic action—empowering digital teams to make smarter, evidence-based decisions.