A
Project Report
on

Big Data Analytics

WEB SERVER LOG ANALYSIS

Submitted by-

Student Name(s)

Viresh Chauhan	220422
Nishant Luhera	220313
Raghav Arora	220370

Under the guidance of

Dr. Yogesh Gupta
Professor

Department of Computer Science and Engineering
SCHOOL OF ENGINEERING AND TECHNOLOGY
BML MUNJAL UNIVERSITY GURGAON-122413, INDIA

Dec, 2024

Acknowledgement

We are profoundly appreciative to all those who assisted us in completing this work. We extend special thanks to Dr. Yogesh Gupta for offering expert guidance, insightful suggestions, and unwavering encouragement throughout this project. Their feedback and perspectives propelled us to uphold a high standard in our work. Our sincere gratitude goes to our families, whose patience, love, and steadfast support provided us the strength to overcome challenges. We also wish to recognize our friends, who inspired us with their words of encouragement and stood by us at every stage of this journey. We are extremely thankful to our peers for their collaborative efforts, stimulating discussions, and exchange of ideas. Their support was instrumental in enabling us to enhance our work and achieve improved results. This achievement is genuinely a reflection of the joint effort and generosity of everyone who contributed throughout the process.

Thanking You
Viresh Chauhan
Nishant Luhera
Raghav Arora

Table of Contents

List of Figures

List of Tables

1. Introduction

Web servers serve as the foundation of contemporary internet applications, processing large volumes of user requests and delivering data instantly. Nonetheless, overseeing and comprehending the vast amounts of data produced by these servers can present considerable challenges. Web server logs, which record essential details about incoming requests, response statuses, user agents, and traffic origins, frequently remain underused due to their complexity and lack of structure. This insufficient analysis may result in missed chances for improving server performance, recognizing trends in user behavior, or addressing potential security threats. The consequences of poor web server data management are evident in multiple situations. For example, unexplained traffic surges may go unnoticed, resulting in server slowdowns or even outages. Likewise, analyzing the distribution of HTTP methods, such as GET and POST, can assist in diagnosing unusual patterns, such as unauthorized attempts to reach secure endpoints. Response codes like 404 and 500 can offer insights into problems impacting user experience. Without actionable insights derived from this data, organizations may find it challenging to make informed decisions regarding resource distribution, expansion, or troubleshooting. Our solution seeks to tackle these issues by thoroughly analyzing web server data and delivering actionable insights in an accessible format. Utilizing an interactive Streamlit-based UI, we visualize key metrics such as the distribution of HTTP methods, response codes, data transfer patterns, and more. For instance, users can examine graphs that display the average number of requests per day or the distribution of country codes accessing the server. Furthermore, our analysis emphasizes browser and operating system preferences, facilitating a deeper understanding of client-side behavior. Through this project, we aim to provide users with a simplified approach to interpret web server data, detect patterns, and make decisions based on data. By merging comprehensive analysis with intuitive visualizations, we aspire to convert raw server logs into useful insights that can improve operational efficiency and user experience.

2. Problem Statement

Web servers produce extensive amounts of data in the form of logs, documenting every request made to the server. Nonetheless, this data is frequently unstructured and challenging to interpret directly, hindering the identification of significant patterns or trends. Without appropriate analysis, organizations may encounter difficulties in:

a) Comprehending user behavior and traffic patterns.
b) Diagnosing server performance problems or causes of downtime.
c) Identifying anomalies or security threats.
d) Enhancing server resources for improved efficiency.

For instance, increases in POST requests may suggest possible server misuse, while a rise in 404 response codes could indicate broken links or absent content. Likewise, without visualizing the distribution of browsers, devices, or countries accessing the server, administrators are unable to adjust services effectively to meet user needs.

2. 1. Problems our project solves

The project addresses the following specific problems:

i. Insufficient visibility into the allocation of HTTP methods such as GET, POST, and others.
ii. Challenges in comprehending the trends of response codes (e. g, 200, 404, 500).
iii. Inability to conveniently monitor the volume of data transferred over time.
iv. Restricted insights into request trends, including average requests per day.
v. Lack of analysis regarding geographical distribution through country codes.
vi. Absence of device, browser, and operating system-specific access trends.
vii. Nonexistence of referrer analysis to determine primary sources of traffic and their distribution.
viii. Unavailability of comprehensive URI type distribution.
ix. Limited understanding of session dynamics and behavior.

2.2. Explanation to how our project solves the problem(s)

This project offers an engaging and interactive approach for examining web server logs through a Streamlit-based user interface, converting unstructured log information into useful insights via visualizations and statistical metrics. The solution features several important elements that tackle typical issues in web server analysis. It illustrates the occurrence of HTTP methods such as GET and POST, enabling the identification of strange patterns or misuse. The distribution of response codes, including 404, 500, and 401, is

presented to aid in recognizing errors or unauthorized access. It monitors the amount of data transmitted over time, providing insights into server load and bandwidth usage. Graphs that display daily request trends emphasize traffic patterns and possible bottlenecks. Country code analysis charts the geographical source of requests, facilitating audience segmentation and identifying unexpected access from certain regions. Furthermore, the project delivers insights into client-side preferences by examining the distribution of devices, browsers, and operating systems. Referrer analysis identifies and depicts primary traffic sources, supporting marketing strategies and referral management. URI type analysis details popular endpoints and potential vulnerabilities, while session analysis explores user interactions and session lengths, offering a thorough overview of web server activity.

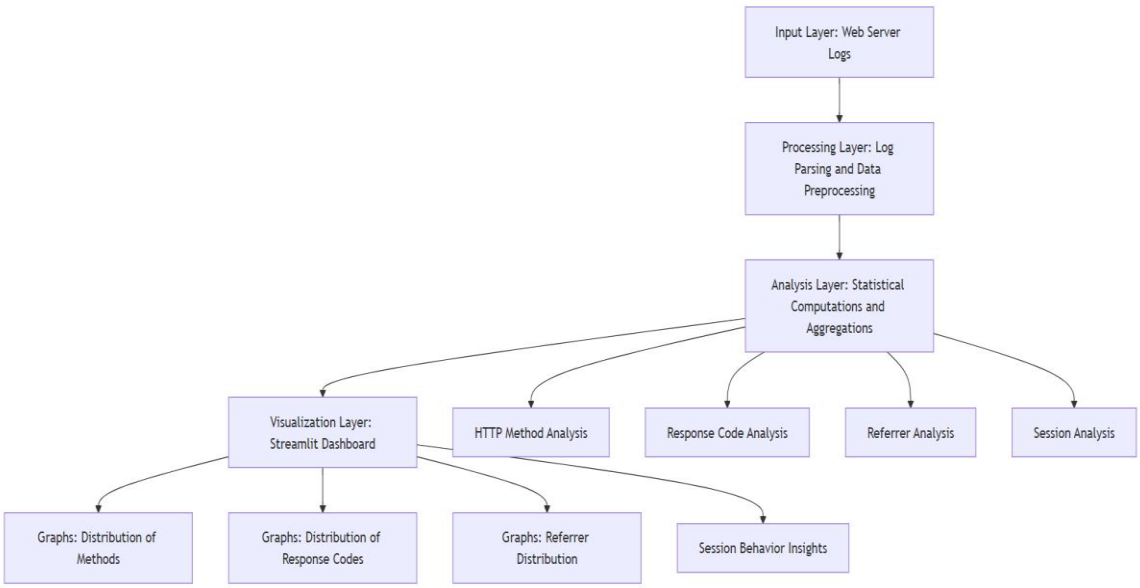

Fig.1. Conceptual Diagram of Web Server Analysis Solution

3. Literature Review

Web server log analysis has been a critical area of study, especially in comprehending user behavior, evaluating system performance, and improving web security. The growing amount of web log data necessitates the use of advanced technologies for data processing and analysis. Multiple studies have concentrated on overcoming challenges such as the rapid speed and large volume of log data, along with the need for effective preprocessing and scalable architectures. [1] conducted a thorough review of log file preprocessing techniques, specifically targeting NASA web logs. They illustrated that by employing data cleaning and preprocessing methods, they achieved a 72. 47% reduction in the size of raw logs while maintaining essential analytical value, facilitating the management of extensive datasets. This method was successful in converting unstructured log data into a more analyzable format. However, a shortcoming of this approach was the potential loss of some valuable data during the cleaning process. [2] introduced a distributed computing strategy utilizing Apache Flume, HDFS, and Pig to process substantial quantities of web log data in real time. This framework proved advantageous for addressing scalability challenges associated with conventional log processing techniques. The research demonstrated notable enhancements in scalability compared to traditional database methods. Nevertheless, the system necessitated a complex infrastructure, which could pose difficulties for smaller setups with limited resources. [3] employed an Information Bottleneck-based clustering technique with MapReduce to derive user profiles from web logs. They attained a remarkable classification accuracy of 98. 6%, underscoring the effectiveness of machine learning techniques in web log analysis. However, this method required substantial computing resources, making it less applicable in environments with restricted computational capabilities. [4] utilized data mining methods on web log data, concentrating on session identification and discovering usage patterns. Their research made significant contributions to the understanding of user behavior and system enhancements. The clustering and classification methods demonstrated in their study exhibited high accuracy rates exceeding 95%, particularly in differentiating between human and bot traffic. A limitation of their methodology, however, was its absence of real-time processing capabilities.

Table 1: Summary of existing research on Web Server Log Analysis

Study	Technology Stack	Key Features	Performance Metrics
Chodak et al.	Log preprocessing	Data cleaning, User identification	72.47% data reduction
Nagdive et al.	Apache Flume, HDFS, Pig	Real-time log ingestion, Distributed processing	Improved scalability
Suchacka & Iwański	Information Bottleneck, MapReduce	User behavior clustering	98.6% classification accuracy
Suneetha & Krishnamoorthi	Data mining	Session analysis, Pattern discovery	Enhanced user profiling

3.1. Existing solutions and Drawbacks

Web server log analysis solutions have progressed with improvements in distributed computing, data mining, and machine learning. These methods have shown varying degrees of effectiveness, but they also have certain limitations that this project seeks to resolve. For instance, [1] concentrated on minimizing log file size and enhancing data analysis through preprocessing; however, this technique may lead to the loss of important data during the cleaning process. Moreover, their method does not facilitate real-time analysis, restricting its usability in dynamic settings where log data is continually updated. [2] introduced a distributed method utilizing Apache Flume, HDFS, and Pig, which enhanced scalability. Nonetheless, the complexity of their system may hinder deployment in resource-limited environments, and the integration of multiple technologies can heighten the risk of system failure. [3] attained high precision in user behavior clustering through machine learning, yet the requirement for significant computational resources constrains the applicability of

the approach in smaller-scale environments. Additionally, these solutions generally do not feature user-friendly interfaces for displaying log analysis outcomes, complicating the interpretation of results for non-technical stakeholders. [4] demonstrated effectiveness in employing data mining for pattern identification and user profiling, achieving high accuracy in differentiating between human and bot traffic. However, their strategy was deficient in real-time data processing, which poses a notable limitation in contemporary web applications that require rapid responses to fluctuating traffic patterns.

Table 2: Existing Approaches and Drawbacks

S. No.	Existing State of Art	Drawbacks in Existing State of Art	How This Project Overcomes
1	Chodak et al.	Loss of valuable data during cleaning, No real-time processing	Real-time processing and visualization using Streamlit
2	Nagdive et al.	Complex infrastructure, Difficult for smaller setups	Simplified, scalable solution with minimal infrastructure requirements
3	Suchacka & Iwański	High computational requirements, Lack of user-friendly visualization	Efficient, scalable processing with a user-friendly interface for analysis results
4	Suneetha & Krishnamoorthi	No real-time processing, Limited scope for dynamic analysis	Real-time log analysis with immediate insights and visualizations

4. Methodology

4.1. Dataset: The dataset is a collection of web server logs from a Poland-based e-commerce website spanning six months (December 1, 2019, to May 31, 2020) and includes 35,157,691 HTTP request records organized into 10 columns. It is stored in CSV format, has a raw size of 9. 1 GB (10. 7 GB after processing), and covers 183 days of user interactions, intended for web traffic analysis, anomaly detection, and behavioral modeling.
Key Features:
IpId: Concealed IP address with a country identifier.
UserId: Differentiates regular users from administrators.
TimeStamp: UTC time noted in 100-nanosecond increments since 1 A. D.
HttpMethod: Type of HTTP request (e. g. , GET, POST).
Uri: Identifier of the target resource (partially redacted for privacy).
HttpVersion: Version of the HTTP protocol (e. g. , HTTP/1. 0, HTTP/1. 1).
ResponseCode: Status indicating server processing (e. g. , 200, 404).
Bytes: Size of the server's response in bytes.
Referrer: The webpage that directed the client (partially masked).
UserAgent: Details of the client software (browser, OS, version).
This dataset is suitable for analyzing user behavior, evaluating website performance, and enhancing e-commerce systems through machine learning or statistical techniques. It facilitates applications such as anomaly detection, session monitoring, and conversion rate assessment.

4.2. Pre-Processing: The raw server logs of an e-commerce platform were converted into an organized dataset for analysis. The TimeStamp column, which was in Windows FileTime format, was converted to an easily interpretable datetime format to facilitate time-related analytics. The IpId column was utilized to derive a CountryCode, offering geographic insights. The UserAgent column was dissected into Browser, OS, and Device_Type, classifying user devices and platforms. The Uri column was sorted into URI_Type, distinguishing page types such as products or categories, while the Referrer column was sorted into Referrer_Type, specifying traffic sources like search engines or

social media. The unnecessary UserId column was eliminated, and missing values were examined for quality assurance. The final refined dataset comprised enriched columns such as CountryCode, Browser, OS, Device_Type, URI_Type, and Referrer_Type, stored in CSV format, prepared for in-depth analysis of user behavior and website performance.

4.3. Working: In this report, we meticulously evaluated web traffic data utilizing Python, employing pandas for data handling and matplotlib and seaborn for graphical representation. The analysis commenced with data preprocessing, where we imported the event log dataset and executed sequential analyses. Initially, we provided a summary of traffic by scrutinizing HTTP methods and response codes, using both linear and logarithmic scales to accommodate the dataset's broad value spectrum. Subsequently, we performed a temporal analysis, changing timestamps to datetime format and consolidating requests by hour to uncover time-related trends, including peak activity times. For geographical analysis, we categorized countries with fewer than 100,000 requests under an "Others" label to streamline visualizations while maintaining clarity. Additionally, we thoroughly examined the usage of browsers, devices, and operating systems, followed by an analysis of URI and referrer trends. In our session analysis, we applied time-based clustering with defined thresholds: sessions were defined by a minimum duration of 2 minutes and a maximum of 120 minutes, with a new session initiated after 30 minutes of inactivity. This session data was scrutinized across variables such as time of day, URI types, referrer types, and geographic distribution, providing significant insights into user behavior and engagement metrics.

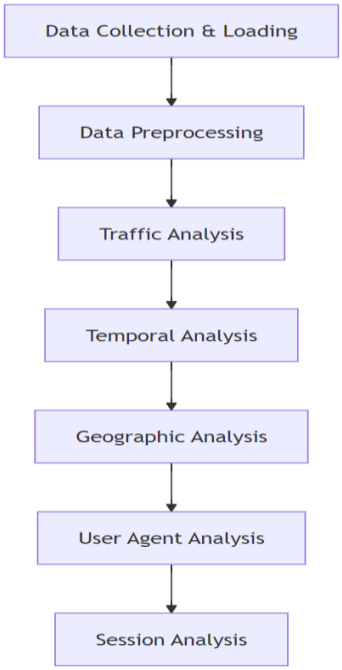

Fig.1. Workflow of the project

The below flowchart illustrates the structured approach to traffic analysis within web log data. Starting with an overview of HTTP methods, response codes, and data volume, the process dives deeper into analyzing request counts, visualizing distributions, and applying log scale transformations for better clarity of patterns. Each analysis segment—HTTP methods, response codes, and data volume—culminates in deriving actionable traffic pattern insights.

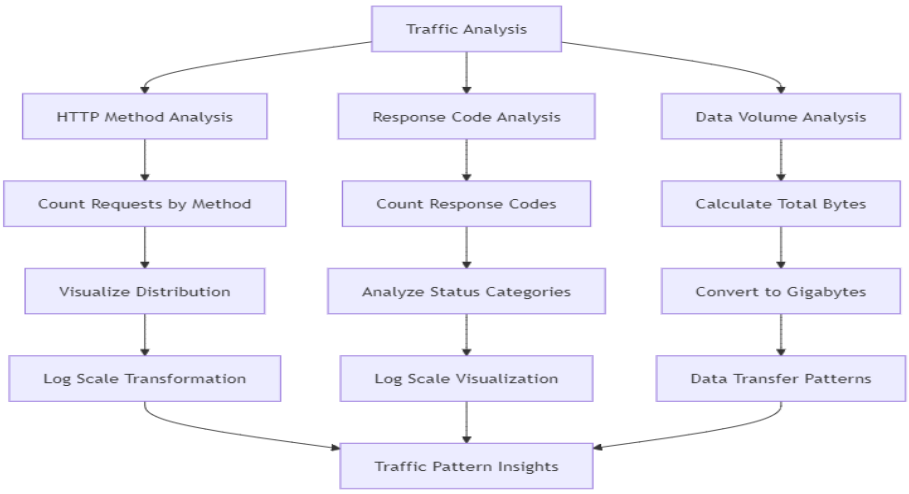

Fig.2. Detailed Workflow of Traffic Analysis in Web Log Data

Fig.3, outlines the step-by-step process for temporal analysis of web traffic data. The workflow begins with preprocessing time data by converting timestamps, extracting hour information, and sorting chronologically. It progresses to aggregating data based on time, including counting hourly requests and calculating averages. Peak analysis identifies high-traffic periods, calculates peak metrics, and examines traffic trends. These components converge to derive comprehensive time pattern insights for understanding temporal user behavior.

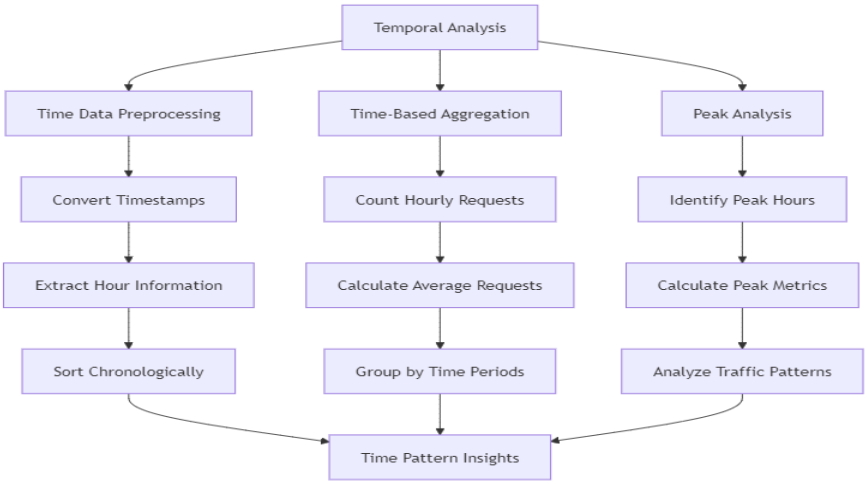

Fig.3. Detailed Workflow of Temporal Analysis in Web Traffic Data

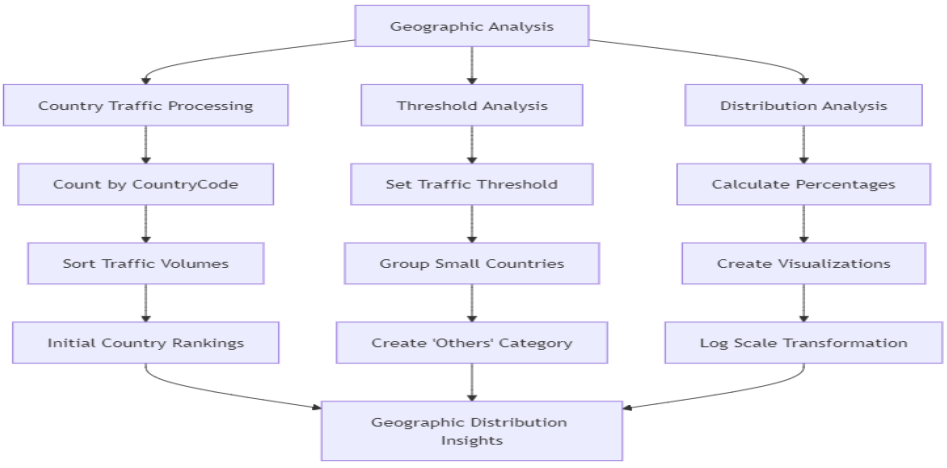

Fig.4. Workflow for Geographic Analysis of Web Traffic Data

This figure 4, represents the step-by-step process of geographic analysis for web traffic data. The workflow begins with processing country traffic by counting requests per country code, sorting traffic volumes, and ranking countries based on their request counts. Threshold analysis identifies smaller countries with low traffic and groups them into an "Others" category to maintain clarity. Distribution analysis calculates traffic percentages, creates visualizations, and applies log scale transformations to better understand traffic distribution patterns. These steps collectively provide insights into user geographic distribution and patterns.

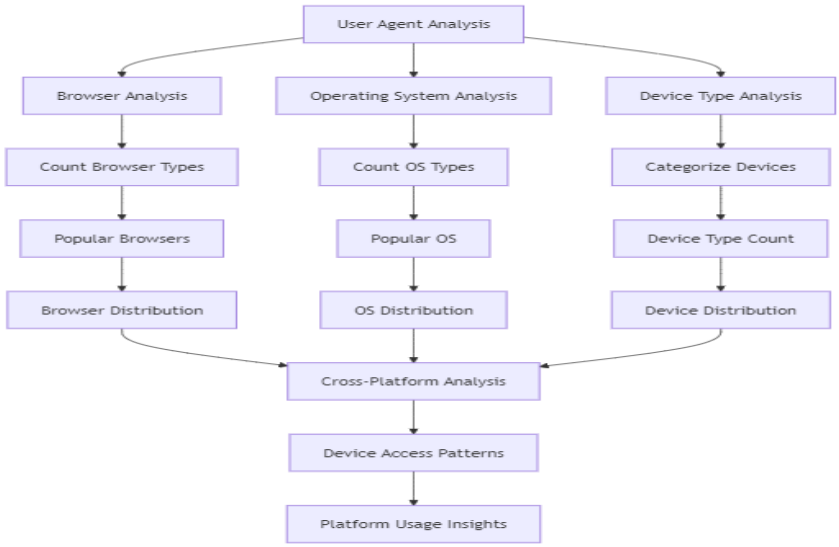

Fig.5. Workflow for Analyzing User Agent Data

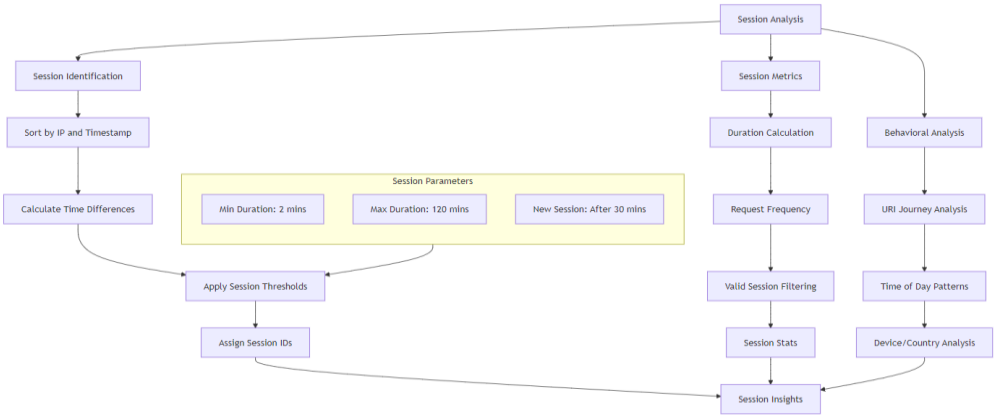

Fig.6. Session Analysis Workflow

Lastly, from Figures 5 and 6, we can conclude that the analysis commenced with User Agent Analysis, concentrating on grasping user behavior through browser types, operating system usage, and device type patterns. In Browser Analysis, various browser types were recognized, prominent browsers were emphasized, and their distribution was visualized. Likewise, Operating System Analysis classified operating systems utilized by users, identifying trends among different platforms. The Device Type Analysis classified user access by device types (mobile, desktop, tablet, etc.), with insights into distribution patterns. These analyses were integrated through Cross-Platform Analysis to explore shared trends across platforms and their impact on user access behavior. Ultimately, the insights were compiled into Device Access Patterns, showcasing platform usage trends and user preferences across various browsers, operating systems, and devices. Additionally, the Session Analysis workflow delivers insights into user engagement. It initiates with Session Identification, where user sessions are created by organizing requests based on IP addresses and timestamps, calculating time differences, and applying specified session thresholds to allocate unique session IDs. Significant session parameters were established to ensure precise session grouping, including a minimum session duration of 2 minutes, a maximum of 120 minutes, and a new session commencing after 30 minutes of inactivity. The Session Metrics component evaluates user engagement by analyzing session durations, request frequencies, and filtering valid sessions to investigate statistical patterns. Concurrently, Behavioral Analysis centers on URI journey patterns and traffic trends, such as time-of-day usage patterns and geographic insights across device types and regions. Insights from these components are merged into Session Insights, offering a detailed understanding of user engagement, access patterns, and behavioral trends across platforms and sessions.

5. Results and Discussion

This section offers a brief overview of the main findings from the analysis of user interactions with the e-commerce store's website. Insights were obtained from the variation of HTTP methods, response codes, user sessions, geographical trends, device and browser categories, session lengths, and referrer types.

5.1. Traffic Analysis

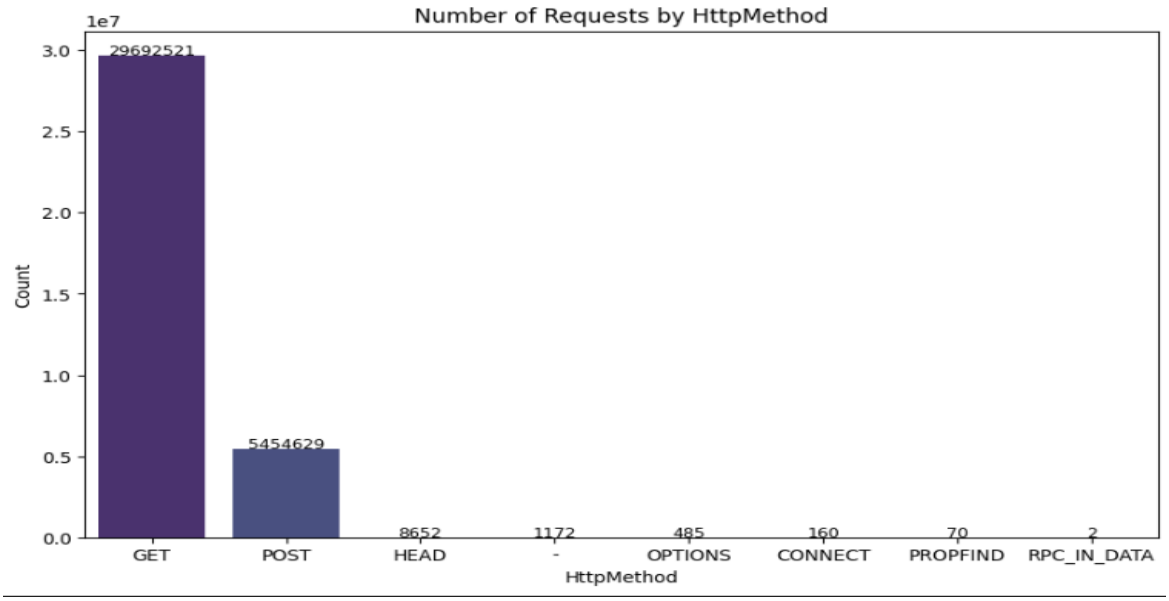

Fig.7 HTTP Method vs Number of Requests

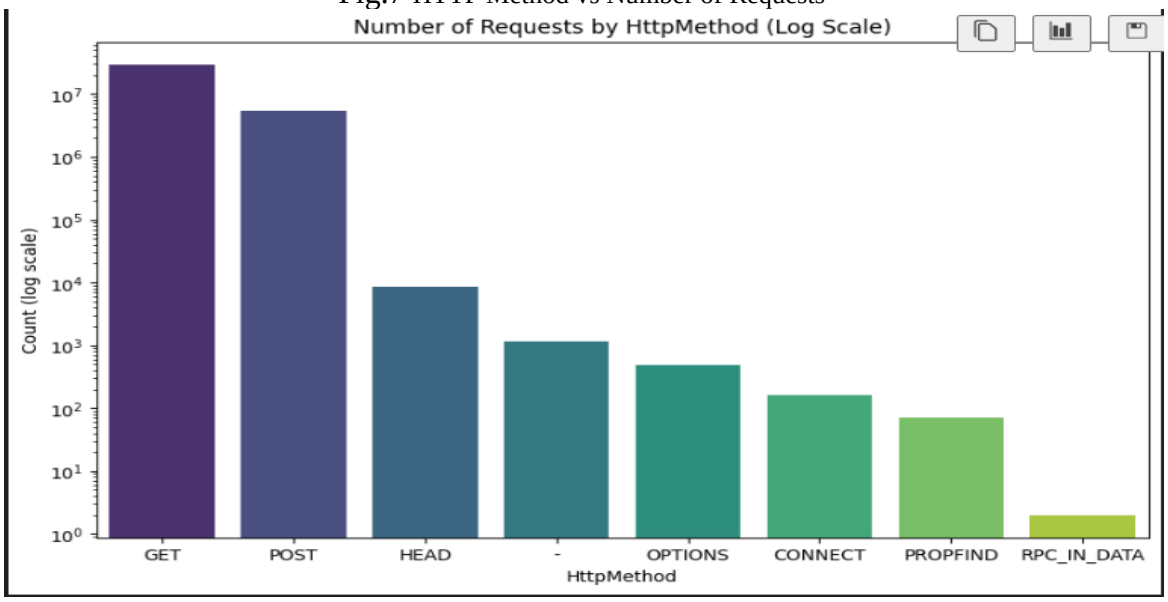

Fig.8. Http Method vs log of Number or requests

Inference of Figure 7 and 8: The graphs offer insights into the allocation of HTTP methods in the e-commerce store dataset. The non-logarithmic graph illustrates the prevalence of GET and POST methods, vital for retrieving data and submitting forms, respectively, while other methods are less frequent but may aid particular functionalities. The logarithmic graph emphasizes the significant frequency disparity between these primary methods and the others, revealing an exponential drop as we progress down the y-axis. These insights can guide optimizations concerning resource loading, form submission efficiency, and possible security threats associated with less common methods.

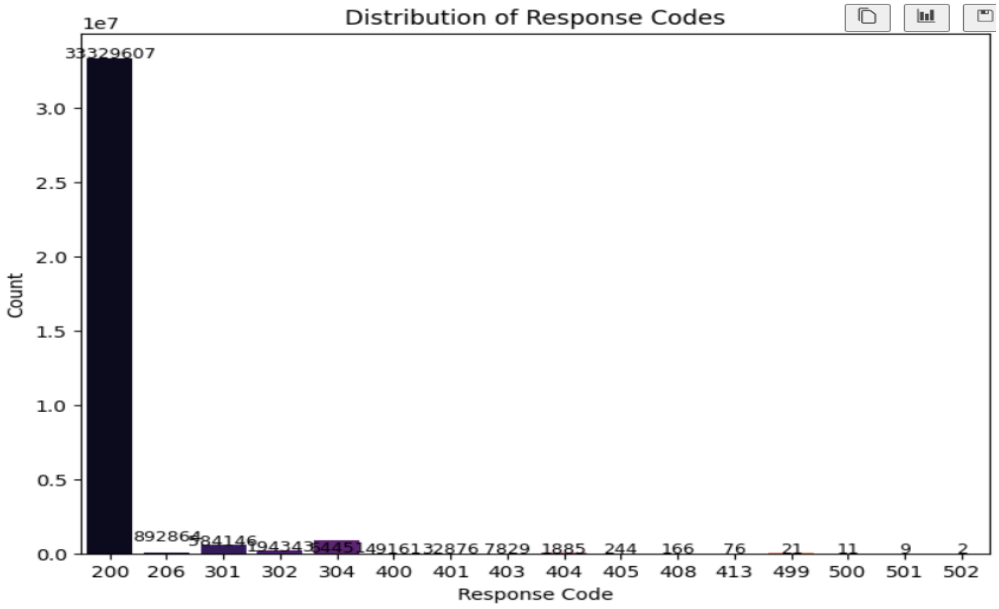

Fig.9. Response Code vs Number of Responses

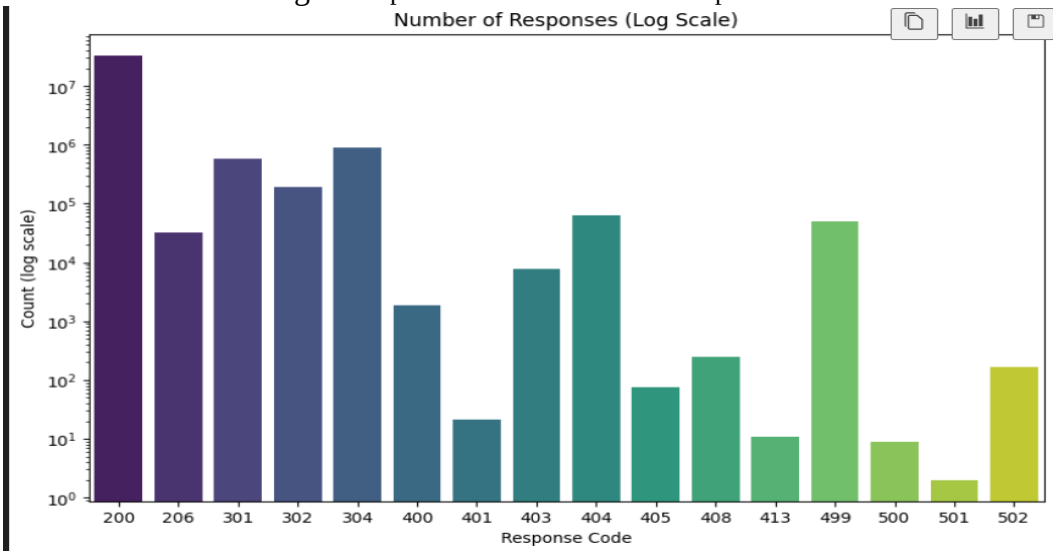

Fig.10. Response code vs Number of Responses (Log Scale)

Inferences of Figure 9 and 10: The graphs illustrate the distribution of HTTP response codes within the e-commerce store. The predominant response is "200 OK," signifying successful requests and indicating the overall performance of the site. Other common codes like "206 Partial Content," "301 Moved Permanently," and "302 Found" are associated with e-commerce activities such as pagination, redirects, and content delivery. Error codes such as "404 Not Found," "403 Forbidden," and "500 Internal Server Error" reveal possible issues with user experience and performance, including broken links, unauthorized access, and server failures. Comprehending these response codes allows for performance enhancement, improved user experience, and increased security. Furthermore, the total data transferred was 519. 92 GB, providing insights into patterns of resource usage.

5.2. Temporal Analysis

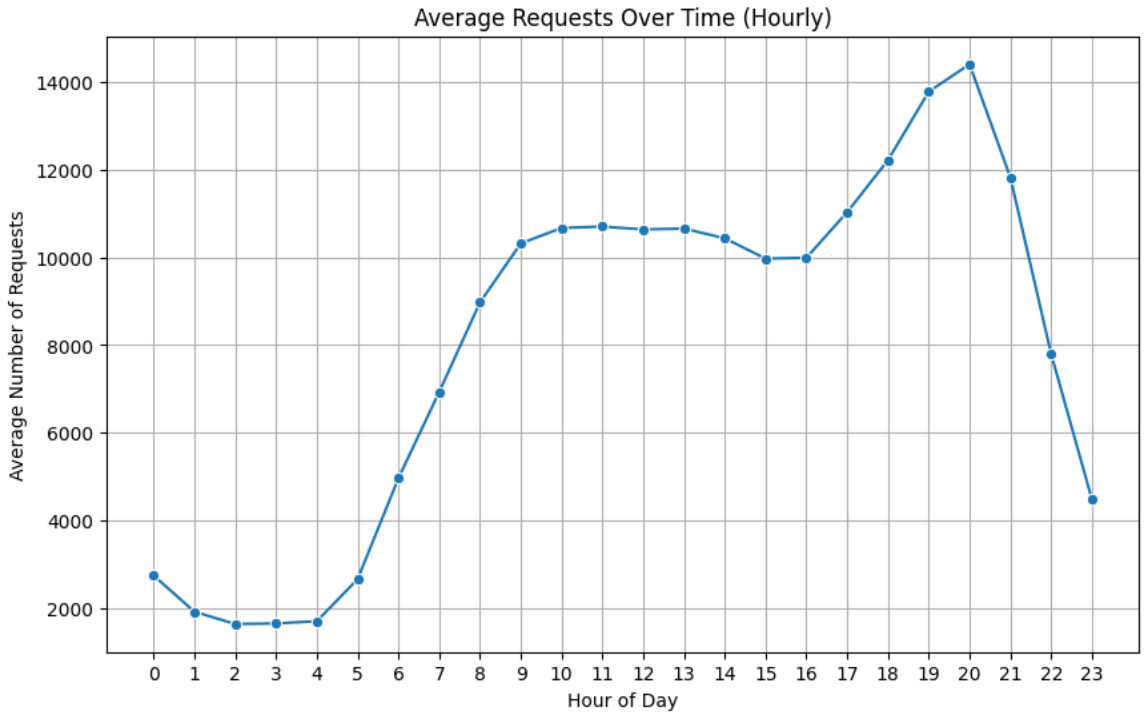

Fig.11. Average Requests for each hour of the day throughout the 183 Days

Inference: The line graph offers a straightforward representation of the average number of requests per hour, illustrating the recurring pattern of user engagement on the e-commerce website. A notable peak occurs between 18:00 and 20:00, indicating that this period is the most active time of day for the website. This data can be crucial for enhancing resource

distribution, planning marketing initiatives, and managing customer support staffing. By recognizing these peak hours, the e-commerce store can effectively allocate resources to manage heightened traffic, ensuring a smooth user experience. Moreover, focused marketing campaigns can be executed during these peak hours to take advantage of user interaction. By aligning staffing schedules with peak traffic periods, the store can deliver prompt customer support and resolve any potential issues. Additionally, examining the off-peak hours can reveal areas for enhancement. By understanding the factors contributing to reduced activity during these times, the store can develop strategies to draw more visitors and boost sales, such as providing exclusive offers or tailored recommendations.

Peak Activity Hour: 20:00 with an average of 14402 requests.

5.3. Geographical Analysis

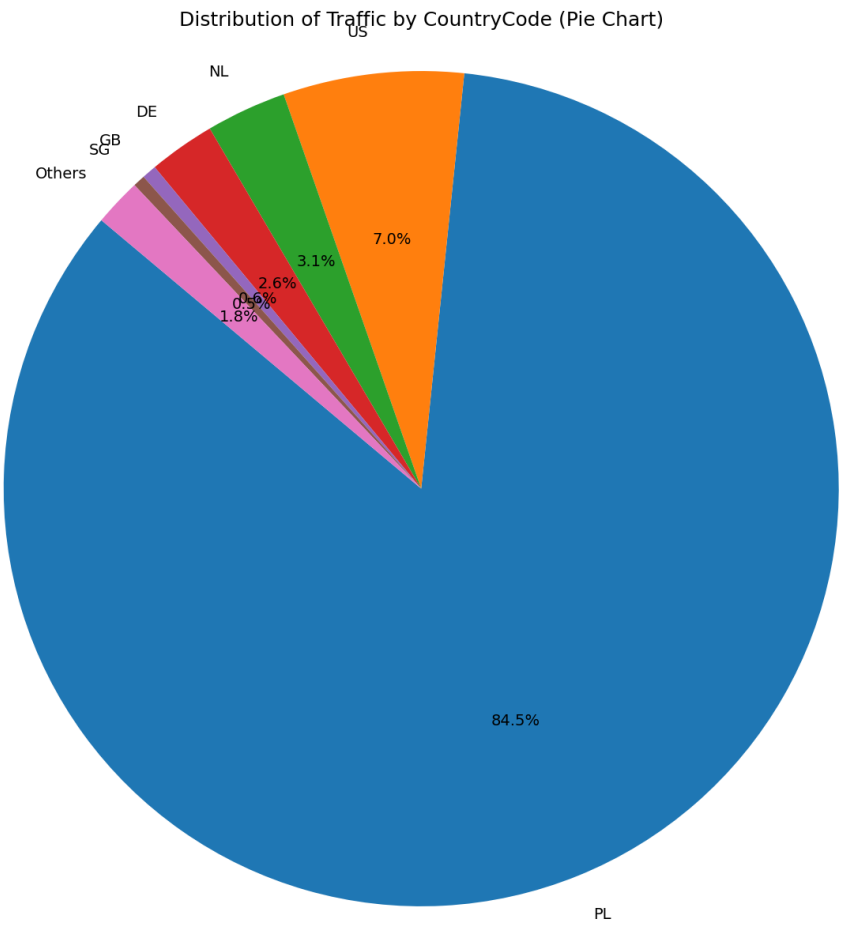

Fig.12. Pie Chart visualizing request share of each country (Other->Number of requests<100,000)

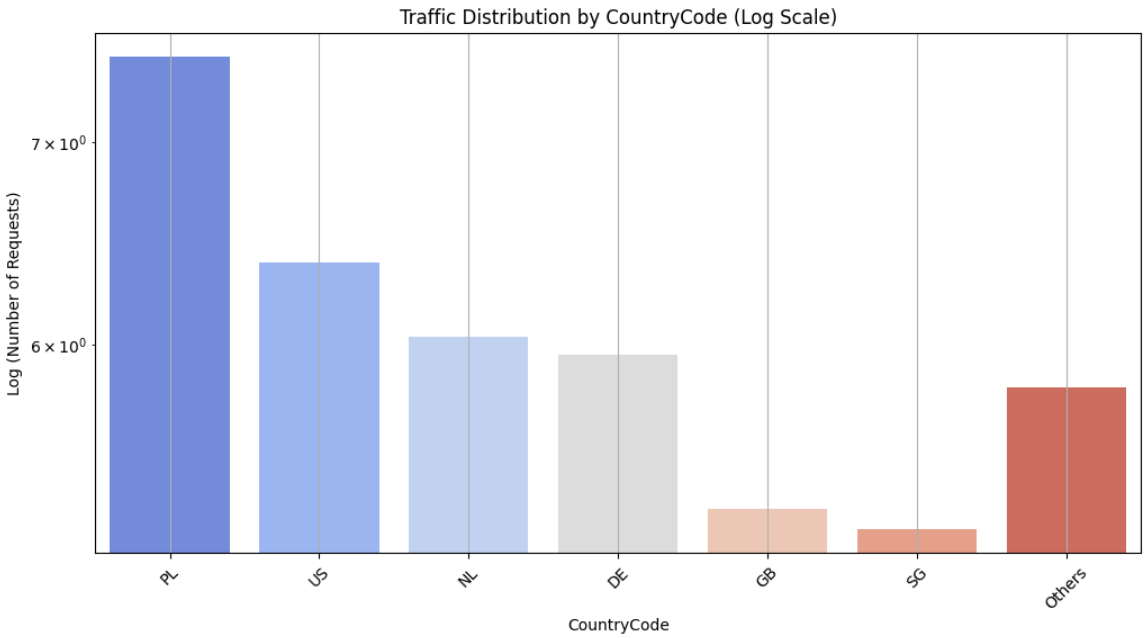

Fig.13. Bar Graph based on Logarithmic Scale showing exponential differences in various countries

Inference of figure 12 and 13: The presented data and visualizations provide a clear overview of the e-commerce store's international traffic distribution. Poland (PL) is the leading market, comprising a large share of the total website visits. The United States (US) ranks as the second-largest contributor, reflecting a considerable international presence for the store. Other significant markets include the Netherlands (NL), Germany (DE), and the United Kingdom (GB). Although Singapore (SG) and the "Others" category hold a smaller percentage, they still add to the website's overall reach. By analyzing this distribution, the e-commerce store can focus marketing initiatives and resource allocation on key markets such as Poland and the US. Furthermore, identifying opportunities in other European nations like the Netherlands, Germany, and the UK can further enhance the store's international presence. While smaller markets like Singapore may need a more focused approach, they still signify potential growth opportunities. The log-scaled bar graph presents a more detailed view of the traffic distribution, emphasizing the considerable gap between the highest-performing countries and the others. Poland (PL) distinctly leads the traffic, with its bar significantly taller than the rest. The US, while substantial, shows to be notably less dominant on the log scale. The log scale effectively condenses the range of values, facilitating the visualization of differences between countries with lower traffic volumes. This indicates that while Poland and the US are the main players, countries like the Netherlands (NL), Germany (DE), and the UK (GB) still make meaningful

contributions to the total traffic, albeit to a lesser degree. By employing a log scale, we gain a better understanding of the relative significance of each country and can pinpoint potential growth opportunities in markets that may not be easily noticeable on a linear scale. This data-informed insight can guide strategic decisions regarding marketing, localization, and resource distribution.

5.4. User Agent Analysis

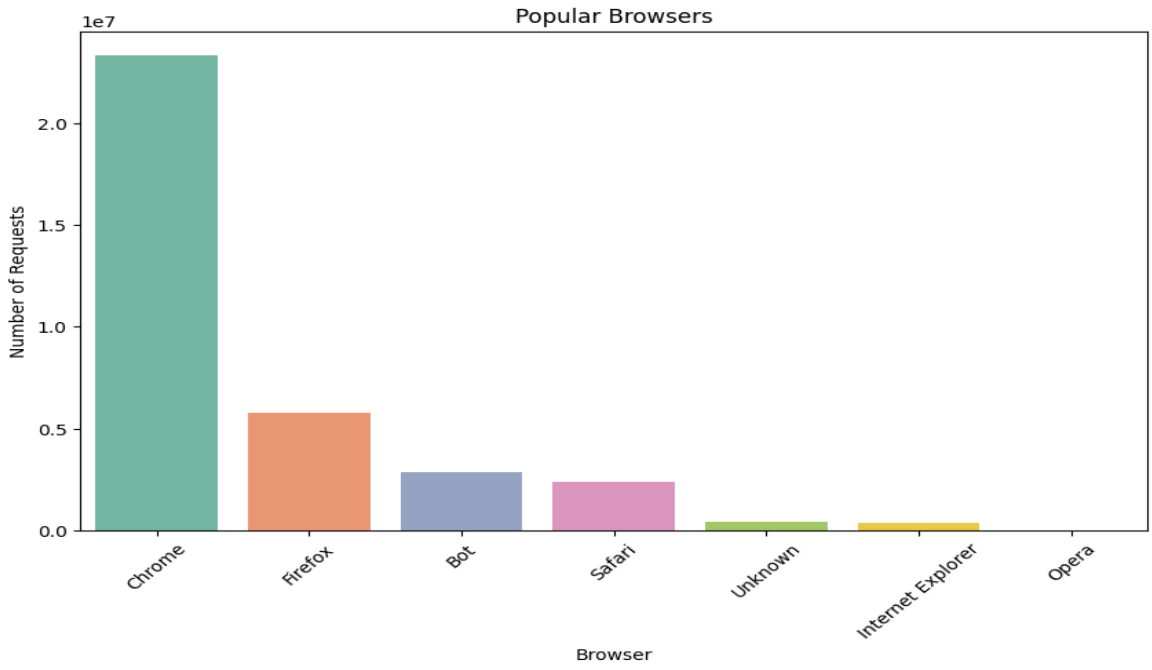

Fig.14. Bar graph of No. of requests vs browser

Inference: The provided bar chart presents important information regarding the distribution of web browsers utilized to access the e-commerce store's website. Chrome stands out as the leading browser, representing a considerable share of the website traffic. This suggests that a significant segment of the user base depends on Chrome for their online shopping experience. Firefox secures the second spot, indicating a noteworthy user base that prefers this browser. Safari's inclusion suggests a segment of users accessing the website via Apple devices (iPhones, iPads, or Macs). The "Bot" category, while important, underscores the presence of automated scripts interacting with the website. These may be search engine crawlers indexing the site or potentially harmful bots. The "Unknown"

category likely signifies users whose browser information could not be identified, possibly due to privacy settings or technical constraints. Internet Explorer and Opera, with their notably smaller user bases, imply that the store's audience is less inclined to utilize these older browsers. By analyzing these browser usage trends, the e-commerce store can make knowledgeable decisions regarding website development, testing, and optimization strategies. Focusing on ensuring compatibility and performance for Chrome, Firefox, and Safari is essential. Moreover, examining bot traffic can assist in identifying potential security threats and enhancing website performance for search engine indexing.

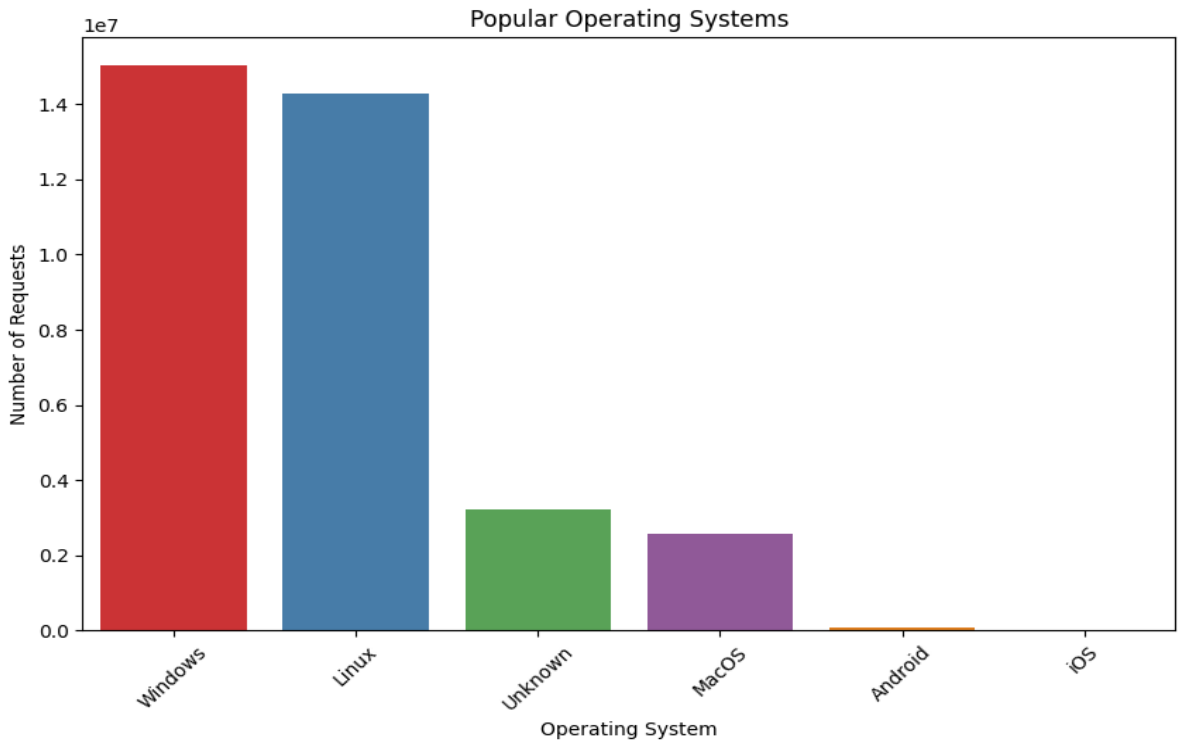

Fig. 15. Bar graph of requests vs operating systems

Inference: The supplied data provides information on the operating systems utilized to access the e-commerce store. Windows and Linux are predominant, representing roughly 92% of the traffic. Windows, comprising 53% of the traffic, is the most widely used, indicating that a substantial segment of the user base visits the website from Windows-based PCs and laptops. Linux, at 40%, reflects a significant tech-savvy user group, likely including developers, IT specialists, or individuals who favor open-source systems. macOS, accounting for 7%, signifies users visiting the website from Apple devices. This

indicates a potentially wealthy demographic with a taste for Apple products. Android and iOS, with collectively less than 1%, signify a comparatively smaller mobile user base. This may stem from various reasons, such as the nature of the products or services provided by the store or the targeted demographic. Recognizing this distribution enables the e-commerce store to enhance its website for Windows and Linux while also taking mobile optimization into account for future expansion. Additionally, the considerable presence of tech-savvy users on Linux presents opportunities for targeting specific niches or offering specialized products.

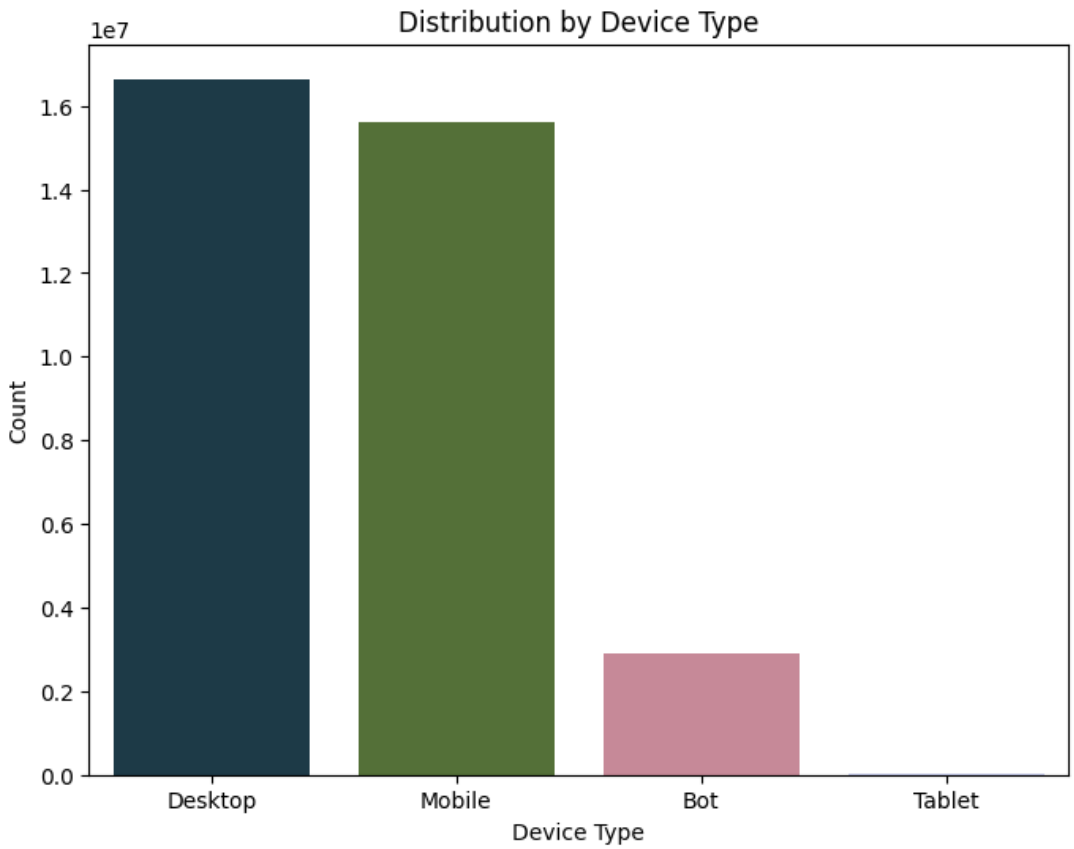

Fig. 16. Bar chart for the distribution of device types

The provided data indicates a notable inclination towards desktop devices among the e-commerce store's user base. Desktops represent 47% of the total traffic, signifying that a considerable segment of users favors traditional browsing experiences. Mobile devices, although slightly less favored, still constitute a notable 44% of the traffic, highlighting the necessity for mobile optimization. Tablets, conversely, account for only 0. 3% of the total

traffic, implying a restricted user base for this device category. Bots, comprising 8%, likely consist of search engine crawlers and other automated scripts engaging with the website. Understanding this distribution of device types enables the e-commerce store to prioritize optimization for desktop and mobile platforms while also acknowledging the potential of tablet users. Moreover, managing and monitoring bot traffic is essential for maintaining website security and performance.

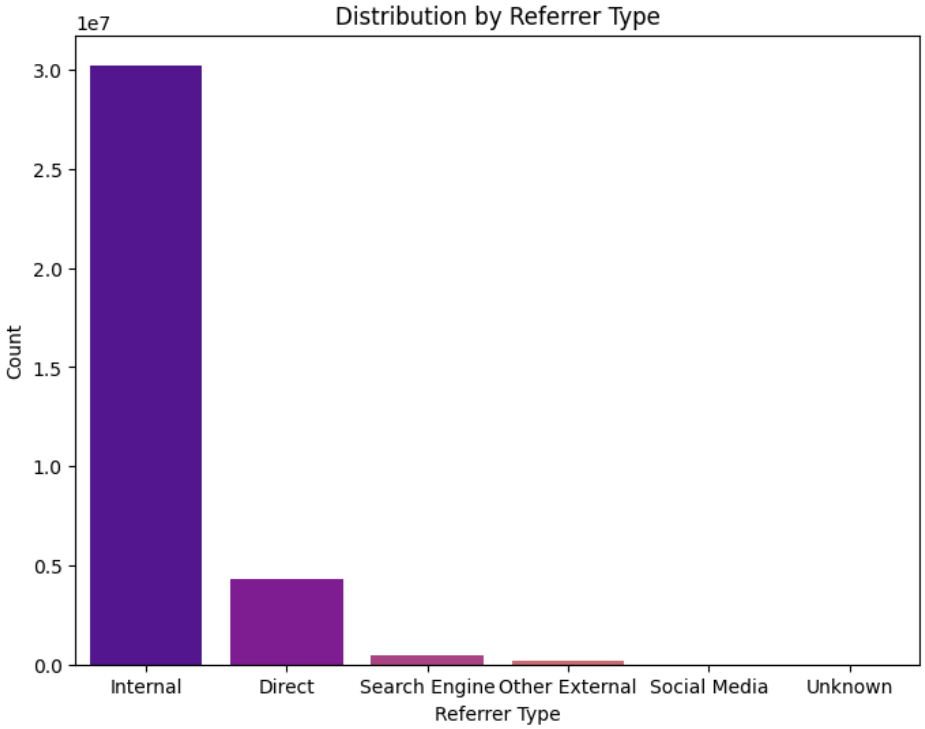

Fig.17. Distribution of referrer types

Internal Traffic: The majority of traffic, around 85%, is generated from internal sources. This demonstrates a significant level of user engagement within the website, with users browsing through various pages and sections. Direct Traffic: About 12% of the traffic arrives directly at the website, skipping search engines or referrals. This may be due to users entering the website's URL directly in their browser's address bar, utilizing bookmarks, or accessing the website via mobile app shortcuts. Search Engine Traffic: Search engines account for approximately 1. 3% of the traffic. This indicates that search engine optimization (SEO) efforts are relatively less effective in generating traffic to the website compared to direct traffic and internal navigation. Other Traffic Sources: Other

external sources, social media, and unidentified sources make up a minimal segment of the total traffic. Understanding these traffic sources can assist the e-commerce store in refining its marketing strategies and website layout. Focusing on internal navigation and user experience can aid in retaining current users and promoting greater engagement. While SEO initiatives can be further enhanced to draw more organic traffic, direct traffic and internal navigation continue to be the main contributors to website traffic.

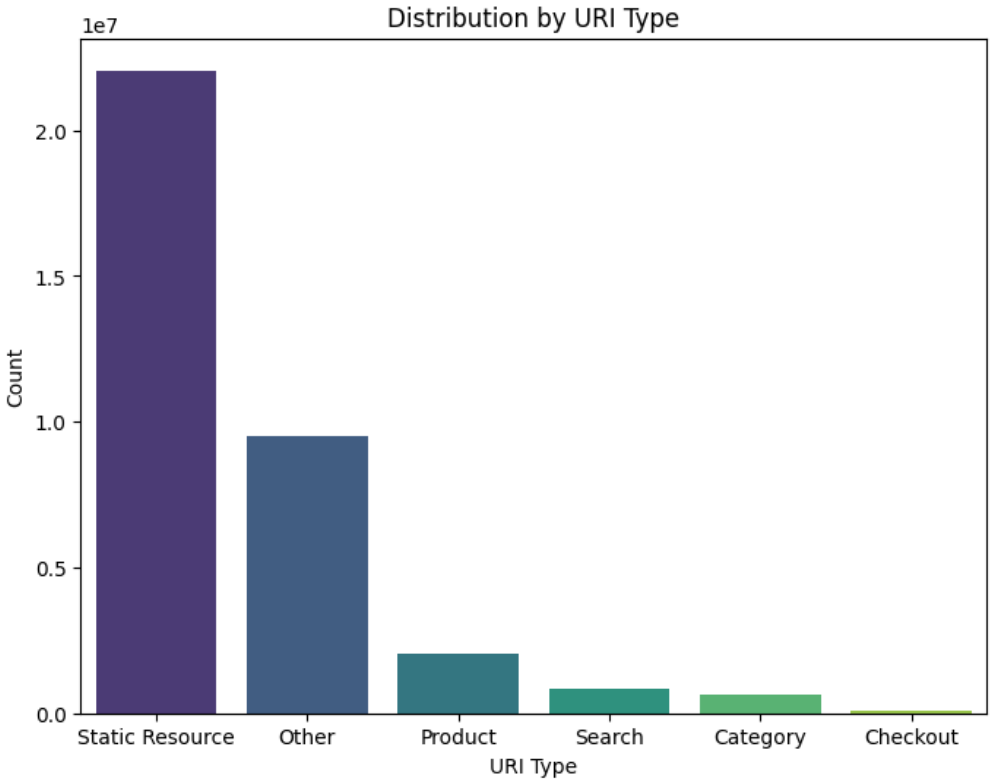

Fig.18. Distribution by URI types

The provided data shows a distinct preference for static resources, which make up a considerable 78% of the overall requests. This suggests that a large part of the website traffic consists of loading images, CSS, JavaScript, and other static files. The "Other" category, which represents 34%, likely includes a variety of resources, such as error pages, API calls, or miscellaneous content. This category requires further analysis to determine specific resource types and their influence on website performance. Product-related pages, making up 7%, are a notable source of user engagement, indicating that a significant portion of the traffic is dedicated to product discovery and exploration. Search and category pages,

accounting for 3% and 2% respectively, underscore the significance of efficient search functionality and product categorization for user navigation. The "Checkout" category, with only 0. 03%, reflects that a relatively small amount of the traffic pertains to the checkout process. This may result from various factors, such as cart abandonment or an efficient checkout process. Understanding this distribution of URI types allows the e-commerce store to enhance website performance, improve user experience, and pinpoint potential areas for conversion enhancement. By focusing on the optimization of static resources, enhancing product page experiences, and refining the search and checkout procedures, the store can markedly boost its overall performance and user engagement.

5.5. Session Analysis

Table 3. URI Type Analysis

URI Type	**Unique Sessions**	**Total Visits**	**Avg Visits Per Session**
Category	1,099	177,457	161.47
Checkout	114	26,792	235.02
Other	1,956	2,384,520	1,219.08
Product	1,581	231,969	146.72
Search	1,107	101,969	92.11
Static Resource	1,885	7,531,886	3,995.70

Table 4. Referrer Type Analysis

Referrer Type	Unique Sessions	Total Visits	Avg Visits Per Session
Direct	2,220	380,373	171.34

Internal	749	9,946,887	13,280.22
Other External	532	42,800	80.45
Search Engine	173	82,354	476.03
Social media	15	2,110	140.67
Unknown	3	69	23.00

Table 5. Top Countries by Average Session Duration

Country Code	Mean	Count
PL	50,754.07	30
SG	4,283.02	769
IE	2,655.01	28
US	1,472.33	587
NL	610.23	64
FR	139.92	742

Table 6. Most Common Session Journeys

URI Type	Count
(Static Resource,)	87
(Static Resource, Other)	38
(Static Resource, Static Resource)	35
(Other,)	32
(Static Resource, Static Resource, Other)	25

Inference: The examination of URI types and user behaviors offers important insights into the performance and engagement patterns of the e-commerce website. Static resources are predominant in user interactions, with the highest average visits per session (3,995. 7) associated with these resources, indicating substantial utilization of cached elements such

as images, stylesheets, and scripts. This implies that the website is resource-heavy, likely containing rich media content. The checkout process displays an intriguing pattern: although there are relatively few unique sessions (114), the average number of visits per session is notably high at 235. 02. This suggests multiple user interactions during the checkout phase, either due to a complicated multi-step process or thorough review behavior, highlighting a potential area for improvement. Referrer analysis indicates strong engagement from internal referrals, with an exceptionally high average of 13,280. 22 visits per session, suggesting extensive internal navigation or single-page application (SPA) routing. Direct traffic leads in unique sessions with 2,220, demonstrating robust brand recognition, while search engine traffic, despite being lower in volume (173 sessions), exhibits substantial engagement with an average of 476. 03 visits per session. Geographically, users from Poland (PL) display very high average session durations of 50,754. 07 minutes, although this is likely an anomaly due to bot activity or session timing issues. Conversely, users from Singapore (SG) (4,283. 02 minutes across 769 sessions) and France (FR) (139. 92 minutes across 742 sessions) exhibit more consistent patterns, reflecting varying engagement across different regions. Device and browser usage patterns also provide insights into user engagement. Desktop users in regions such as the US and France have longer average session durations (25. 9 and 26. 0 minutes, respectively), indicating deeper exploration. In contrast, mobile user sessions are significantly shorter (under 8. 4 minutes), suggesting potential improvements to the mobile user experience. Furthermore, session journeys reveal that many users primarily engage with static resources, with 87 sessions exclusively involving requests for static resources. Patterns such as "Static Resource > Other" (38 sessions) suggest user flows from initial page loading to interactive features, indicating effective caching strategies and opportunities to retain users after the initial resource loading. These observations highlight a complex e-commerce website with a heavy reliance on static resources, diverse traffic sources, and regional variations in user engagement. The results indicate a need to focus on optimizing mobile interactions, streamlining the checkout process, and addressing obstacles to user engagement beyond visits to static resources to enhance the overall user experience.

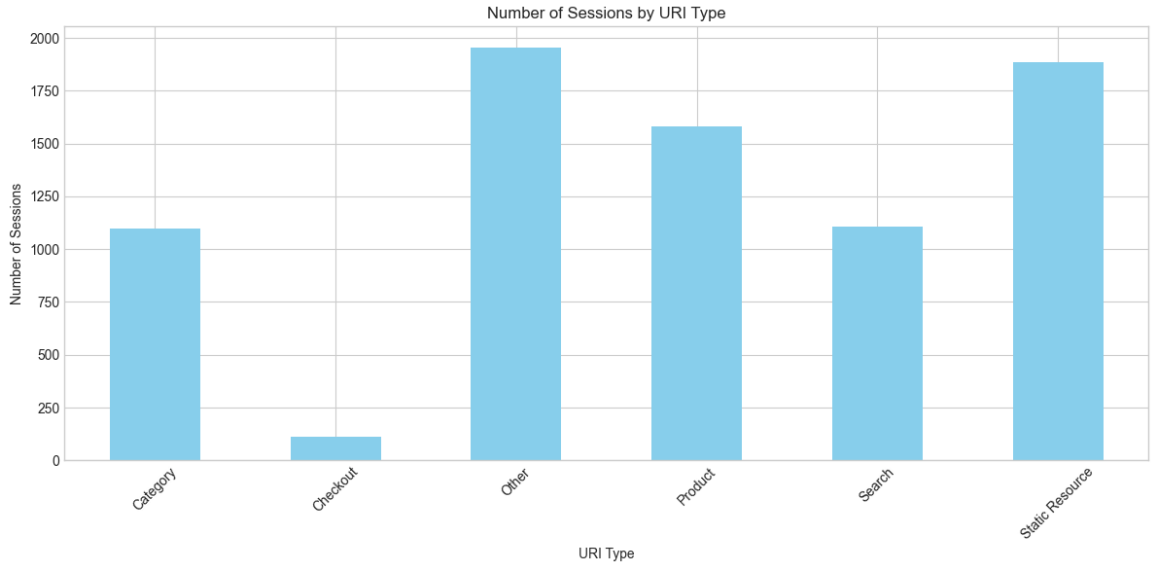

Fig.19. Number of sessions by URI type

From figure 19, we can see that "Other" and "Static Resource" URIs lead the traffic with nearly 1,900 sessions each. Product pages garner around 1,500 sessions, whereas Category and Search pages each draw roughly 1,100 sessions. Checkout pages exhibit the least amount of traffic with about 120 sessions. This distribution indicates that users invest significant time exploring different resources and general pages prior to selecting specific products or finalizing purchases.

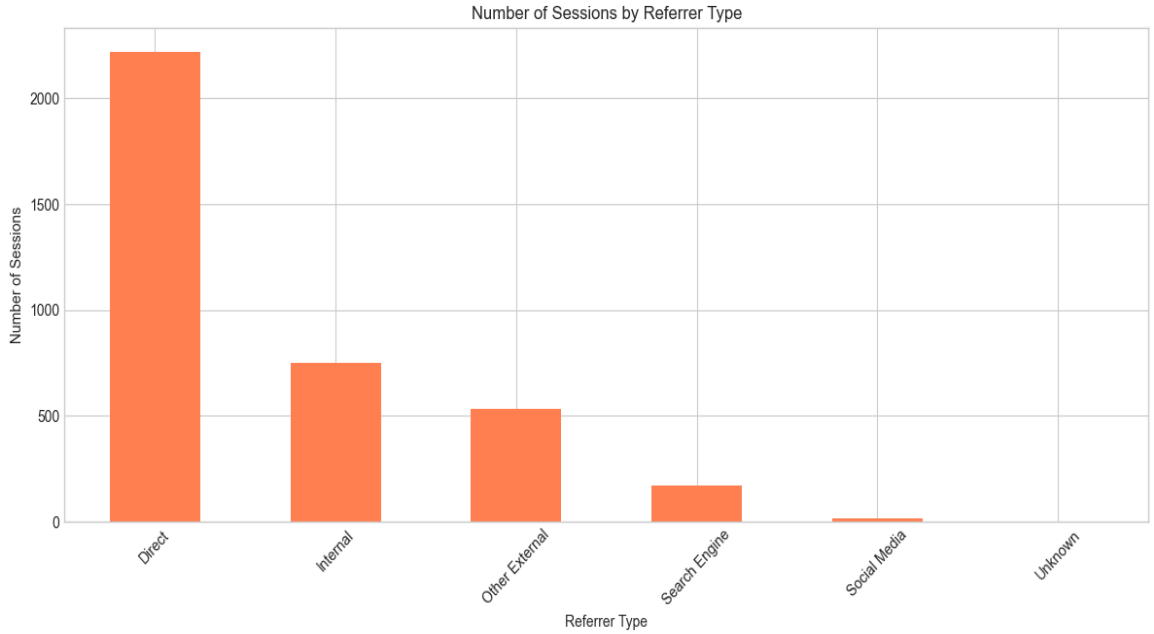

Fig.20 Referrer distribution

Figure 20, shows a notable trend where Direct traffic dominates with roughly 2,200 sessions, followed by Internal referrals at approximately 750 sessions. Other External sources comprise about 500 sessions, while Search Engine traffic provides around 200 sessions. Social media and Unknown sources display minimal impact with fewer than 50 sessions combined. This distribution suggests that a majority of users are either entering the URL directly, utilizing bookmarks, or clicking on untagged links, indicating high brand recognition or frequent returning visitors.

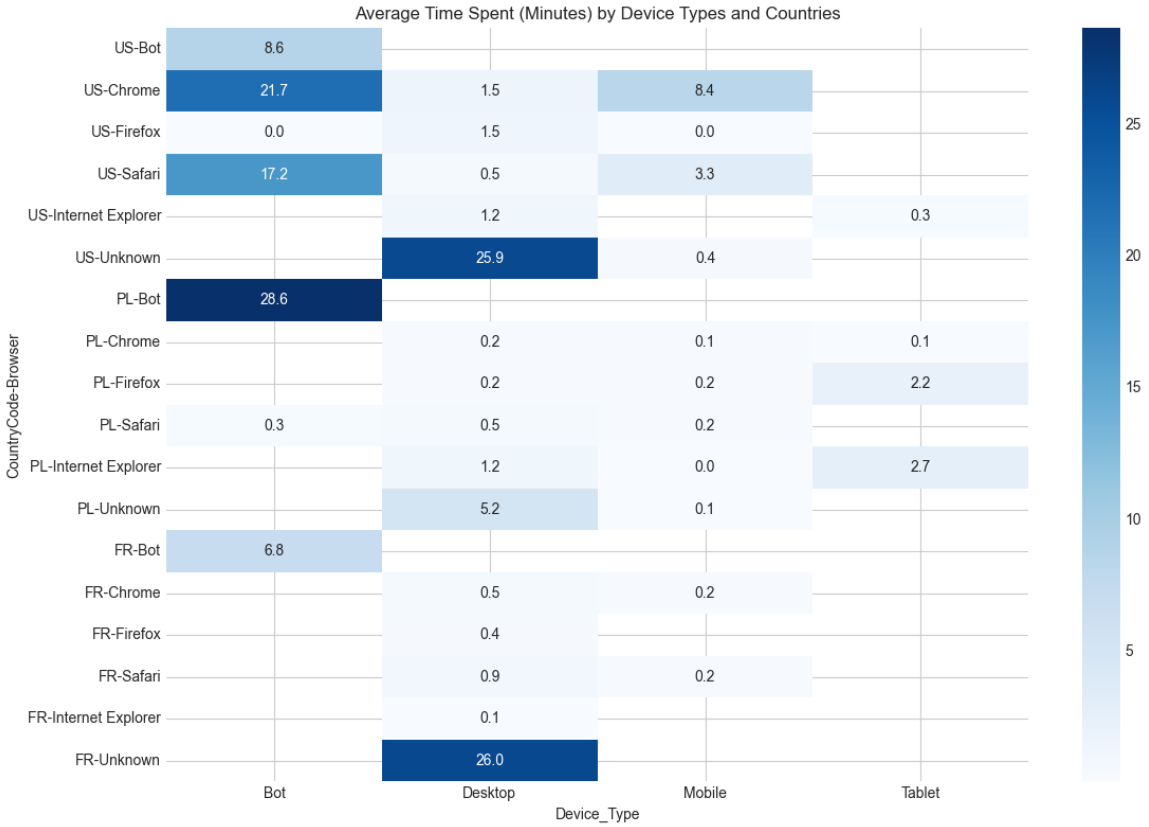

Fig.21. Device Types and Countries

Figure 21 presents a heatmap illustrating the average duration of engagement across various combinations of countries (US, PL, FR) and device types (Bot, Desktop, Mobile, Tablet). The analysis identifies several important trends: users on US-Unknown device types demonstrate the highest level of engagement, averaging 25. 9 minutes, whereas PL-Bot reveals robust activity at 28. 6 minutes, suggesting extended engagement likely due to automated interactions. Likewise, FR-Unknown device indicates considerable engagement

at 26. 0 minutes. Mobile and tablet users typically exhibit significantly lower engagement times, with averages below 8. 4 minutes, indicating shorter browsing durations. Engagement from Chrome and Safari browsers remains moderate across the regions, whereas Internet Explorer displays negligible usage, highlighting its diminished popularity. These results emphasize significant variations in user engagement influenced by device type, geographical location, and browser selection.

Figure 22. Browser Types and Countries

Figure 22 offers additional insights into user engagement trends based on browser and device utilization. It illustrates that unidentified browsers in desktop settings have the longest average session durations, ranging from 25 to 26 minutes. Chrome maintains steady, moderate usage across different regions, especially in the US, whereas Safari shows significant usage in the US market. In contrast, Firefox and Internet Explorer reflect low engagement across all regions. Moreover, bot activity differs by location, with Poland displaying the highest rates of bot engagement. This examination reveals a website

characterized by substantial direct traffic and varied device usage across nations. The elevated bot activity and existence of unidentified browsers indicate a necessity for enhanced traffic filtering and user agent identification. Additionally, the consistently lower engagement times for mobile users across regions suggest potential avenues for improving the mobile user experience.

6. CONCLUSONS AND FUTURE WORK

The detailed examination of over 35 million HTTP requests from this e-commerce platform has uncovered complex user behavior and system performance patterns across various dimensions. The traffic analysis indicated a clear prevalence of GET methods and 200 response codes, with a significant total data transfer of 519. 92 GB, reflecting a generally sound system. Temporal analysis identified peak user activity between 18:00-20:00, with the highest average of 14,402 requests at 20:00, offering important insights for resource allocation and marketing timing. Geographic distribution revealed strong concentrations in Poland and the United States, with notable European presence in the Netherlands, Germany, and the United Kingdom. The user agent analysis indicated Chrome as the leading browser, while the operating system distribution showed Windows (53%) and Linux (40%) as the main platforms, with an unexpectedly high Linux usage implying a tech-savvy user base. Device type analysis demonstrated a nearly equal division between desktop (47%) and mobile (44%) usage, although mobile sessions exhibited significantly shorter durations, suggesting potential user experience challenges on mobile platforms. The session analysis offered particularly useful insights, revealing intricate user journey patterns, with static resources dominating at an average of 3,995. 70 visits per session, while checkout sessions, though fewer in count (114), displayed high engagement (235. 02 visits per session). The referrer analysis showed that 85% of traffic was internal, with direct traffic making up 12%, indicating strong brand recognition but potential prospects for increasing external traffic sources. Future efforts should concentrate on several key areas: mobile optimization to remedy shorter session durations; advanced analytics implementation including predictive models and machine learning algorithms for anomaly detection; performance enhancement through static resource optimization and checkout process refinement; geographic expansion with market-specific strategies; security and bot management featuring sophisticated detection systems; and cross-platform integration encompassing multiple data sources. These improvements will develop a more robust, user-friendly platform capable of accommodating an increasingly diverse and demanding user base while upholding security and performance standards. The execution of these recommendations, founded on the comprehensive analysis of user behavior patterns and

system performance metrics, will position the platform for ongoing growth and enhanced user engagement across all segments and regions.

REFERENCES

[1] G. Chodak, G. Suchacka, and Y. Chawla, "HTTP-level e-commerce data based on server access logs for an online store," Computer Networks, vol. 183, p. 107589, 2020.

[2] A.S. Nagdive, R.M. Tugnayat, G.B Regulwar, and D. Petkar, "Web Server log Analysis for Unstructured data Using Apache Flume and Pig," International Journal of Computer Sciences and Engineering, vol. 7, no. 3, pp. 220-225, 2019.

[3] G. Suchacka and J. Iwański, "Identifying legitimate Web users and bots with different traffic profiles — an Information Bottleneck approach," Knowledge-Based Systems, vol. 197, p. 105875, 2020.

[4] K. R. Suneetha and R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File," IJCSNS International Journal of Computer Science and Network Security, vol. 9, no. 4, pp. 327-332, 2009.

…….. End of Report ……

BDA-Project-Report.pdf

ORIGINALITY REPORT

1%	1%	0%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	<1%
2	Submitted to South Bank University Student Paper	<1%
3	www.khznoise.com Internet Source	<1%
4	websitetrafficdata61604.post-blogs.com Internet Source	<1%
5	Wu, Tai-Chi. "Definition, analysis, and an approach for discrete-event simulation model interoperability", Proquest, 20111109 Publication	<1%

Exclude quotes	On	Exclude matches	< 6 words
Exclude bibliography	On		