Big Data User And Retention Analysis

BDA mini project submitted in partial fulfillment of the requirements
of the degree of

B. E. Computer Engineering

By

Reuel Amin	**01**	**202004**
Suzanne Corda	**08**	**202021**
Elvina Fernandes	**19**	**202042**

Name of the Guide: Ms. Jayashree Mittal
Designation: Assistant Professor

Department of Computer Engineering
St. Francis Institute of Technology
(Engineering College)

University of Mumbai
2023-2024

CERTIFICATE

This is to certify that the mini project entitled "**Big-Data-user-and-retention-analyzing**" is a bonafide work of **Reuel Amin 01, Suzanne 08, Elvina Fernandes 19** submitted to the University of Mumbai in partial fulfillment of the requirement for the BDA subject in the final year of Computer Engineering.

Ms. Jayashree Mittal

Guide

i. Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Reuel Amin 01

Suzanne Corda 08

Elvina Fernandes 19

Date:

ii. Abstract

This project is a comprehensive analysis of user engagement and retention on the BagiKopi.id website, a popular online coffee platform. The main objective is to gain insights into user behaviors and visit patterns to improve the website's performance and user experience. The analysis includes the examination of user visits to different pages, tracking daily user visits, understanding user retention by page, and analyzing hourly user visits. It aims to provide actionable insights for the BagiKopi.id team to enhance user engagement and retention strategies, guiding decisions on content, marketing, and user experience improvements for the continued growth and success of the platform. The project endeavors to furnish the BagiKopi.id team with actionable insights that will guide their efforts in enhancing user engagement and retention.

These insights, drawn from the in-depth analysis of user behavior, have the potential to influence strategic decisions regarding content creation, targeted marketing campaigns, and user experience improvements.

iii. Contents

List of Figures

List of Abbreviations

Sr. No.	Abbreviation	Expanded form
1	BDA	Big Data Analysis
2	CRM	Customer Relationship Management
3	DFD	Data Flow Diagram

Chapter 1

Introduction

1.1 Description

The BagiKopi.id website serves as a hub for coffee enthusiasts, providing a wide array of content, resources, and community engagement opportunities related to coffee culture. As user preferences and expectations in the digital space evolve rapidly, it is imperative for the platform to continuously assess and improve its user engagement and retention strategies. This project seeks to delve deep into the wealth of data available from user interactions on the website to derive actionable insights that can drive these enhancements.

1.2 Problem Formulation

The problem at hand revolves around the need to understand user behavior, visit patterns, and the effectiveness of user retention strategies on the BagiKopi.id website. Users come to the website seeking information, community, and experiences related to their passion for coffee. However, it is crucial to identify which aspects of the platform are resonating with users and where improvements can be made. The project aims to formulate data-driven answers to these critical questions.

1.3 Motivation

The motivation for this project lies in the desire to provide a superior user experience on BagiKopi.id. A well-engaged and retained user base is not only a testament to the platform's value but also the lifeblood of its growth and success. Understanding how users navigate the website, which pages they frequent, and what drives them to return is key to refining content strategies and enhancing overall user satisfaction.

1.4 Proposed Solution

The proposed solution involves a multifaceted data analysis approach. We will explore and visualize user visits by page to identify which pages are attracting the most attention. Daily user visit trends will be tracked to uncover temporal patterns. User retention by page will provide insights into the effectiveness of content and engagement on different sections of the site. Finally, we will delve into the hourly distribution of user visits to identify peak hours, offering opportunities for targeted content updates. These analyses will collectively inform strategies to enhance user engagement and retention on BagiKopi.id.

1.5 Scope of The Project

The scope of this project encompasses the analysis of historical user data from the BagiKopi.id website, focusing on user engagement and retention patterns. It includes the extraction, cleaning, and exploration of data, followed by the application of data visualization and analysis techniques to extract meaningful insights. The project does not involve direct website modifications but aims to provide actionable insights that can guide future decisions for optimizing user engagement and retention strategy.

Chapter 2
Review of Literature

[1] This paper focuses on the application of predictive modeling techniques for customer retention in the e-commerce sector, utilizing big data analytics to forecast customer behavior and optimize retention strategies. The study emphasizes the importance of leveraging advanced data analytics and machine learning algorithms to identify key patterns and trends contributing to customer churn and engagement in the dynamic e-commerce landscape.

The research delves into the process of data preprocessing, feature engineering, and model development, highlighting the significance of data-driven decision-making in understanding customer preferences, purchase behavior, and product affinity. Gupta and Singh demonstrate the effectiveness of predictive models in forecasting customer retention rates, enabling e-commerce businesses to implement proactive customer engagement initiatives and personalized marketing campaigns tailored to individual customer segments.

The paper underscores the role of customer relationship management (CRM) and targeted marketing strategies in fostering customer loyalty and long-term engagement. By integrating big data analytics into the customer retention framework, Gupta and Singh emphasize the potential for e-commerce companies to optimize customer experiences, reduce churn rates, and drive revenue growth through personalized recommendations and tailored retention programs.

Overall, "Predictive Modeling for Customer Retention in E-commerce Using Big Data Analytics" by Gupta and Singh contributes valuable insights into the pivotal role of predictive modeling and data-driven analytics in the e-commerce sector, emphasizing the importance of customer-centric approaches and proactive retention strategies in enhancing customer satisfaction and fostering long-term customer relationships.

[2] This paper provides a comprehensive analysis of user engagement in the online realm, focusing on the understanding and prediction of user behavior within digital platforms. The study employs an extensive dataset and utilizes various machine learning algorithms to decipher the factors influencing user engagement and retention in the online domain.

The research emphasizes the significance of personalized user experiences, targeted content delivery, and data-driven analytics in driving user engagement and enhancing user satisfaction. It highlights the importance of user segmentation and behavior analysis in identifying key patterns and trends associated with user engagement, enabling organizations to tailor their online strategies and content offerings to meet user preferences and expectations effectively.

The paper underscores the role of predictive modeling and data-driven decision-making in optimizing user experiences and fostering long-term user engagement. By leveraging advanced analytics and machine learning techniques, Wang et al. demonstrate the potential for organizations to predict user behavior, anticipate user preferences, and proactively address user needs, thereby enhancing user retention rates and overall online engagement.

Overall, "Understanding and Predicting User Engagement in an Online World" by Wang et al. contributes valuable insights into the complex dynamics of user engagement within the digital landscape, emphasizing the pivotal role of data analytics and predictive modeling in driving user-centric strategies and fostering sustainable user relationships in the online domain.

[3]This paper offers an extensive and insightful analysis of user retention strategies in the context of mobile applications. The study provides a comprehensive overview of various retention techniques and approaches employed by mobile app developers to enhance user engagement and prolong user retention rates.

The research emphasizes the significance of personalized user experiences, effective onboarding processes, and targeted push notification campaigns in fostering long-term user relationships and improving overall user satisfaction. Liu and Chen delve into the intricacies of user segmentation, user behavior analysis, and user feedback integration, highlighting the importance of data-driven decision-making and user-centric strategies in optimizing user retention rates and minimizing user churn.

The paper underscores the role of user engagement metrics, user journey mapping, and A/B testing methodologies in evaluating the effectiveness of different retention strategies and facilitating continuous improvement in mobile app user retention. Liu and Chen demonstrate the potential for personalized recommendations, in-app rewards, and tailored user experiences in driving user satisfaction and loyalty, thereby contributing to increased user retention rates and enhanced app performance.

Overall, "A Comprehensive Review of User Retention Strategies in Mobile Apps" by Liu and Chen serves as a valuable resource for mobile app developers and industry practitioners, providing actionable insights and best practices for devising effective user retention strategies and fostering sustainable user engagement in the competitive mobile app market.

Chapter 3

System Analysis

3.1 Functional Requirement

1. Data Cleaning and Preprocessing: The system must perform data cleaning and preprocessing to handle missing or erroneous data, ensuring data quality for analysis.

2. User Engagement Analysis: The system should allow for the analysis of user engagement by providing insights into the most visited pages, frequent user interactions, and the distribution of user activities across different website sections.

3. User Retention Analysis: The system should support the analysis of user retention by page, tracking how often users return to specific pages, and assessing the effectiveness of content and engagement on the website.

4. Temporal Analysis: The system should enable the analysis of user activities over time, including daily and hourly trends, to identify patterns and potential seasonality in user visits.

5. Data Visualization: The system must offer data visualization capabilities, including bar charts, line charts, and other relevant visualizations to present the results of the analysis.

6. Data Extraction and Integration: The system should be able to extract and integrate user data from the BagiKopi.id website, including user visits, timestamps, and page interactions.

3.2 Non Functional Requirements:

1. Data Security: Ensure the security of user data by implementing appropriate data protection measures to safeguard user privacy and comply with data protection regulations.

2. Scalability: The system should be scalable to handle large volumes of data as the website's user base grows, without a significant decrease in performance.

3. Usability: The user interface of the system should be user-friendly, making it accessible to non-technical stakeholders and data analysts.

4. Reliability: The system should be reliable, with minimal downtime, ensuring that users can access and analyze data when needed.

3.2.1 Performance Requirements

1. Data Processing Time: The system should process and analyze the data efficiently, with a reasonable response time to ensure a smooth user experience.

2. Real-time Data Updates: If real-time data analysis is a requirement, the system should provide up-to-date insights, especially for time-sensitive decision-making

3.2.2 Software Quality Attributes

1. Accuracy: The analysis results should be accurate, reflecting the true user engagement and retention patterns on the website.

2. Maintainability: The system's code and data processing pipelines should be maintainable to allow for updates, enhancements, and troubleshooting as needed.

3. Scalability: The system should be designed to handle increased data volumes as the website's user base grows, without a significant loss of performance.

4. Interoperability: The system should be able to integrate with other data sources or tools that the BagiKopi.id team may use for a comprehensive analysis.

5. Robustness: The system should be resilient to errors and able to handle unexpected situations, ensuring the stability of the analysis process.

3.3 Specific Requirements:

Hardware :

- Intel Pentium III/800 MHz or higher (or compatible).
- 512 MB of memory minimum.
- 1 GB of memory is recommended.
- 120 MB of available hard disk space for installation.
- 1024x768 minimum screen resolution.

Software :

- Windows 7 or above operating system.
- Visual Studio Code
- Required libraries installed.
- Python
- GUI- Python Tkinter

3.4 Use-Case Diagrams and description

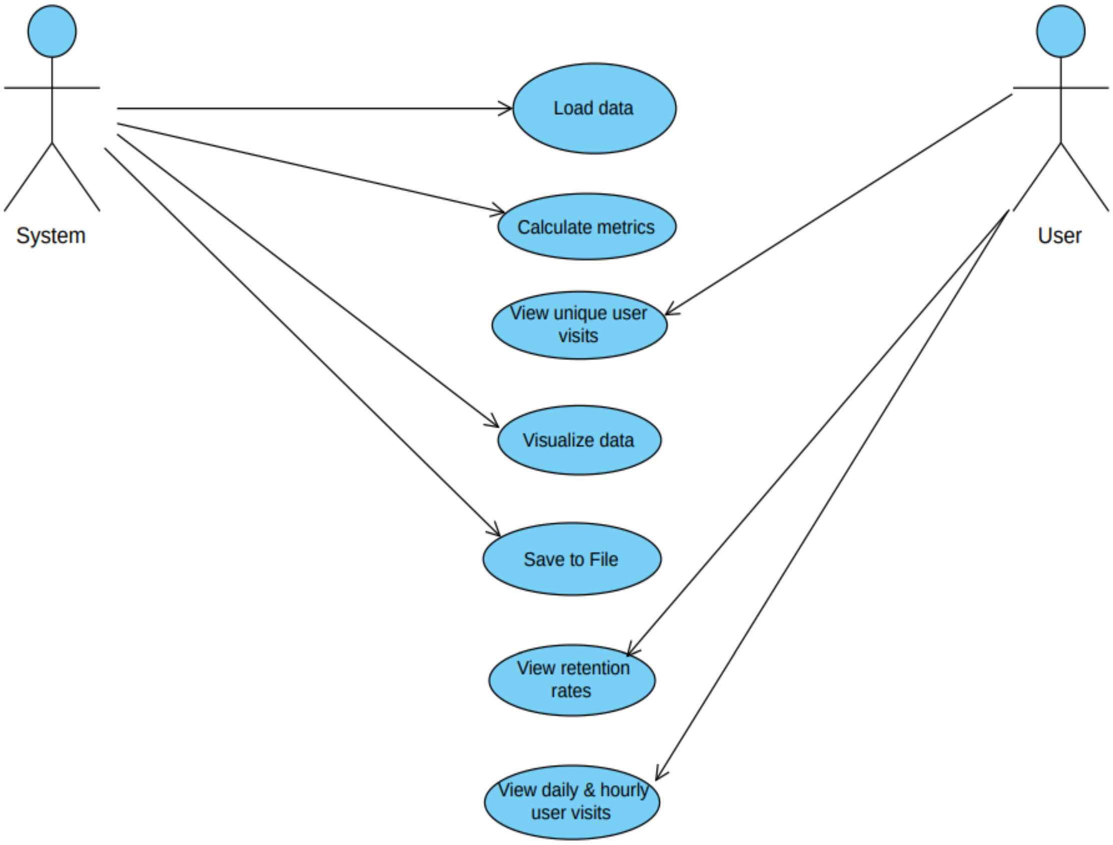


Fig. 3.1 Use Case Diagram

Actors: User and system are the actors that interact with each other.

Use cases: Load data,calculate metrics,view unique user visits,save to file, view retention rates and view daily & hourly user visits are use cases.

The system can load data and calculate metrics and the user will be able to view the number of unique user visits, retention rate and daily and hourly visits of the users in the form of visualizations.

Chapter 4

Analysis Modeling

4.1 Class Diagram and Activity Diagram

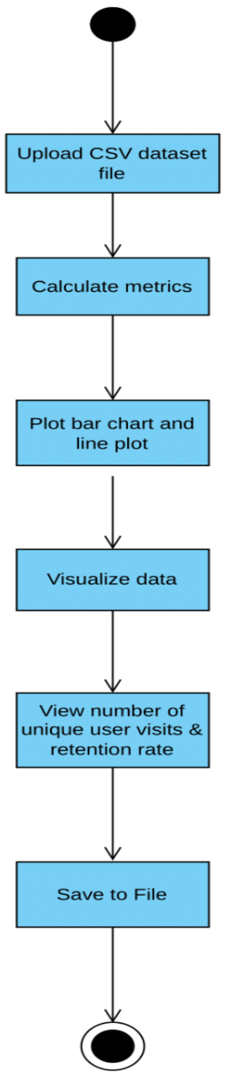

Upload CSV dataset file
Calculate metrics
Plot bar chart and line plot
Visualize data
View number of unique user visits & retention rate
Save to File

Fig. 4.1 Activity Diagram

In the above figure, the csv file is taken as input and number of unique users and retention rate are calculated as metrics and bar chart and line plot are used for data visualization and then the file is saved in the form of png.

4.2 Functional Modelling

Data Flow Diagram

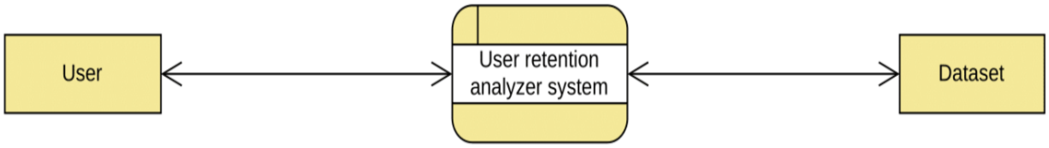


Fig. 4.2 : DFD Level 0 Diagram

In the above figure, the user will communicate with the user retention analyzer System . This system will communicate with the user as well as refer to the dataset and generate output.

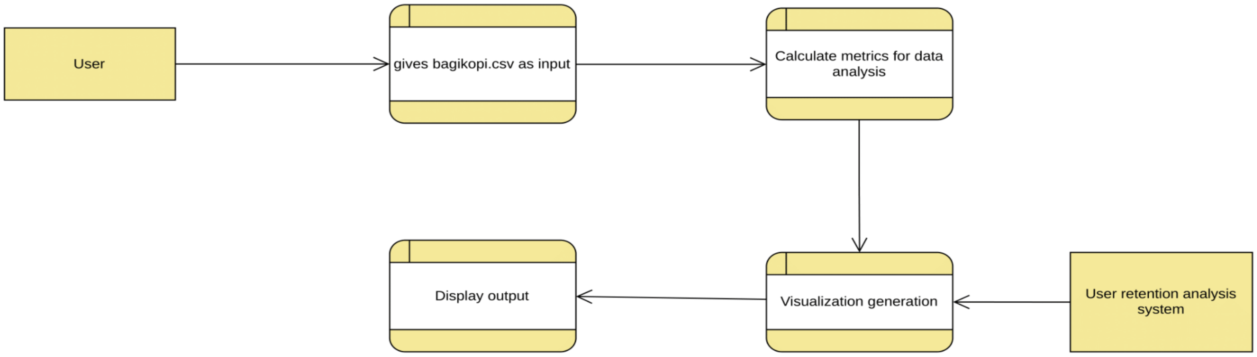

User
gives bagikopi.csv as input
Calculate metrics for data analysis
Display output
Visualization generation
User retention analysis system

Fig. 4.3 DFD Level 1 Diagram

In the above figure, the user provides the dataset of bagikopi website containing different pages and user_ids as input, the system processes it, generates visualizations, and presents the results for user viewing and interaction.

Chapter 5

Design

5.1 Architectural Design

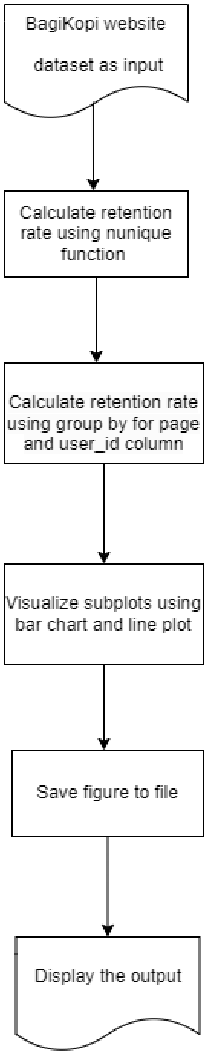
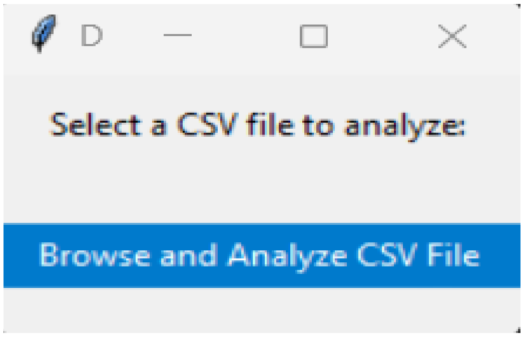

BagiKopi website dataset as input
Calculate retention rate using nunique function
Calculate retention rate using group by for page and user_id column
Visualize subplots using bar chart and line plot
Save figure to file
Display the output

Fig 5.1 Architectural Design

In the above figure, the bagikopi website dataset is given as input using which the system calculates retention rate and number of unique users and provides visualization of subplots using bar charts and line plots to the user and saves the subplot figure to a file.

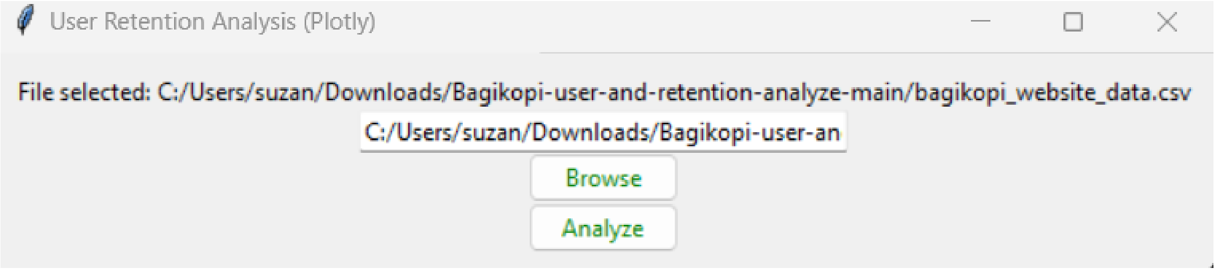
5.2 User Interface Design



D

Select a CSV file to analyze:

Fig 5.2 User interface 1 (Initially)



User Retention Analysis (Plotly)

File selected: C:/Users/suzan/Downloads/Bagikopi-user-and-retention-analyze-main/bagikopi_website_data.csv

C:/Users/suzan/Downloads/Bagikopi-user-an

Fig 5.3 User Interface 2 (While giving input)

Chapter 6
Implementation

6.1 Working of the project

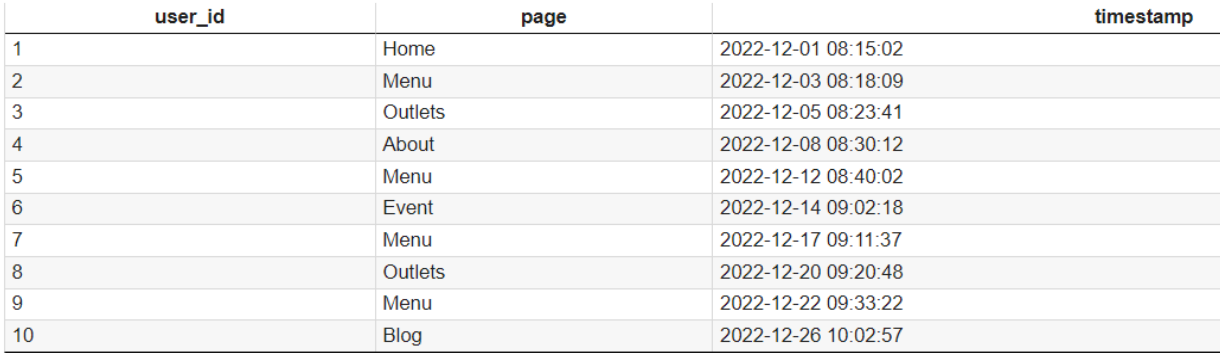
Code:

bagikopi_analysis.py

```
import pandas as pd
import matplotlib.pyplot as plt
import tkinter as tk
from tkinter import filedialog

# Create a function to handle data analysis and visualization
def analyze_and_visualize_data():
    # Prompt the user to select the dataset CSV file
    file_path = filedialog.askopenfilename(filetypes=[("CSV Files", "*.csv")])

    if not file_path:
        return  # User canceled the file dialog

    # Load the website data into a pandas DataFrame:
    df = pd.read_csv(file_path)

    # Calculate the number of unique users who visited each page:
    unique_users = df.groupby('page')['user_id'].nunique()

    # Calculate the retention rate for each page:
    retention_rate = df.groupby('page')['user_id'].apply(lambda x: x.nunique() / x.count())

    # Visualize the results using a bar chart and line plot:
    fig, ax1 = plt.subplots()
    ax1.bar(unique_users.index, unique_users.values)
    ax1.set_ylabel('Unique Users')
    ax2 = ax1.twinx()
    ax2.plot(retention_rate.index, retention_rate.values, color='r')
    ax2.set_ylabel('Retention Rate')
    plt.title('User Views and Retention on bagikopi.id')
    plt.show()

    # Save the figure to a file:
    fig.savefig('user_views_and_retention.png')

# Create a GUI window
root = tk.Tk()
root.title("Data Analysis and Visualization")

# Create a label for instructions
instructions = tk.Label(root, text="Select a CSV file to analyze:")
instructions.pack(pady=10)

# Create a button with custom styling
analyze_button = tk.Button(root, text="Browse and Analyze CSV File",
command=analyze_and_visualize_data, bg="#007acc", fg="white", relief="flat", padx=10)
analyze_button.pack(pady=20)

# Run the GUI main loop
root.mainloop()
```

dashboardui.py

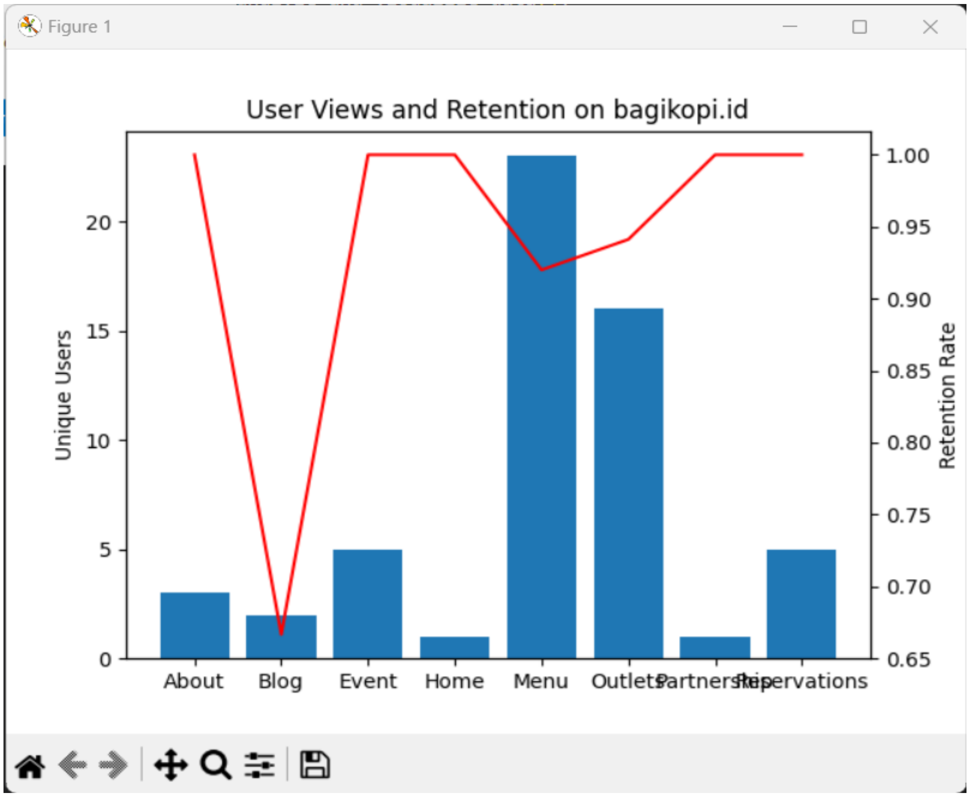
```
import pandas as pd
import tkinter as tk
from tkinter import filedialog
from tkinter import ttk
import plotly.express as px
from plotly.subplots import make_subplots
import plotly.graph_objs as go

def analyze_retention_plotly(file_path):
    # Load your data and perform user retention analysis using Plotly
    df = pd.read_csv(file_path)

    # Convert 'timestamp' column to datetime data type
    df['timestamp'] = pd.to_datetime(df['timestamp'])

    # Create a figure with multiple subplots
    fig = make_subplots(rows=2, cols=2, subplot_titles=("User Visits by Page", "Daily User Visits", "User Retention by Page", "Hourly User Visits"))

    # Add trace for user visits by page
    page_visits = df.groupby('page').count()['user_id']
    fig.add_trace(go.Bar(x=page_visits.index, y=page_visits.values), row=1, col=1)

    # Add trace for daily user visits
    daily_visits = df.groupby(df['timestamp'].dt.date).count()['user_id']
    fig.add_trace(go.Scatter(x=daily_visits.index, y=daily_visits.values, mode='lines+markers'), row=1, col=2)

    # Add trace for user retention by page
    page_retention = df.groupby(['page', df['timestamp'].dt.date]).nunique()['user_id'].reset_index()
    fig.add_trace(go.Scatter(x=page_retention[page_retention['page'] == 'Menu']['timestamp'], y=page_retention[page_retention['page'] == 'Menu']['user_id'], mode='lines+markers', name='Menu'), row=2, col=1)
    fig.add_trace(go.Scatter(x=page_retention[page_retention['page'] == 'Outlets']['timestamp'], y=page_retention[page_retention['page'] == 'Outlets']['user_id'], mode='lines+markers', name='Outlets'), row=2, col=1)

    # Add trace for hourly user visits
    hourly_visits = df.groupby(df['timestamp'].dt.hour).count()['user_id']
    fig.add_trace(go.Scatter(x=hourly_visits.index, y=hourly_visits.values, mode='lines+markers'), row=2, col=2)

    # Update figure layout
    fig.update_layout(title="User Analysis Dashboard",
                      xaxis_title="Page/Dates/Hour of Day",
                      yaxis_title="Number of User Visits",
                      height=700,
                      width=1000)

    # Display the Plotly dashboard
    fig.show()

def open_file_dialog(label, entry):
    file_path = filedialog.askopenfilename(title="Select CSV file")
    if file_path:
        entry.delete(0, tk.END)
        entry.insert(0, file_path)
        label.config(text="File selected: " + file_path)

def analyze_data():
    file_path = file_entry.get()
    if file_path:
        analyze_retention_plotly(file_path)

root = tk.Tk()
root.title("User Retention Analysis (Plotly)")

# Create a frame to hold the content
content_frame = ttk.Frame(root)
content_frame.pack(padx=10, pady=10)

# File selection section
file_label = ttk.Label(content_frame, text="Select a CSV file:")
file_label.pack()
file_entry = ttk.Entry(content_frame, width=40)
file_entry.pack()
browse_button = ttk.Button(content_frame, text="Browse", command=lambda: open_file_dialog(file_label, file_entry))
browse_button.pack()

# Analyze button with custom styling
analyze_button = ttk.Button(content_frame, text="Analyze", command=analyze_data, style="TButton")
analyze_button.pack()

# Style the button with green text and maintain the color
style = ttk.Style()
style.configure("TButton", foreground="green")

# Run the GUI main loop
root.mainloop()
```



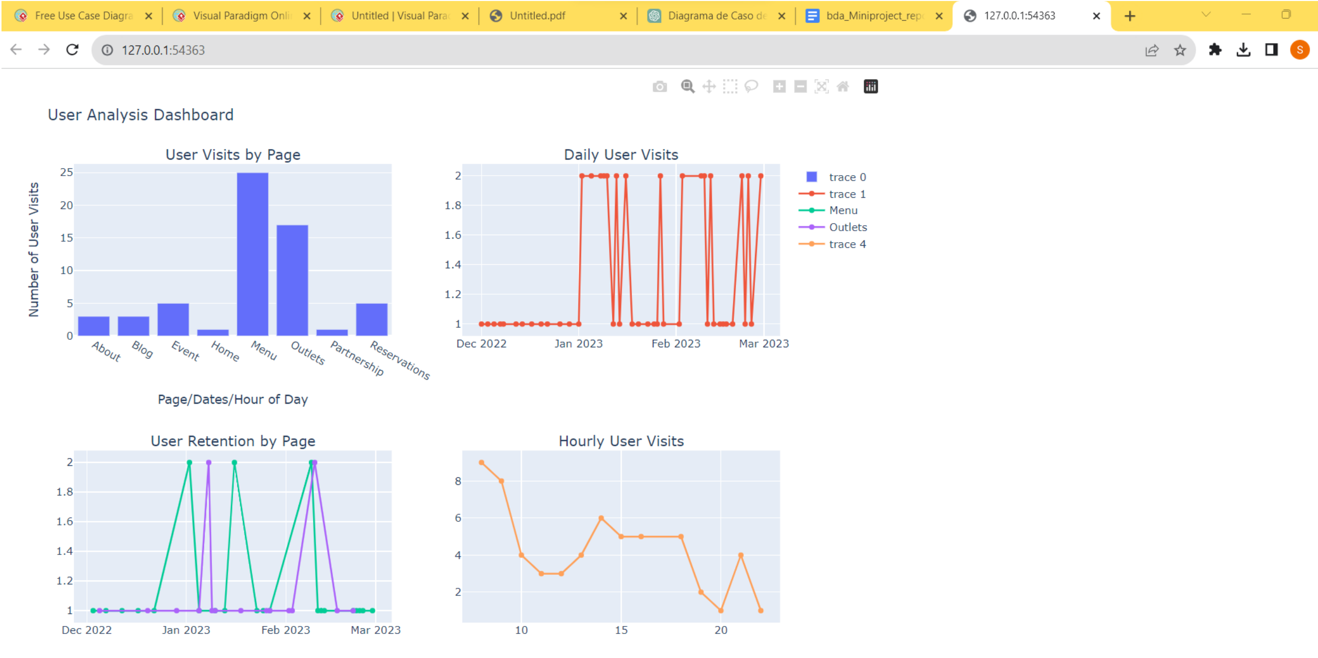
user_id	page	timestamp
1	Home	2022-12-01 08:15:02
2	Menu	2022-12-03 08:18:09
3	Outlets	2022-12-05 08:23:41
4	About	2022-12-08 08:30:12
5	Menu	2022-12-12 08:40:02
6	Event	2022-12-14 09:02:18
7	Menu	2022-12-17 09:11:37
8	Outlets	2022-12-20 09:20:48
9	Menu	2022-12-22 09:33:22
10	Blog	2022-12-26 10:02:57

6.1 bagikopi_website_data.csv

Output:



6.2 User views and Retention



6.3 User Analysis Dashboard

Chapter 7

Conclusion

In conclusion, this project has undertaken a comprehensive analysis of user engagement and retention on the BagiKopi.id website. Through a multifaceted approach, we have examined user visits by page, tracked daily user visit trends, assessed user retention by page, and analyzed hourly user visit distributions. The insights derived from this analysis will play a pivotal role in guiding the BagiKopi.id team in optimizing their user engagement and retention strategies.Understanding which pages are most frequented by users, discerning temporal visit patterns, and evaluating the effectiveness of content on different sections of the website are invaluable for enhancing the overall user experience. By focusing on these aspects of user behavior and retention, we aim to provide actionable insights that can drive future decisions regarding content creation, targeted marketing campaigns, and user experience improvements.

References

[1] “Predictive Modeling for Customer Retention in E-commerce Using Big Data Analytics" by Gupta and Singh (2020).

[2] “Understanding and Predicting User Engagement in an Online World" by Wang et al. (2017).

[3] “A Comprehensive Review of User Retention Strategies in Mobile Apps" by Liu and Chen (2019).

[4]https://userpilot.com/blog/retention-analysis/

[5]https://www.revenera.com/blog/software-monetization/what-is-user-retention/

Acknowledgements

We would like to express our deepest appreciation to our academic mentors, especially principal Sincy George, for her vision and encouragement, HOD Dr. Kavita Sonawane, for her academic insights and direction, and Guide Jayashree Mittal, for her expertise and mentorship. Their dedication and knowledge have greatly enriched the quality of this project.We are also immensely grateful to the BagiKopi.id team for providing essential access to the data necessary for this analysis. Their collaboration and support have been instrumental in the successful execution of this project. Furthermore, we express our gratitude to the developers, data analysts, and stakeholders who have generously contributed their expertise, insights, and time to this project. Their invaluable input has been a key factor in the success of this analysis.