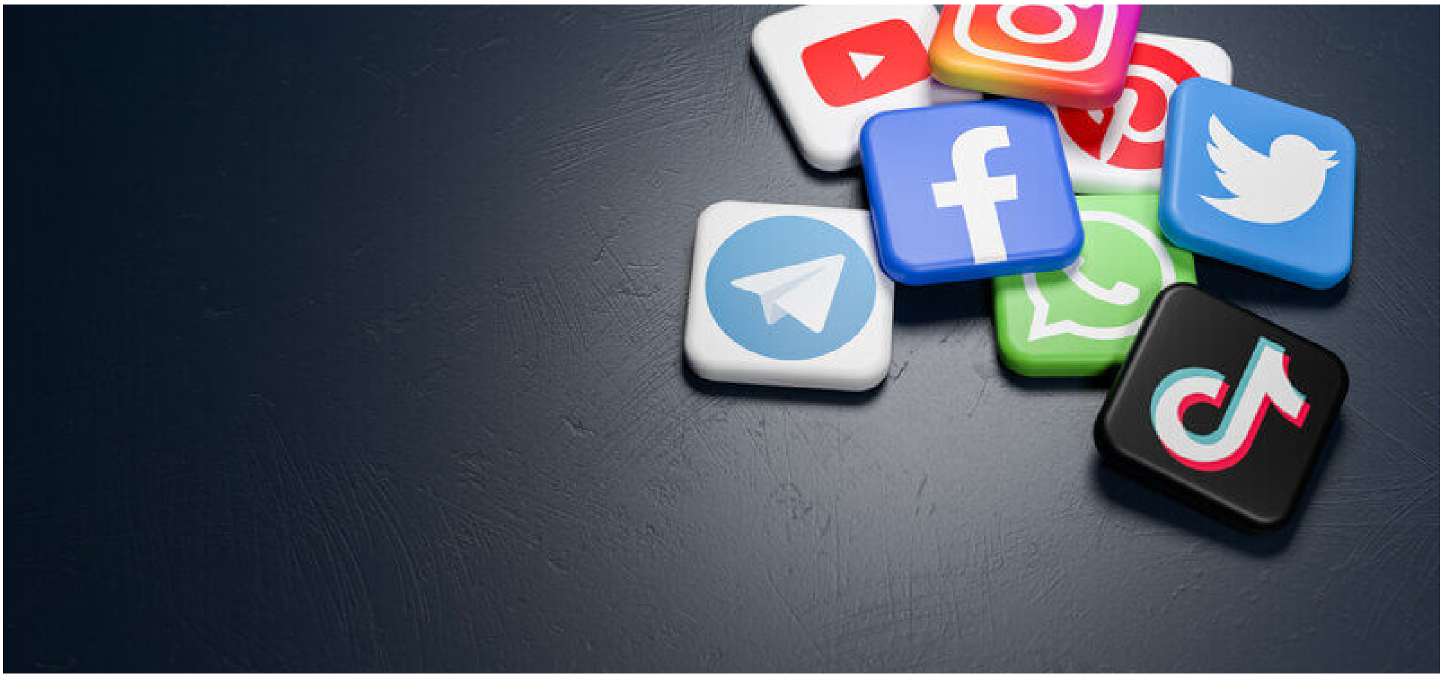


Socialinsider Exploratory Data Analysis

1. Introduction and Objectives

1.1 Overview of the Project

📌 Our project analyzes user interaction events on Socialinsider's website—such as pages viewed and modules used—to predict purchase likelihood using machine learning classification methods.

1.2 Goals of the EDA

- **Understand** the typical user journey of a customer on Socialinsider
- **Identify** key user behavior patterns
- **Assess** feature importance for predictive modeling
- **Prepare** the data for modeling and future analysis

2. Data Overview

2.1 Data Source

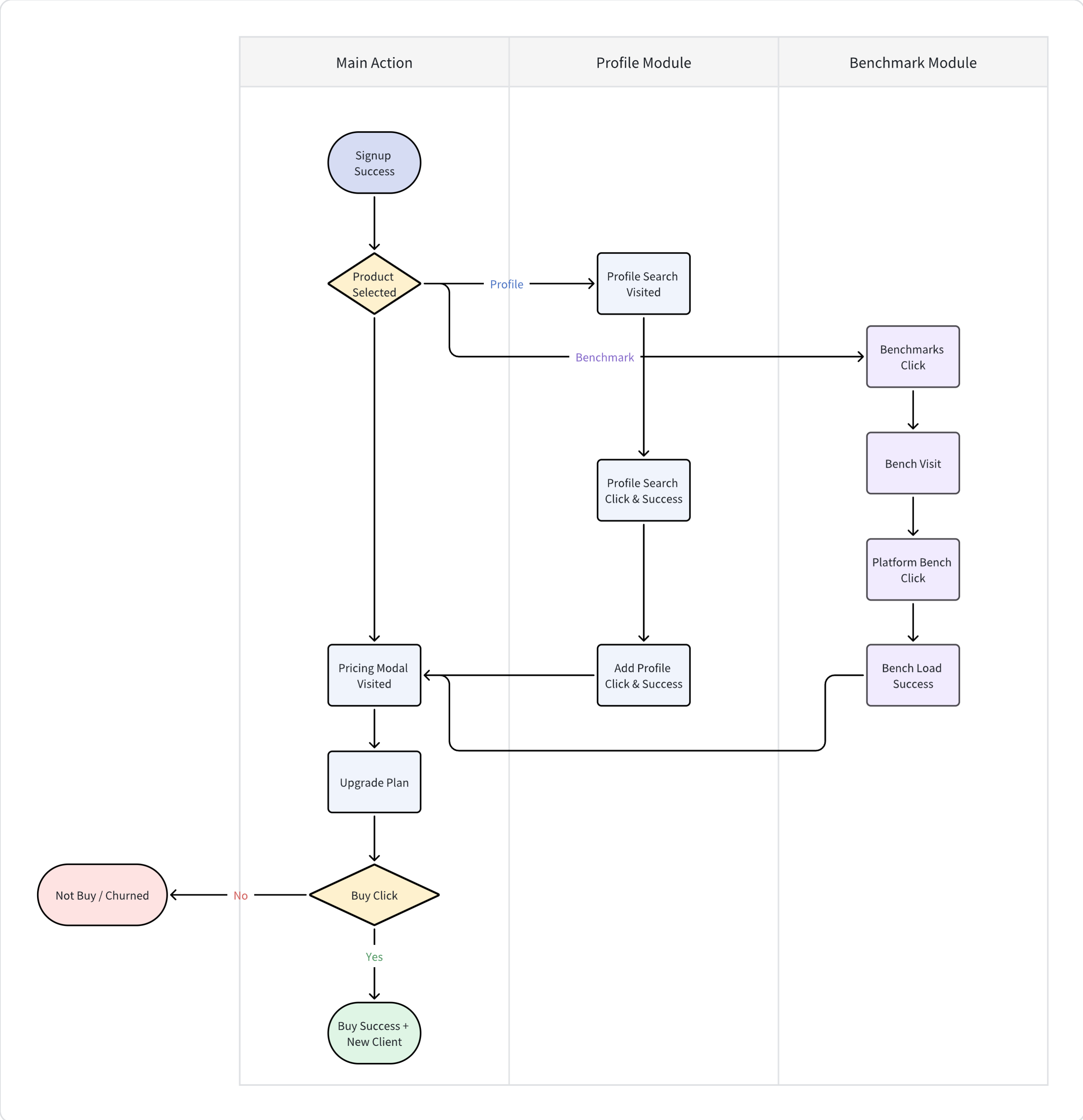
The data was collected and sent to us by the client.

2.2 Data Description

Name	Value
Timeframe	4/30/2024 to 9/17/2024
Column Names	event_name, user_id, time_created, user_type, time_zone, country, view, platform, report_type, load_time
Total Events	970,517
Total Unique Users	10,608
Total Converted Users	103
Overall CVR	0.971%
Total Unique Countries	143
Total Unique Timezones	218
Unique Views	'profile', 'projecthome', 'settings', 'hashtag', 'upgradeplan', 'benchmark', 'campaigns', 'reports', 'postsfeed', 'brands', 'bench', 'add', 'addprofiles', 'page', 'proj', 'ads', 'search', 'connect', nan
Unique Platforms	'ig', 'instagram', 'tw', 'twitter', 'tk', 'tiktok', 'yt', 'youtube', 'meta', 'fb', 'facebook', 'xch', 'cross-platform', 'li', 'brbench', 'linkedin', 'hashtags', 'showFacebook', 'all', nan
Unique Report Types	'ppt_new', 'pptx', 'pdf_new', 'pdf', 'xls', 'xlsx', 'csv', 'ppt', nan

3. Data Insights & Visualization

3.1 Example User Journey



3.2 Data Visualizations

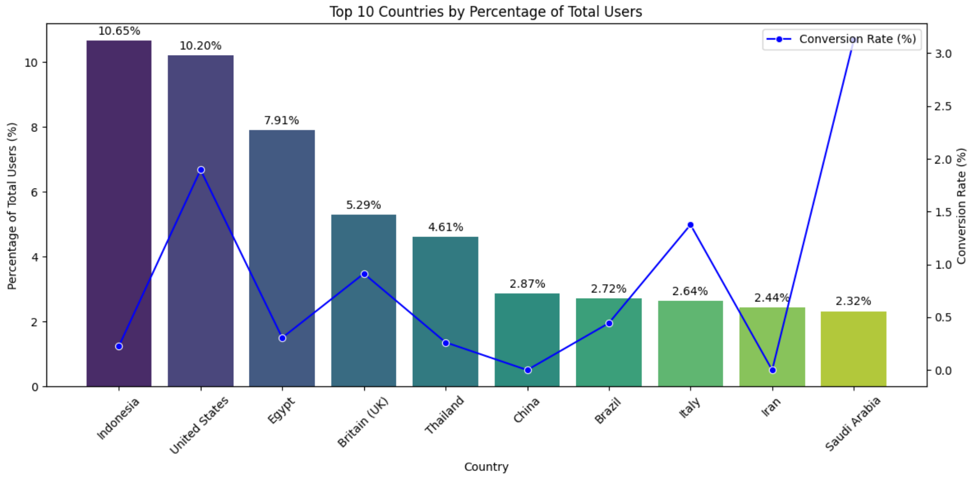
> We have selected the following meaningful visualizations from our analysis.
>
> Note: We have taken India out of this analysis since it is not a target audience for potential new clients

Description	Visualization
User & CVR Breakdown by Country	

India leads with 17.22% of users, followed by Indonesia and the U.S.

Wealthier countries like the US, Italy, and Saudi Arabia have higher conversion rates than the rest of the world.

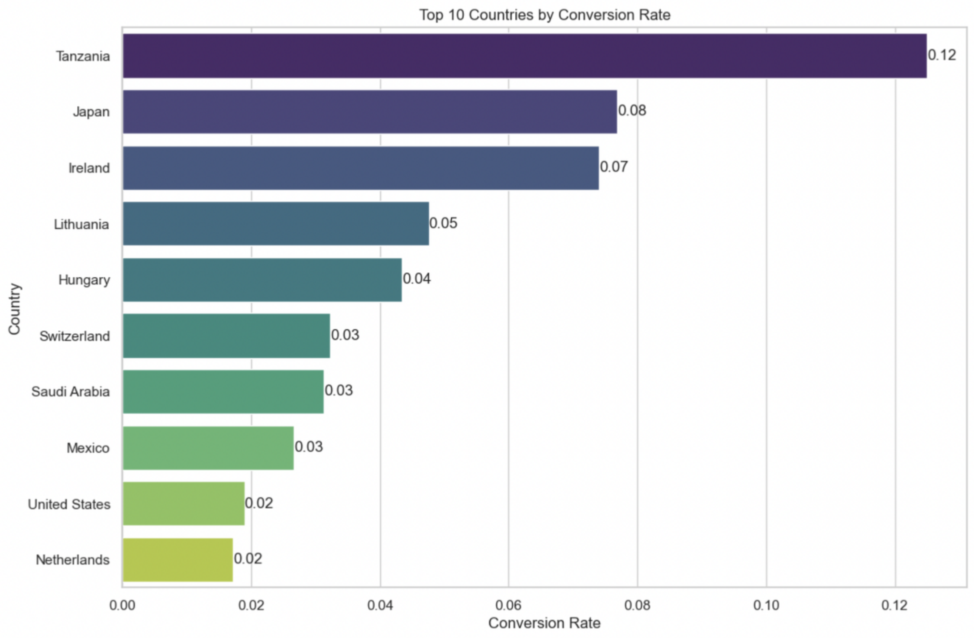
> 💡 **Country-specific factors** could play a crucial role in user behavior, making **"country" a potentially influential feature** for predictive models.



Top 10 Counties that have the most sign up users become to buy users

Tanzania, Japan and Ireland have relatively high conversion rates among the top 10.

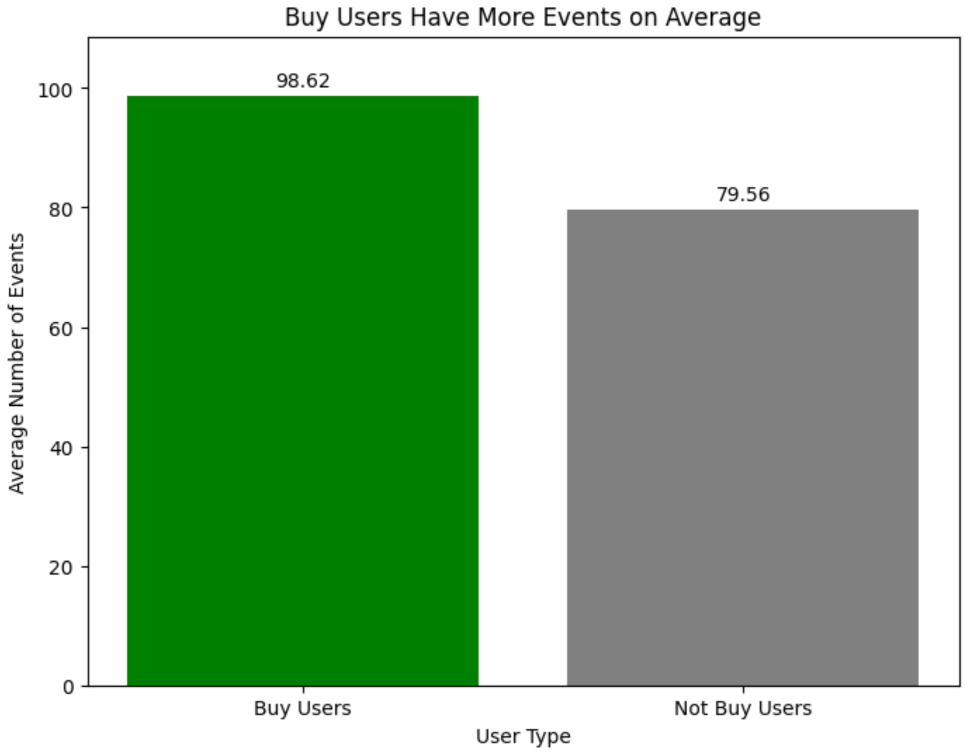
> 💡 We can **advertise** more in **Tanzania, Japan and Ireland**.



# of Total Events for Buy Users and Not Buy Users

Buy users exhibit a slightly higher average number of events (104.73) compared to non-buy users (91.36).

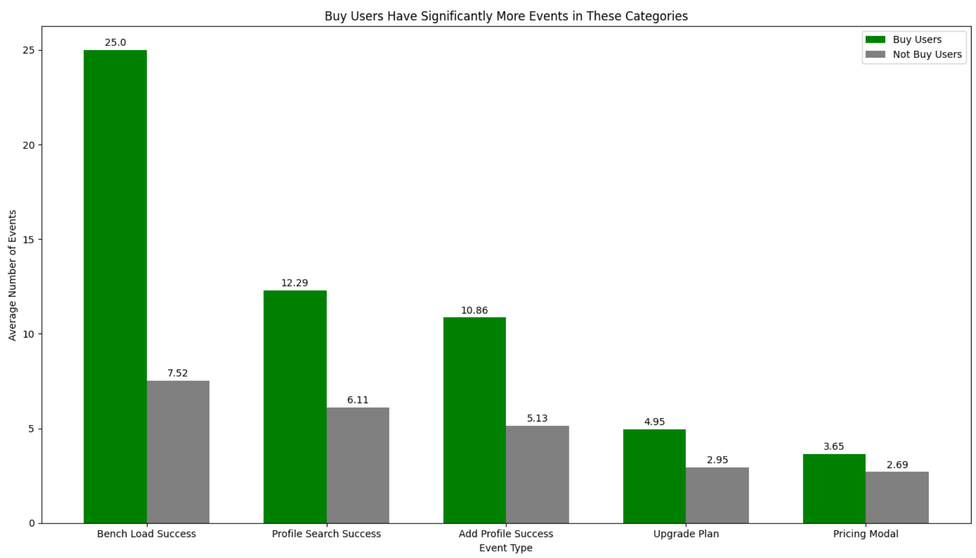
> 💡 This alone might not be a strong predictor. Further analysis of the **specific types of events and user interactions** is needed to determine which event types have a **more significant influence** on conversion.



Event Counts for Buy Users and Not Buy Users (Positive Correlation)

Buy users engage significantly more in certain event categories, especially in "Bench Load Success" and "Profile Search Success".

> 💡 These event types could be **strong indicators** of user intent to purchase.

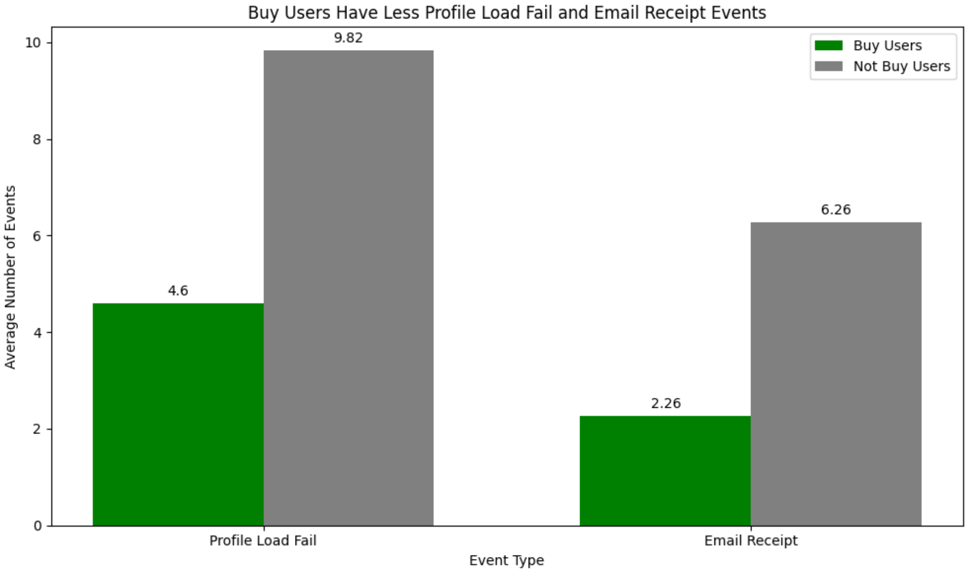


Event Counts for Buy Users and Not Buy Users

(Negative Correlation)

Buy users have significantly fewer "Profile Load Fail" and "Email Receipt" events compared to non-buy users.

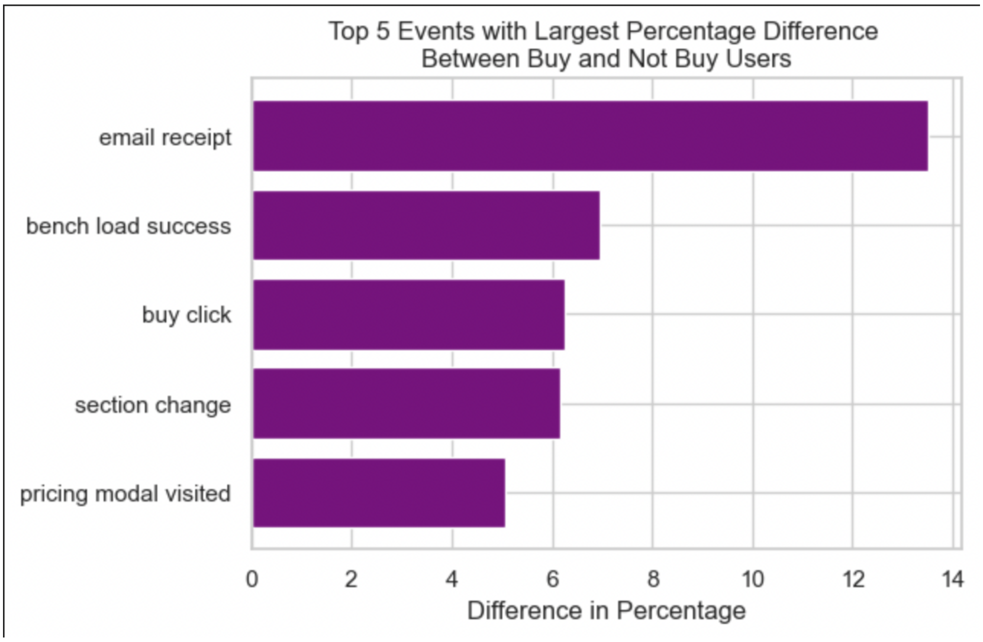
💡 These two events may **hinder user experience** and lead to **lower conversion**. They might also be influential predictors with **negative correlation** to conversion rate.



Top 5 Events that have the Largest Difference in Percentage of the Event Frequency of Buy and Not Buy Users

"Email receipt" is the event that has the most significant difference.

💡 The event **"email receipt"** can be used as a factor to **differentiate users**. The **more it happens**, the more possible the user will **subscribe**.

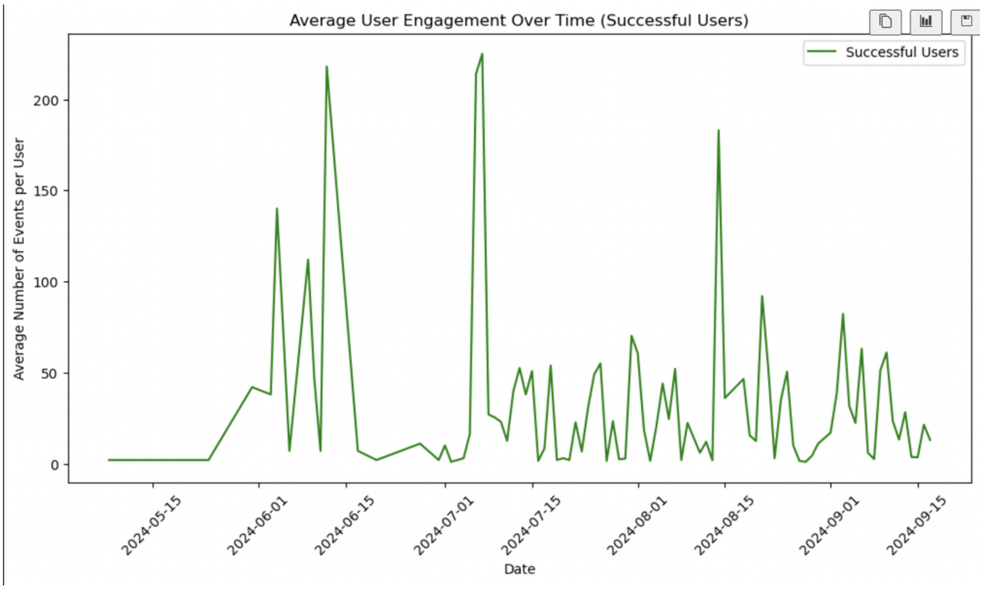


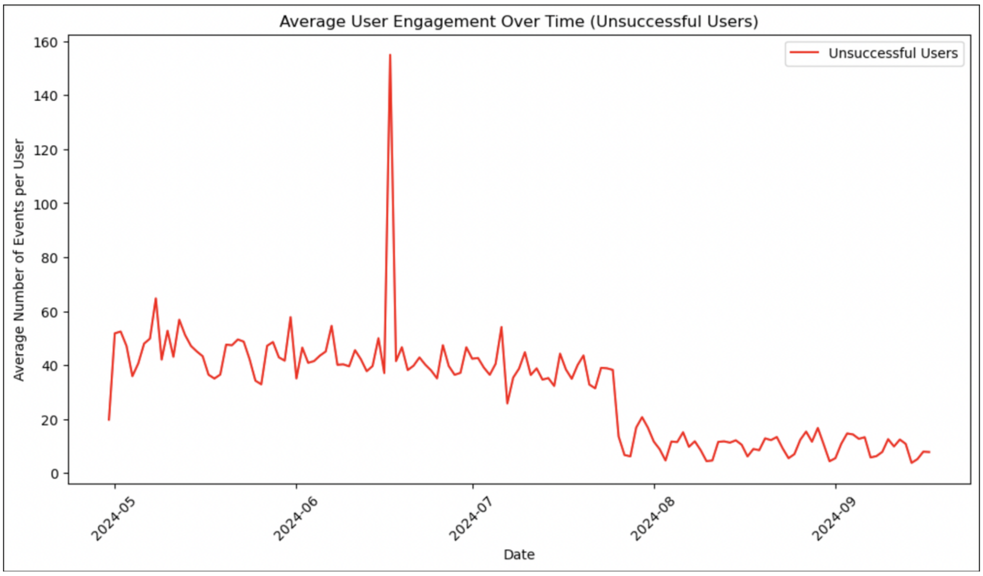
Time Series Plot of User Engagement by Counting

Average Events Numbers Happened per User.

Buy users consistently engage more on average than not buy users, and the declining trend in general is more obvious in the plot of not buy users.

💡 The **more engagement(average number of events)** the user shows, the more possible he or her will **subscribe**. We could do further study on peaks to check what happened on the website during that time, for example launching new functions.

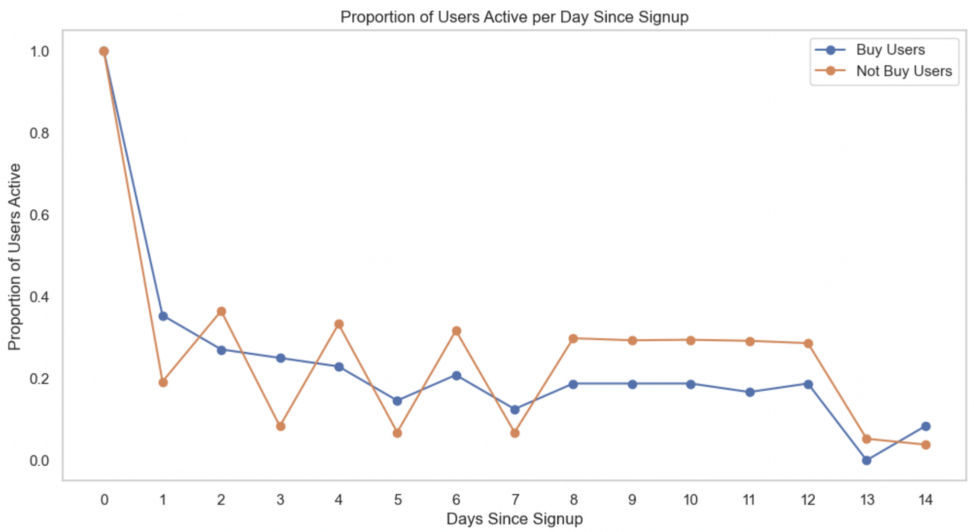




Proportion of Users Active per Day Since Signup (14 Days active status)

After signing up, the number of active users all decline in buy and not buy users. However, buy users show a more stable active users' proportion.

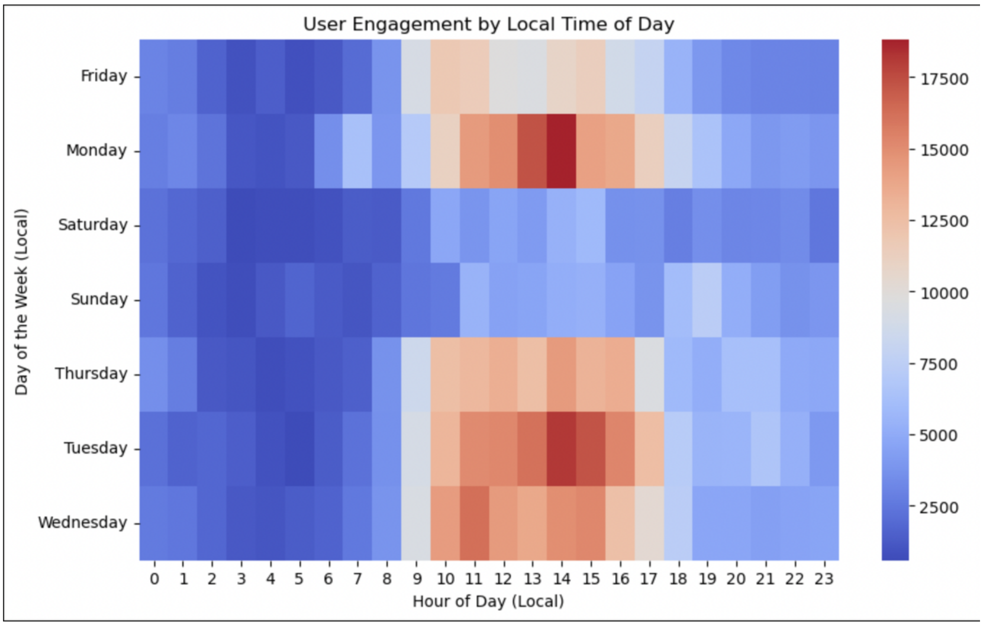
💡 Users that show a **more consistent**



active pattern tend to subscribe since they may have need everyday.

User Hourly Engagement (Number of Total Events) in a Week

💡 People use the web the most on Monday 2pm and usually use it a lot from **9am-5pm on weekdays**.

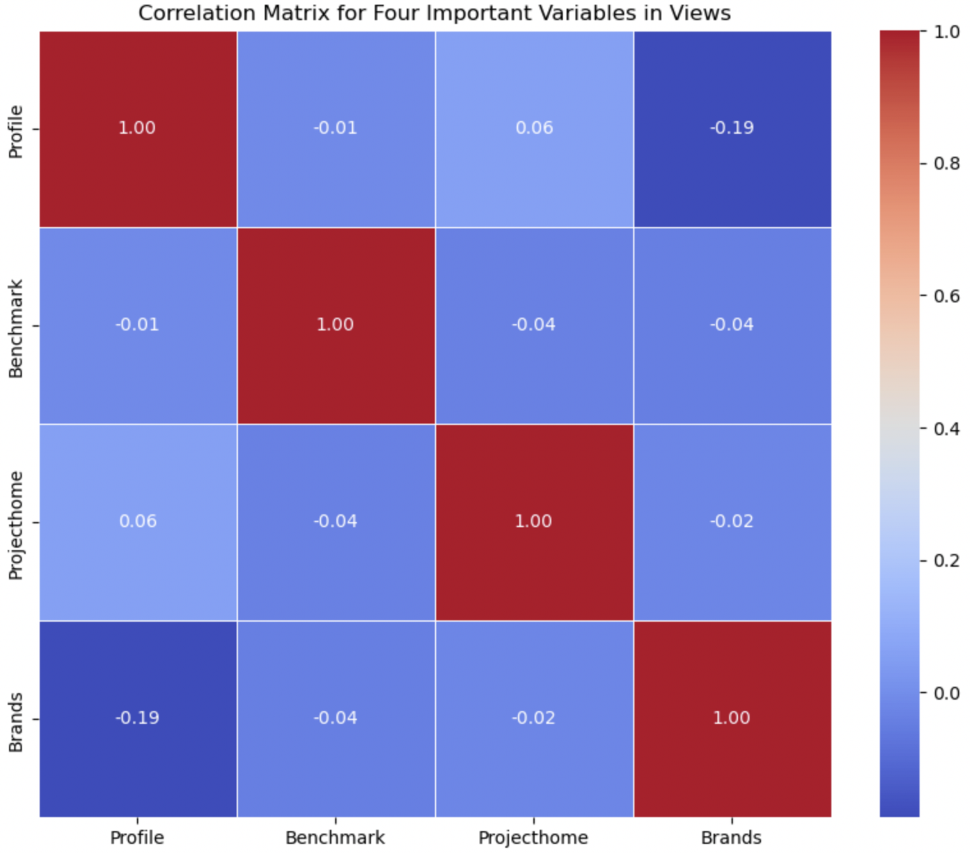


Correlation Matrix of Profile, Benchmark, Projecthome and Brands under View

We noticed that there is obvious difference between these four kinds of views between buy and not buy users.

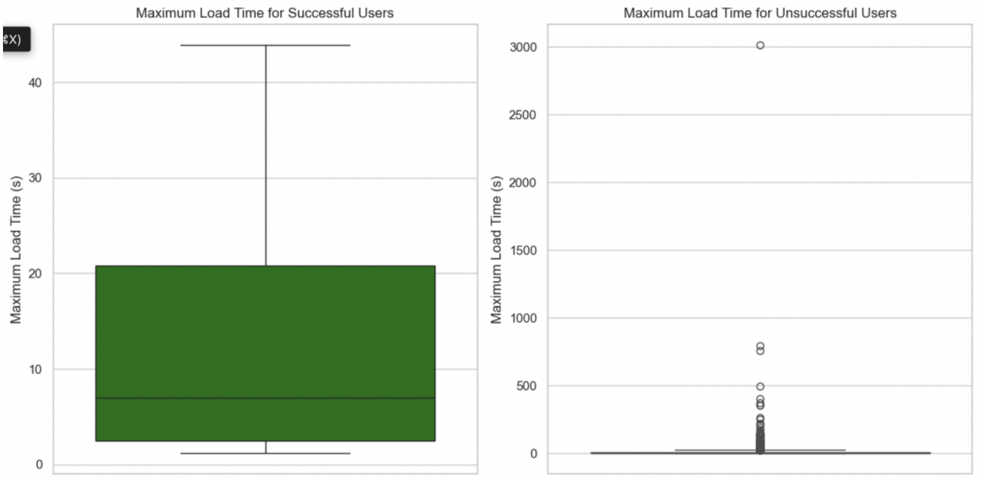
💡 **No linearity** between these four. We may use them as factors according to models' results

View	Buy Success	Not Buy Success
Profile	20.92%	38.79%
benchmark	20.62%	9.11%
projecthome	10.56%	7.11%
brands	6.04%	2.62%
postsfeed	2.20%	2.89%



Boxplots of Maximum Load Time for Buy and Not Buy Users

For the plot of buy users, most of the data falls within a relatively narrow range, with a median near the lower quartile. No outlier. For the plot of not buy users, it shows a much wider range of load times, with a significant number of high outliers. The median is higher than that of buy users, and the spread of data points is broader.



> 💡 Buy users generally experience **quicker load times**

4. Data Cleaning & Feature Engineering

Transformed from event-level data (each row represents a web event) to user-level data (each row represents a unique user.)

Original Data

By event:

	Event Name	User ID	Time Created	User Type	Country	View	Platform	Report Type	Load Time	Successful
Event #1										
Event #2										
Event #3										

Transformed Data:

By user:

User ID	Conversion	Country	Average Load Time	Maximum Load Time	Total Number of events	Count of Certain Events	Total Number of Platforms	Count of Certain Platform	Count for Each Type of View
User 1									
User 2									
User 3									

Transformed Features:

- Conversion
- Country--select the first country shown at the event

- Aggregated Load Time
  - Average Load time
  - Maximum load time
- Count of events for each user (events that can potentially distinguish whether users can convert)
  - bench load success
  - profile search success
  - add profile success
  - pricing modal visited
  - profile load fail
  - email receipt
- Count of each platform--combined categories
  - Facebook & fb -----> fb -----> platform_fb_count
  - Twitter & tw -----> tw -----> platform_tw_count
  - Instagram & ig -----> ig -----> platform_ig_count
  - Youtube & yt -----> yt -----> platform_yt_count
  - Linkedin & li -----> li -----> platform_li_count
  - Tiktok & tk -----> tk -----> platform_tk_count
  - Cross-platform & xch -----> xch -----> platform_xch_count
  - Total number of platform -----> platform_total_count
- Count for each type of view (19 categories)
  - Profile -----> view_profile
  - Projecthome -----> view_projecthome

5. Next Steps

Student Team

- [ ] Revise data visualizations and data pipeline
- [ ] Start building the model

Socialinsider

- [ ] Social Insider Events Q&A Spreadsheet