

# 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

## 2.3 Data Cleaning & Feature Engineering

Transformed from event data (each row represents a web event) to user data (each row represents a user)

### Original Data:

By event:

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

Event 3									
---------	--	--	--	--	--	--	--	--	--

### Transformed Data:

By user:

	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									

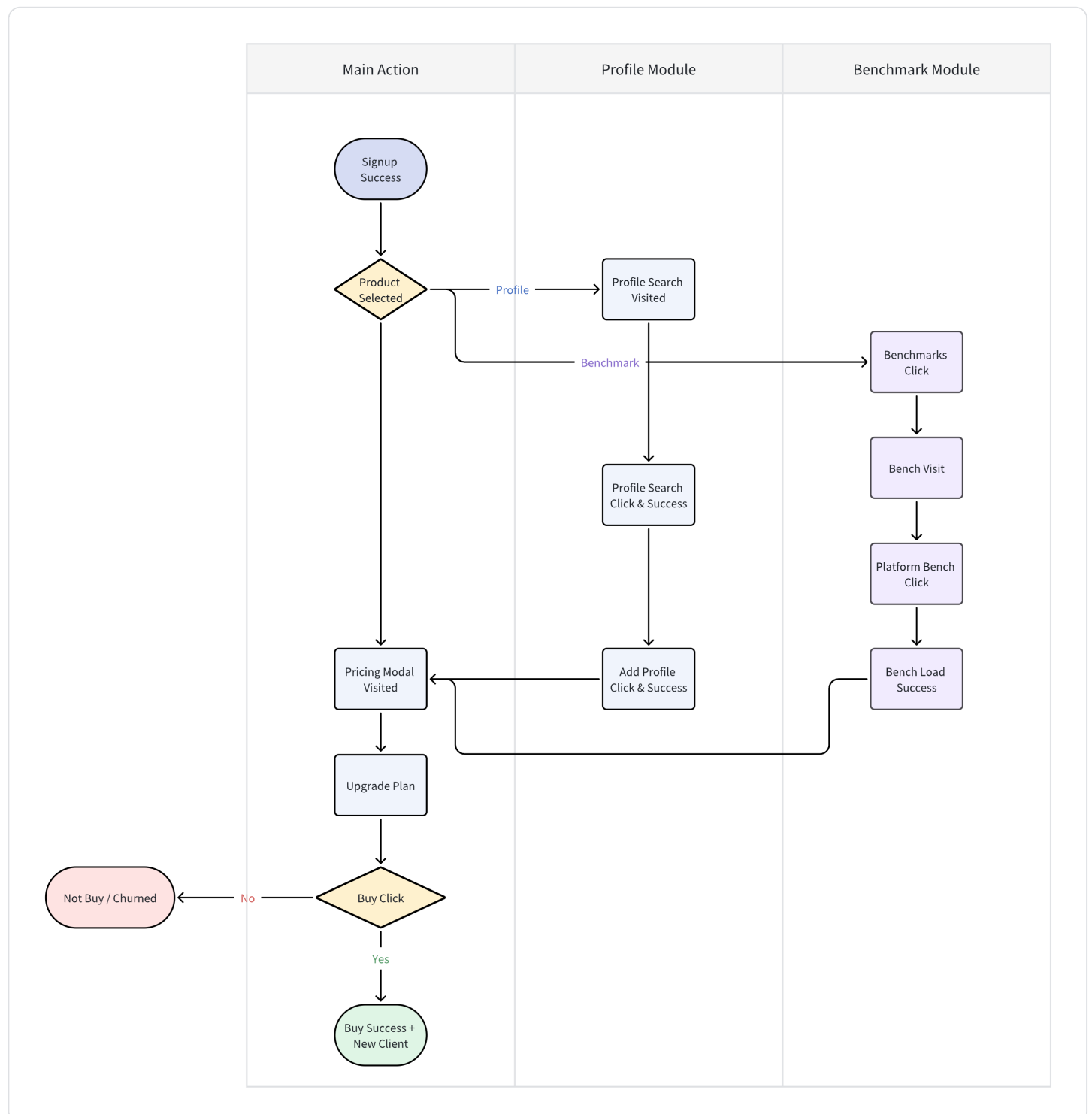
### Transformed Features:

- 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--combine categories
  - fb
  - tw


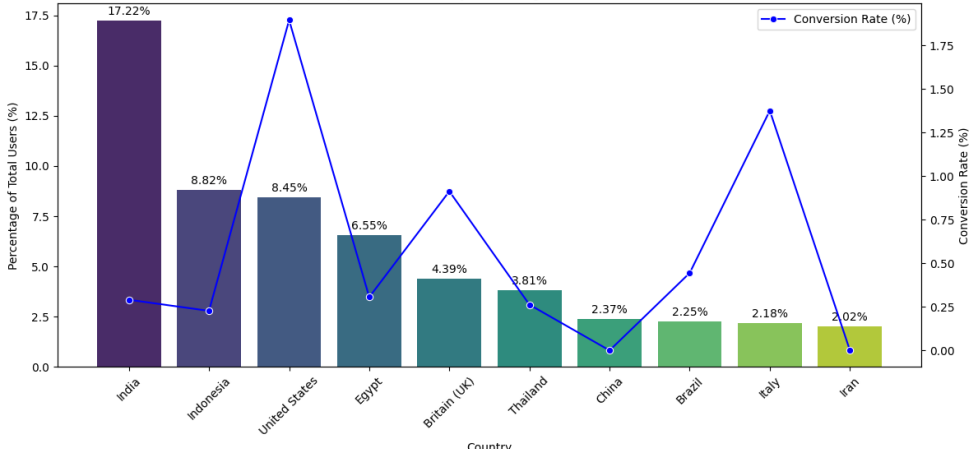

- ig
- yt
- li
- tk
- Count for each type of view (19 categories)

## 3. Data Insights & Visualization

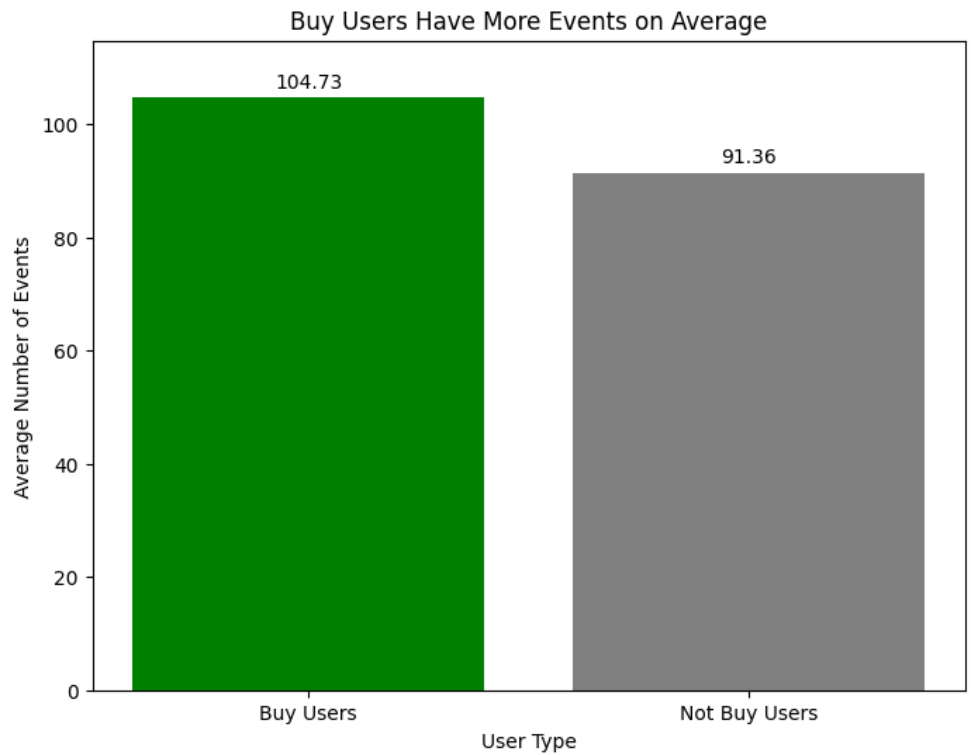
### 3.1 Example User Journey



### 3.2 Data Visualizations

Description	Visualization																																	
<div><div>User &amp; CVR Breakdown by Country</div><div>India leads with 17.22% of users, followed by Indonesia and the U.S.</div><div>Western countries like the US, UK, and Italy have higher conversion rates than rest of the world.</div></div> <div><div> <b>Country-specific factors</b> could play a crucial role in user behavior, making <b>“country”</b> a <b>potentially influential feature</b> for predictive models.</div></div>	<div><div>Top 10 Countries by Percentage of Total Users</div><table><tr><th>Country</th><th>Percentage of Total Users (%)</th><th>Conversion Rate (%)</th></tr><tr><td>India</td><td>17.22%</td><td>0.25</td></tr><tr><td>Indonesia</td><td>8.82%</td><td>0.25</td></tr><tr><td>United States</td><td>8.45%</td><td>1.50</td></tr><tr><td>Egypt</td><td>6.55%</td><td>0.25</td></tr><tr><td>Britain (UK)</td><td>4.39%</td><td>0.85</td></tr><tr><td>Thailand</td><td>3.81%</td><td>0.40</td></tr><tr><td>China</td><td>2.37%</td><td>0.10</td></tr><tr><td>Brazil</td><td>2.25%</td><td>0.45</td></tr><tr><td>Italy</td><td>2.18%</td><td>1.30</td></tr><tr><td>Iran</td><td>2.02%</td><td>0.10</td></tr></table></div>	Country	Percentage of Total Users (%)	Conversion Rate (%)	India	17.22%	0.25	Indonesia	8.82%	0.25	United States	8.45%	1.50	Egypt	6.55%	0.25	Britain (UK)	4.39%	0.85	Thailand	3.81%	0.40	China	2.37%	0.10	Brazil	2.25%	0.45	Italy	2.18%	1.30	Iran	2.02%	0.10
Country	Percentage of Total Users (%)	Conversion Rate (%)																																
India	17.22%	0.25																																
Indonesia	8.82%	0.25																																
United States	8.45%	1.50																																
Egypt	6.55%	0.25																																
Britain (UK)	4.39%	0.85																																
Thailand	3.81%	0.40																																
China	2.37%	0.10																																
Brazil	2.25%	0.45																																
Italy	2.18%	1.30																																
Iran	2.02%	0.10																																
<div><div># of Total Events for Buy Users and Not Buy Users</div><div>Buy users exhibit a slightly higher average number of events (104.73) compared to non-buy users (91.36).</div></div> <div><div> This alone might not be a strong predictor. Further analysis of the <b>specific types of events</b> and <b>user interactions</b> is</div></div>																																		

needed to determine which event types have a **more significant influence** on conversion.

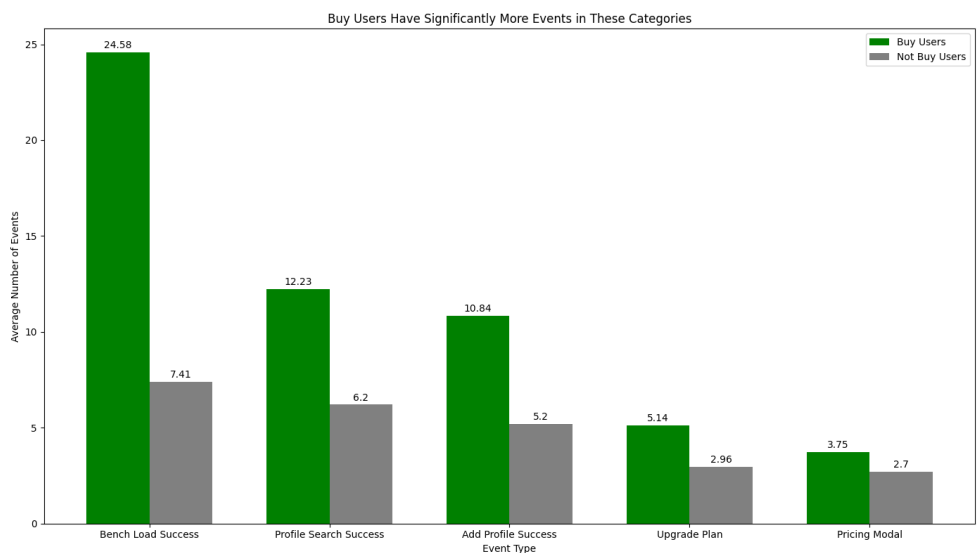


### Event Counts for Buy Users and Not Buy Users (Positive Correlation)

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.



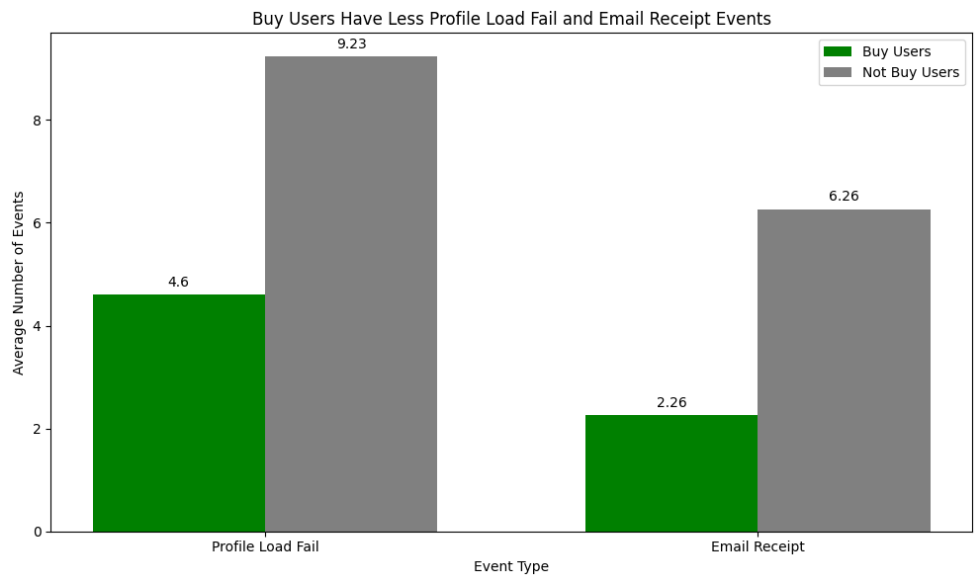
### Event Counts for Buy Users and Not Buy Users (Negative Correlation)

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.

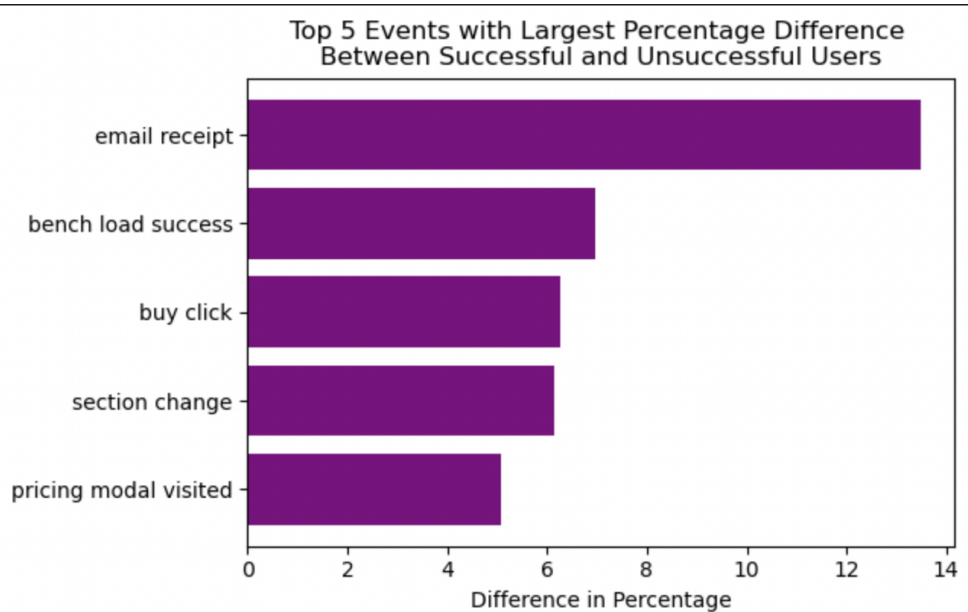


### 5 Events that have the Largest Difference in Percentage of the Event Frequency of Successful and Unsuccessful Buyers

"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**.

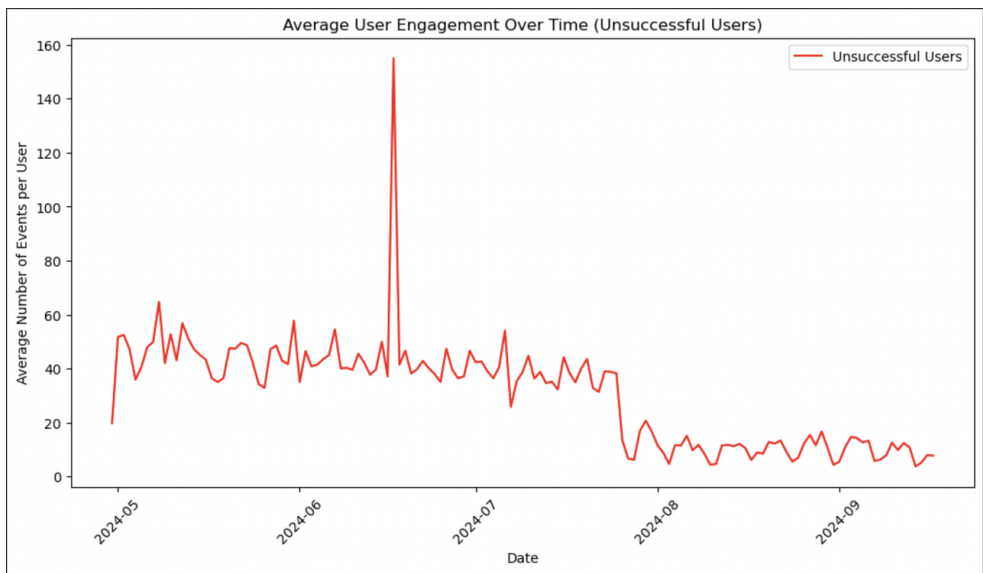
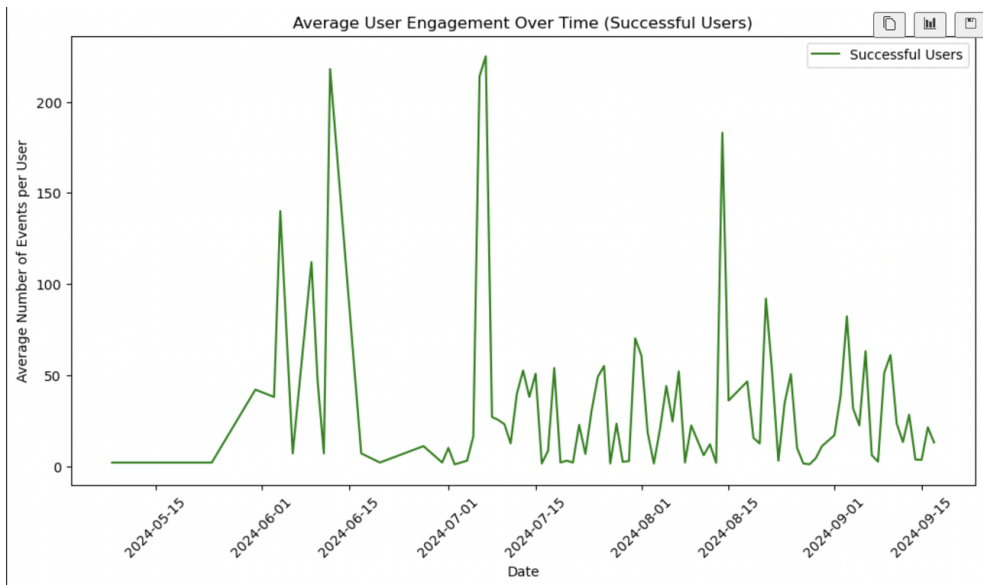


**Time Series Plot of User Engagement by Counting Average Events Numbers Happened per User.**

Successful users consistently engage more on average than unsuccessful users, and the declining trend in general is more obvious in the plot of unsuccessful users.



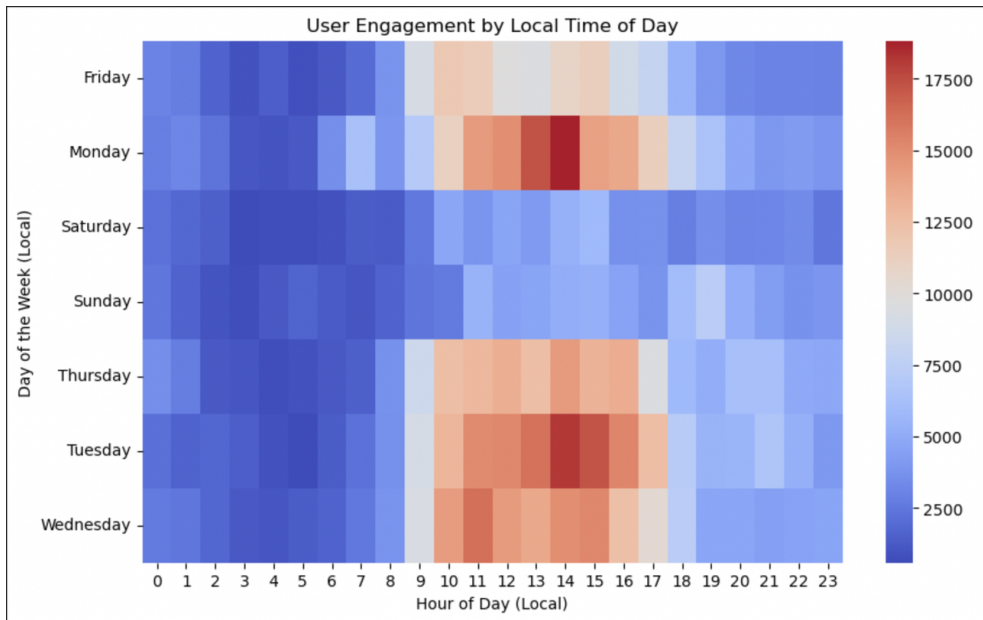
The **more engagement (average number of events)** the user shows, the more possible he or her will **subscribe**. We could do further study on peaks to check what happened on the website during that time, for example launching new functions.



### 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**.





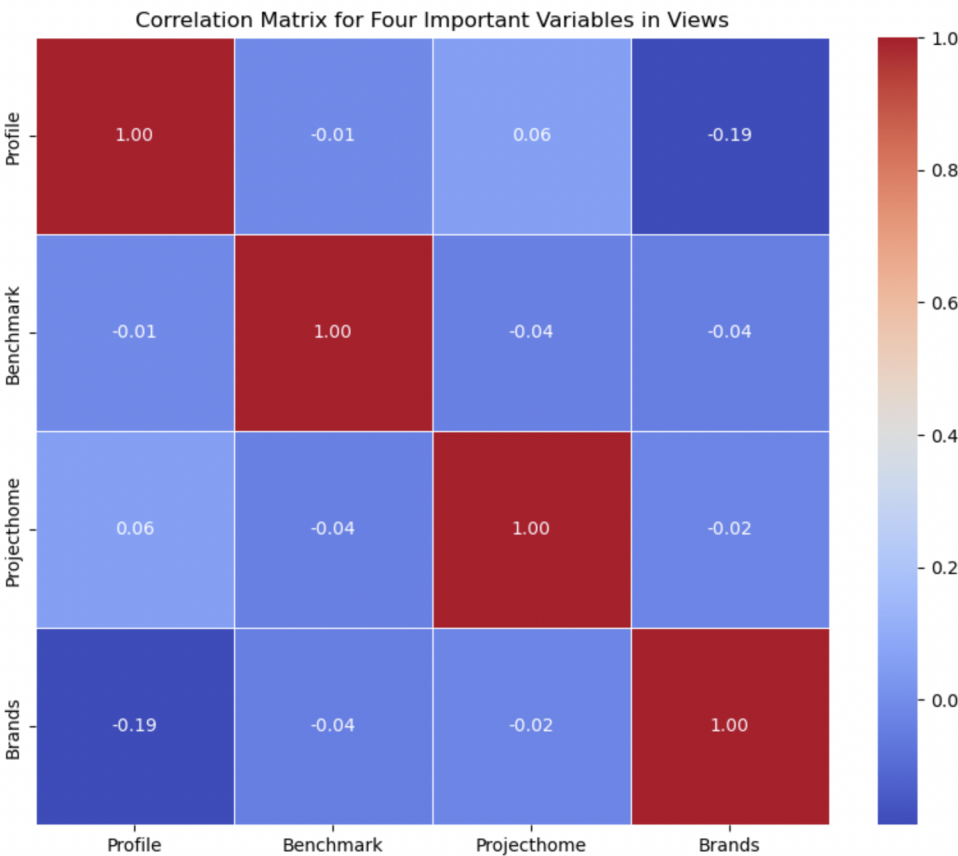
Correlation Matrix of Profile, Benchmark, Projecthome and Brands under View

We noticed that there is obvious difference between these four kinds of views between subscribers and non-subscribers.



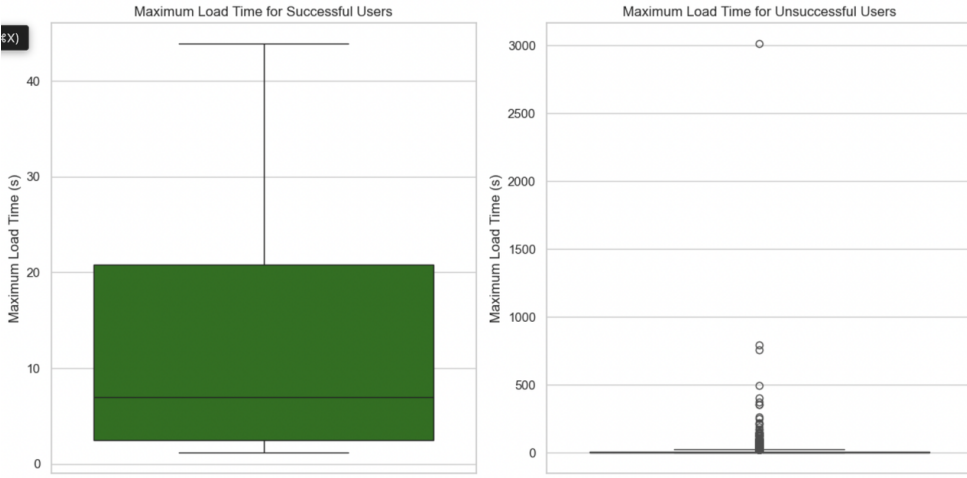
**No linearity** between these four. We may use them as factors according to models' results

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



Boxplots of Maximum Load Time for Subscribers and Non-subscribers

For the plot of subscribers, most of the data falls within a relatively narrow range, with a median near the lower quartile. No outlier. For the plot of non-subscribers, it shows a



much wider range of load times, with a significant number of high outliers. The median is higher than that of successful users, and the spread of data points is broader.



#### **Subscribers**

generally  
experience  
**quicker load  
times**

## 4. Next Steps

### **Student Team**

- ☐ Revise data visualizations and data pipeline
- ☐ Start building the model

### **Socialinsider**

- ☐ Social Insider Events Q&A Spreadsheet