YEAR 2024

SOFTWARE USER ANALYSIS

PREPARED BY:
TRAN XUAN TIEN

INTRODUCTION

In today's dynamic digital landscape, Software as a Service (SaaS) has emerged as a pivotal model for delivering software solutions to users worldwide. Our company is at the forefront of this paradigm shift, offering innovative SaaS solutions designed to streamline workflows, enhance productivity, and empower businesses of all sizes.

Understanding the importance of providing a seamless user experience, we provide prospective customers with the opportunity to trial our products before making a commitment. However, to ensure fair usage and maintain the integrity of our services, heavier workloads trigger a prompt for users to transition to a paid subscription model.

To efficiently manage user interactions and transactions, we meticulously record these events in our database, facilitating insights into usage patterns, subscription conversions, and user engagement metrics. Additionally, to maintain transparency and accountability, user information, including login activity and account statuses, are securely stored in our user_info table.

In this report, we delve into the intricate interplay between user events, subscription transitions, and account management within our SaaS ecosystem. By analyzing these datasets, we aim to extract valuable insights that inform strategic decision-making, drive product enhancements, and ultimately foster sustained growth and customer satisfaction.

CURRENT DATA

Our dataset comprises three primary tables: user_info, events, and geography.

events: Captures a detailed log of user interactions, including session data and time per session. Each entry contains event timestamps, user IDs, and feature usage information, enabling comprehensive analysis of user behavior and engagement patterns. This table only contrain events occured in 2 months January and Febuary of 2023.

id	user_id	date	datetime	platform	volume	fee
2673	608	2023-01-01	2023-01-01 17:55:00 UTC	web	10.67715783461...	0.008548565
2674	608	2023-01-01	2023-01-01 09:23:00 UTC	web	8.161172104975...	0.006534165
2675	608	2023-01-01	2023-01-01 19:25:00 UTC	web	4.129510288990...	0.003306253
2676	608	2023-01-01	2023-01-01 23:56:00 UTC	web	4.854240296068...	0.003886501
6875	1695	2023-01-01	2023-01-01 14:24:00 UTC	android	0.379275240848...	0.00030342

user_info: This table stores essential user data, including user IDs, platform (web, mobile,..), city, operating system (macOS, Win10, iOS13, etc.), country code, and features used. It serves as a central repository for user profiles, facilitating personalized experiences and targeted feature analysis. Additionally, it records last login, feature, session, time per session summary only if the account is suspended, otherwise it will be null.

id	user_id	platform	city	os	mp_country_code	created_date	last_login	feature	session	time_per_session
10	11	mobile	null	Android	BD	2022-09-25	null	null	null	null
18	15	mobile	null	Android	ID	2022-12-13	null	null	null	null
62	28	mobile	null	Android	PK	2022-12-08	null	null	null	null
64	28	mobile	null	Android	NL	2022-12-10	null	null	null	null
85	39	mobile	null	Android	BD	2022-09-21	null	null	null	null

geography: Stores geographical data related to user locations, such as city and country code. This table complements user_info by providing additional context for user demographics and geographical distribution.

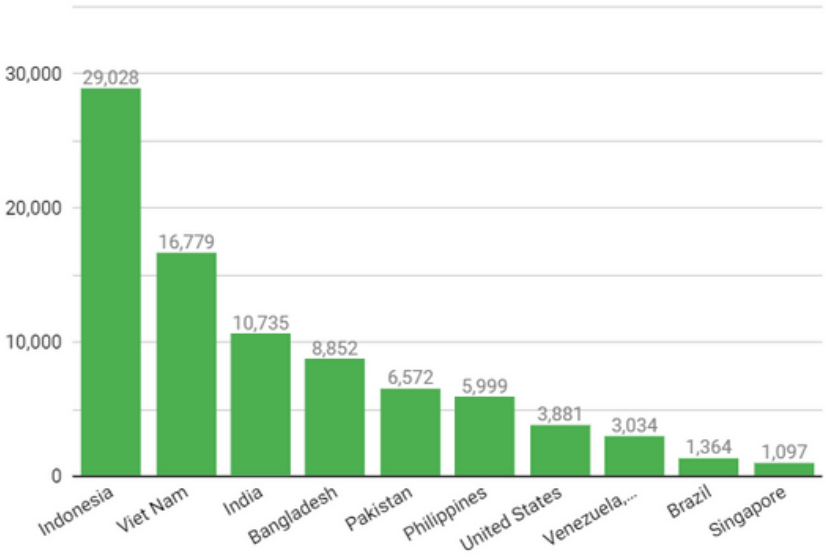
Country	Code	Latitude	Longitude
Congo, the Democratic Republi...	CD	0.0	25.0
Equatorial Guinea	GQ	2.0	10.0
Ecuador	EC	-2.0	-77.5
Rwanda	RW	-2.0	30.0
Malaysia	MY	2.5	112.5

Together, these tables provide a comprehensive view of user behavior, platform usage, and geographical trends within our SaaS platform, enabling data-driven decision-making and optimization efforts.

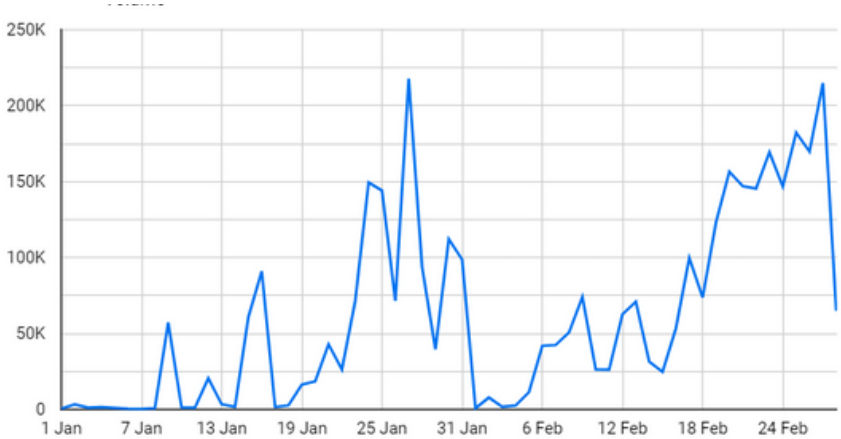
OVERVIEW ANALYTICS

Based on the information provided, the first graph illustrates the distribution of registered users across the top 10 countries with the largest customer base. The graph indicates that the top 3 countries with the highest number of users are Indonesia, Vietnam, and India. Specifically, Indonesia has **29,000** users, Vietnam has **16,000** users, and India has **10,000** users.

These top 3 countries collectively account for more than **56%** of all users of the company, indicating that they are the primary contributors to the customer base. This information highlights the significance of these countries in terms of user engagement and market presence for the company's services.

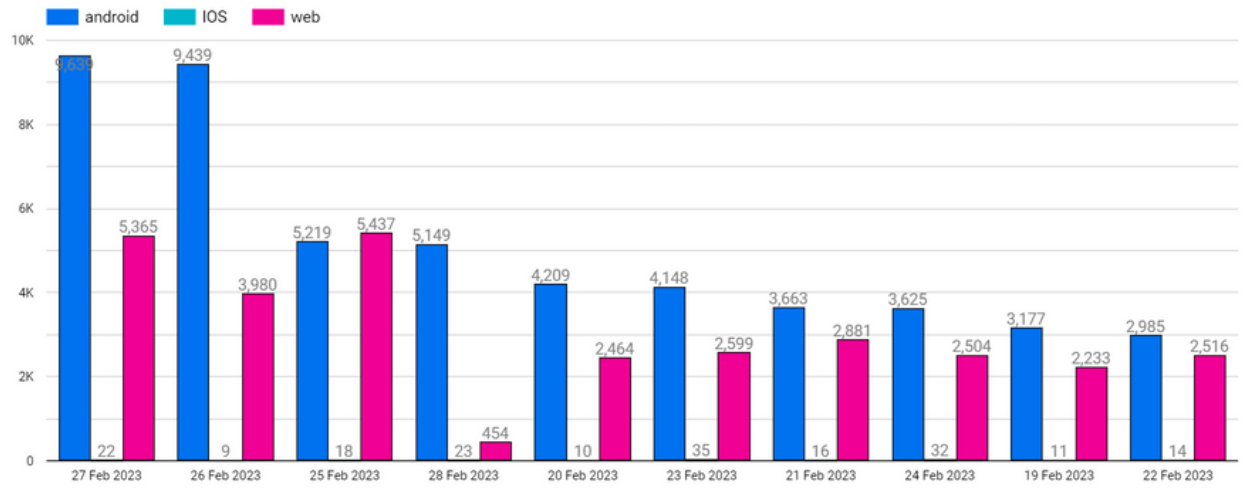


The second chart displays the usage patterns of our software over the duration of data collection. It reveals that the highest usage occurs between the **25th and 28th of each month**, with usage peaking at over **200,000 units**. Conversely, the lowest usage is observed between the **6th and 10th of each month**, with usage ranging from just a few hundred units to less than **50,000 units per day**. This suggests a trend where our software is predominantly utilized towards the end of each month. Based on this observation, we can prioritize optimizing the performance of our software during these peak periods.



OVERVIEW ANALYTICS

The final graph in this section depicts the daily user activity across three distinct platforms: iOS, web, and Android. A clear trend emerges, indicating that the majority of our users prefer the Android operating system, while the number of iOS users is considerably smaller. This observation suggests that our application is primarily favored by Android users, while there exists a significant untapped user base among iOS users. Consequently, our next developmental focus should prioritize optimizing our software to better support iOS devices, thereby expanding our user base and attracting new users from this platform.



USER RETENTION RATE

In this section, we compute the user retention rate by examining whether accounts are suspended after 7 days since registration. The process involves designating the registration date as day 0 and tallying the number of accounts created on that day. Subsequently, starting from day 0, we return after intervals of 7 days to assess if the accounts registered on that day have been suspended. The retention rate is calculated by dividing the remaining number of accounts by the number created on day 0. This process is iterated for each day, with each day serving as day 0 and having its own set of registered accounts. We evaluate the retention rate 7 times, corresponding to intervals of 7 days, with each evaluation day denoted as day 1, day 2, day 3, and so forth, up to day 7.

created_date	day0_retention	day1_retention	day2_retention	day3_retention	day4_retention	day5_retention	day6_retention	day7_retention
01/09/2023	100%	3.47%	3.47%	3.47%	3.47%	3.47%	3.47%	3.47%
01/10/2023	100%	3.98%	3.98%	3.98%	3.98%	3.98%	3.98%	3.98%
01/11/2023	100%	5.04%	5.04%	5.04%	5.04%	5.04%	5.04%	0%
01/12/2023	100%	2.23%	2.23%	2.23%	2.23%	2.23%	2.23%	0%
01/13/2023	100%	2.99%	2.99%	2.99%	2.99%	2.99%	2.99%	0%
01/14/2023	100%	2.51%	2.51%	2.51%	2.51%	2.51%	2.51%	0%
01/15/2023	100%	6.16%	6.16%	6.16%	6.16%	6.16%	6.16%	0%
01/16/2023	100%	3.77%	3.77%	3.77%	3.77%	3.77%	3.77%	0%
01/17/2023	100%	4.59%	4.59%	4.59%	4.59%	4.59%	4.59%	0%
01/18/2023	100%	1.77%	1.77%	1.77%	1.77%	1.77%	0%	0%
01/19/2023	100%	5.12%	5.12%	5.12%	5.12%	5.12%	0%	0%

created_date	day0_retention	day1_retention	day2_retention	day3_retention	day4_retention	day5_retention	day6_retention	day7_retention
10/03/2022	100%	100%	100%	100%	100%	100%	100%	100%
10/04/2022	100%	100%	100%	100%	100%	100%	100%	100%
10/05/2022	100%	100%	100%	100%	100%	100%	100%	100%
10/06/2022	100%	100%	100%	100%	100%	100%	100%	100%
10/07/2022	100%	100%	100%	100%	100%	100%	100%	100%
10/08/2022	100%	100%	100%	100%	100%	100%	100%	100%
10/09/2022	100%	100%	100%	100%	100%	100%	100%	100%
10/10/2022	100%	100%	100%	100%	100%	100%	100%	100%
10/11/2022	100%	100%	100%	100%	100%	100%	100%	100%
10/12/2022	100%	100%	100%	100%	100%	100%	100%	100%
10/13/2022	100%	100%	100%	100%	100%	100%	100%	100%

I randomly selected 10 days from both 2022 and 2023 to compare the retention rates of users registered during these time periods. Notably, it was observed that **100%** of users registered in the 10 selected days of **2022** continued to use our software after 7 checking intervals (49 days). However, the retention rates for users registered in **2023** were markedly lower, **ranging from 2% to 5%**.

This substantial disparity in retention rates between users from the two years prompts us to examine the quality of our software. It raises questions about whether our software is up to date compared to our competitors, if there are any new features offered by competitors that we lack, or if our billing policy is less satisfactory than that of our competitors. These considerations underscore the importance of assessing and potentially improving various aspects of our software and services to enhance user retention and competitiveness in the market. Additionally, it leads us to question whether users are intentionally suspending their accounts to exploit our trial policy. The next section will delve into the topic of identifying and preventing cheating users, including strategies to identify the most serious offenders and measures to mitigate their impact.

IDENTIFYING SERIOUS CHEATING INCIDENTS

In this section, we'll discuss how we detect individuals who exploit glitches by repeatedly suspending their accounts and creating new ones after using up their trial usage quota. This behavior results in one user having multiple account histories. We'll also explain why it's important to identify and address this type of cheating.

One straightforward method to identify potential exploitation of our trial program is by tracking instances where users unsubscribe or suspend their accounts within the initial 3 days following registration. Such actions indicate an effort to circumvent payment for official subscriptions by prematurely discontinuing the trial. To calculate these indicators, we'll measure the date difference between the created_date and last_login for user IDs that have been suspended. This will provide valuable insights into potential misuse of our trial program.

id	user_id	mp_country_code	created_date	last_login	os	timediff
15109	5177	IN	2023-01-01	2023-01-01	Android	0
19451	6563	ID	2023-01-01	2023-01-01	Android	0
32558	10760	ID	2023-01-01	2023-01-01	Android	0
32660	10778	BD	2023-01-01	2023-01-01	Android	0
32935	10868	MY	2023-01-01	2023-01-01	Windows	0
37609	12518	IN	2023-01-01	2023-01-01	Android	0

In the provided table, we've computed the time difference between the **created_date** and **last_login** dates of accounts that have been suspended. Our next step is to identify accounts where this time difference is **less than 3 days**. Such instances indicate users who attempted to avoid paying for subscriptions by discontinuing their trial prematurely.

Additionally, we can identify potential cheaters by examining the **user's region and device**. If the same user ID is assigned to a user in the same country with the same device type (Android, Apple, web) multiple times, there is a high likelihood that it is being re-granted to the same individual attempting to cheat our program.

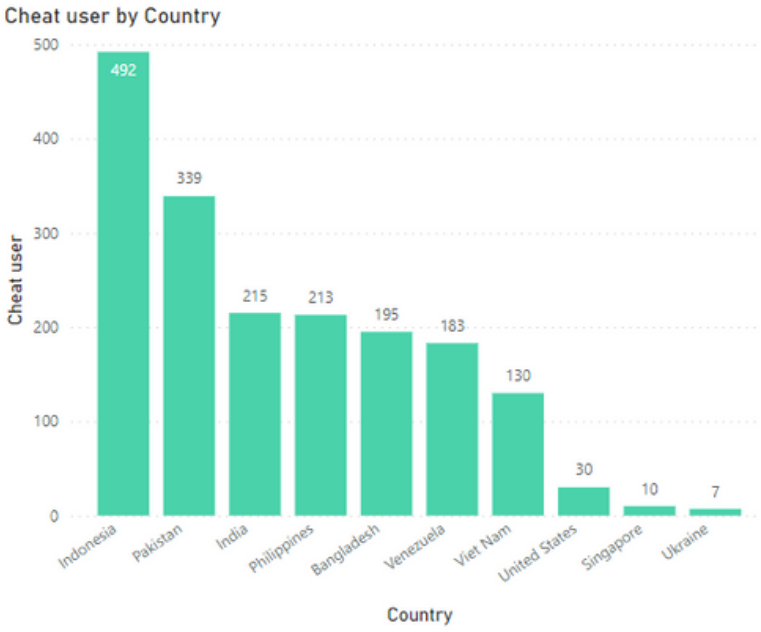
To effectively indentify the serious cheaters and protect the integrity of our trial program, we're taking steps to identify the most serious cheaters. Our algorithm flags users who repeatedly unsubscribe or suspend their accounts within the first 3 days after registration, then proceed to create new accounts with the same device in the same region. If a user exhibits this behavior more than 5 times, they are marked as serious cheaters.

IDENTIFYING SERIOUS CHEATING INCIDENTS

user_id	mp_country_code	os	number_create
6563	ID	Android	36
10760	ID	Android	16
10778	BD	Android	11
13289	ID	Android	9
14497	ID	Android	49
14987	ID	Android	19
15472	PK	Android	24
17284	ID	Android	25
18021	ID	Android	6

In the table above, we've compiled a summary of user IDs that have prematurely ended their subscriptions before the trial interval concludes, and have been re-granted more than 5 times to the same region and devices. The "number_created" column indicates the frequency of re-grants for each user ID. These findings highlight instances where users may be attempting to exploit our trial program through repeated account suspensions and re-creations, warranting further investigation and potential action.

Next, we will proceed to count the number of flagged user IDs in each region and identify which region poses the highest risk of exploiting our system. Based on the results of this analysis, we can develop tailored policies to address the specific challenges posed by regions with a high risk of cheating users. By implementing special measures for these unique regions, we can effectively mitigate the risks associated with fraudulent activity and safeguard the integrity of our trial program.



This chart illustrates the number of user IDs flagged as having a high risk of cheating in each country. Indonesia and Pakistan emerge as the top two countries with the highest number of cheaters, totaling 492 and 339 respectively. Additionally, India, the Philippines, Bangladesh, and Venezuela each have approximately 200 cheaters.

NEW CALCULATING METHOD

In this chapter, we address the issue of certain users exploiting our app by excessively accessing the service. To tackle this problem, our managers have implemented a recalibration of access time. Specifically, if a user re-accesses the service within a 6-hour window, it will still be counted as a single access event, regardless of the actual number of accesses made. However, the volumes used by customers will continue to be calculated normally. This recalibration aims to provide a fair and accurate representation of user access patterns while mitigating the impact of excessive usage on our system.

Original Calculation:

id	user_id	date	datetime	platform	volume	fee
0	4	2023-01-29	2023-01-29 23:17:00 UTC	web	27.56083379560...	0.0
1	4	2023-01-29	2023-01-29 06:46:00 UTC	web	27.59173313921...	0.0
2	4	2023-01-29	2023-01-29 05:20:00 UTC	web	27.57637711699...	0.0
3	4	2023-01-29	2023-01-29 17:03:00 UTC	web	27.59614718833...	0.0
4	4	2023-01-29	2023-01-29 07:48:00 UTC	web	0.931729129240...	0.0
5	4	2023-01-29	2023-01-29 07:15:00 UTC	web	27.57849299143...	0.0
6	4	2023-01-29	2023-01-29 11:32:00 UTC	web	27.57407594296...	0.0
7	4	2023-01-29	2023-01-29 14:03:00 UTC	web	27.61173985474...	0.0
8	4	2023-01-29	2023-01-29 17:15:00 UTC	web	27.60056323702...	0.0

New Calculation:

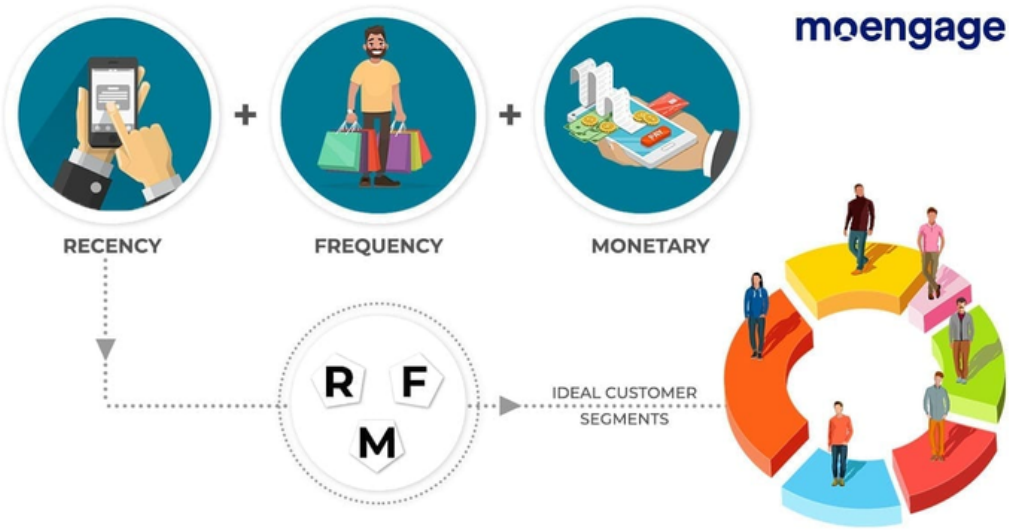
id	user_id	date	datetime	platform	volume	fee
1	4	2023-01-29	2023-01-29 00:19:00 UTC	web	137.8907440202...	0.0
2	4	2023-01-29	2023-01-29 06:46:00 UTC	web	221.6284192917...	0.0
3	4	2023-01-29	2023-01-29 14:03:00 UTC	web	165.5418132231...	0.0
4	4	2023-01-29	2023-01-29 22:07:00 UTC	web	55.16371121484...	0.0
5	4	2023-02-22	2023-02-22 08:02:00 UTC	web	358.9788637838...	0.0
6	4	2023-02-22	2023-02-22 14:22:00 UTC	web	307.7589502589...	0.0
7	4	2023-02-22	2023-02-22 21:53:00 UTC	web	51.31465993574...	0.0
8	4	2023-02-23	2023-02-23 00:45:00 UTC	web	314.9554925837...	0.0

Expanding the access event interval to 6 hours will likely decrease the number of access events attributed to cheaters. With only 12 potential access events over a 3-day period, the frequency of access for these users will appear lower compared to genuine users who adhere to normal usage patterns. This adjustment enhances fairness in customer segmentation by ensuring that users who engage in genuine, sustained usage are not overshadowed by those who artificially inflate their access metrics through frequent, but brief, interactions with the system.

By minimizing the number of access events per day, we can better differentiate between users who utilize the system for prolonged periods versus those who simply access it frequently but briefly. This helps to accurately assess the level of user engagement and prioritize resources and support accordingly. Ultimately, this approach fosters a more equitable and transparent evaluation of user activity, benefiting both genuine users and the overall integrity of our subscription model.

CUSTOMER SEGMENTATION

RFM analytics introduction



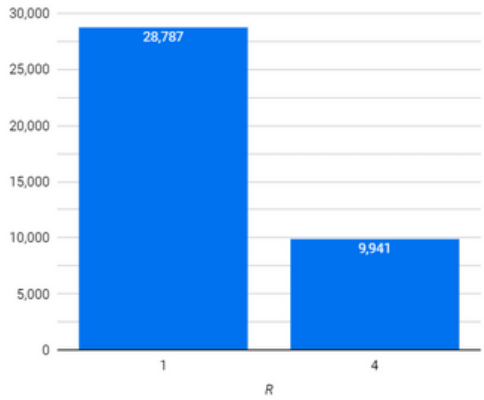
- **Recency** is the score calculated based on how long is it since the last time the customer use our software
- **Frequency** is determined by how frequently a user using our software on average. To calculate this, we sum the access numbers of a user and devided by his/her total subcription interval . If the account subcription duration is ongoing and hasn't ended, we use the date of the report's creation, in this case, **2023-04-01**, to calculate the contract's length.
- **Monetary** calculated by sum up the fee charged by the system based on the features, and time they used on our software and then divide this sum by the length of their subcription interval. Same as Frequency, if the subcription is still under effects, we utilize the report's date of creation (**2023-04-01**) to determine the contract's length.

We score the value of R,F,M by deviding the customer base into 4 range using interquartile range:

- Range 1: Min Value - First Quartile
- Range 2 :First Quartile (Q1) - Median
- Range 3: Median - Third Quartile (Q3)
- Range 4: Third Quartile (Q3) - Max Value

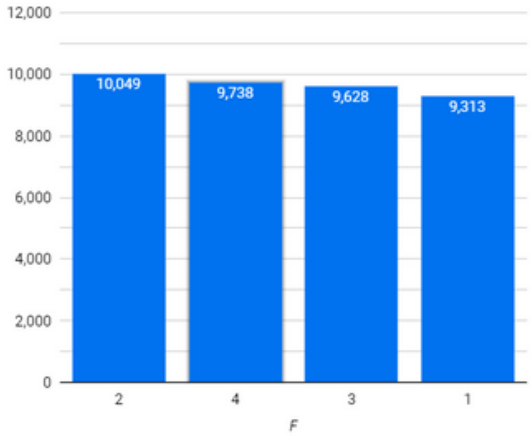
If the values of R,F,M of customers fall within each of these range, the R,F,M scores will be assign following the range they belong to. For example: 111,123,444,132,444,... The first number is Recency score, second is Frequency and then Monetary.

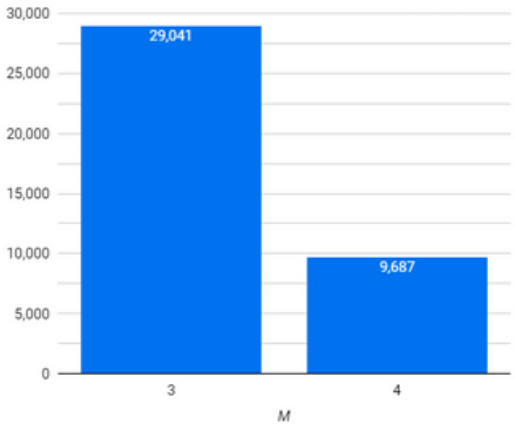
CUSTOMER SEGMENTATION



The chart illustrates our customer base across two main Recency groups: Group 1, representing users who haven't used our software for an extended period, and Group 4, consisting of recent users. **Group 1** dominates with over **28,000** users, which is approximately three times the number of users in **Group 4**, totaling around **10,000**. This significant disparity highlights a substantial attrition trend, with over **70% of users inactive** for a considerable time. Understanding this trend is crucial for devising strategies to re-engage dormant users and ensure the sustained growth of our software platform.

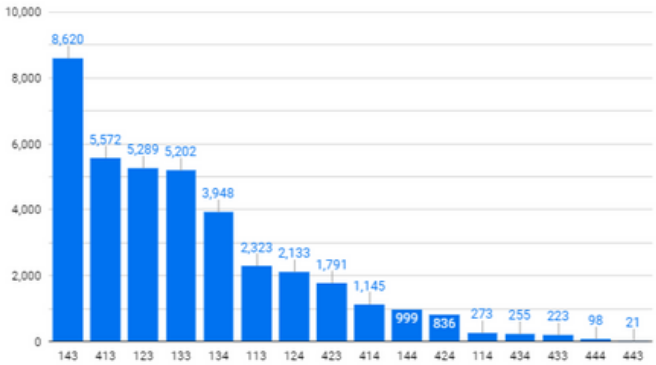
The chart depicts the distribution of users across four frequency groups, where Group 1 represents users with minimal access and Group 4 represents those with the highest access frequency. Notably, this distribution pattern showcases a striking **equalization among the populations of these groups**. The number of users in each group **ranges from 9,300 to 10,000**, with no significant gaps observed. This balanced distribution suggests a relatively **uniform engagement level** among users across different access frequency groups.





Regarding the distribution of customers in the Monetary group, we observe a **similar trend to the Recency groups**, with only two distinct groups: 1 and 4. Group 1 represents users who contribute smaller fees to our revenue, while Group 4 comprises users who make larger contributions. Interestingly, approximately **70% of users** fall into **Group 1**, while the **remaining users** are classified into **Group 4**. This distribution underscores the significant disparity in contribution levels among our user base, with a majority of users generating lower fees compared to a smaller subset contributing substantially higher amounts.

Combining the Recency, Frequency, and Monetary clustering of customers, we have derived RFM groups. The population of each RFM group is depicted in the bar chart alongside. Notably, approximately **5-6 RFM groups dominate 70%** of the customer base. Leading the pack is **Group 143**, comprising o**ver 8,000 users**, followed closely by **Groups 413, 123, and 133**, each with nearly **5,300 users**. This distribution reveals several RFM groups with similar characteristics, indicating a need to consolidate these groups into a segmentation. The algorithm for segmenting the customer base is presented in the following pages.



CUSTOMER SEGMENTATION

The customer segmentation method is developed based on BGC matrix below

Now we apply the dimensions of RFM analytics into this matrix to segment the customer base. Firstly, we will just apply Recency to the vertical axis and Frequency to the horizontal axis of the matrix.

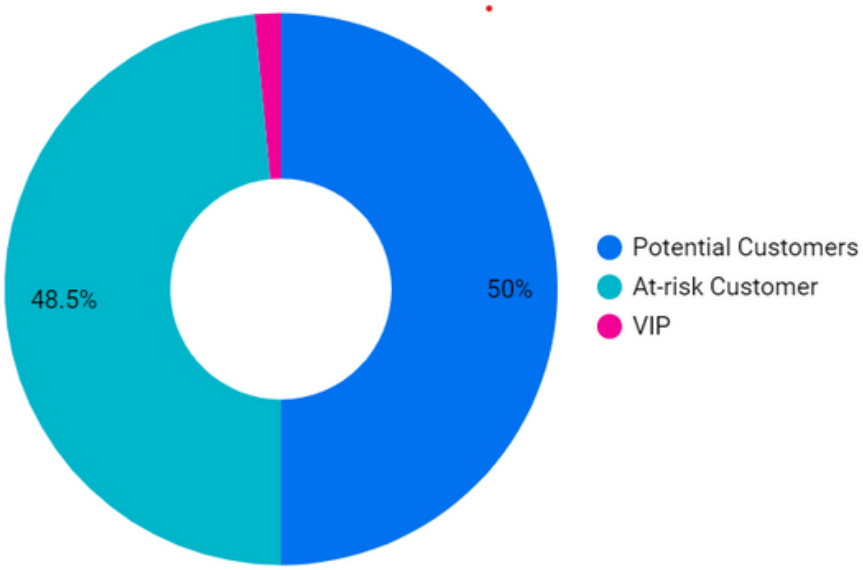
Then we can devide the customer base into 4 quarters:

- Low Recency (R = 1 - 2) & Low Frequency (F = 1 - 2) (Dog): Walk-in guests
- Low Recency (R = 1 - 2) & High Frequency (F = 3 - 4) (Cash Cow): At-risk Customers
- High Recency (R = 3 - 4) & Low Frequency (F = 1 - 2) (Questrion mark): New Customers
- Hight Recency (R = 3 - 4) & High Frequency (F = 3 - 4) (Star): Regular customers

Then we apply the third dimension which is Monetary into our current result:

- Walk-in guest & Low Monetary (M = 1 - 2): **Walk-in guests**
- Walk-in guest & High monerary (M = 3 - 4): **Potential Customer**
- At-risk customer & Any monetary score (M = 1 - 2 - 3 - 4): **At-risk Customers**
- New Customer & Low Monetary (M = 1 - 2): **Walk-in guests**
- New Customer & High Monetary (M = 3 - 4): **Potential Customer**
- Regular Customers & Low Monetary (M = 1 - 2): **Regular Customers**
- Regular Customers & High Monetary (M = 3 - 4): **VIP Customers**

CUSTOMER SEGMENTATION



The donut chart above provides an overview of our company's customer base, segmented into three categories: VIP customers, At-risk customers, and Potential customers. Notably, Potential and At-risk customers together make up approximately 98.5% of our customer base, with VIP customers representing only a small minority.

A significant concern is the high number of customers categorized as 'At risk,' comprising nearly half of our customer base. This indicates a substantial risk of losing a significant portion of our customer base. Additionally, the proportion of VIP customers is alarmingly low at 1.5%.

To address these challenges, we need to conduct a thorough review of our software to identify any shortcomings compared to our competitors. Additionally, we must take proactive steps to convert Potential customers into VIP customers by offering them additional benefits and incentives. By addressing these issues, we can work towards improving customer retention and loyalty, ultimately ensuring the long-term success and sustainability of our business.