



Take Home Exercise

SR Data Analyst
Bruno Bucalon Serra

INDEX

1. Problem Definition
 - 1.1 Dataset Attributes
 - 1.2 Objective and Key Questions
 - 1.3 Technologies and Programming Language
2. First Steps
3. Data GAP
 - 3.1 Campaign ID
 - 3.2 User ID
 - 3.3 Nulls, NaNs and Missings
 - 3.4 Wrong Campaign
 - 3.5 Inconsistence in Window Number
 - 3.6 Missing PK
 - 3.7 N:N Relationship
 - 3.8 Non Attributed Channel
4. Data Analysis
 - 4.1 Valuable Users
 - 4.2 Channels & Valuable Users
 - 4.3 Conversion Rate
 - 4.4 Recommendations
5. Recommendations

1. PROBLEM DEFINITION

1. PROBLEM DEFINITION

1.1 DATASET ATTRIBUTES

Acquisitions	Visits	Campaign (DIM)	Purchases
userID	userID	SEASON	userID
ACQUISITION_DATE	SEASON	CAMPAIGN_ID	CAMPAIGN_ID
ORDER_SEASON	CAMPAIGN_ID	CAMPAIGN_TITLE	CAMPAIGN_TITLE
PLAN_CODE	WINDOW_NUMBER	CAMPAIGN_TYPE	SKU
SAME_DAY_REFUND	CAMPAIGN_TITLE		CART_QUANTITY
CURRENT_SUBSCRIPTIONS_STATE	CAMPAIGN_TYPE		MSRP
AGE	WINDOW_START_TS_PST		COGS
USER_REGION	WINDOW_END_TS_PST		SALE_PRICE
LTV14	CHANNEL		CATEGORY
	NEW_OR_RETURNING_MEMBER		
	VISIT_CNT		

1. PROBLEM DEFINITION

1.2 OBJECTIVES AND KEY QUESTIONS

Who are our most valuable members? How are they different from our less valuable members?

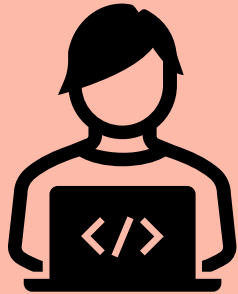
Which channels are best at driving these members to sales?

Where are purchase conversion rates strong?

Based on your analysis, provide actionable recommendations to improve the conversion rate from visits to purchases. Explain the rationale behind each recommendation.

1. PROBLEM DEFINITION

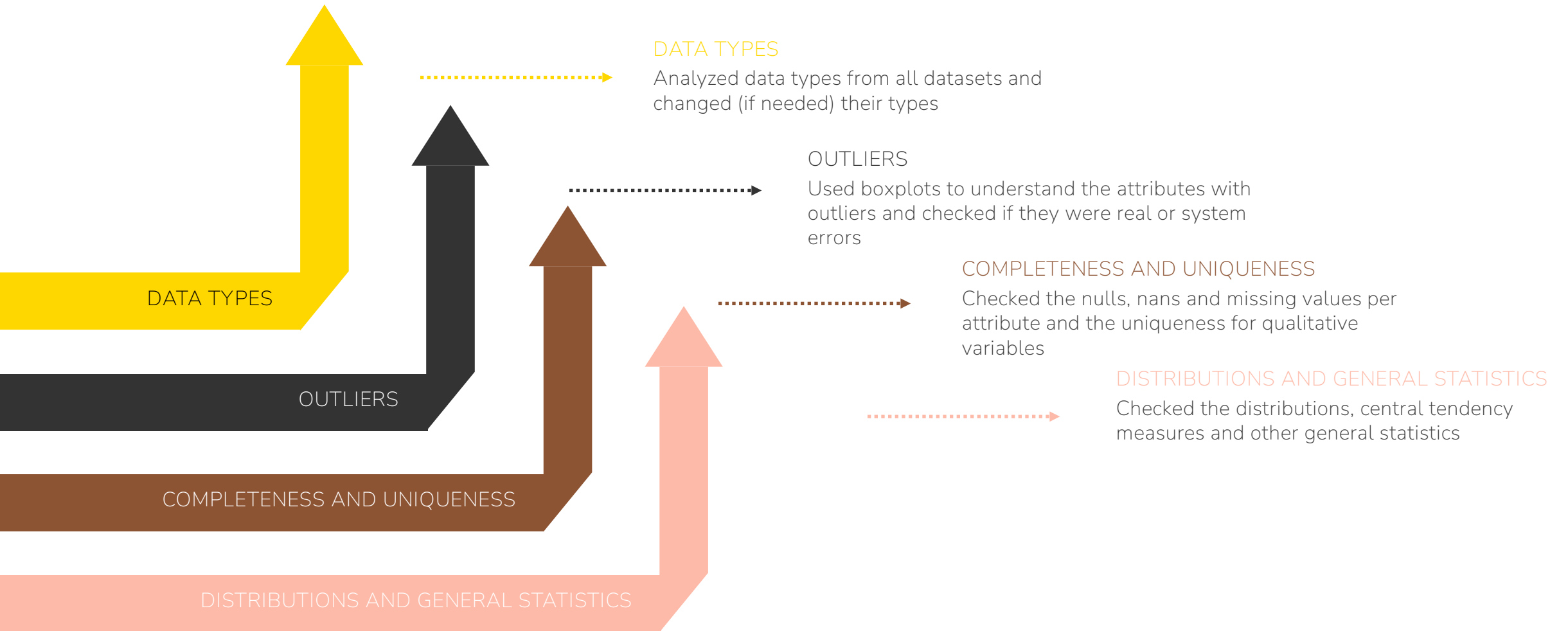
1.3 TECHNOLOGIES AND PROGRAMMING LANGUAGE



- Environment using Anaconda
- All analysis performed using Python
- Used libraries: NumPy, Pandas, Matplotlib, Sklearn
- Visual Studio Code used as IDE

2. FIRST STEPS

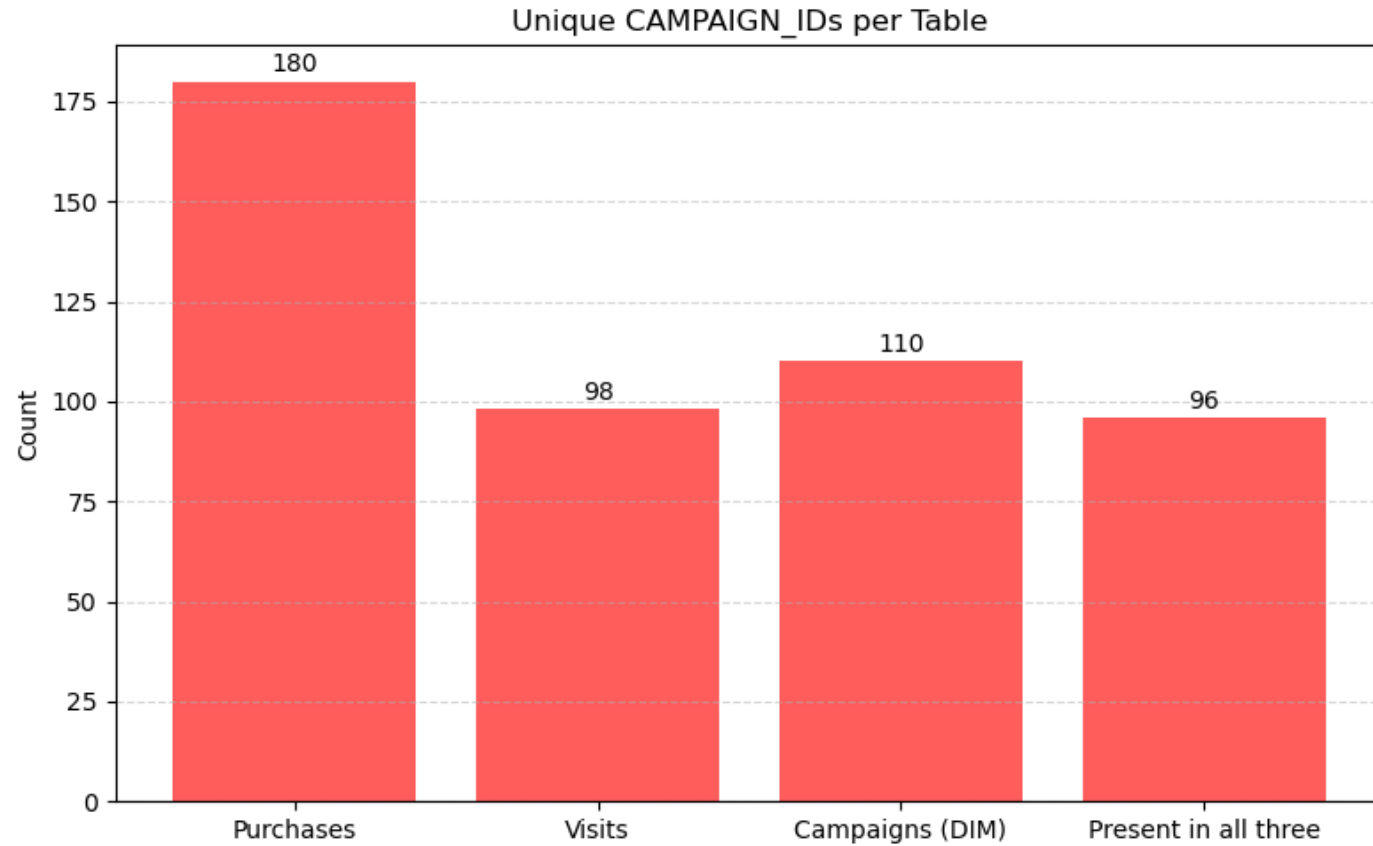
2. FIRST STEPS



3. DATA GAPS

3. DATA GAP AND INCONSISTENCIES

3.1 CAMPAIGN ID

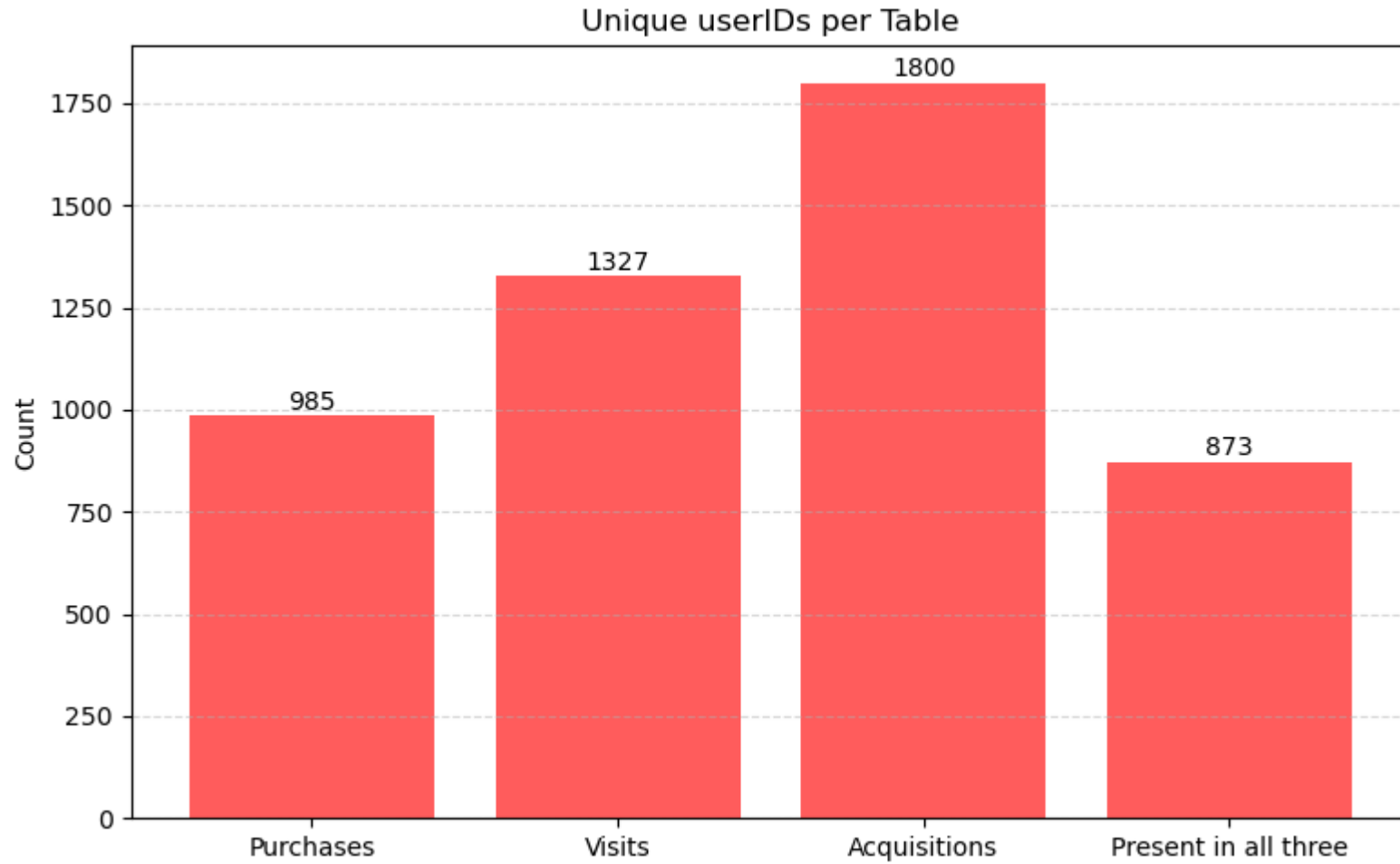


INSIGHTS

Only 96 campaigns are available in the 3 datasets.

3. DATA GAP AND INCONSISTENCIES

3.2 USER ID



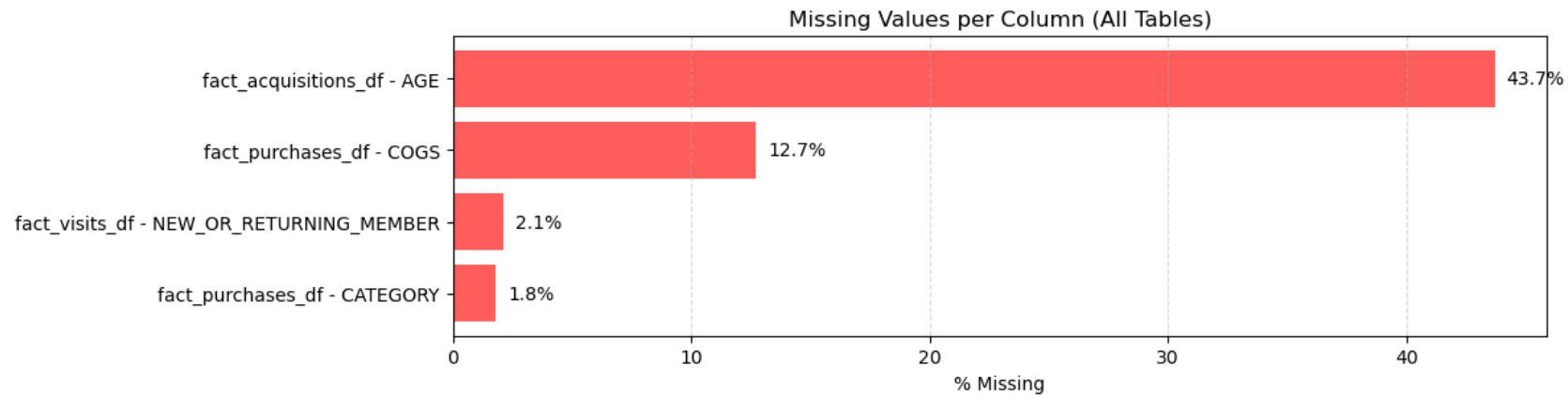
INSIGHTS

Only 873 users are available in the 4 datasets.

fabfitfun

3. DATA GAP AND INCONSISTENCIES

3.3 NULLS, NaNs and Missings



INSIGHTS

Ages has the higher quantity of Nulls, preventing its usage for deeper insights about users

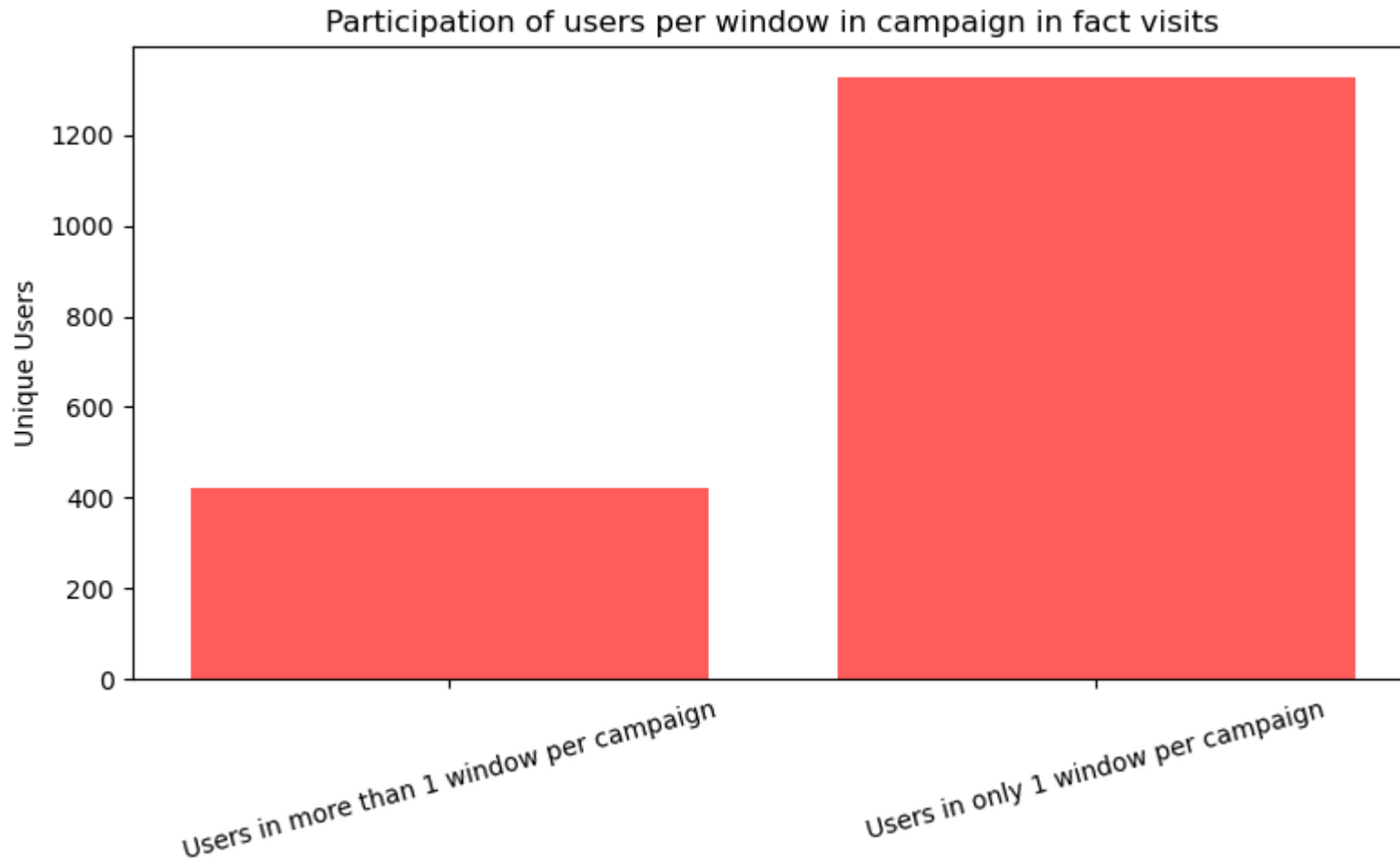
3. DATA GAP AND INCONSISTENCIES

3.4 MISSING PRIMARY KEY

Campaign 184 (12 Days of Deals 2022) has 12 different start and end dates.

3. DATA GAP AND INCONSISTENCIES

3.5 INCONSISTENCE WINDOW NUMBER



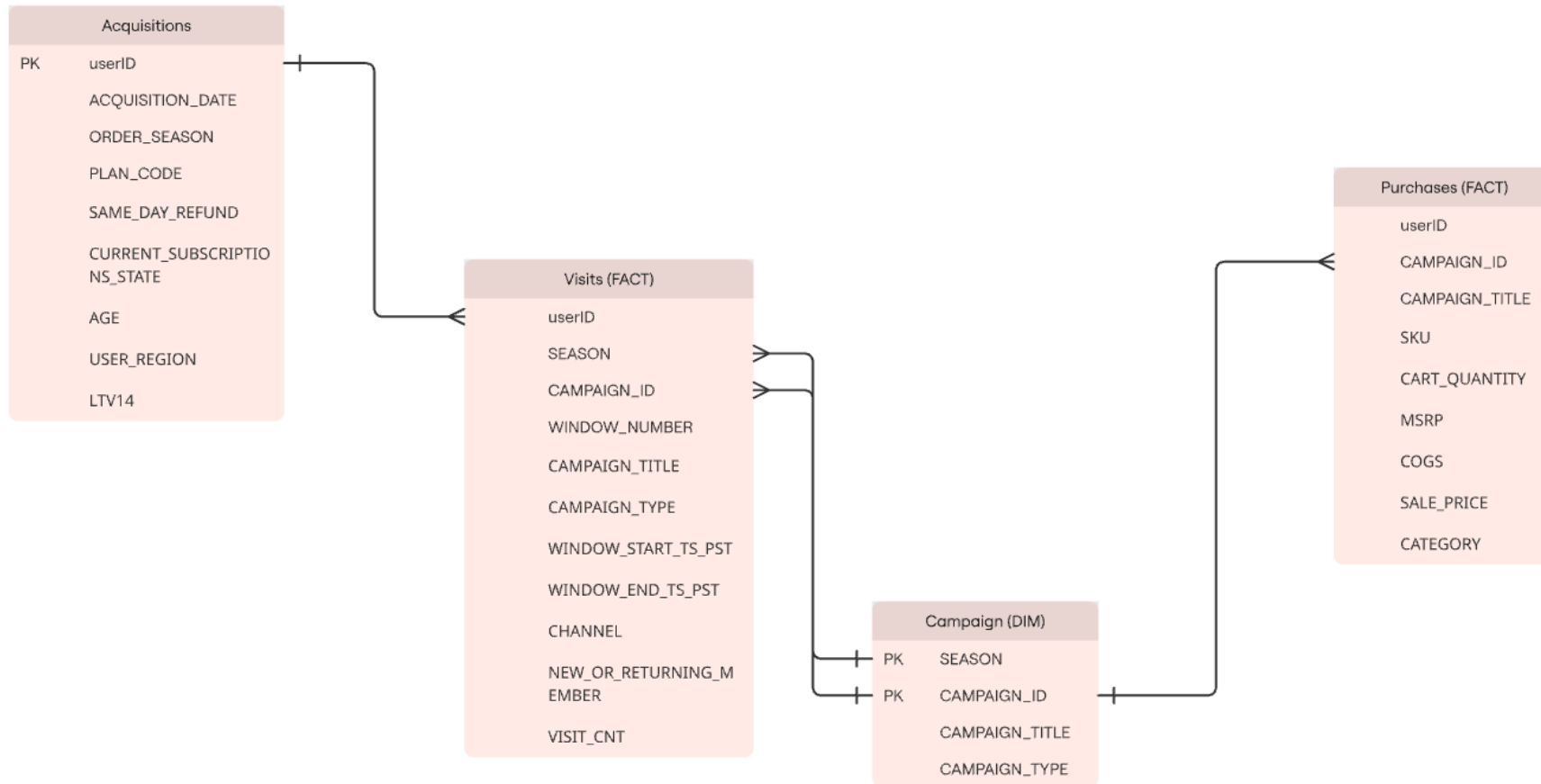
INSIGHTS

ASSUMPTION:

- Subscription plan changed during campaign
- Error in data ingestion

3. DATA GAP AND INCONSISTENCIES

3.6 MISSING PRIMARY KEY

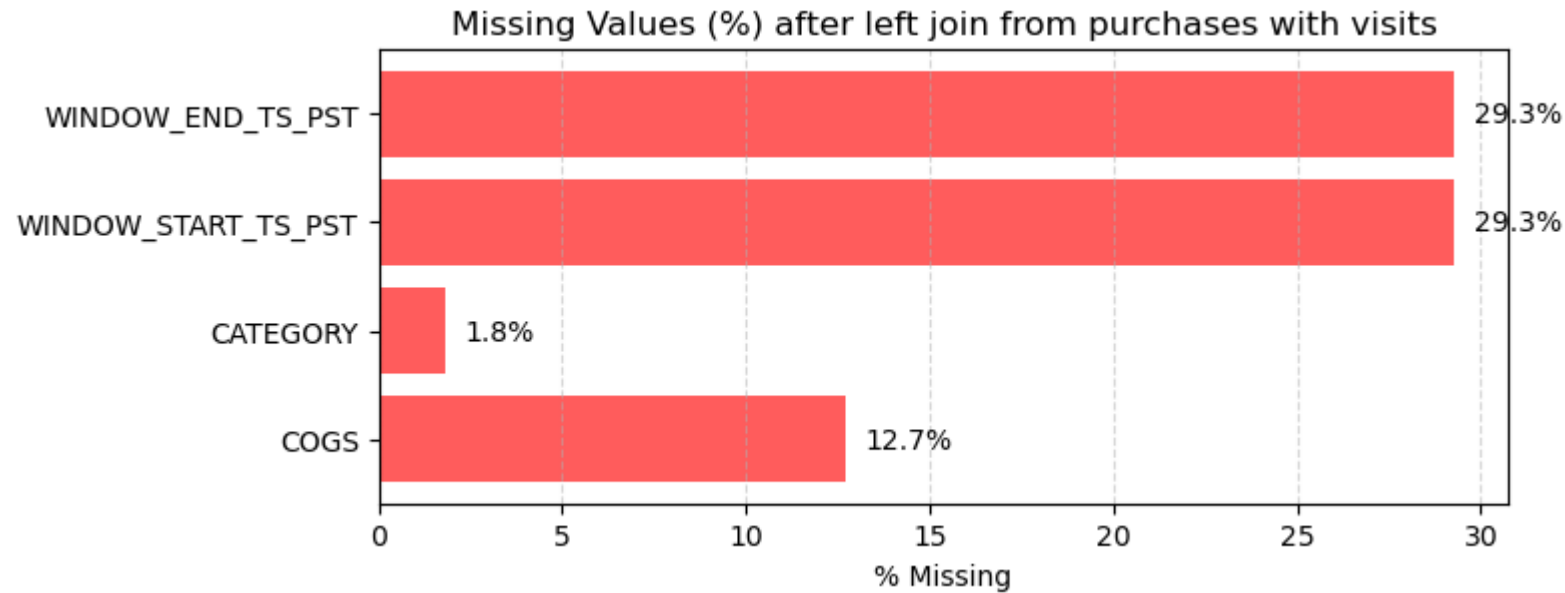


INSIGHTS

- Relationship between Visits and Purchases table is N:N
- No PK in Visits and Purchases
- There's no Window Number in Purchases, therefore there's no way to identify which window it was converted

3. DATA GAP AND INCONSISTENCIES

3.7 JOIN BETWEEN VISITS AND PURCHASES



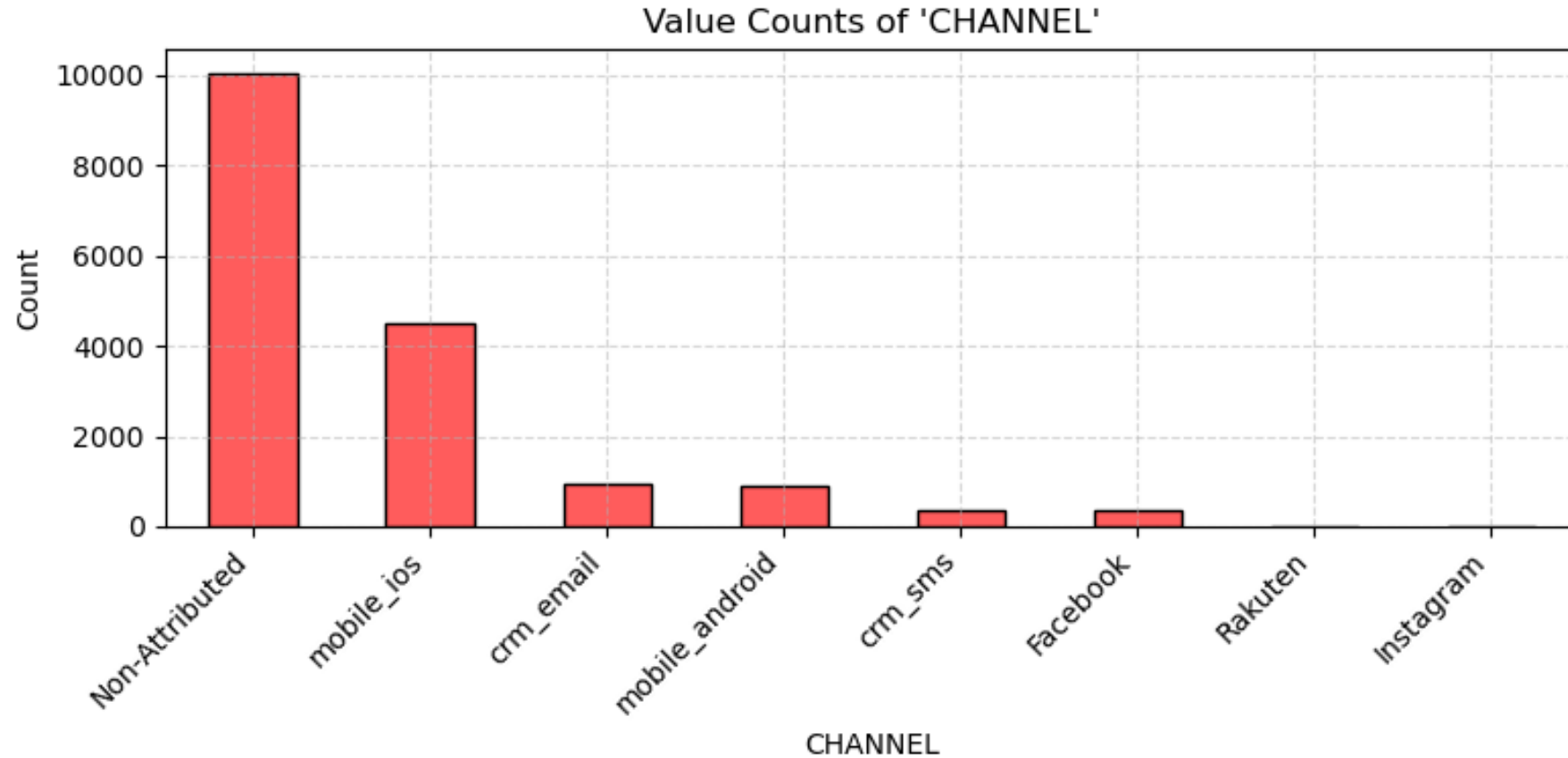
INSIGHTS

By joining visits and purchases using User ID and Campaign ID, we'll have a GAP of 29.3% of missing values for Window Start and Window End.

In addition, there's no window number in purchases.

3. DATA GAP AND INCONSISTENCIES

3.8 NON-ATTRIBUTED IN CHANNELS



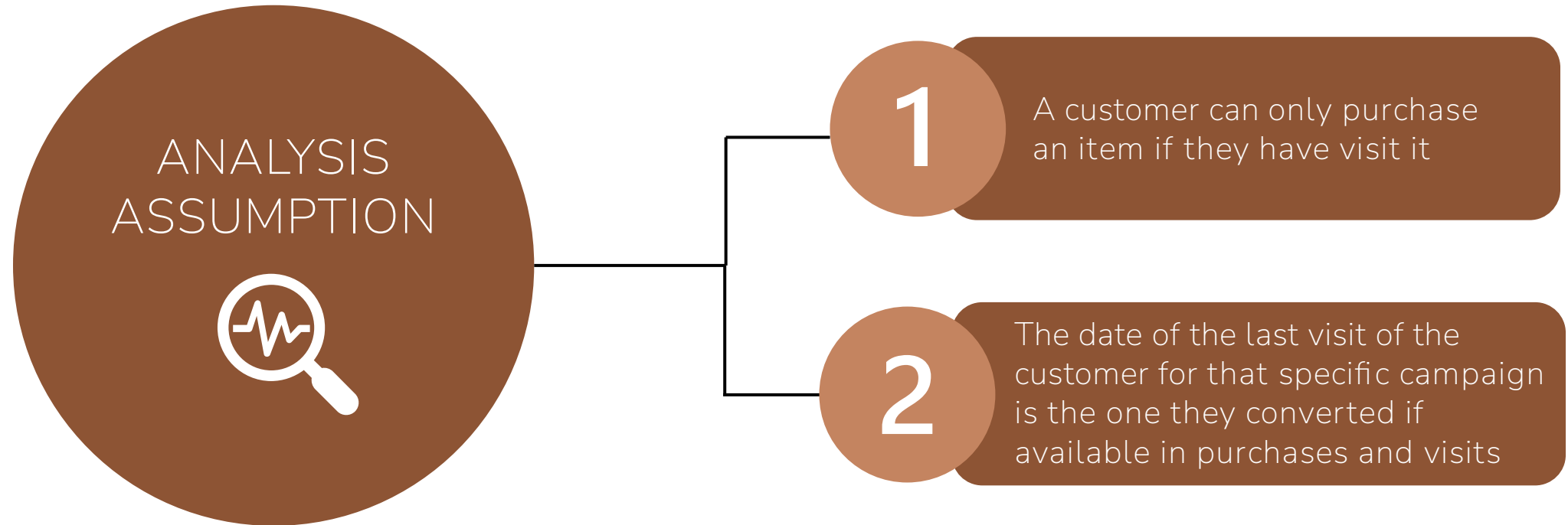
INSIGHTS

Over 10000 rows with non attributed channel.

4. DATA ANALYSIS

4. DATA ANALYSIS

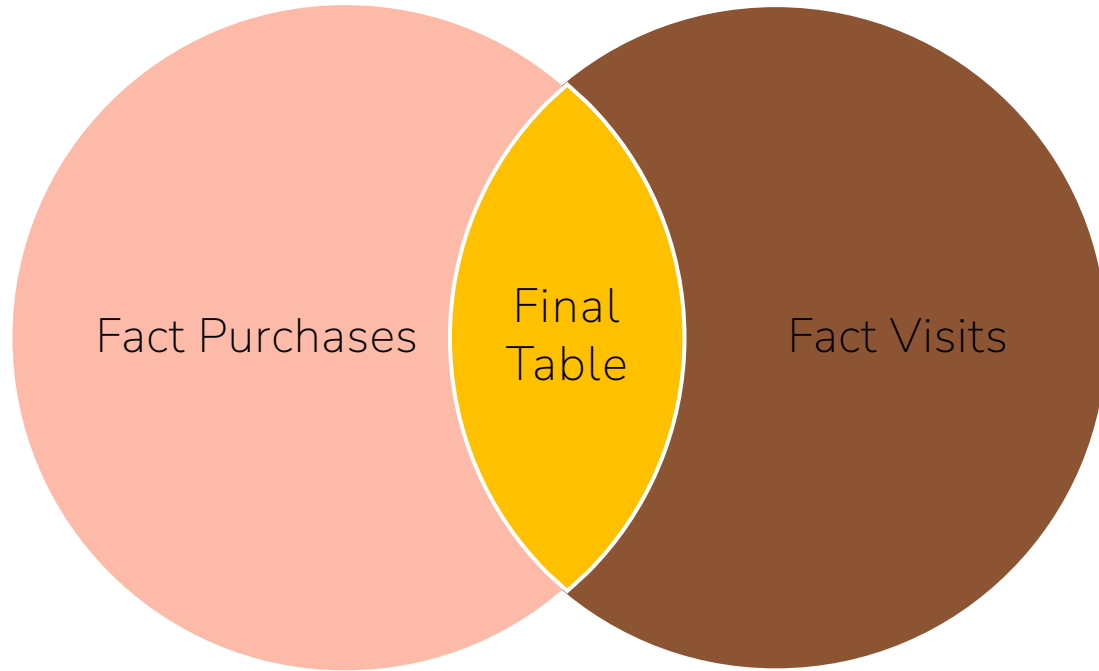
4.1 VALUABLE USERS



By performing this join, we kept 70.71% of the rows from Fact Purchases.

4. DATA ANALYSIS

4.1 VALUEABLE USERS



- Sort values by *WINDOW_END_TS_PST* and drop duplicates of *userID* and *CAMPAIGN_ID*
- Inner join Fact Purchases with Fact Visits using of *userID* and *CAMPAIGN_ID*
- Calculate *TOTAL_REVENUE* by performing $CART_QUANTITY \times SALE_PRICE$

By performing this join:

- Kept 70.71% of the rows from Fact Purchases
- Kept 77.77% of the users from Fact Purchases
- Kept 53.33% of the campaigns from Fact Purchases

4. DATA ANALYSIS

4.1 VALUABLE USERS

Recency

How recently a customer made a purchase (in days).
Used a *time window*¹ between:
2023-08-06 and 2024-08-06
per user.

Frequency

How often or how many times the customer makes a purchase.
Used a *time window*¹ between:
2023-08-06 and 2024-08-06
per user.

Monetary

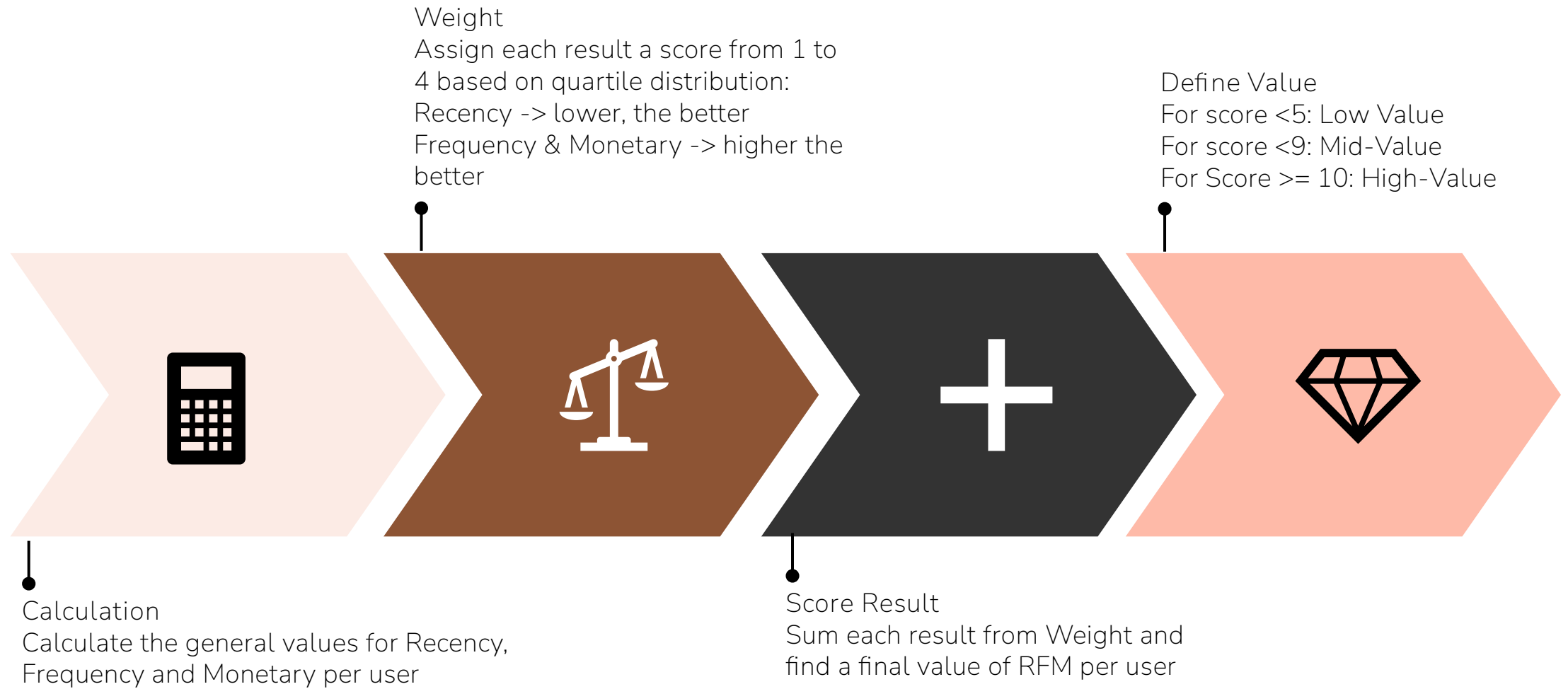
How much money the customer has spent.
Used a *time window*¹ between:
2023-08-06 and 2024-08-06
per user.

*Time window*¹:
BETWEEN
MAX(WINDOW_END_TS_PST)
AND
MAX(WINDOW_END_TS_PST) - 1 YEAR

fabfitfun

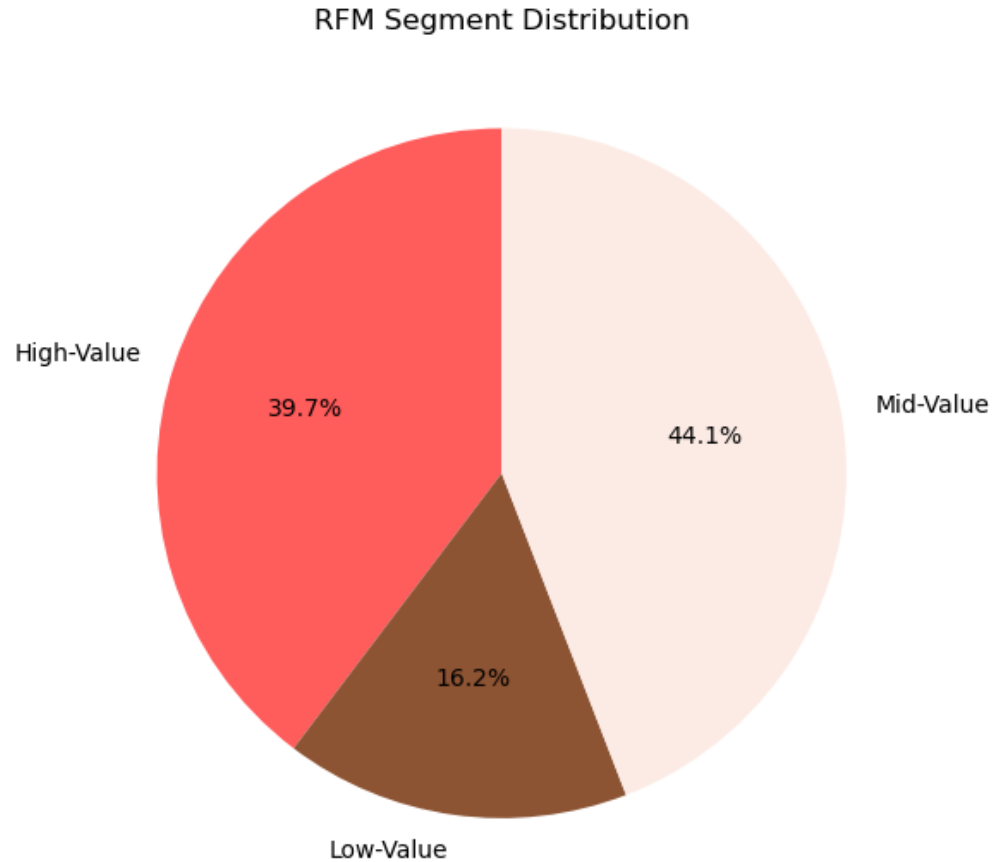
4. DATA ANALYSIS

4.1 VALUABLE USERS



4. DATA ANALYSIS

4.1 VALUEABLE USERS



Most valueble users represents 39.7%.

INSIGHTS

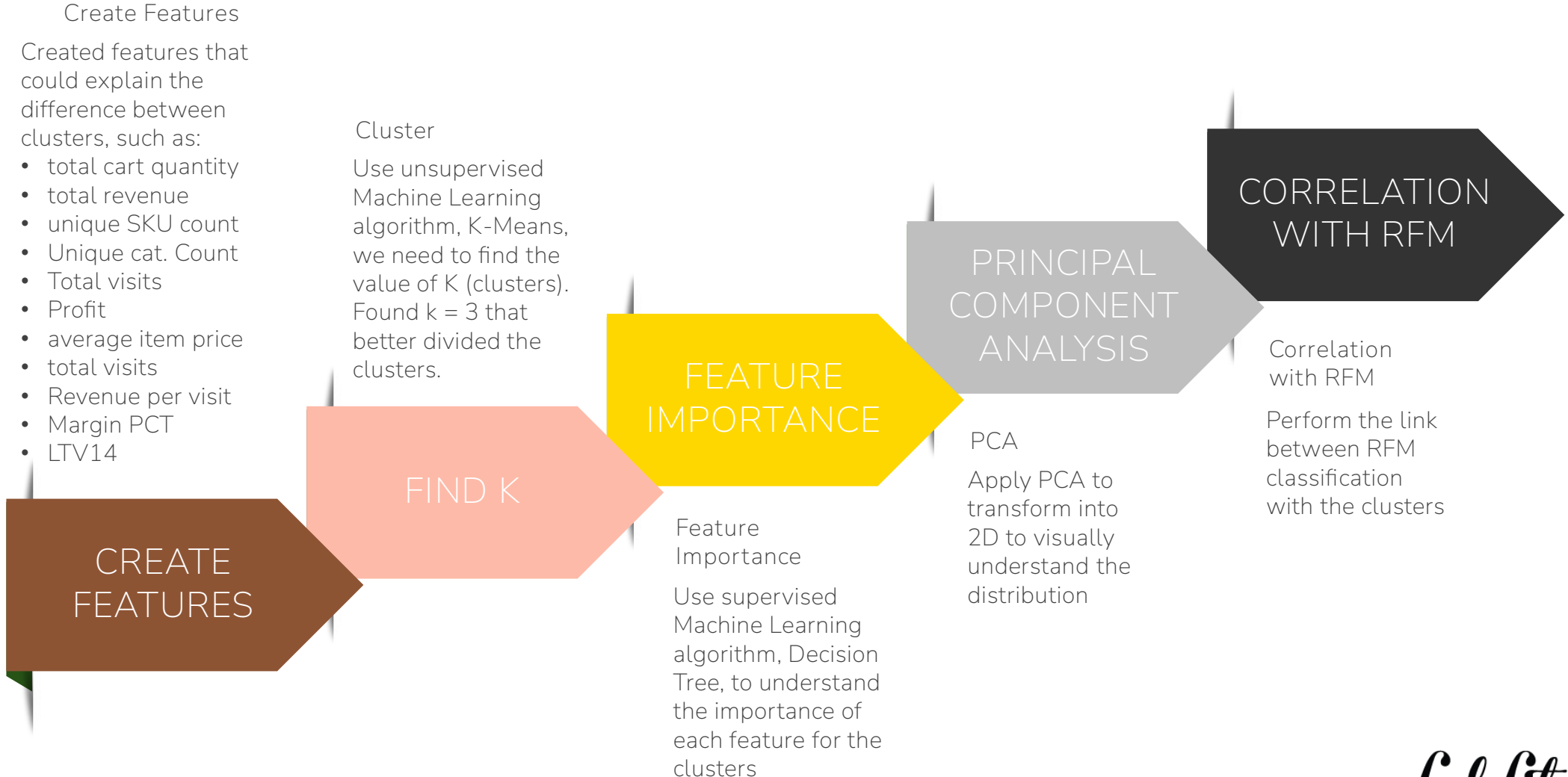
From 451 customers from 2023-08-06 to 2024-08-06:

- 39.7% have High Value
- 44.1% have Mid Value
- 16.2% have Low Value

fabfitfun

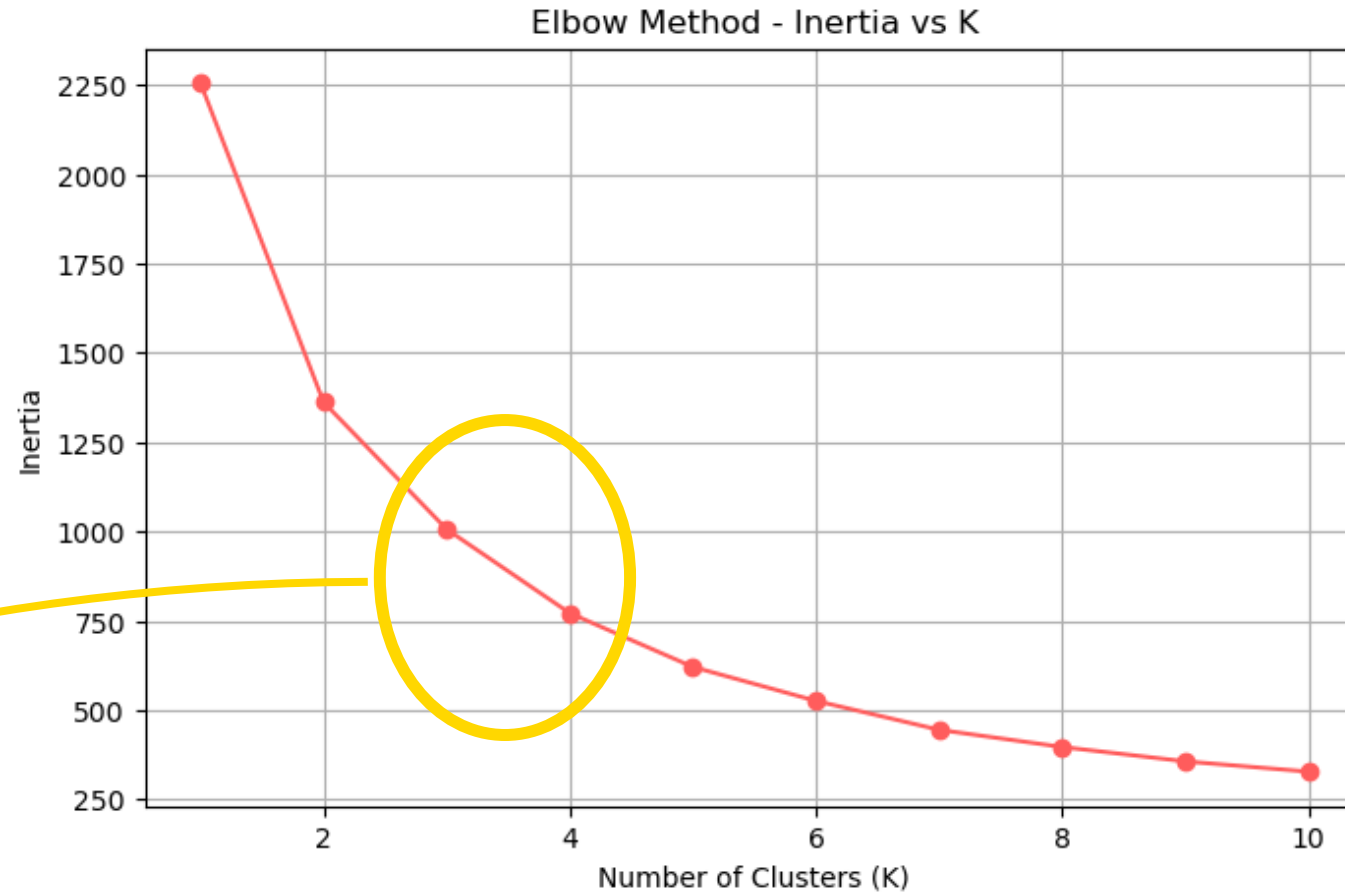
4. DATA ANALYSIS

4.2 DIFFERENCE BETWEEN VALUABLE USERS AND OTHER USERS



4. DATA ANALYSIS

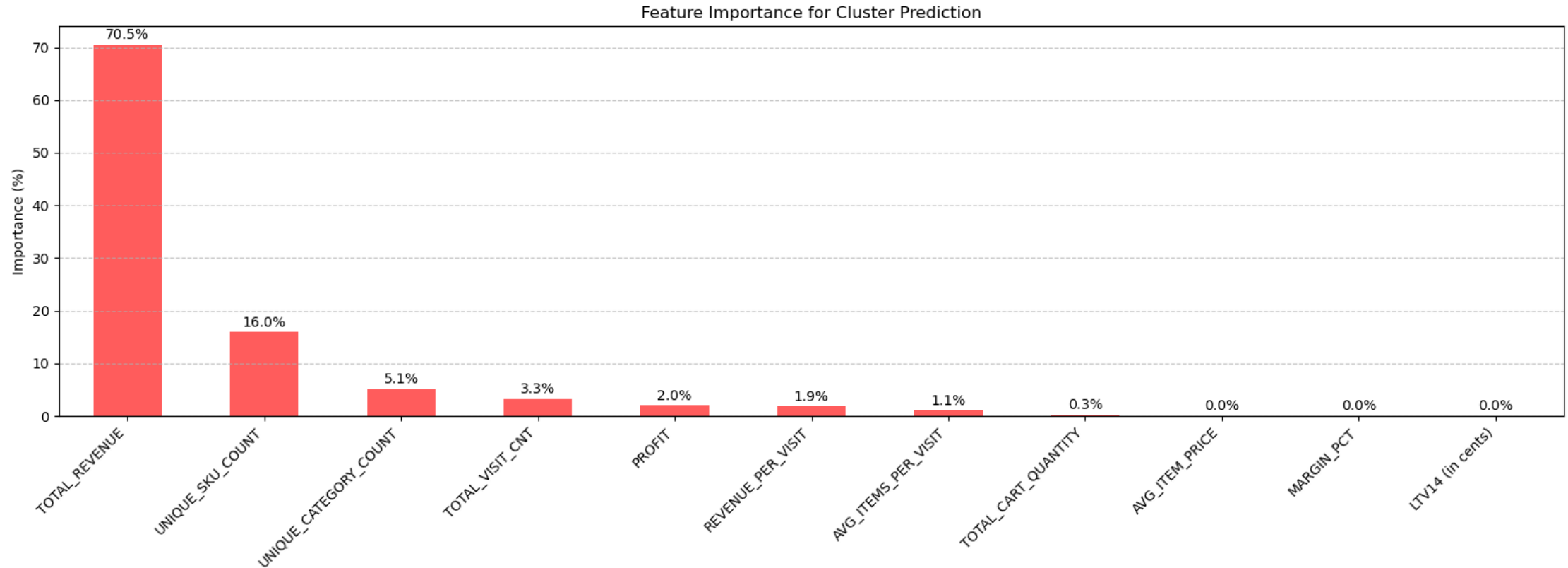
4.2 DIFFERENCE BETWEEN VALUABLE USERS AND OTHER USERS



Inertia stops to fall aggressively around $k = 3 \sim 4$.
We'll choose 3.

4. DATA ANALYSIS

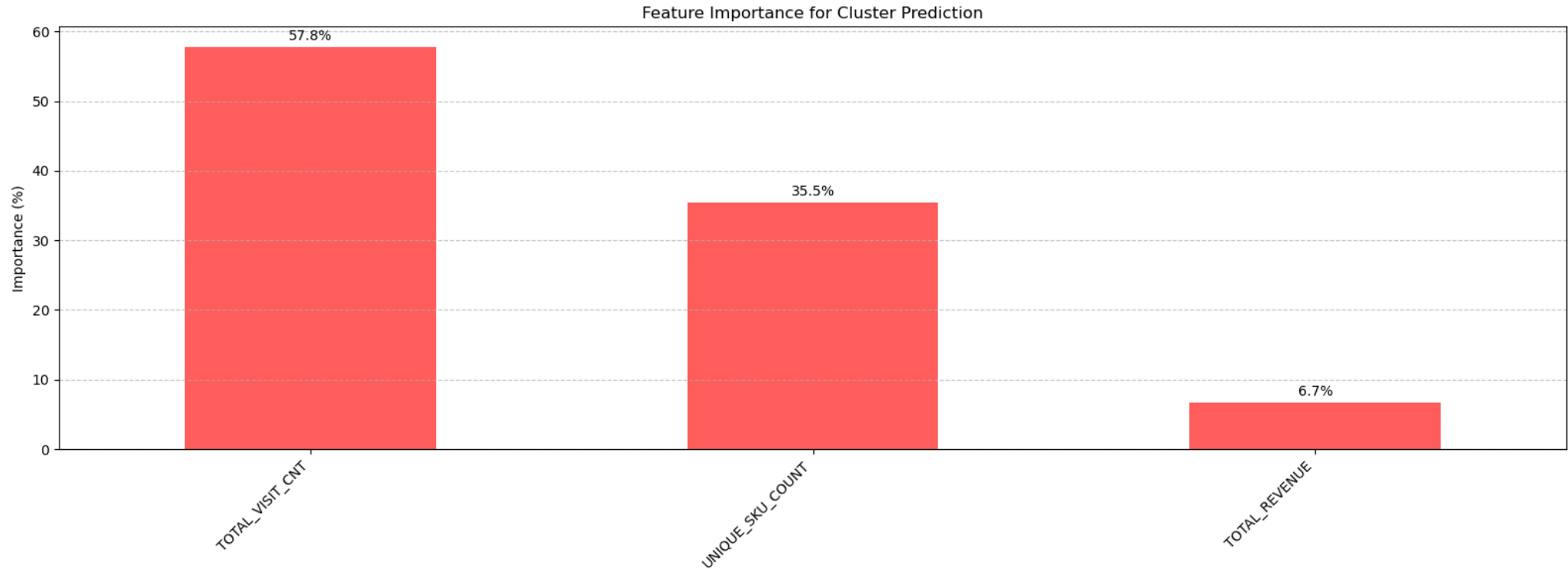
4.2 DIFFERENCE BETWEEN VALUABLE USERS AND OTHER USERS



Most of the features doesn't represent much of the differences between clusters. Let's create a threshold for over 3% to consider as an important feature

4. DATA ANALYSIS

4.2 DIFFERENCE BETWEEN VALUABLE USERS AND OTHER USERS



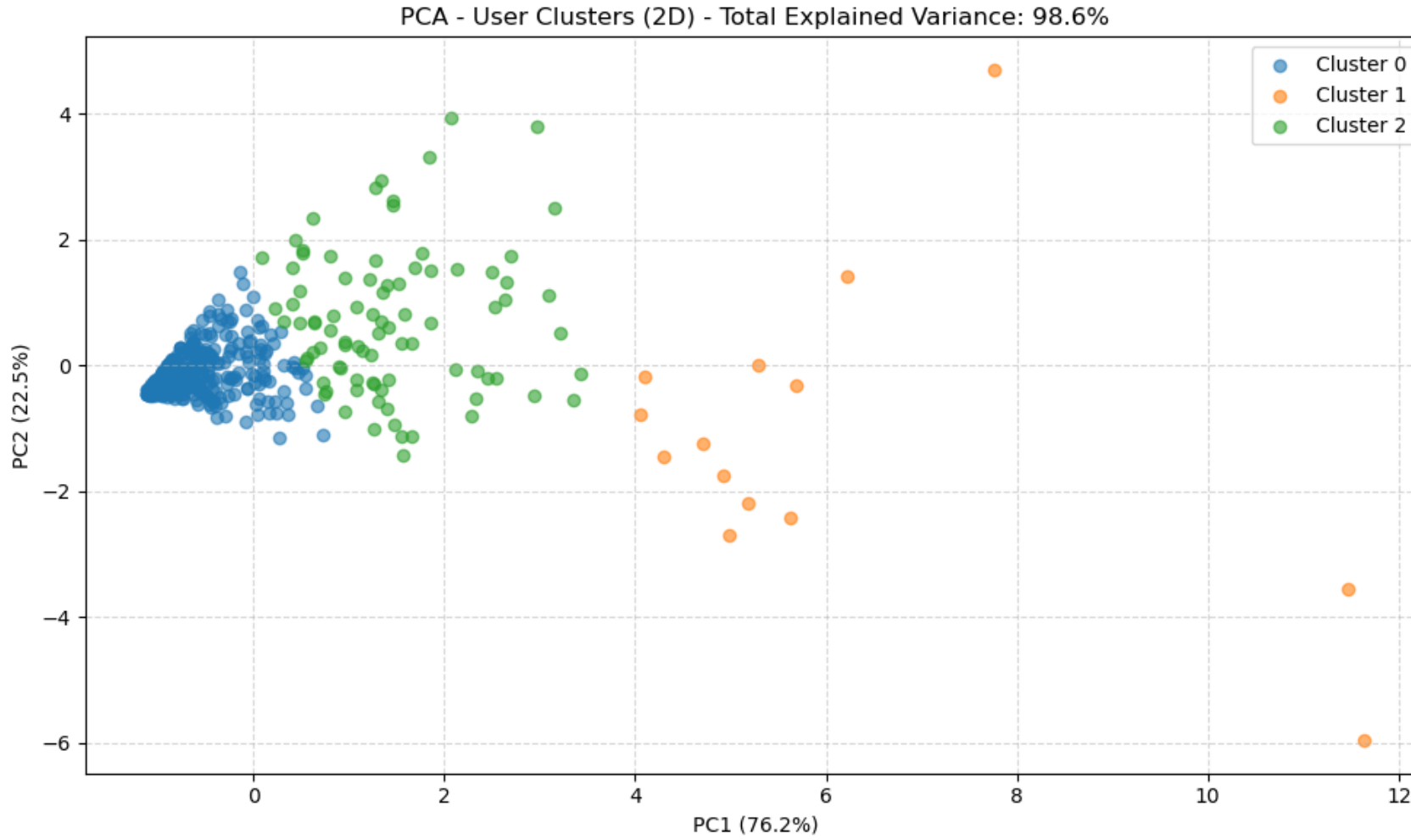
Recalculated the new feature importances for the algorithm.

Total revenue is not the main attribute that best divides the cluster, it's the visits and unique skus, meaning: **DISCOVERABILITY (visits) and EXPLORABILITY (unique skus)**



4. DATA ANALYSIS

4.2 DIFFERENCE BETWEEN VALUABLE USERS AND OTHER USERS



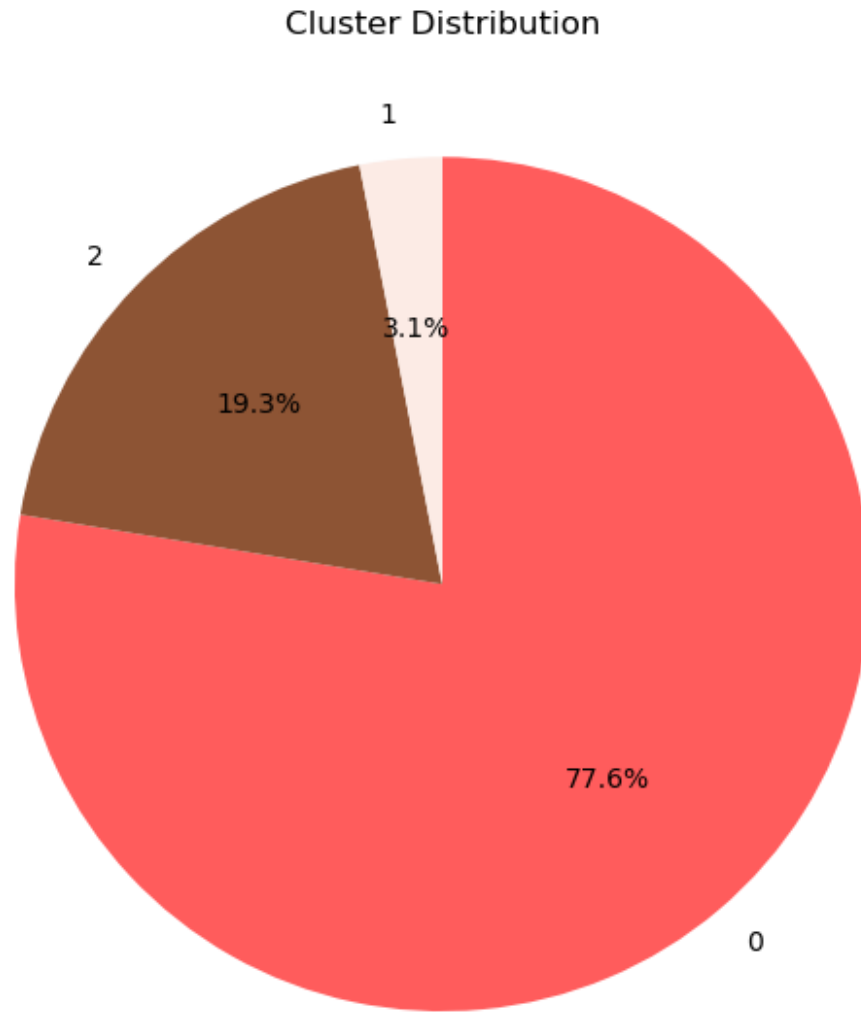
INSIGHTS

Accumulated Variance: 98.6%

fabfitfun

4. DATA ANALYSIS

4.2 DIFFERENCE BETWEEN VALUABLE USERS AND OTHER USERS

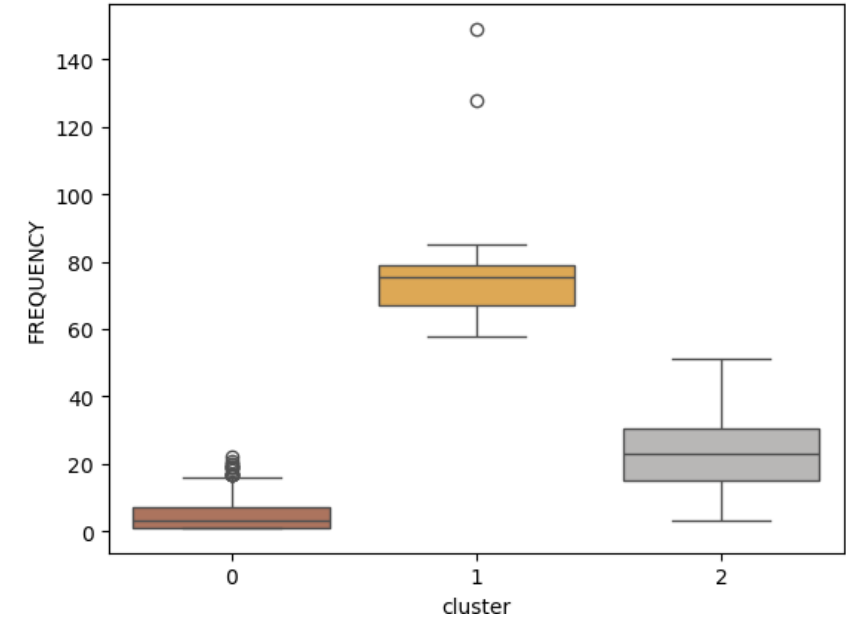
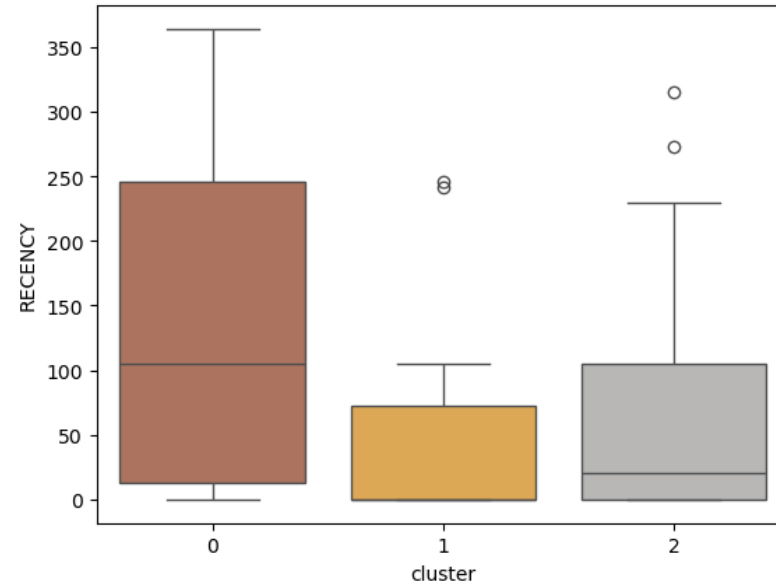
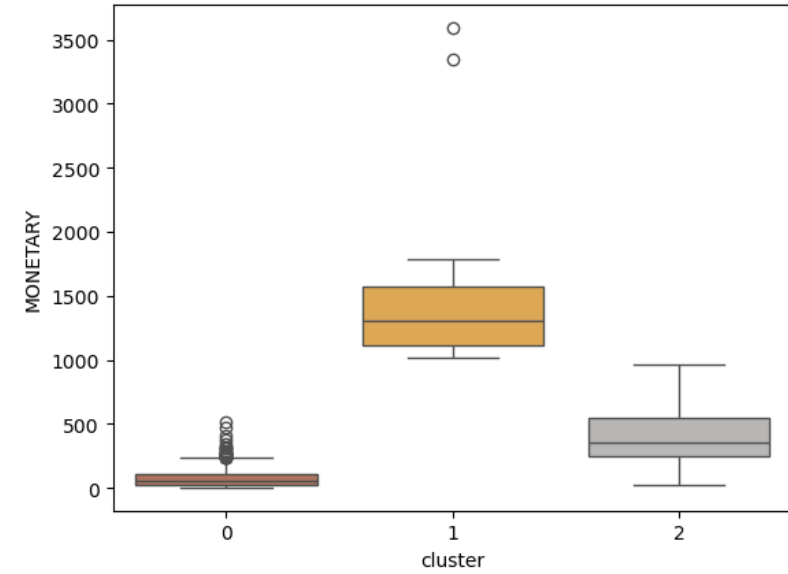


From 451 customers from 2023-08-06 to 2024-08-06:

- 77.61% are in cluster 0
- 03.10% are in cluster 1
- 19.29% are in cluster 2

4. DATA ANALYSIS

4.2 DIFFERENCE BETWEEN VALUABLE USERS AND OTHER USERS



4. DATA ANALYSIS

4.2 DIFFERENCE BETWEEN VALUABLE USERS AND OTHER USERS

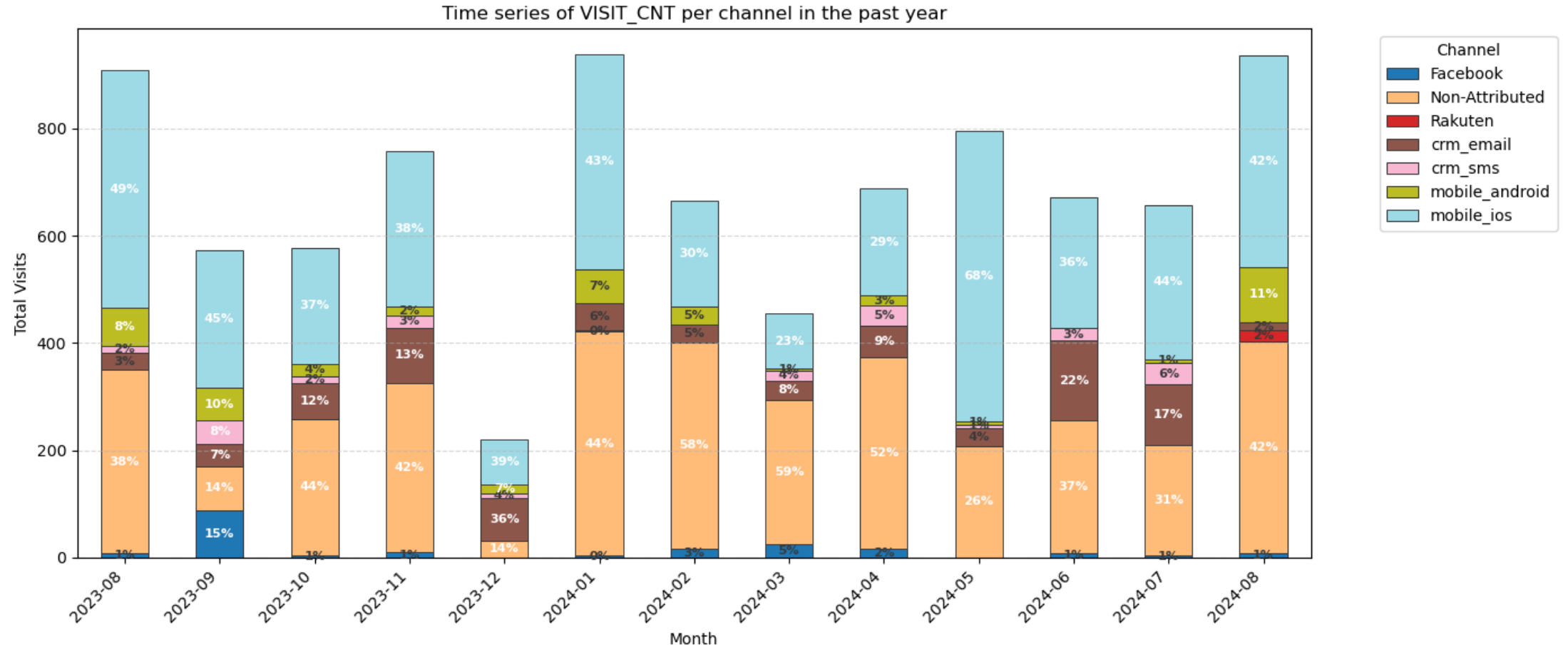
CLUSTER	RFM SCORE	MEAN TOTAL REVENUE	MEAN UNIQUE SKU COUNT	MEAN VISIT CNT	% Of High Value Users	% Of Mid Value Users	% Of Low Value Users
0	6.65	81.10	4.92	11.24	25.71%	20.86%	53.43%
1	11.29	1599.75	79.07	52.93	100%	0.00%	0.00%
2	10.67	398.96	22.56	44.21	86.21%	0.00%	13.79%



Most valuable users are mainly in cluster 1. They differ from cart quantity, total revenue, unique skus and visits.

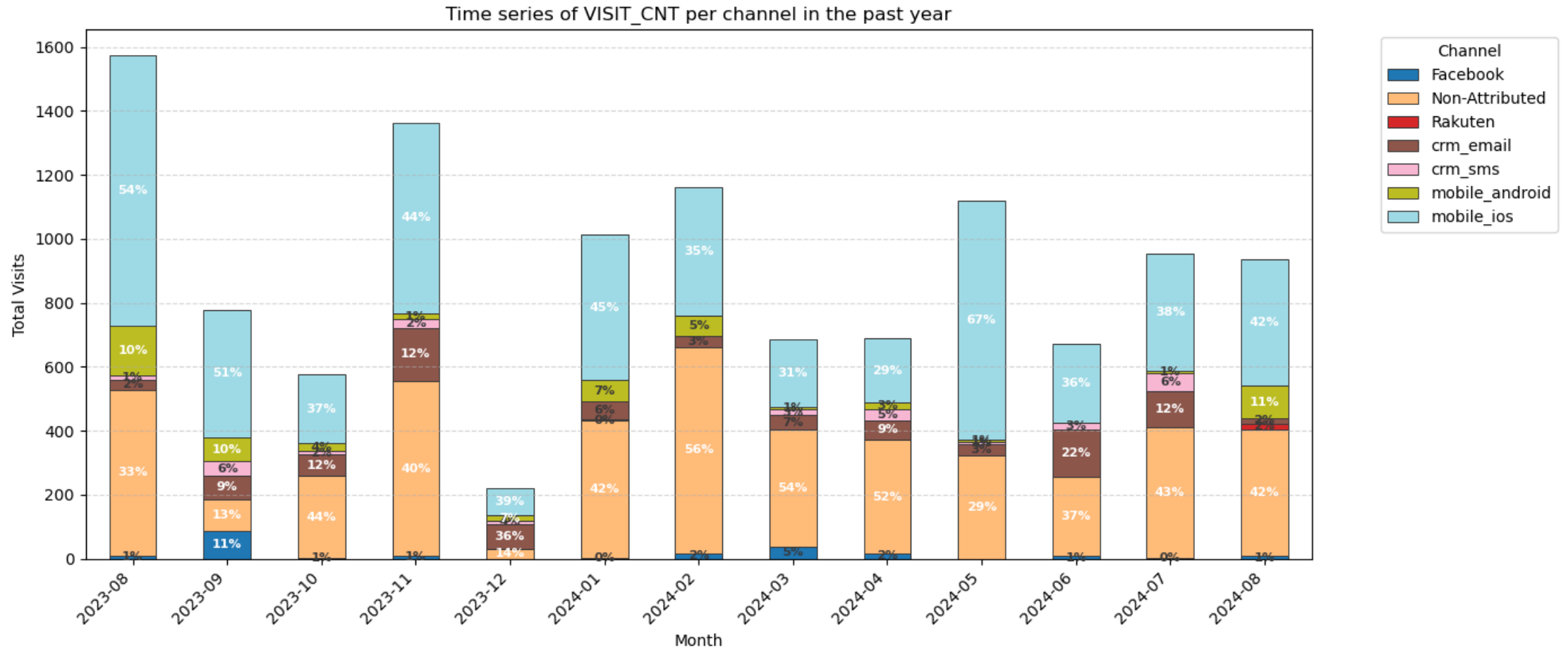
4. DATA ANALYSIS

4.2 CHANNELS – USING ONLY LAST VISIT



4. DATA ANALYSIS

4.2 CHANNELS – USING LAST VISIT + EQUAL CHANNELS PREVIOUSLY ACCESSED



4. DATA ANALYSIS

4.2 CHANNELS

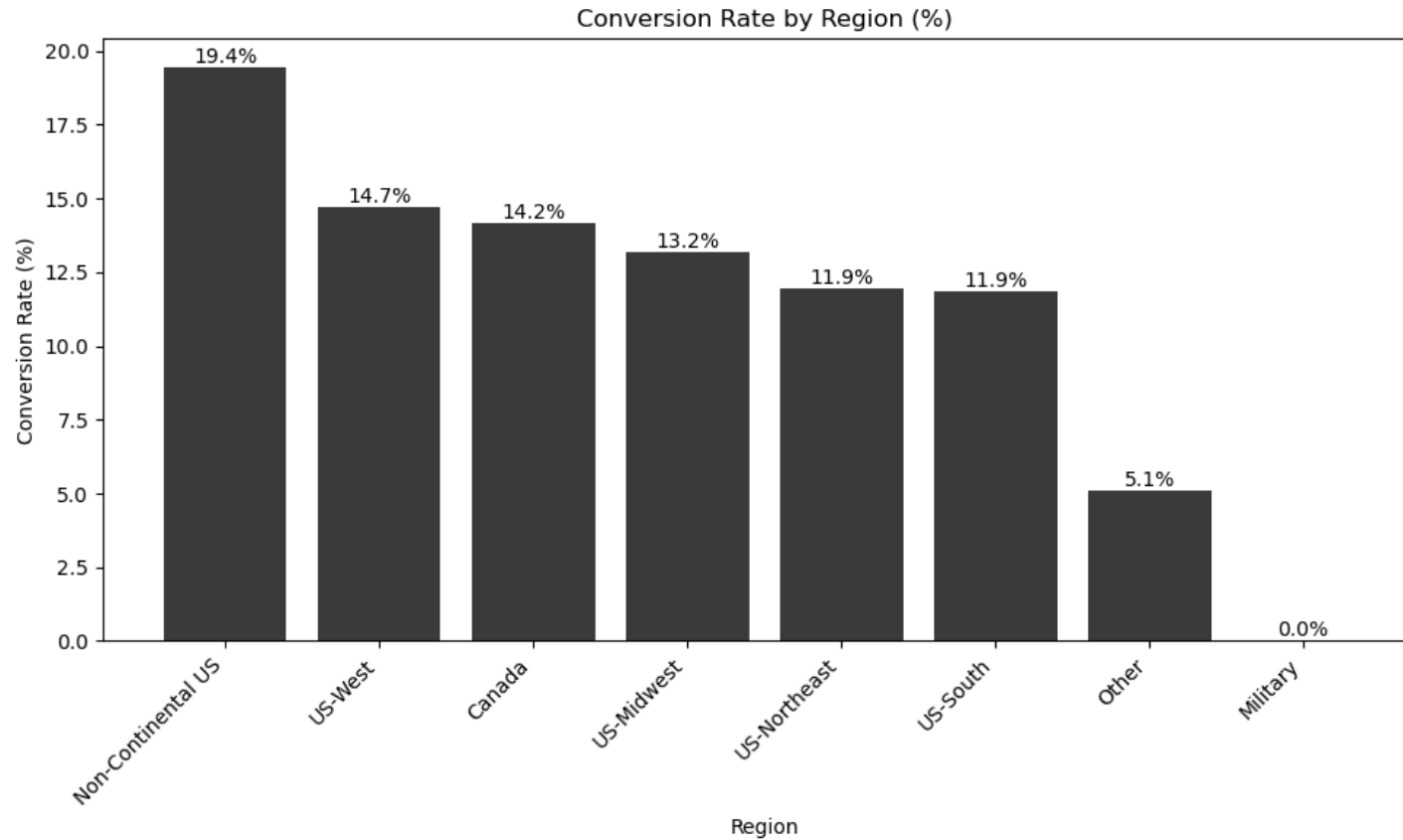


The channel that best drives these members to sales is Mobile IOS in all months.

* Considering time window of 1 year, same users and that the sum of visit cnt that has the same channel for each user + campaign.

4. DATA ANALYSIS

4.3 STRONG CONVERSION RATE



The locations with a stronger conversion rates are, respectively: Non-continental us (19.4%), us-midwest (14.7%)

5. RECOMMENDATIONS

5. RECOMMENDATIONS



According to the RFM and K-Means, the focus should be on discoverability and explorability.

Personalized homepage*
using machine learning
models to reorder properly the
SKUs according to users.
*E.g. Algorithm: Matrix
Factorization*

* In e-commerce, homepage is mainly used to improve discoverability for users.

fabfitfun

5. RECOMMENDATIONS

4.3 STRONG CONVERSION RATE

CLUSTER 0 (high volume, less conversion (**REVENUE ÷ VISITS**)):

- *Objective: focus on transform visits into purchases, than improve campaigns*
- Discounts in first purchases
- Artificial urgency (“Promotion valid for an hour”, “Last unities on stock”)

CLUSTER 1 (small group, high conversion):

- *Objective: extract more value without pressure*
- Cross-sell in checkout (“Also take this item...”)
- Rewards for volumes (“Earn one extra product after buying 3”)
- Faster checkout (reducing quantity of clicks, save preferences)

CLUSTER 2 (mid size group, mostly with high conversion):

- *Objective: convert more in the visits they already perform*
- Improve session (visit) performance (“Recently viewed” with easy click to buy button)
- Reduce session (visit) pressure (highlight product rate, buy directly from home)

The logo for 'fabfitfun' is written in a black, cursive script font. The letters are connected and have a fluid, handwritten appearance.