Data Project

SR Data Analyst
Bruno Bucalon Serra

INDEX

1. PROBLEM DEFINITION

1. PROBLEM DEFINITION

1.1 DATASET ATTRIBUTES

Acquisitions
userID
ACQUISITION_DATE
ORDER_SEASON
PLAN_CODE
SAME_DAY_REFUND
CURRENT_SUBSCRIPTIONS_STATE
AGE
USER_REGION
LTV14

Visits
userID
SEASON
CAMPAIGN_ID
WINDOW_NUMBER
CAMPAIGN_TITLE
CAMPAIGN_TYPE
WINDOW_START_TS_PST
WINDOW_END_TS_PST
CHANNEL
NEW_OR_RETURNING_MEMBER
VISIT_CNT

Campaign (DIM)
SEASON
CAMPAIGN_ID
CAMPAIGN_TITLE
CAMPAIGN_TYPE

Purchases
userID
CAMPAIGN_ID
CAMPAIGN_TITLE
SKU
CART_QUANTITY
MSRP
COGS
SALE_PRICE
CATEGORY

fabfitfun

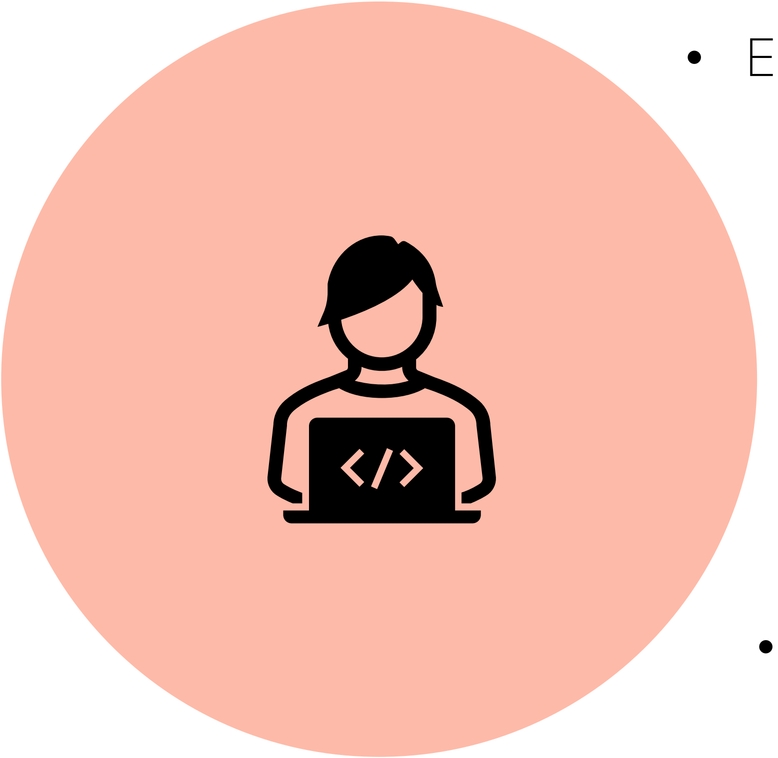
1. PROBLEM DEFINITION

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

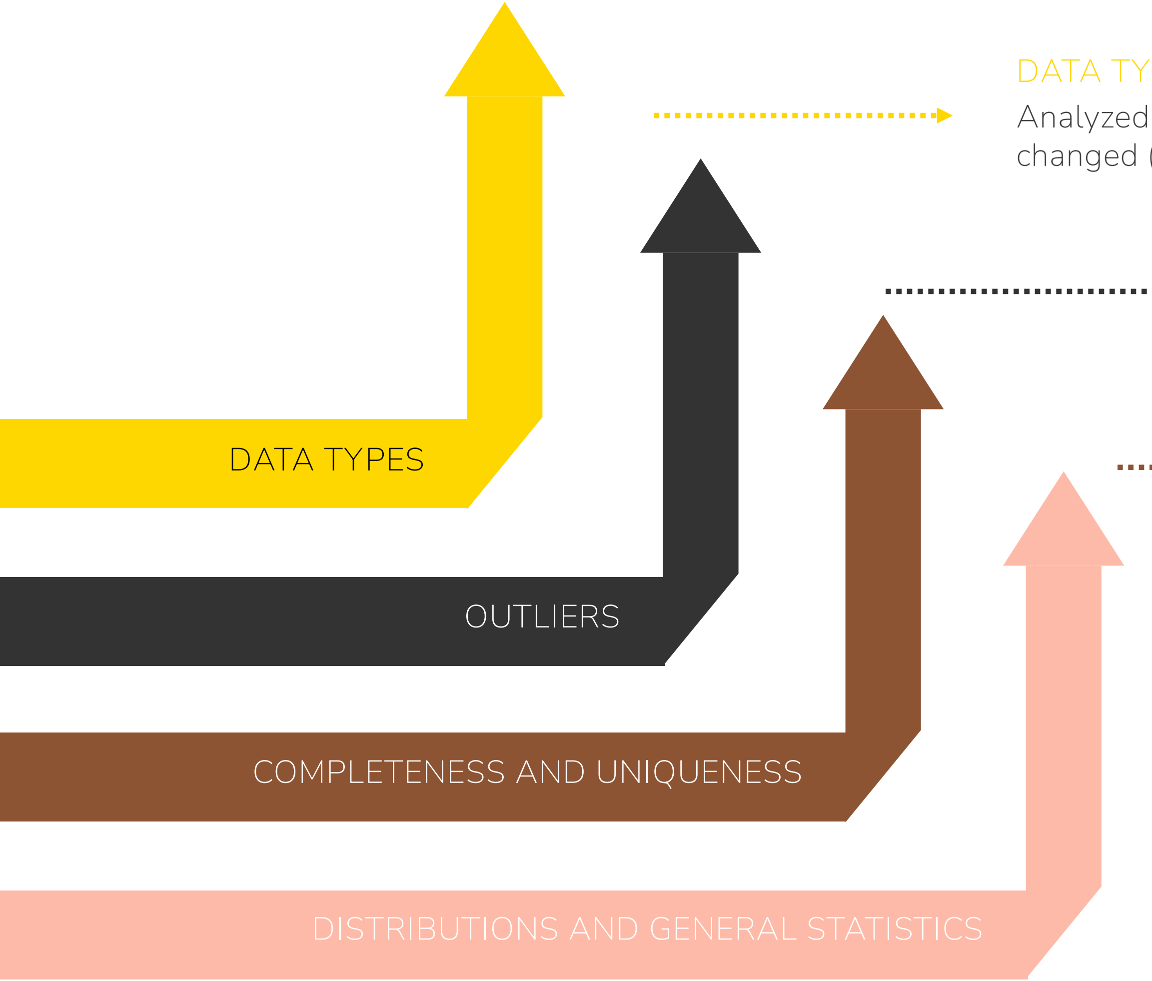
1.3 TECHNOLOGIES AND PROGRAMMING LANGUAGE



- 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



DATA TYPES

Analyzed data types from all datasets and changed (if needed) their types

OUTLIERS

Used boxplots to understand the attributes with outliers and checked if they were real or system errors

COMPLETENESS AND UNIQUENESS

Checked the nulls, nans and missing values per attribute and the uniqueness for qualitative variables

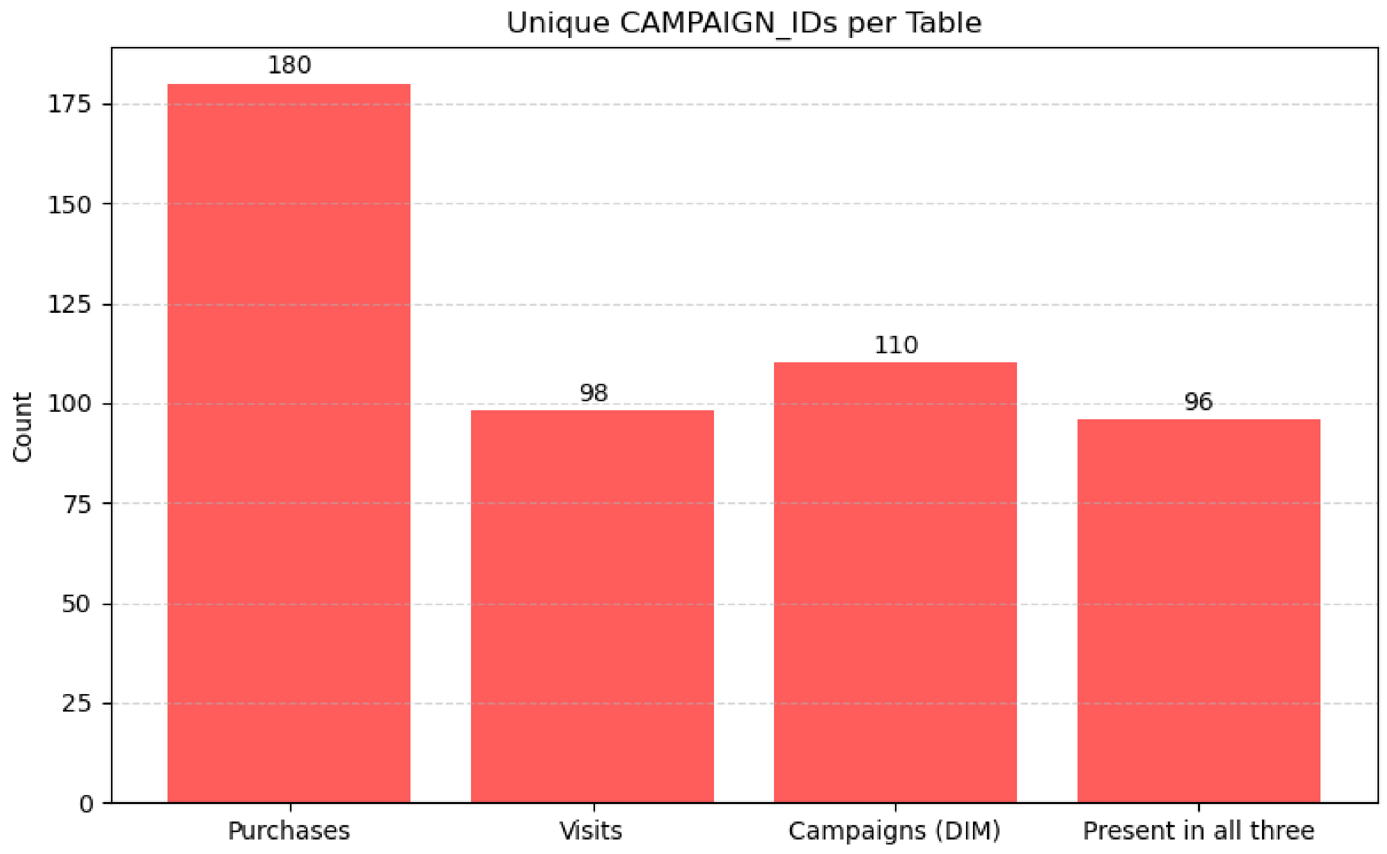
DISTRIBUTIONS AND GENERAL STATISTICS

Checked the distributions, central tendency measures and other general statistics

3. DATA GAPS

3. DATA GAP AND INCONSISTENCES

3.1 CAMPAIGN ID

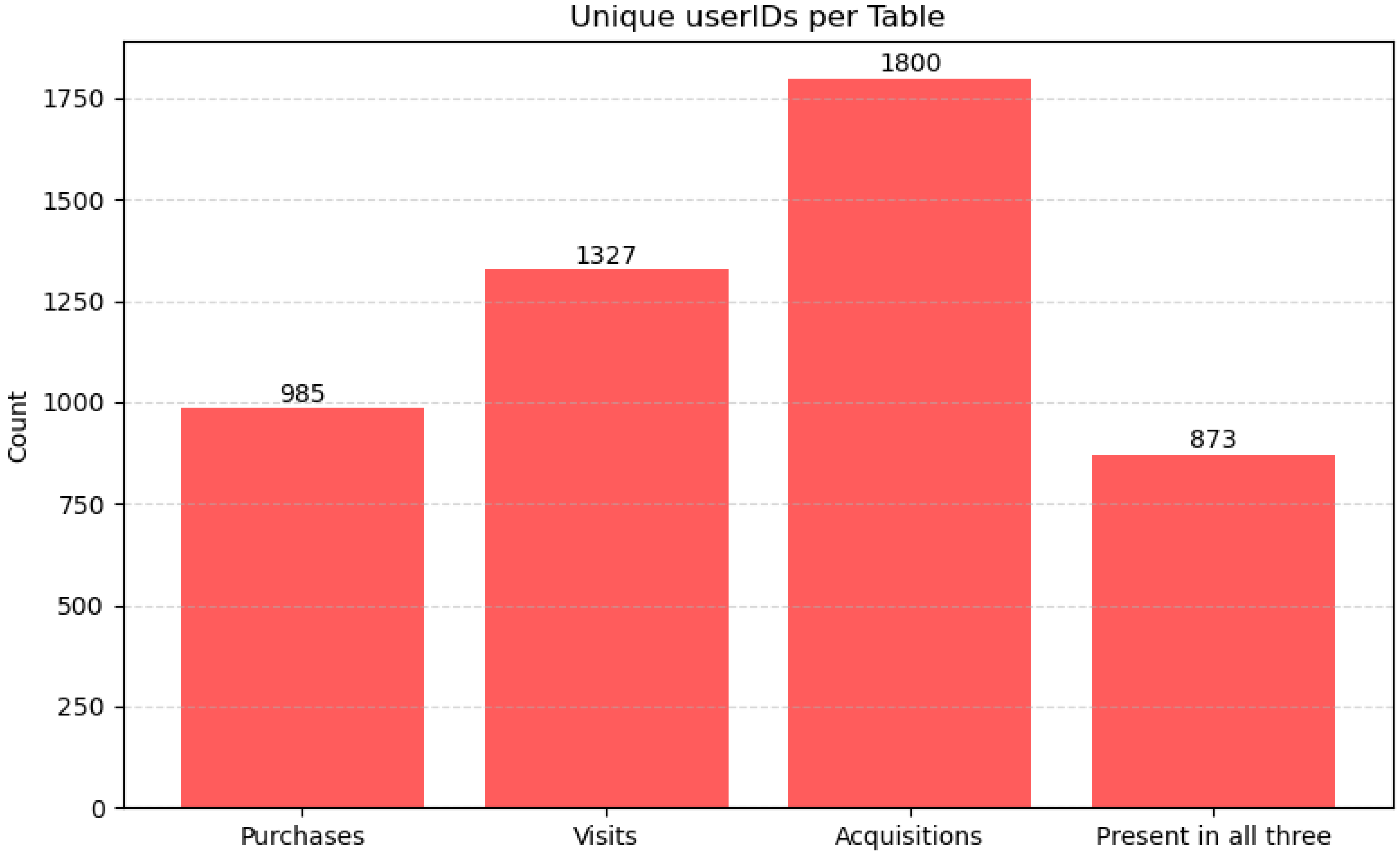


INSIGHTS

Only 96 campaigns are available in the 3 datasets.

3. DATA GAP AND INCONSISTENCES

3.2 USER ID

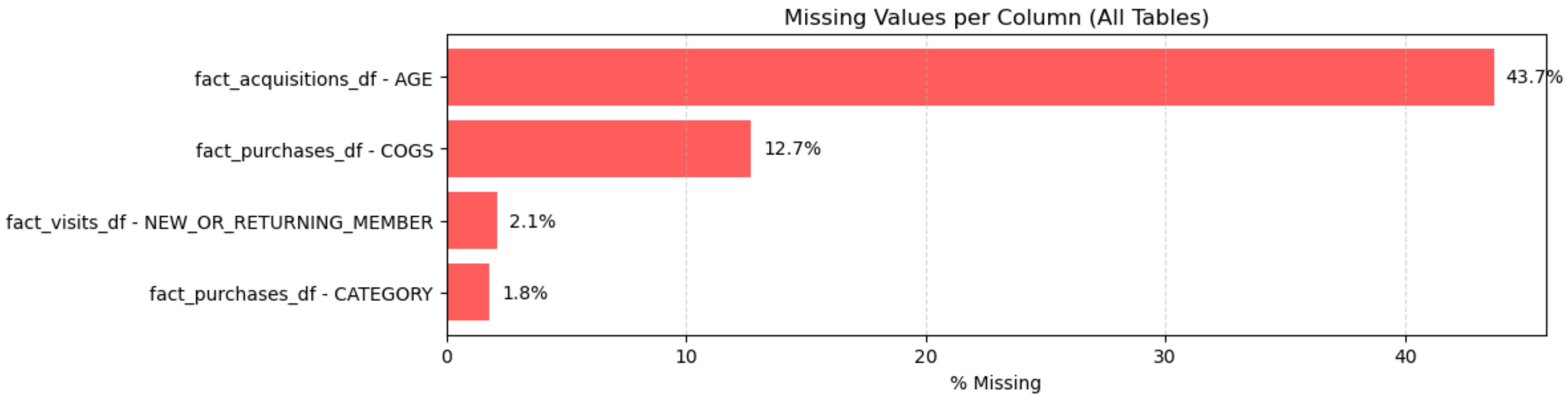


INSIGHTS

Only 873 users are available in the 4 datasets.

3. DATA GAP AND INCONSISTENCES

3.3 AGE



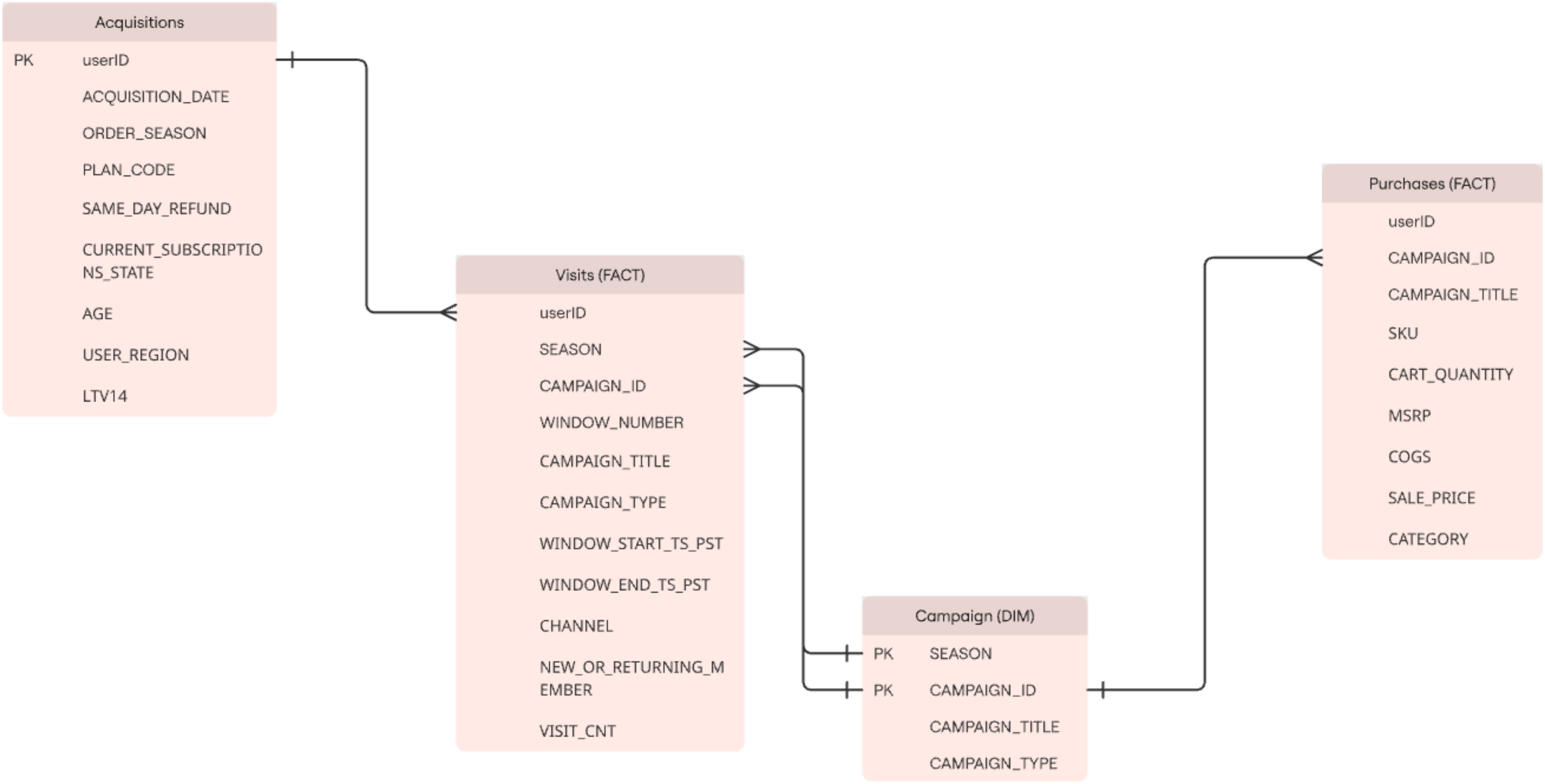
INSIGHTS

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

fabfitfun

3. DATA GAP AND INCONSISTENCES

3.4 MISSING PRIMARY KEY



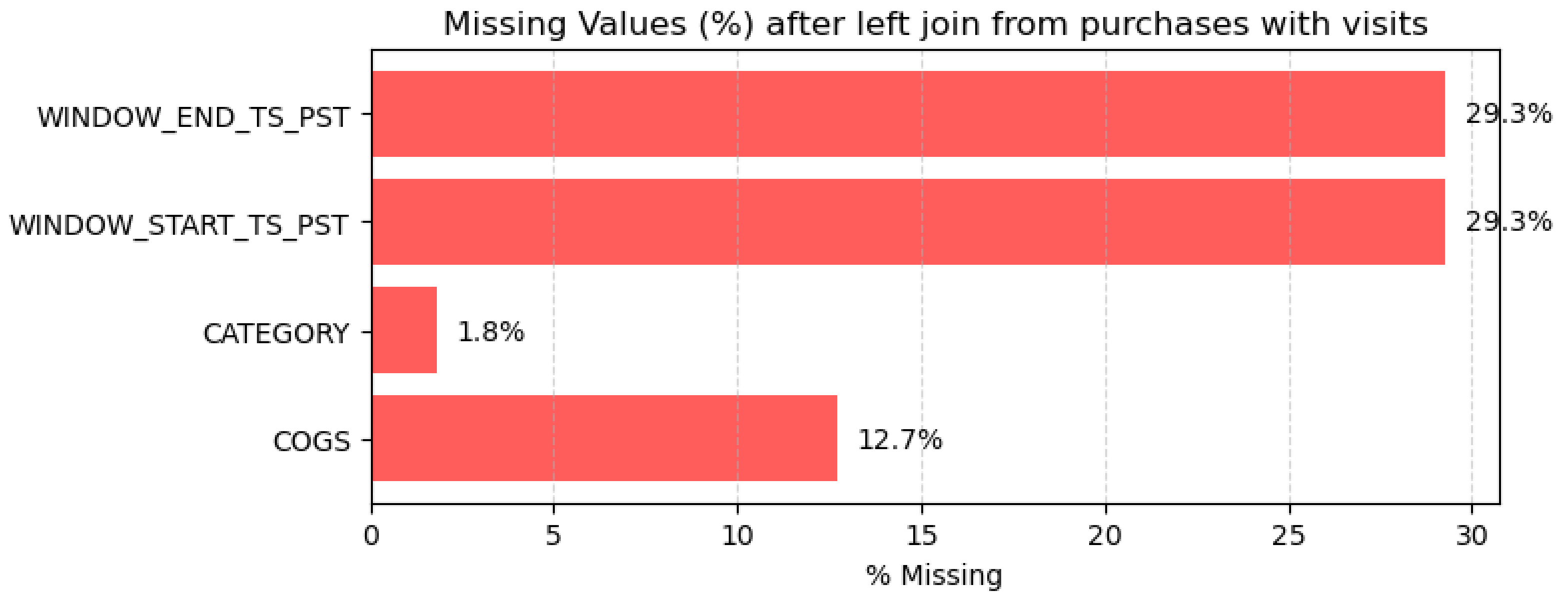
INSIGHTS

i. Relationship between Visits and Purchases table is N:N
ii. No PK in Visits and Purchases

fabfitfun

3. DATA GAP AND INCONSISTENCES

3.5 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.

4. DATA ANALYSIS

4. DATA ANALYSIS

4.1 VALUEABLE USERS

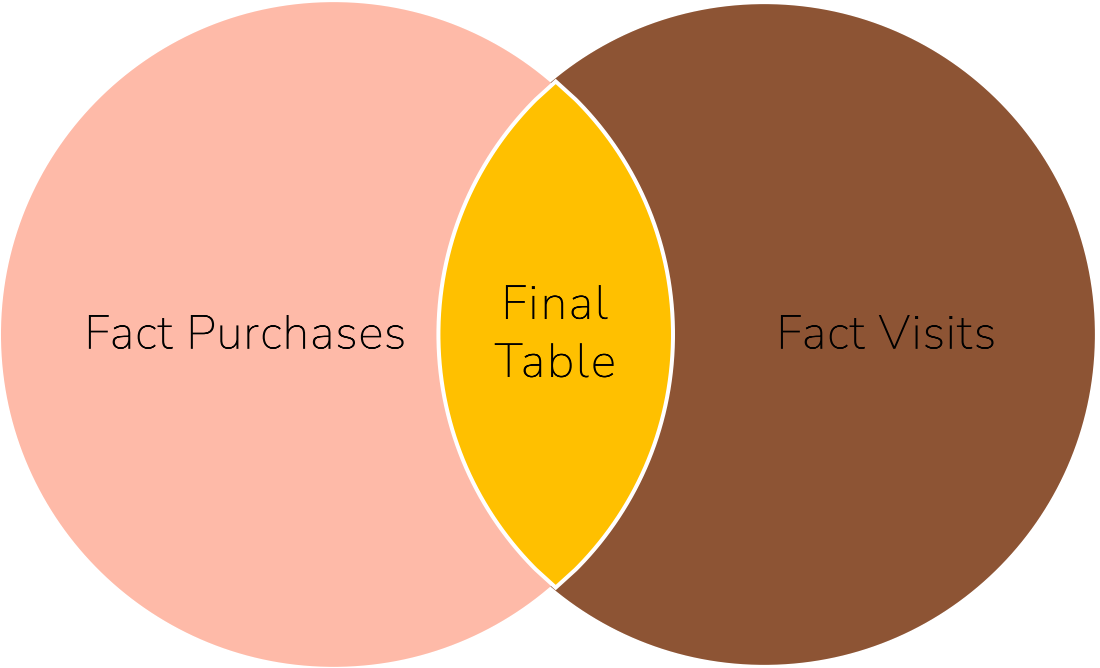
ANALYSIS ASSUMPTION

1. A customer can only purchase an item if they have visit it
2. The date of the last visit of the customer for that specific campaign is the one they converted

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* x *SALE_PRICE*

By performing this join:

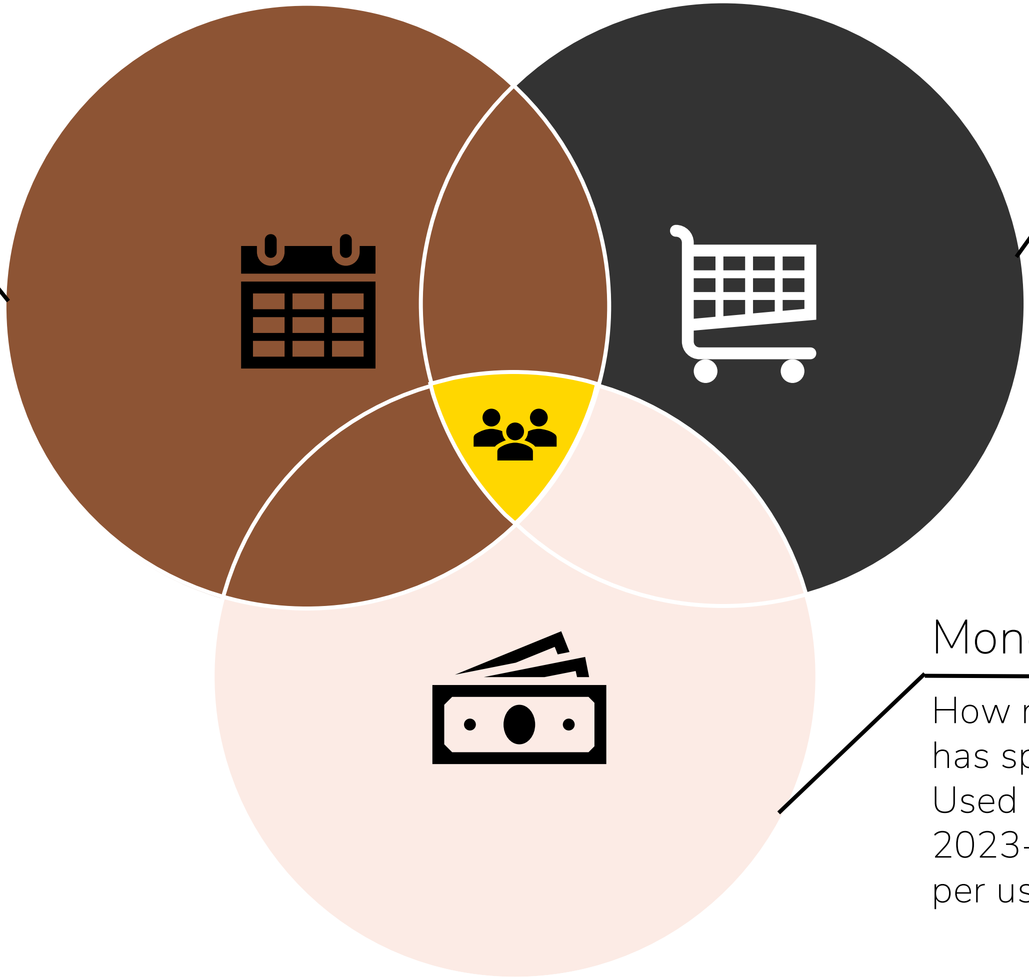
- Kept 70.71% of the rows from Fact Purchases
- Kept 77.77% of the users from Fact Purchases
- Kept 53.33% of the campaigns from Fact Purchases

4. DATA ANALYSIS

4.1 VALUEABLE 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 VALUEABLE USERS

Weight
Assign each result a score from 1 to 4 based on quartile distribution:
Recency -> lower, the better
Frequency & Monetary -> higher the better

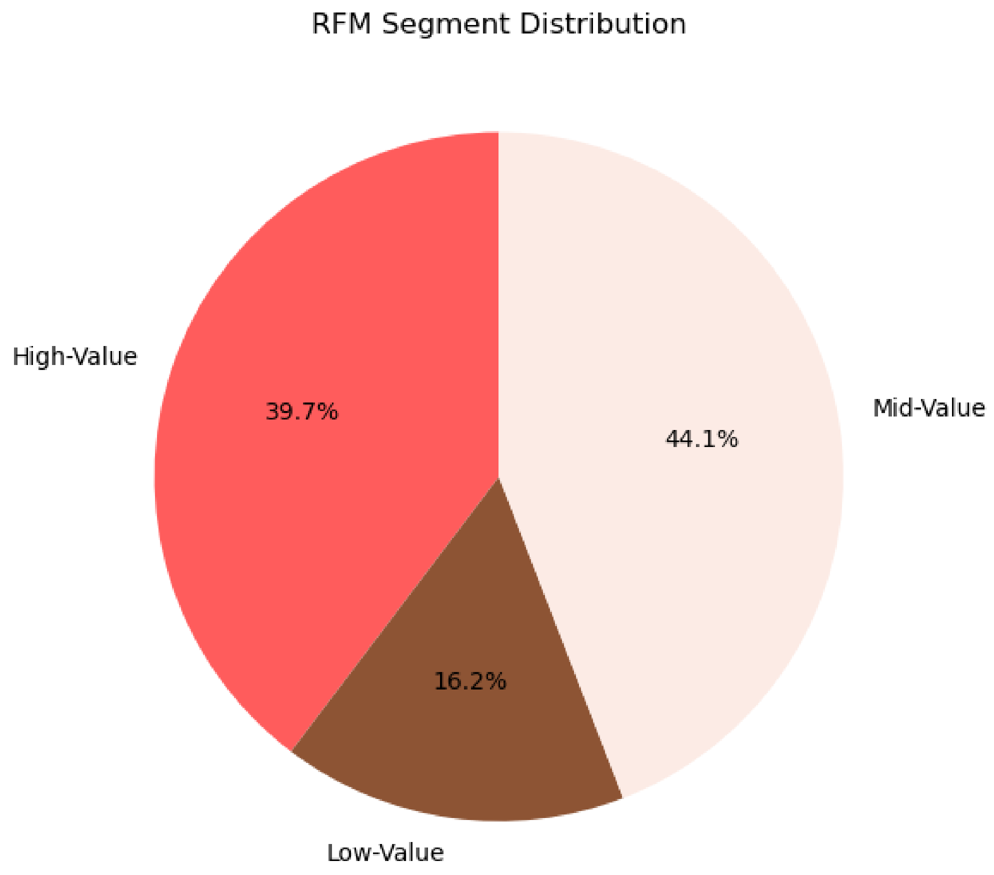
Define Value
For score <5: Low Value
For score <9: Mid-Value
For Score >= 10: High-Value

Calculation
Calculate the general values for Recency, Frequency and Monetary per user

Score Result
Sum each result from Weight and find a final value of RFM per user

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 VALUEABLE USERS AND OTHER USERS

CREATE FEATURES

Create Features

Created features that could explain the difference between clusters, such as:

- total cart quantity
- total revenue
- unique SKU count
- average item price
- total visits

FIND K

Cluster

Use unsupervised Machine Learning algorithm, K-Means, we need to find the value of K (clusters). Found k = 3 that better divided the clusters.

FEATURE IMPORTANCE

Feature Importance

Use supervised Machine Learning algorithm, Decision Tree, to understand the importance of each feature for the clusters

Principal Component Analysis

PCA

Apply PCA to transform into 2D to visually understand the distribution

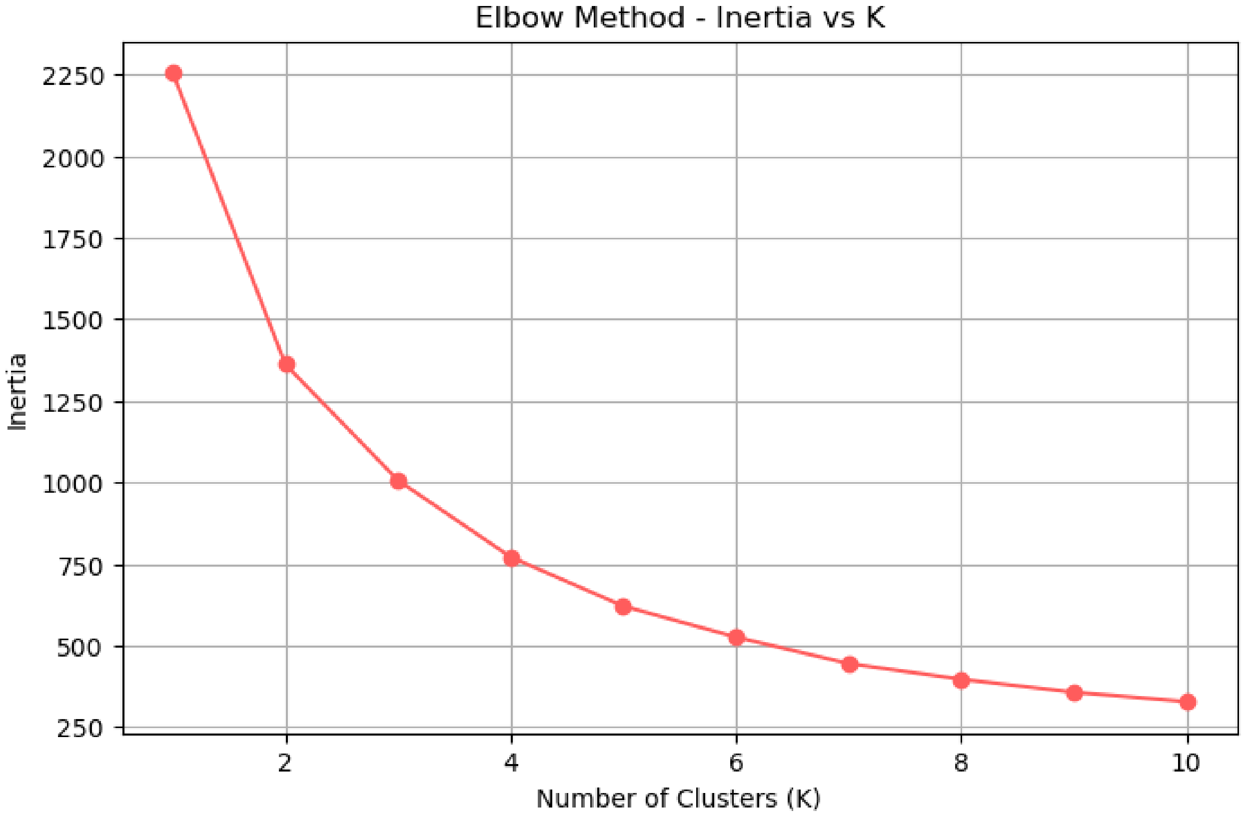
CORRELATION WITH RFM

Correlation with RFM

Perform the link between RFM classification with the clusters

4. DATA ANALYSIS

4.2 DIFFERENCE BETWEEN VALUEABLE USERS AND OTHER USERS

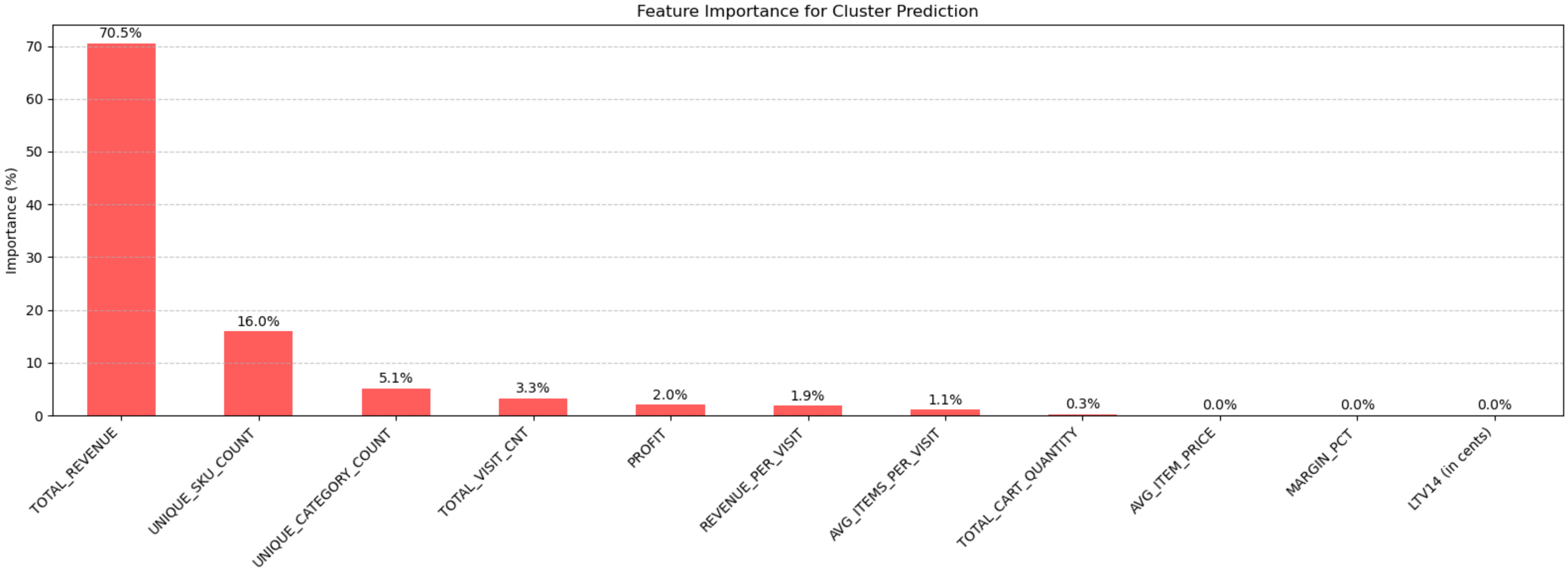


INSIGHTS

fabfitfun

4. DATA ANALYSIS

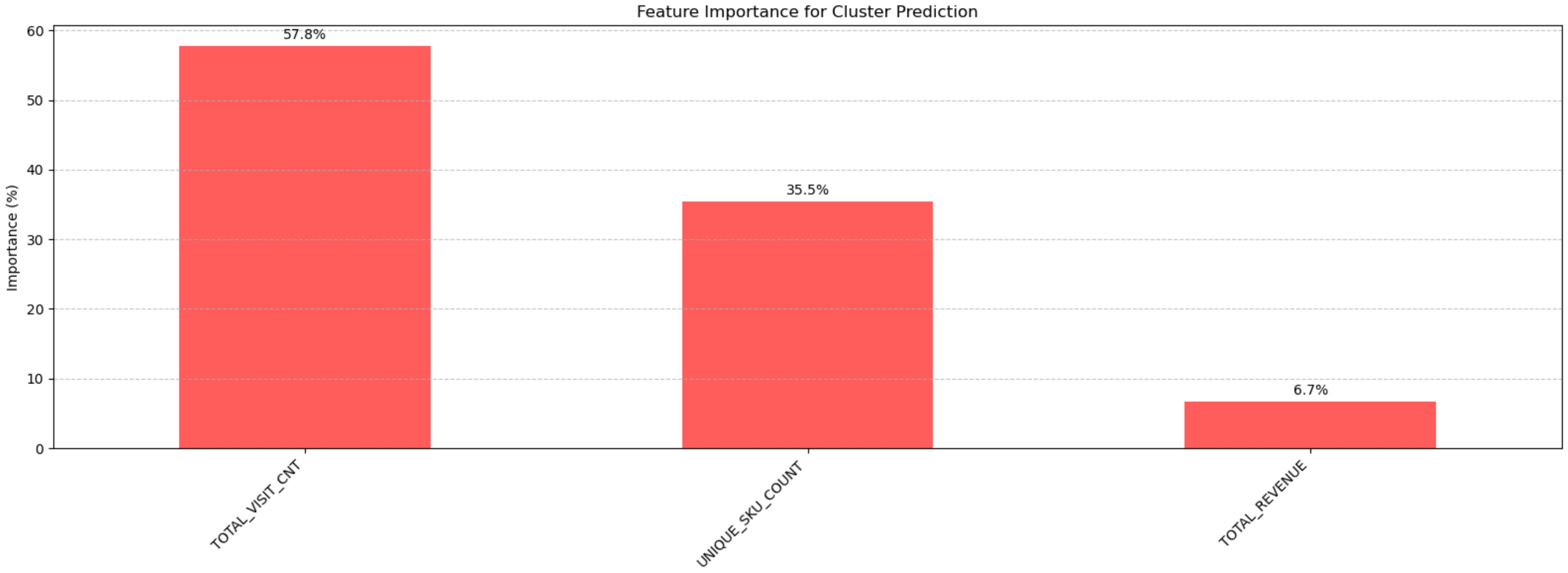
4.2 DIFFERENCE BETWEEN VALUEABLE USERS AND OTHER USERS



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

4. DATA ANALYSIS

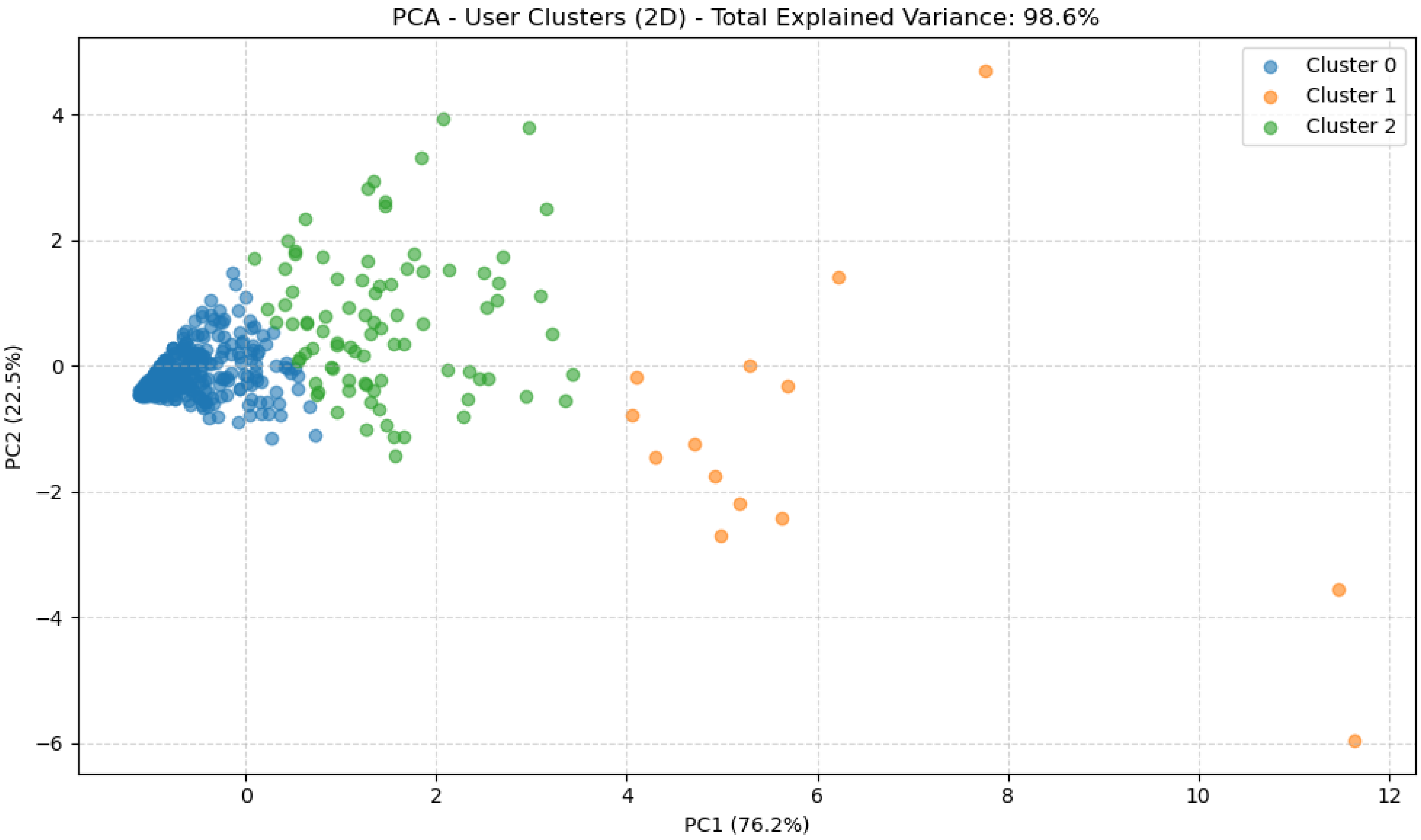
4.2 DIFFERENCE BETWEEN VALUEABLE USERS AND OTHER USERS



Recalculated the new feature importances for the algorithm.

4. DATA ANALYSIS

4.2 DIFFERENCE BETWEEN VALUEABLE USERS AND OTHER USERS



INSIGHTS

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

fabfitfun

4. DATA ANALYSIS

4.2 DIFFERENCE BETWEEN VALUEABLE USERS AND OTHER USERS

		MEAN					
CLUSTER	RFM SCORE	TOTAL REVENUE	UNIQUE SKU COUNT	TOTAL 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 valueble users are in cluster 0 and 2. They differ from cart quantity, total revenue, unique skus and visits.

4. DATA ANALYSIS

4.2 CHANNELS

CHANNEL	High-Value	Low-Value	Mid-Value
Facebook	158	3	18
Non-Attributed	1644	67	405
Rakuten	20	0	1
crm_email	550	9	75
crm_sms	143	1	46
mobile_android	166	3	29
mobile_ios	1339	13	198

The channels that best drives these members to sales are: Mobile IOS and E-mail

4. DATA ANALYSIS

4.3 STRONG CONVERTION RATE

user_region	users_campaign_visited	users_campaign_bought	conversion_rate
canada	536	150.0	27.99
military	3	0.0	0.0
non-continental us	22	7.0	31.82
other	44	5.0	11.36
us-midwest	1040	253.0	24.33

The locations with a stronger convertion rates are, respectively: Non-continental us (31.82%), Canada (27.99%), us-midwest (24.33%).

INSIGHTS

The locations with a stronger convertion rates are, respectively:

- Non-continental us = 31.82%
- Canada = 27.99%
- Us-midwest = 24.33%