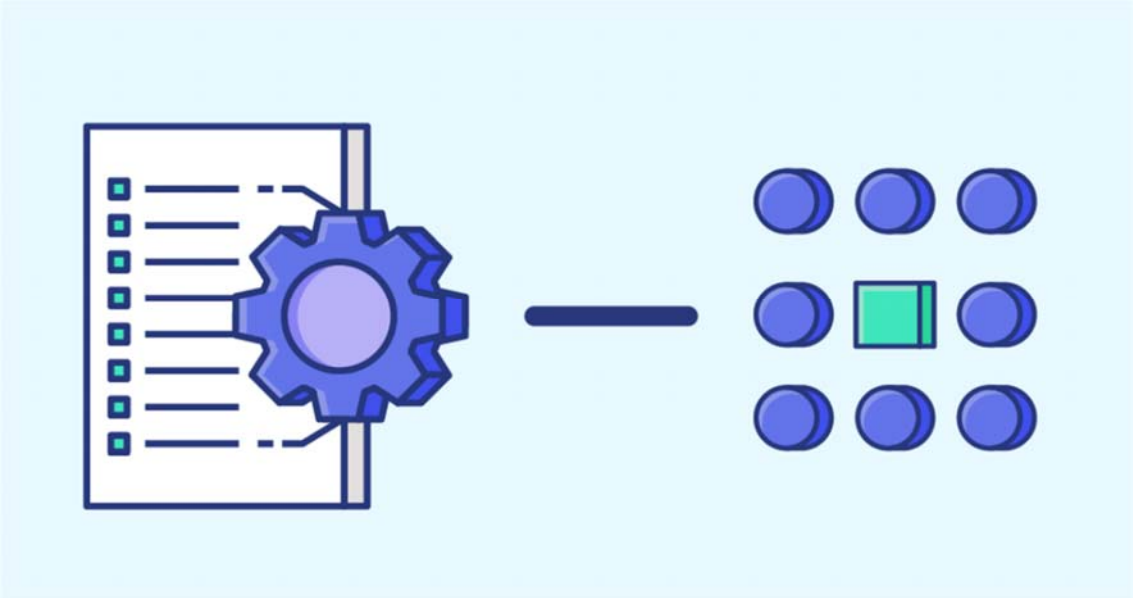
Feature Engineering and Business Understanding

By : Intania



Objective

The goal of this project is to derive meaningful insights into customer behavior based on transaction data. This process involves defining business problems, setting goals, and outlining the approach and methods to be employed.

Business Problems

The primary business problems addressed in this project include:

- **Understanding Customer Behavior:** Identifying how different user segments behave regarding spending, preferences, and transaction patterns.

Goals

The goals of this project are:

1. To create features that facilitate effective segmentation for targeted marketing.
2. To develop a clear understanding of user segments based on transaction data.
3. To provide actionable insights that the Marketing Department can implement to improve customer acquisition and retention strategies.

Approach and Method

The approach consists of several key steps:

1. **Data Preprocessing:** Cleaning and transforming the raw transaction data to prepare it for analysis.
2. **Feature Engineering:** Creating relevant features based on user demographics and transaction behaviors.
3. **Segmentation Analysis:** Utilizing clustering techniques to group users into distinct segments based on the engineered features.

Evaluation Method

The effectiveness of the segmentation will be evaluated using the following methods:

- **Elbow Method:** Identifying the optimal number of clusters by analyzing the variance explained as a function of the number of clusters.
- **Insight Generation:** Analyzing the segments to extract valuable insights that inform better strategies and business decisions, helping to align offerings with customer needs and preferences.

Data Understanding and Cleaning

In this phase, I examined the dataset to ensure its readiness for analysis. The following steps were taken:

- Changed data types for certain features.
- Analysed the time period of the data.
- Generated a statistical summary of the dataset.
- Checked for missing values and outliers.
- Conducted simple exploratory data analysis (EDA).
- Removed outliers to focus on normal transactions, improving the clustering process.

Data Preparation

The data preparation involved several key actions:

1. Translate Merchant Categories: I translated the merchant_category_id feature into its respective category, corresponding to Visa Merchant Categories Code, and grouped them into broader categories such as Food and Beverage, Retail, Health & Wellness, Home and Lifestyle, and Travel.

```python
# Dictionary mapping MCC to categories
mcc_to_category = {
    '5732': 'Electronics Stores',
    '5812': 'Eating Places and Restaurants',
    '5411': 'Grocery Stores, Supermarkets',
    '5814': 'Fast Food Restaurants',
    '5964': 'Direct Marketing - Catalog Merchants',
    '5311': 'Department Stores',
    '5999': 'Miscellaneous and Specialty Retail Stores',
    '5912': 'Drug Stores and Pharmacies',
    '5541': 'Service Stations (with or without ancillary services)',
    '5942': 'Book Stores',
    '7230': 'Beauty and Barber Shops',
    '5661': 'Shoe Stores',
    '5712': 'Furniture, Home Furnishings, and Equipment Stores',
    '5651': 'Family Clothing Stores',
    '8220': 'Colleges, Universities, and Professional Schools',
    '5412': 'Convenience Stores',
    '5941': 'Sporting Goods Stores',
    '8211': 'Elementary and Secondary Schools',
    '4722': 'Travel Agencies and Tour Operators',
    '5251': 'Hardware Stores',
    '7994': 'Video Game Arcades/Establishments'
}
```

```python
# Create broad category
mcc_to_broad_category = {
    '5732': 'Retail',
    '5812': 'Food and Beverage',
    '5411': 'Food and Beverage',
    '5814': 'Food and Beverage',
    '5964': 'Retail',
    '5311': 'Retail',
    '5999': 'Retail',
    '5912': 'Health and Wellness',
    '5541': 'Food and Beverage',
    '5942': 'Retail',
    '7230': 'Health and Wellness',
    '5661': 'Retail',
    '5712': 'Home and Lifestyle',
    '5651': 'Retail',
    '8220': 'Education',
    '5412': 'Food and Beverage',
    '5941': 'Retail',
    '8211': 'Education',
    '4722': 'Travel',
    '5251': 'Home and Lifestyle',
    '7994': 'Entertainment'
}
```

Figure 1 : Mapping MCC to Categories and Broad Categories

2. User Agent Analysis: I translated the user agent information into three columns:
    - OS Type: Identifying whether the device is Android or iOS.
    - OS Version: Assessing whether the version is old or new.
    - Device Type: Determining if the device is a tablet or phone.

    After visualizing this data, I found that there was no significant difference in preferences based on device type, so I decided to explore combining this feature with other factors rather than prioritizing it.

3. User Feature Table Creation: I derived key insights into user behaviour and preferences from transaction data using various aggregation techniques. The following features were generated for each user:
    - Recency of Last Transaction: Days since the last transaction.
    - Frequency of Transactions: Total number of transactions made by each user.
    - Total Spending: Sum of transaction amounts to understand overall spending behavior.
    - Average Transaction Amount: Mean of transaction amounts to identify typical spending patterns.
    - Merchant Category Preferences: Most frequently transacted merchant category for each user.
    - Loyalty Program Participation: Proportion of transactions indicating loyalty program participation.
    - Promo Usage: Ratio of transactions utilizing promotions and total promotional amounts.
    - Payment Method Preferences: Proportion of transactions paid using the balance.
    - Merchant Rating: Average rating given by users for merchants.
    - Refund Frequency: Frequency of refunds as a proportion of total transactions.
    - Operating System Type: Proportion of users utilizing Android devices.
    - Device Type: Understanding the number of users accessing the platform via mobile phones.
    - Device Age: Proportion of users with new devices.

All these metrics were compiled into a structured DataFrame named user_features, enabling further analysis and segmentation of users based on their behavior, preferences, and engagement levels. This information is crucial for informing business strategies, enhancing customer engagement, and optimizing marketing efforts.

Feature Engineering

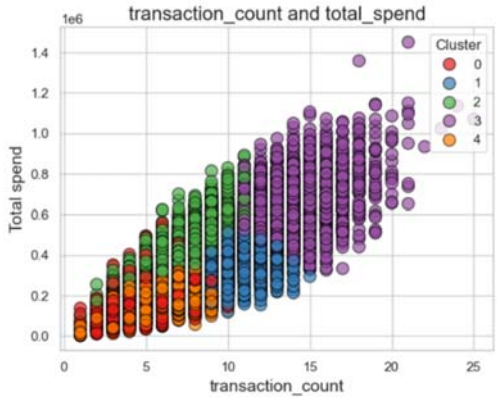
Income Level

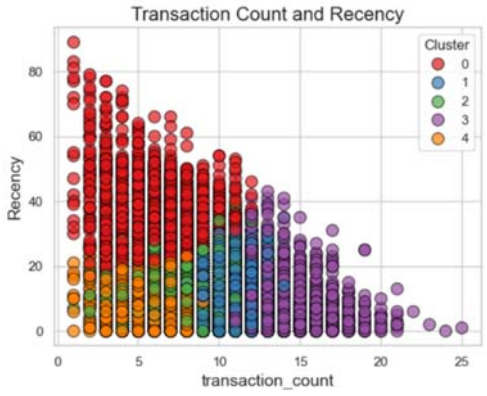
I conducted clustering analysis by performing Principal Component Analysis (PCA) to identify the optimal number of components based on the cumulative sum variance graph. I determined that the best number of components is 8.

The following features were identified as the most important based on their coefficient variance:

Feature	**Coefficient**
Transaction Count	0.500350
Total Spend	0.447805
Promo Usage	0.437287
Food and Beverage	0.396544
Recency	0.319332
Retail	0.222093
Total Promo	0.126128
Average Transaction	0.110822
Home and Lifestyle	0.098244
Health and Wellness	0.095840

Recognizing the significance of Transaction Count and Total Spend as the top two features, I decided to utilize these metrics to define income levels for clustering. I then applied the K-Means algorithm and, based on the elbow method score, concluded that the data should be clustered into 5 groups.





```
# Explore how each feature behaves in the clusters
feature_columns = ['Food and Beverage', 'Health and Wellness', 'Home and Lifestyle',
       'Retail', 'Entertainment', 'Education', 'Travel', 'recency',
       'transaction_count', 'total_spend', 'avg_transaction','preferred_payment']
cluster_analysis = merged_df.groupby('cluster_1')[feature_columns].median()
cluster_analysis
```

✓ 0.0s Python

cluster_1	Food and Beverage	Health and Wellness	Home and Lifestyle	Retail	Entertainment	Education	Travel	recency	transaction_count	total_spend	avg_transaction	preferred_payment
0	3.0	1.0	0.0	1.0	0.0	0.0	0.0	32.0	6.0	250100.0	40433.333333	0.800000
1	6.0	1.0	0.0	2.0	0.0	0.0	0.0	6.0	10.0	375300.0	37000.000000	0.800000
2	5.0	1.0	0.0	2.0	0.0	0.0	0.0	8.0	8.0	441800.0	55416.666667	0.800000
3	8.0	1.0	0.0	3.0	0.0	0.0	0.0	4.0	13.0	601800.0	46613.333333	0.812500
4	4.0	1.0	0.0	1.0	0.0	0.0	0.0	7.0	7.0	222950.0	34269.047619	0.833333

1. Cluster 0

- Spending Patterns: Moderate total spend (250,100) with an average transaction of 40,433.33.
- Categories: Higher spending in Food and Beverage (3) and Retail (1) suggests a balanced lifestyle with some focus on dining out.
- Income Level: This cluster might represent a lower-middle to middle-income level, as they are spending moderately but not extensively across categories.

2. Cluster 1

- Spending Patterns: Total spend (375,300) and an average transaction of 37,000.
- Categories: Similar to Cluster 0, with higher transactions in Food and Beverage (6) and Retail (2).
- Income Level: Likely represents a middle-income level, as their spending indicates more disposable income for discretionary purchases.

3. Cluster 2

- Spending Patterns: Total spend of 441,800, with an average transaction of 55,475.00.
- Categories: Focused on Food and Beverage (5) and Retail (2), indicating a preference for quality over quantity.
- Income Level: This cluster appears to represent a middle to upper-middle-income level, with a willingness to spend more per transaction, suggesting higher disposable income.

4. Cluster 3

- Spending Patterns: Highest total spend (601,800) and an average transaction of 46,613.33.
- Categories: Highest activity in Food and Beverage (8) and Retail (3).
- Income Level: Likely indicates an upper-middle to upper-income level, reflecting significant discretionary spending and a lifestyle that values experiences and products.

5. Cluster 4

- Spending Patterns: Lower total spend (222,950) with an average transaction of 34,271.43.
- Categories: Moderate spending in Food and Beverage (4) and Retail (1).
- Income Level: This cluster could represent a lower-middle-income level, as their total spend and transaction frequency suggest a more cautious approach to spending.

Summary

- Lower-Middle Income: Cluster 0 and Cluster 4

- Middle Income: Cluster 1

- Middle to Upper-Middle Income: Cluster 2

- Upper-Middle to Upper Income: Cluster 3

Vehicle Ownership

In the user_features data, code 5541 refers to "Service Stations (with or without ancillary services)," typically including establishments like gas stations that may offer fuel, convenience items, and services like car washes or repairs. We can therefore infer that users in this category likely have vehicle ownership.

Age

I categorized age groups following the statistics from OJK (Otoritas Jasa Keuangan) on financial literacy and inclusion rates across different age groups in Indonesia:

Age Group (Years)	**Financial Literacy (%)**	**Financial Inclusion (%)**
15-17	51.70	57.96
18-25	70.19	79.21
26-35	74.82	84.28
36-50	71.72	81.51
51-79	52.51	63.53

Additionally, I referenced information from *Kontan* indicating that the majority of Dana's users are in the **25-34** age group.

Process Overview:

1. **Data Preparation:** Created a DataFrame with the following key features:
    - transaction_count
    - recency
    - total_transaction for each merchant
2. **PCA Analysis :** Similar to previous clustering, I ran a PCA on the dataset to reduce dimensionality and identify key components.
3. **Clustering :** I plotted a scatter plot comparing transaction_count and recency. From the analysis, the optimal number of clusters was found to be **5**, representing the different age groups.

Cluster Mapping:

Each cluster is mapped to an age group:

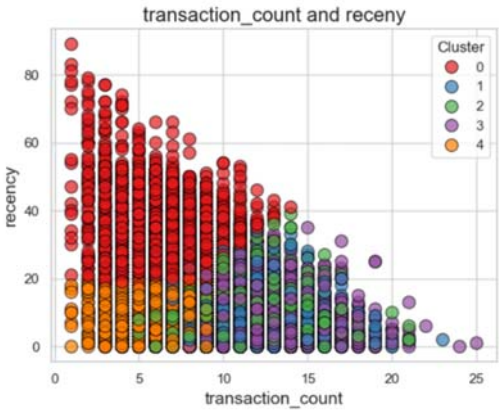
- Cluster 0: Age group **51-79**
- Cluster 1: Age group **18-25**
- Cluster 2: Age group **15-17**
- Cluster 3: Age group **26-35**
- Cluster 4: Age group **36-50**

```
# Explore how each feature behaves in the clusters
feature_columns_age = ['recency', 'transaction_count', 'preferred_payment',
       '4722', '5251', '5311', '5411', '5412', '5541', '5651', '5712',
       '5732', '5812', '5814', '5912', '5941', '5942', '5964', '5999', '7230',
       '7994', '8211', '8220']
cluster_analysis = df_group_age.groupby('cluster_age')[feature_columns_age].mean()
cluster_analysis
```

✓ 0.0s

cluster_age	recency	transaction_count	preferred_payment	4722	5251	5311	5411	5412	5541	5651
0	32.386868	6.433724	0.801278	0.099877	0.183107	0.147041	0.595561	0.305795	0.302096	0.098644
1	7.646859	10.557042	0.800846	0.146972	0.258789	0.195963	0.810841	0.404173	0.431617	0.136085
2	7.575861	10.339849	0.800141	0.152700	0.271683	0.192280	2.444906	0.419497	0.425116	0.130711
3	6.647362	11.493653	0.801830	0.174931	0.338754	0.241571	0.580127	0.595795	0.598969	0.153511
4	7.725209	7.005443	0.796361	0.123722	0.236957	0.169919	0.428780	0.365060	0.353246	0.119342

Based on the feature behavior table, I decided to compare recency, transaction count, and preferred payment. My assumption is that a higher recency value indicates older users, while a higher transaction count suggests younger, more tech-savvy users. For preferred payment, values closer to zero imply a greater use of credit cards. Young generations typically do not have access to credit cards, whereas older generations often rely on them to help finance their needs. Using these assumptions, I created the cluster mapping as described above.



Education Level

I performed clustering using the features income_level, age, and merchant_id that has a high probability of defining this feature, and the results are categorized into two broad education categories: **High School** or **University**.

Clustering Approach:

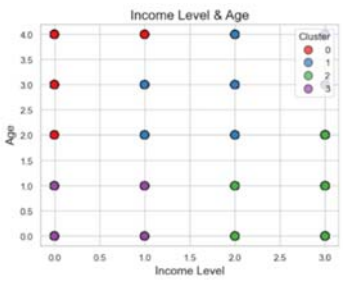
The clustering process resulted in 4 clusters. However, based on the analysis, I simplified it into two main categories for clarity:

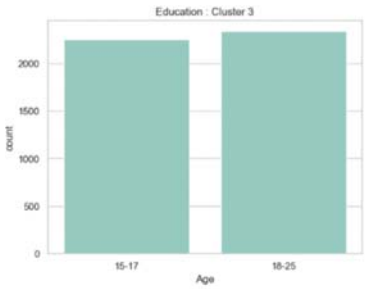
1. High School Students: Clusters with younger age groups, typically 15-17 or 18-25.
2. University Students: Clusters with older individuals, typically ranging from 26-35 and above.

Analysis:

When plotting the clusters using scatter plots and bar charts, a clear pattern emerged:

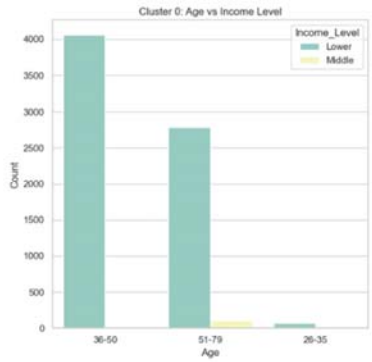
- Cluster 3 has a high probability of being high school students, as their age consistently falls in the 15-25 range.
- When visualizing this in a bar plot, Cluster 3 is the only group where the age ranges between 0-1, strongly indicating that it represents high school students.

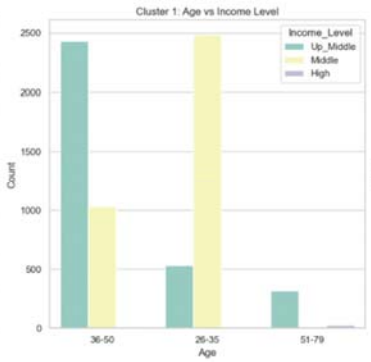


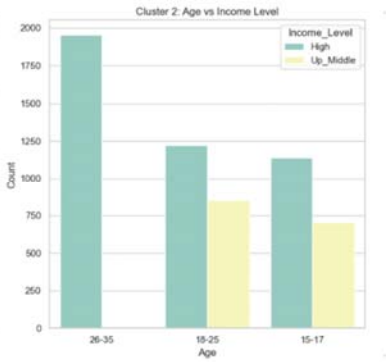


o The remaining clusters have a broader range of ages, from 0-4, likely representing university students or older.

By using these insights, I classified each cluster into either High School or University categories, simplifying the overall education-level analysis.







Homeownership

For the homeownership analysis, I used the features income level, age, and two specific merchant IDs: 5712 and 5251. These merchant IDs are relevant because they represent categories commonly associated with home-related purchases:

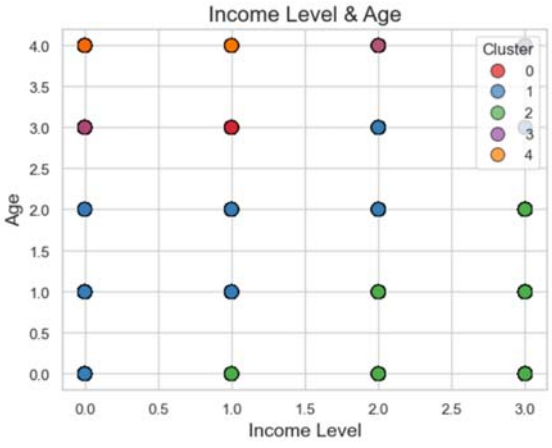
- 5712: Refers to "Furniture, Home Furnishings, and Equipment Stores."
- 5251: Refers to "Hardware Stores."

These features were chosen as indicators of potential homeownership-related spending.

Homeownership Insights:

Based on the clustering analysis using income level, age, and the specific merchant IDs (5712 and 5251 related to home-related purchases), I derived the following insight:

- Lower income individuals (income level 0-1) tend to purchase or own homes later in life, typically when they are older than 35 years (age group 3-4).
- In contrast, higher income individuals (income level 2-4) have a higher likelihood of purchasing or owning homes at a younger age, even before they reach 35 (age group 1-3).



This suggests that higher income provides more opportunities for early homeownership, while lower-income individuals often need to wait until later in life.

Gender

This process uses user transaction patterns across various merchant categories to perform gender classification. Here are the key steps involved:

1. Data Standardization

   The transaction data is standardized using RobustScaler to handle outliers and normalize the dataset, making it ready for analysis.

2. Merchant Category Weights

   Predefined weights are assigned to each merchant category based on assumed spending preferences of male and female users. For example, Electronics Stores (MCC 5732) are weighted 0.7 for males and 0.3 for females, while Beauty and Barber Shops (MCC 7230) have a weight of 0.2 for males and 0.8 for females.

3. Score Calculation

   For each transaction, both a male score and a female score are calculated by multiplying the transaction amount by the corresponding gender weight for that merchant category.

4. User-Level Aggregation

   These individual transaction scores are then aggregated for each user.

5. Gender Classification

   Users are classified as either male or female based on whichever aggregated score (male or female) is higher.

The final output is a DataFrame containing user IDs, their respective male and female scores, and their predicted gender based on transaction behavior.

Parental Status

To define the parental status of users, the process follows a similar approach to gender classification, using specific merchant category IDs combined with user features like income level and age. Here's how the process works:

Key Steps in the Process:

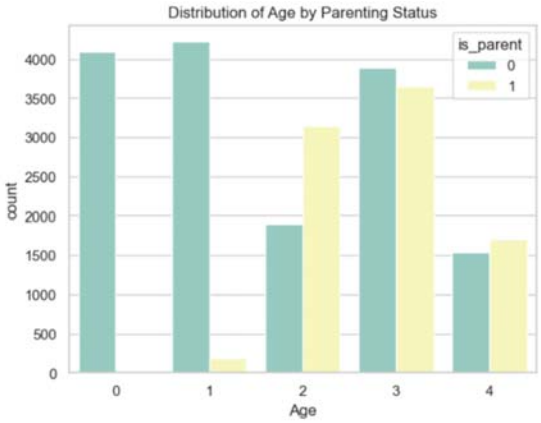
1. Age-Based Weights:
   The code assigns higher weights to users based on their age group:
   - Age 0: No score (assumed not to be parents).
   - Age 1 (18-25 years old): A score of 1 (low probability of being parents).
   - Age 2, 3, and 4 (>2 years old): A score of 2 (higher probability of being parents).
2. Merchant Category Code (MCC) Transaction Weights:
   Transactions at specific merchants related to family needs are also considered:
   - Grocery Stores (MCC 5411): Weighted at 0.5.
   - Furniture Stores (MCC 5712): Weighted at 0.3.
   - Family Clothing Stores (MCC 5651): Weighted at 0.2.
   - Colleges/Universities (MCC 8220) and Elementary/Secondary Schools (MCC 8211): Both weighted at 0.4 (indicating possible education expenses).

- Hardware Stores (MCC 5251): Weighted at 0.1.

Final Score and Parental Status Classification:

After calculating these scores based on the user's transactions, age, and income level, the total score is evaluated. The **central tendency** (mean or median) of the scores is used as a threshold. Users with a score **above the central tendency** are classified as having a higher probability of being parents.

The process combines user age, income level, and specific spending patterns to estimate the likelihood of a user being a parent based on a calculated score.



Big Five Personality

Approach to Defining Personality Traits

1. **Openness**:
   Individuals high in openness tend to be curious, seek new experiences, and appreciate variety in their lives. To identify users with high openness, I count the number of unique merchant category IDs they interact with. If a user has a total number of unique merchants greater than the central tendency, they are classified as very open and experienced.

2. **Conscientiousness**:
   Conscientious individuals are known for their self-regulation and impulse control. This trait affects their ability to set and maintain long-term goals, deliberate over choices, and fulfill obligations. Users exhibiting high conscientiousness are likely to be more aware of discounts and may actively seek them out as part of their careful planning and financial management strategies.

3. **Extraversion**:
   Extraversion measures how energetic, sociable, and friendly a person is. To assess this trait, I create a pivot table that counts transactions in specific merchant categories (MCCs):

   - **5812**: Eating Places and Restaurants
   - **5814**: Fast Food Restaurants
   - **5941**: Sporting Goods Stores
   - **5411**: Grocery Stores, Supermarkets
   - **5311**: Department Stores
   - **5999**: Miscellaneous and Specialty Retail Stores
   - **5661**: Shoe Stores
   - **7230**: Beauty Shops
   - **4722**: Travel Agencies and Tour Operators
   - **7944**: Charitable and Social Service Organizations
     Users with a transaction count higher than the central tendency are classified as extroverted.

4. **Agreeableness**:
   This trait reflects kindness and compassion toward others. To evaluate agreeableness, I analyze transactions with the features of is_refunded and merchant ratings. Users who have transactions marked as refunded but still give high ratings (e.g., 4 stars) are assumed to be displaying agreeableness, indicating they may not be fully satisfied with the product but choose to give a favorable rating.

5. **Neuroticism**:
   Individuals with high neuroticism scores are more prone to experiencing negative emotions such as anxiety, worry, and frustration. To identify neurotic users, I calculate the standard deviation of each user's transaction amounts. Users with a standard deviation higher than the average are categorized as neurotic.

Health Status

To define **health status**, I focus on user transactions at pharmacies, specifically analyzing the recency of their transactions at merchant category code **5912** (Drug Stores and Pharmacies). Here's a detailed breakdown of the process:

Steps to Define Health Status

1. **Data Manipulation**:
   - Analyze the transaction data to identify the **last transaction date** and the **count of transactions** for each user at the merchant categorized as **5912**.
   - Extract relevant data for each user, including the last transaction date and the total number of transactions made at this merchant.
2. **Central Tendency Calculation**:
   - Calculate the **central tendency** (mean or median) for both recency (days since the last transaction) and transaction count at the pharmacy.
   - This will provide a baseline to understand typical user behavior and identify outliers.
3. **Classification**:
   - A user is classified as **unhealthy** (coded as **1**) if they meet the following criteria:
     - They have made a transaction at merchant **5912** within the **last 17 days**.
     - They have completed **more than one transaction** at this merchant.
   - Users who do not meet these criteria are classified as **healthy** (coded as **0**).

Summary

This approach allows for a data-driven assessment of users' health status based on their purchasing behavior related to pharmacy transactions. By focusing on recency and transaction frequency at drug stores, I can identify users who may require health-related support or intervention based on their shopping patterns. This classification can help tailor health programs, promotional offers, or targeted communication to encourage healthier choices or increase awareness of health resources.

Home and Work Location

The dataset provides latitude and longitude for each transaction. Using this data, I aim to predict the home and work locations for each user. Here's my approach:

First, I utilize the transaction_date feature to determine whether a transaction occurred on a weekday or a weekend. I create a new column where weekend = 1 and weekday = 0.

Next, I group the data by clusters based on the is_weekend column and calculate the mean value for each group. If the transactions occur predominantly on weekdays, it's likely the location represents their workplace. If the transactions mostly occur on weekends, it's likely to be their home or somewhere near their home.

Result

The dataset contains a feature indicating is_weekend, where 0 represents weekdays and 1 indicates weekends. The mean latitude and longitude were calculated based on this feature to represent home and work locations.