Medicine Database Optimization and Clustering

Objective

This project aims to preprocess a medicines dataset, handle missing values and inconsistencies, and prepare it for clustering analysis. The steps include data cleaning, filling in missing values, checking for duplicates, and standardizing column names. Finally, the cleaned dataset is used for clustering.

Data Preparation and Preprocessing

1. Initial Dataset Information

- Dataset Shape: (1448, 10)
- Duplicate Rows: None found in the dataset.

2. Column Cleaning

- Renamed columns to lowercase for consistency.
- Removed special symbols and spaces from column names (e.g., name, packing_form, salts, manufacturer, etc.).

3. Handling Missing Values

- Manufacturer Column:
    - It contained 2 missing rows, which were dropped due to negligible impact.
- Retail Price and Discount Price Columns:
    - Used the median to fill missing values as it is robust against outliers.
- Packing Form Column:
    - Filled missing values using references from the name column.
- Salts Column:
    - Missing values were imputed using the mode, which is optimal for categorical data.

4. Additional Modifications

- Quantity Column:
    - Updated the column to include measurement units such as "ml" and "tablet" for better clarity and usability.
- Added Column:
    - Created quantity_numeric column by parsing numeric values from the quantity column.

5. Dropped Columns

- Dropped the quantity_numeric column as its information was incorporated into the updated quantity column.

6. Final Dataset

- Stored the updated dataset as updated_dataset_final.csv.

Clustering Analysis

The dataset was prepared for clustering using the K-Prototypes algorithm, designed for mixed numerical and categorical data. Below is the summary of the workflow:

1. Dataset Size After Cleaning:

- Shape: (1448, 9)

2. Preprocessing

- Feature Selection:
    - Selected features: name, salts, packaging_form, quantity, retail_price, discounted_price, and manufacturer.
- Categorical Data Handling:
    - Columns salts, packaging_form, and manufacturer were converted to string format for encoding.
- Quantity Parsing:
    - Parsed numeric values and units (e.g., "500 ml" → 500). Invalid or missing entries were replaced with NaN and removed.
- Normalization:
    - Normalized numerical columns (retail_price, discounted_price, quantity_numeric) using MinMaxScaler.

3. Clustering Process

- Algorithm: K-Prototypes
    - Initialization: Huang’s method.
    - Random Seed: 42.
    - Number of Clusters (k): 5.
- Data Preparation:
    - Combined normalized numerical features and encoded categorical features.
    - Converted categorical features to indices.
- Cluster Assignment:
    - Each data point was assigned to one of five clusters.

4. Results

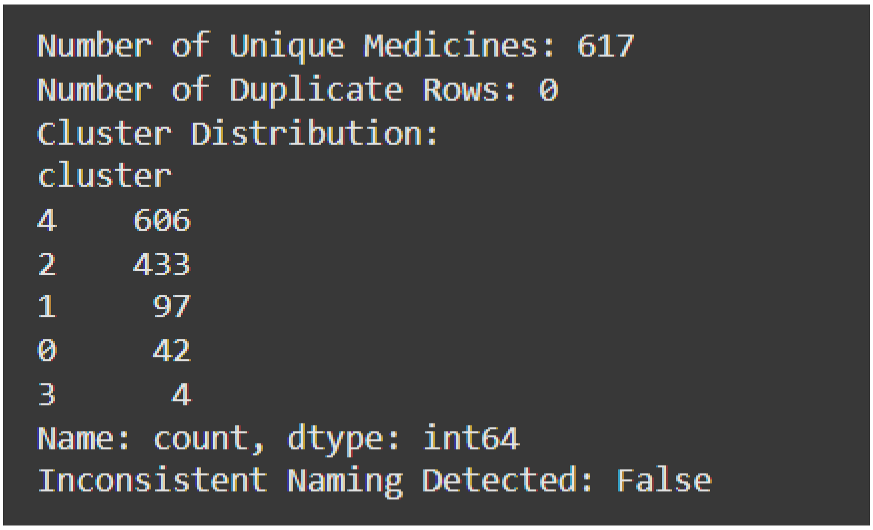
- The clustered dataset was saved as clustered_medicines_final.csv.

- Cluster summaries were generated to review sample data points in each cluster.

5. Upload to Google Sheets

- The clustered dataset was successfully uploaded to Google Sheets as "Medicine Data1."
- The data includes all clustering results for further analysis and sharing.

Final Data Analysis



```
Number of Unique Medicines: 617
Number of Duplicate Rows: 0
Cluster Distribution:
cluster
4    606
2    433
1     97
0     42
3      4
Name: count, dtype: int64
Inconsistent Naming Detected: False
```

Summary

This project successfully preprocessed and clustered a medicines dataset using K-Prototypes, resulting in meaningful groupings of medicines based on salts, packaging form, quantity, and pricing. The results can help optimize inventory management, marketing strategies, and product analysis.