Project Report: Medicine Clustering
Objective: To cluster similar medicines based on their name, salts, and prescription requirements.

Methodology:

1. **Data Loading and Preprocessing:**

- Loaded data and performed basic data exploration (describe, info, unique values, null values) using pandas.
- Extracted numerical quantity from the 'quantity' column.
- Handled missing values using KNN imputation for numerical features and filled categorical features with 'Unknown'.
- Preprocessed text data (name and salts) using tokenization, stop word removal, and lemmatization using nltk.

2. **Feature Engineering:**

- Experimented with TF IDF, fuzzywuzzy and huggingface for treating textual data then finalised on using huggingface embeddings.
- Generated embeddings for medicine names and salts using SentenceTransformer ('all-MiniLM-L6-v2').
- Scaled the embeddings using RobustScaler.
- Combined embeddings, prescription requirement (encoded as 0 or 1) features into a final feature set.

3. **Dimensionality Reduction:**

- Experimented with dimensionality reduction techniques such as PCA, TSNE and UMAP
- Applied UMAP to reduce the dimensionality of the feature space for visualization and clustering.
- Experimented with UMAP parameters (n_neighbors, min_dist, metric) for optimal clustering.

4. **Clustering:**

- Performed hyperparameter tuning using GridSearchCV with silhouette score as the metric.
- Experimented with different clustering algorithms, finalising Hierarchical DBSCAN based on a 0.70 silhouette score

5. **Cluster Analysis:**

- Analyzed cluster sizes and identified the most frequent medicine within each cluster.
- Created a cluster summary table including cluster size, number of unique manufacturers, and the most frequent medicine.
- Visualized the clusters using scatter plots UMAP embeddings and a bar graph showing number of entries per cluster.

6. **Database design**

    - Saved the updated csv file with the imputed values and a cluster summary with the cluster number, count of members and the most frequent medicine.
    - Then designed a schema for the efficient storage and querying of the database using PostgresQL known for its scalability and being SQL based optimising querying

Issues Encountered:

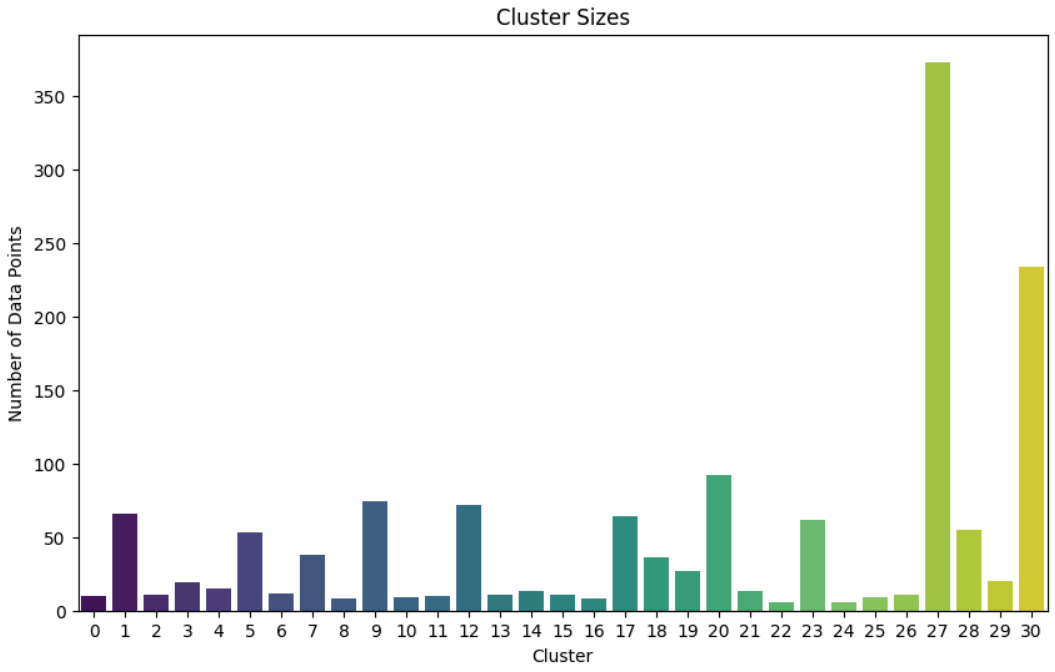
1. **Feature Selection and Engineering**: Selecting the best features for clustering and then using the correct feature engineering techniques was important. Tried TF IDF and fuzzywuzzy matching but finalised on Huggingface (Transformers) embeddings for capturing semantic relationships between the features.

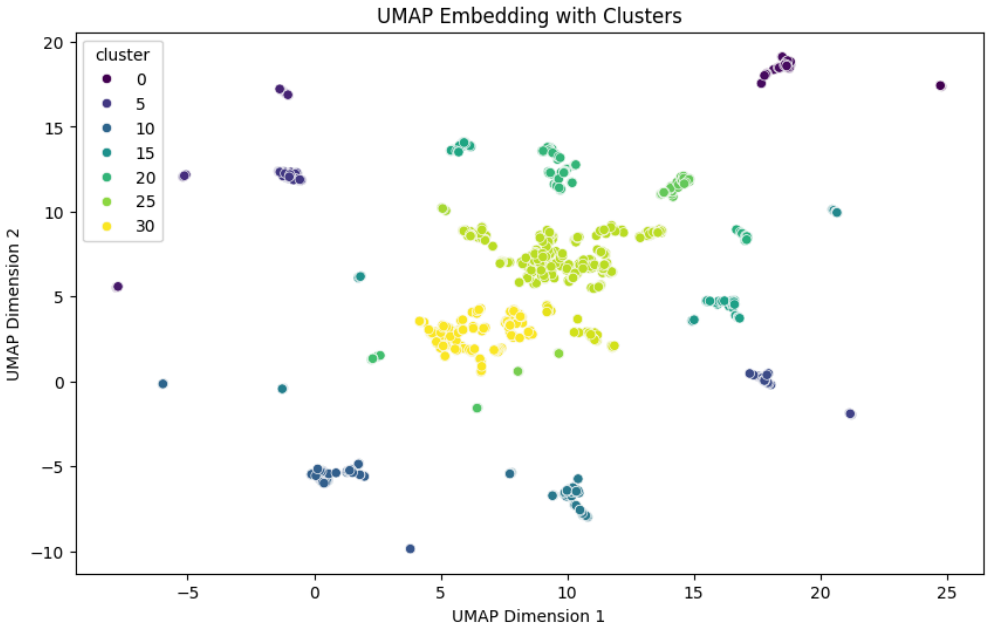
2. **Dimensionality Reduction:** Using PCA and TSNE didn't yield optimum results, so explored more techniques such as UMAP and tweaked with its parameters(n_neighbours and min_dist and metric) to get best results .

3. **Clustering Algorithm Selection and Hyperparameter Tuning**: Choosing the appropriate clustering algorithm and optimizing its hyperparameters significantly influence the quality of clusters. Initially struggled with outliers and incorrect clustering with KMeans and DBSCAN, so went with Hierarchical DBSCAN for outlier protection and optimum clustering. Still the issue of some clusters being more crowded than the others(27 and 30 below) can be improved.

Results

Number of clusters found: 30
Best Silhouette score: 0.70





Potential Improvements:

1. Explore different imputation strategies for missing values, such as using domain knowledge or more advanced imputation techniques.

2. Experiment with alternative text preprocessing methods and feature engineering techniques to better capture medicine similarity.

3. Evaluate different clustering algorithms and explore a wider range of hyperparameters.

4. As per the new clustering and data values, better database structures can be used

5. Incorporate external knowledge or data sources to validate and interpret the clusters.