Exploring Apriori Association Analysis and K-Means Clustering

Arshia Sharifi - 501158323

Background

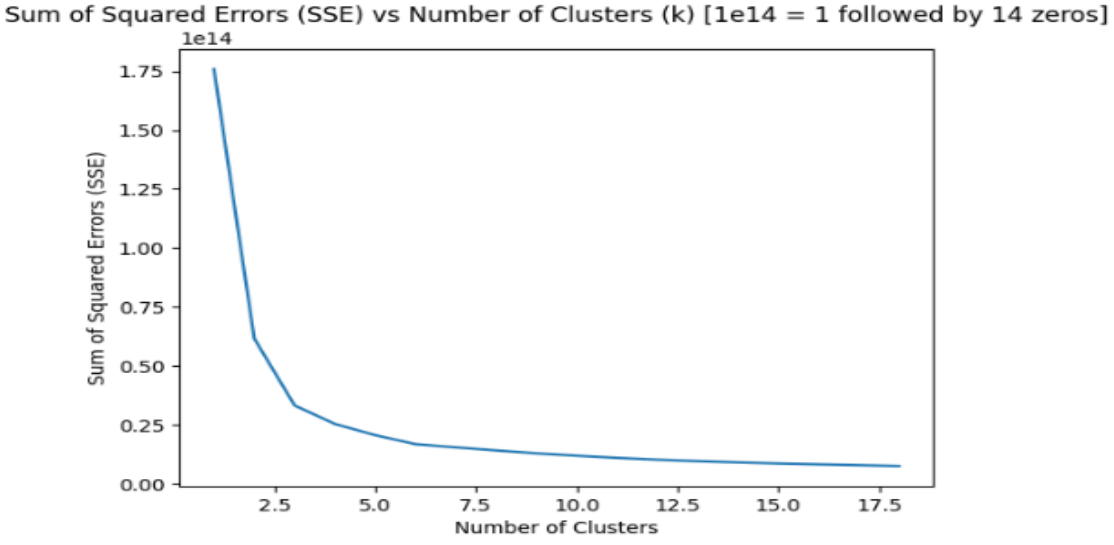
In this study, I explored two fundamental data analysis techniques: association analysis using the Apriori algorithm and clustering analysis using the K-Means algorithm. I leveraged two distinct datasets obtained from Kaggle to address practical problems through these techniques.

For the clustering analysis, I utilized the Toronto Home Price Index dataset from the MLS (Multiple Listing Service), sourced from Kaggle [1]. This dataset consists of home sale records and associated attributes such as property features, prices, and locations for the year 2021. The dataset contains 5093 records and 17 columns.

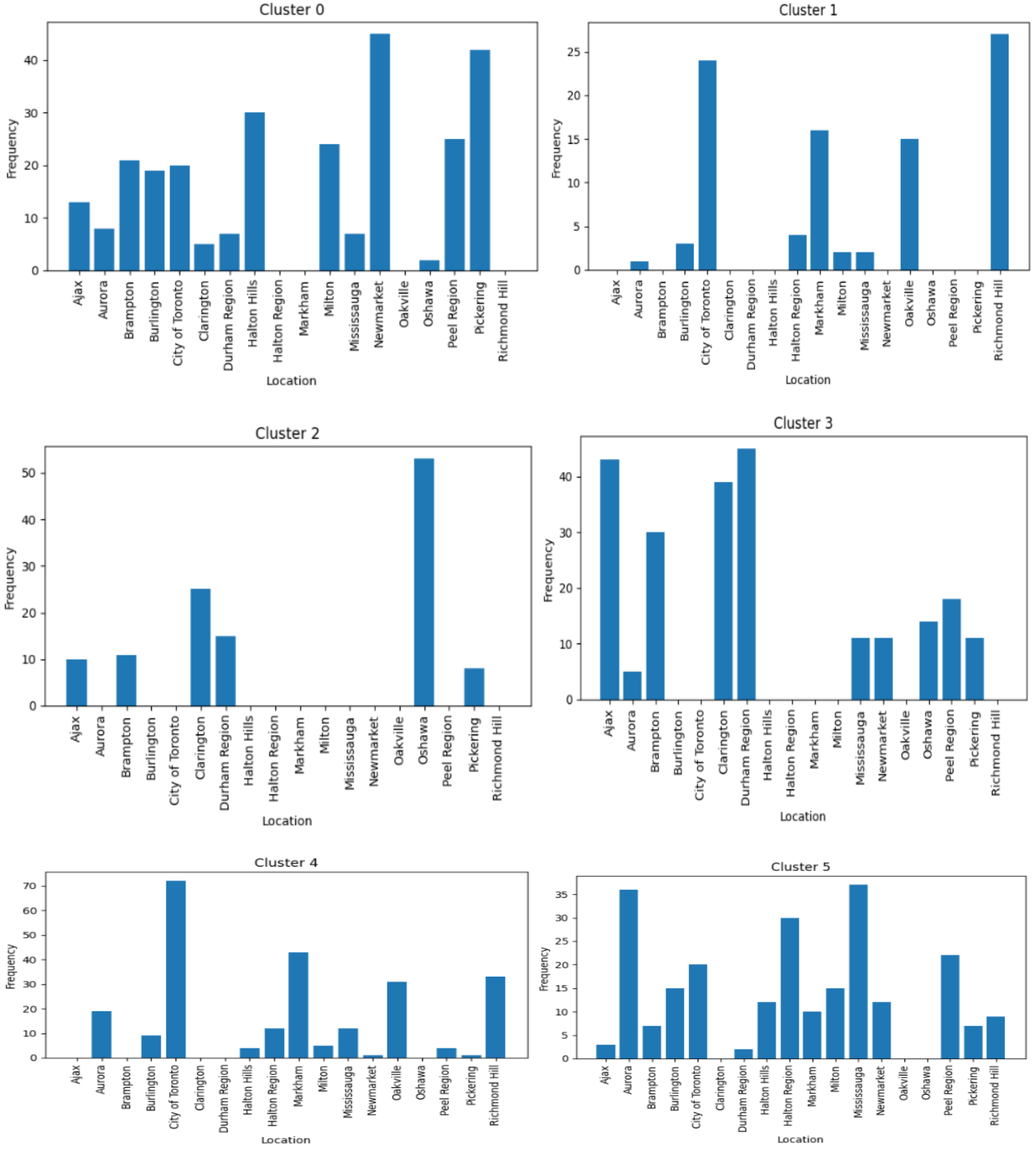
For association analysis, I employed the Groceries Market Basket Dataset, also sourced from Kaggle [2]. This dataset contains 9835 transactions by customers shopping for groceries, comprising a total of 169 unique items. Each transaction represents a list of items purchased by a customer. The categorical transaction data in this dataset is ideal for uncovering patterns such as item associations and frequent itemsets. By analyzing customer purchase patterns, businesses can derive valuable insights to optimize product placement, promotions, and cross-selling strategies.

Methods & Results - Clustering

Firstly, I aimed to identify the optimal number of clusters by plotting a graph to visualize the relationship between the number of clusters (K) and the inertia values obtained from the K-Means algorithm. The inertia measures the sum of squared distances of samples to their closest cluster center. Upon plotting the graph, I observed that the inertia values begin to decrease linearly at k = 6, hence why it was selected as the elbow or the optimal cluster number.



Next, I iterated through each unique KMeans cluster value to plot bar charts illustrating the distribution of the classes. The attribute used for this analysis was the "Location," aiming to identify the hottest city in terms of housing prices. However, to reduce the complexity of computation and avoid overlapping regions, certain cities were manually dropped, particularly partitions of Toronto.



Methods & Results - Association

For association analysis using the Apriori algorithm, the minimum support was set to 2%, and the minimum confidence to 45% after experimentation to determine optimal values. These parameters were selected based on random manual experimentations to find the most optimal values.

One discovered association suggested that customers purchasing butter are likely to buy whole milk as well, indicating a common shopping pattern. The 'Items' attribute containing numeric values was dropped as it was incompatible with the Apriori algorithm, which requires categorical data.

```
Rules:
['butter'] --> ['whole milk']
Support: 0.02755465175394001 Confidence: 0.4972477064220184 Lift: 1.9460530014566455

['curd'] --> ['whole milk']
Support: 0.026131164209456024 Confidence: 0.4904580152671756 Lift: 1.9194805332879712

['domestic eggs'] --> ['whole milk']
Support: 0.029994916115912557 Confidence: 0.47275641025641024 Lift: 1.8502026640954214

['other vegetables', 'root vegetables'] --> ['whole milk']
Support: 0.023182511438739197 Confidence: 0.4892703862660944 Lift: 1.9148325702057454

['yogurt', 'other vegetables'] --> ['whole milk']
Support: 0.02226741230299949 Confidence: 0.5128805620608898 Lift: 2.0072345116867694
```

Conclusions

In the clustering analysis, the elbow method suggested selecting 6 clusters despite the presence of 15 distinct classes in the dataset. This discrepancy indicates that some classes may be underrepresented in the clusters due to the limitation of choosing a smaller number of clusters based on the elbow method. As a complement to the clustering analysis, a plot was generated for each unique cluster to illustrate the distribution of how each city compares to the overall pattern determined by the cluster.

Regarding association analysis, it's important to note that this method is practical only for categorical attributes and may encounter challenges when working with numeric or continuous data. One notable aspect of this analysis is the usage of the Apriori algorithm to uncover associations among grocery items. By setting appropriate minimum support and minimum confidence thresholds, various rules were discovered, such as the association between purchasing butter and buying whole milk.

References

[1] https://www.kaggle.com/datasets/alankmwong/toronto-home-price-index

[2] https://www.kaggle.com/datasets/irfanasrullah/groceries

[3] https://realpython.com/k-means-clustering-python/