Data Mining - Final Project, Matwej Novgorodtsev, ID: 244580IV

Project description

This project demonstrates the application of skills acquired during the course to practical solutions addressing real-world challenges. For this purpose, two datasets were selected. The first dataset pertains to sales data from a furniture store, which was used to create a linear regression model to predict store revenue and to perform association pattern mining to identify which items are most frequently sold together.

The second dataset contains data from a shopping mall, including customer details such as Gender, Age, Annual Income (in thousands of dollars), and Spending Score (ranging from 0 to 100). The dataset provides insights into the spending habits of different genders and age groups, making it valuable for profiling and clustering analysis. The goal was to perform clustering on these customers.

1 Furniture store dataset

1.1 Data preprocessing

Data from this dataset was presented in form like this. Column nr14 was added by me by grouping sales of all items in each month of each year to use it in linear regression model later.

Num	Column names
1	Row.ID
2	Order.ID
3	Order.Date
4	Ship.Mode
5	Customer.ID
6	Country
7	City
8	Product.ID
9	Category
10	Sub.Category
11	Product.Name
12	Sales
13	Quantity
14	Month_Year

Table 1: Data presentation in dataset

1.2 Assocation pattern mining

To explore patterns within the dataset, we initially focused on analyzing the relationships between Product.Name and Sub.Category. The process began with data preparation, where product names were cleaned and standardized for consistency. Due to the very long and specific names of furniture items, a unique ProductID was created to simplify visualization and analysis. Transactions were then created by grouping products under each Order.ID, forming a structured dataset for mining.

The Apriori algorithm was applied with parameters set to a minimum support of 0.0001 and confidence of 0.7, ensuring the discovery of significant and reliable association rules. The rules were sorted by lift to highlight the strongest associations. Key results were visualized using graphs and bar charts, providing insights into item relationships and patterns within the dataset.
As shown in the chart, the strongest initial associations are identical, likely indicating that these products are sold together in the same package or promotion.

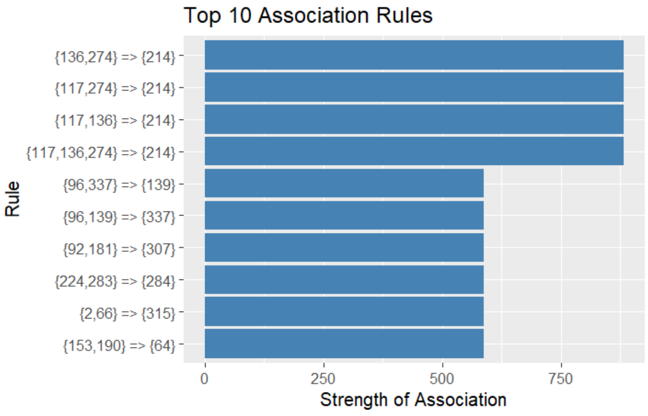

Figure 1: Top 10 stongest assocation rules

Product Name	ID
Global Comet Stacking Armless Chair	117
Tensor Computer Mounted Lamp	136
Eldon Econocleat Chair Mats for Low Pile Carpets	214
Flat Face Poster Frame	274

Table 2: Names of products with strongest assocation

Graph was generated to visually represent the association rules and their relationships within the dataset.

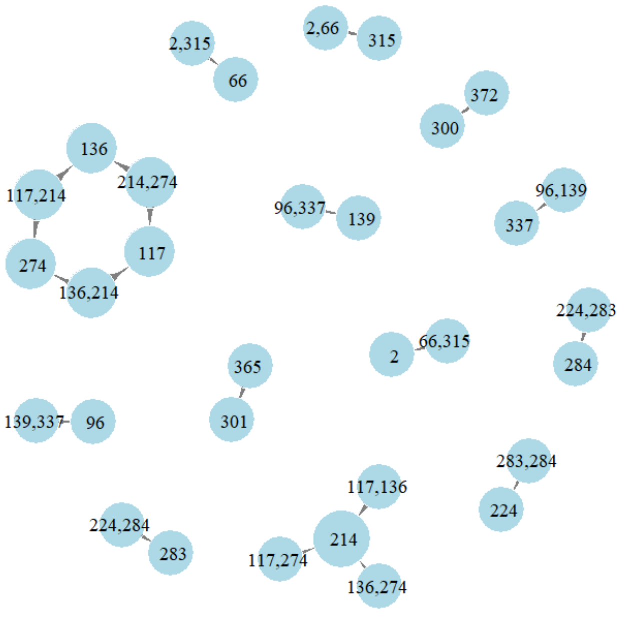

Figure 2: Graph presentation

1.3 Linear regression model

The Month-Year column was created to aggregate sales data, facilitating the preparation of data for building a linear regression model. Given the nature of the data, creating a monthly-based model was deemed the most logical approach.

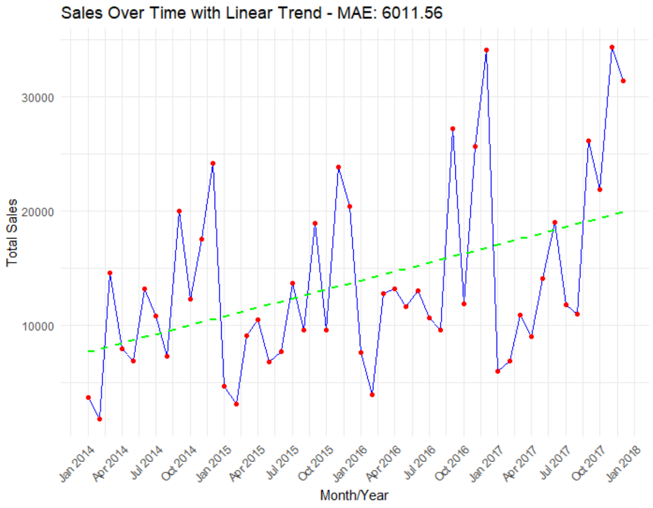

Figure 3: Linear regression model presentation

Additionally, the statsmodels library was added to calculate the MAE. How we can see Linear Model is not the best option to this type of problem, since we have seasonal spikes occurying and non linear models would fit better

2 Mall data customer

2.1 Data preprocessing and outlier detection

In this dataset, the preprocessing phase was already completed, only the "Genre" column was originally represented as a categorical variable with string values "Male" and "Female", was transformed into a binary numerical format with "Male" encoded as 0 and "Female" as 1.

Num	Column names
1	CustomerID
2	Genre
3	Annual_Income_k
4	Spending_Score

Table 3: Data presentation

```
dataframe <- dataframe %>%
  mutate(Genre = ifelse(Genre == "Male", 0, 1))
```

Then all columns were scaled to normalize the data, ensuring that features were on comparable scales to make the outlier detection. In outlier analysis and detection two columns were taken: Annual Income and Spending Score. To detect an outlier DBScan was chosen and tested with different parameters. The final parameters selected for DBSCAN were epsilon = 0.5 and minPts = 8, they were performing the best.

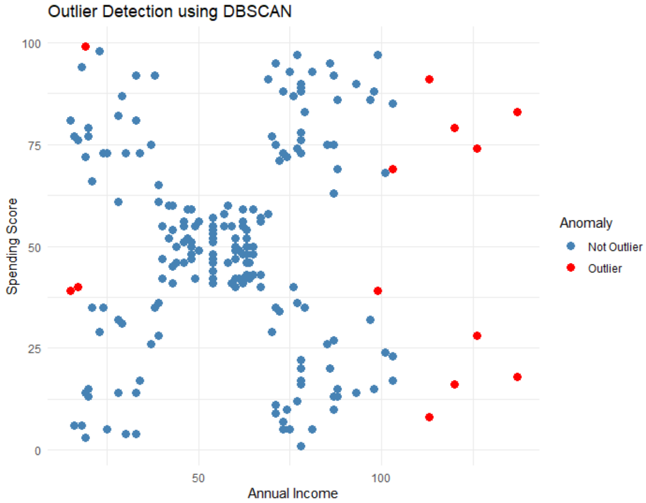

Figure 4: DBScan, eps=0.5, minPts=8

2.2 Clustering

K-means clustering was applied using 3 and 4 centroids to identify the optimal solution. The results indicate that k=3 provides a superior fit, as the silhouette plot demonstrates well-defined clusters without a "knife-like" shape. To mitigate sensitivity to initial centroid seeds, the parameter `n_start`=25 was added, ensuring the algorithm ran 25 iterations and selected the most appropriate solution. The final cluster visualization was presented in a 3D plot to illustrate the data distribution and clustering structure.

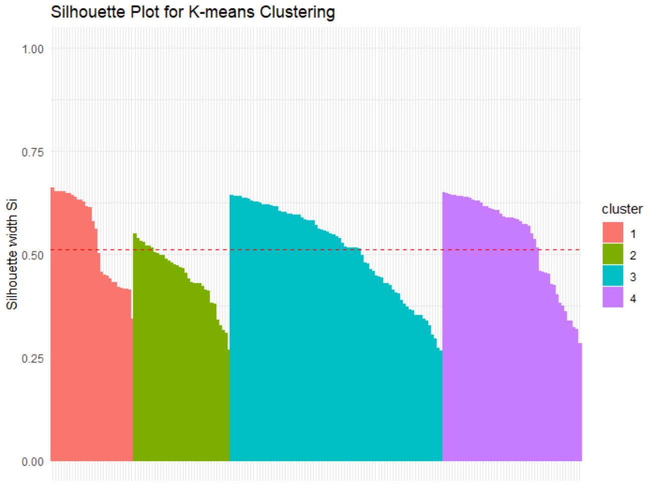

Figure 5: Silhouette coefficient for k=4

Since the clusters initially formed had a long, narrow "knife-like" shape, the number of clusters (k) was adjusted to 3. This change led to better results, shown by a higher silhouette coefficient, meaning the clusters were more distinct and better grouped.

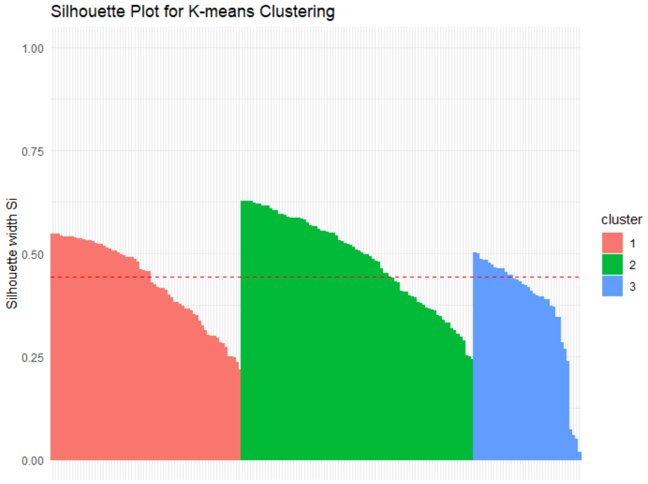

Figure 6: Silhouette coefficient for k=3

The final clusters appear to be well-separated and clearly defined when using k=3. This configuration demonstrates effective grouping of data points, minimizing overlap between clusters.

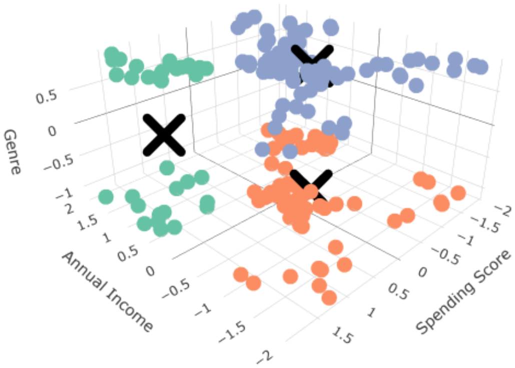

Figure 7: Clustering for k=3