

Group 2

Eugenia Cozma, 13016

eugenia.cozma26@gmail.com

Francisco Lourenço, 13018

Fmplourenco94@gmail.com

João Valente Martins, 13020

Joao.valente.work@gmail.com

**ABSTRACT**

This project has the main purpose of exploring unsupervised learning in 3 distinct datasets with different business coverage – bike sharing, bakery industry and online groceries. More specifically, two different clustering algorithms (hierarchical and partitioning) were applied, as well as pattern and association rules mining.

All datasets were submitted to preprocessing steps such as feature engineering, encoding of categorical features and one-hot encoding, outlier and feature relevance/redundancy studies, before implementing non supervised descriptive learning.

In all datasets we clustered using hard clustering approaches, specifically K-Means and Agglomerative. In dataset 1, using k-means, three clusters were found, mostly described by ride duration, ride distance and weather conditions. In dataset 2, three primary clusters were found, mostly described by the baked goods customers bought, and how much they’ve spent in the bakery in that purchase, as well as datetime preferences, parallel to dataset 3. Lastly, in dataset 3 we were able to come up with three clusters through K-Means that were able to explain differences in user buying behavior regarding buying frequency, quantity.

Lastly, we applied pattern mining with association rules. In dataset 1, several frequent item sets were found with good support mainly combining ride day and weather conditions. However, none was unexpected or relevant for comparison between member and casual user rides. In the second dataset, frequent patterns were mostly related to customers who buy jam that are more inclined to also buy bread. Finally, in dataset 3, similarly to the second dataset, a market basket analysis was done, and most top patterns were around fruits and vegetables which is not surprising as it is a popular essential good. However, it was interesting to find some associations between some specific fruits and dairy products like yogurt or packaged cheese.

# DATA PROFILING

Dataset 1 has bike rides data from Citi bike sharing system in New York. It contains information on the rides such as bike type (electrical/classic), user type (member/casual) and start/end stations and time.

Dataset 2 is related to a South Korean small-sized bakery, where each observation is the invoice of a sale made through a digital platform. There is information about the “total” amount spent in that sale, what the customer bought (product) and respective quantity, the address of the customer, and the datetime of the transaction.

Dataset 3 represents Instacart’s online groceries order ids with Product Detail, where each observation is the product id purchased by user id in that order id (basket). There is information about order of the purchased items as well as if it was (re)ordered previously by the user, days since the last order and order id’s day of week and time.

## Statistical analysis

All datasets received dimensionality, variable type, missing value and distribution research. All datasets have datetime features as well as numeric and symbolic that can be used for feature engineering. All datasets have particular features with impactful outliers that require special treatment for descriptive learning. Dataset 1 and 3 reach millions of observations while Dataset 2 has a few thousands. Aggregation can also be obtained do clustering.

Dataset 1: ~3 million observations, 21 initial variables, as seen in Fig 1 (11 numeric, 2 binary and 7 symbolic). In preprocessing, new time variables were created. Based on stations names columns, we were able to get also start and end borough (Manathan, Bronx...). With latitude and longitude of the stations, new columns like ‘ride distance km’ and ‘ride avg speed’ were created.

Additionally, meteorological data from New York City was also included to analyze the impact of the weather conditions on ride behavior. Several variables had significant outliers with different scaling values, so outlier and scaling treatment should be considered. Study of numeric and symbolic variables showed that 80% of the rides in Citi bike are taken from members (users with subscriptions), while the remaining 20% are from casual users (single ride and 1-day pass). Electrical bikes are the preferred choice among Citi bike users. Most rides occur on weekdays, with 70% starting and ending in the borough of Manathan.

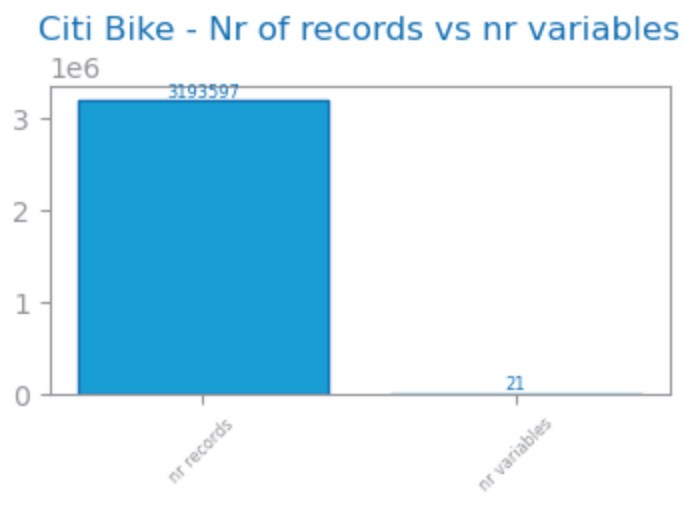


Figure 1: Number of records and variables in dataset 1

Dataset 2: extracted from [kaggle](https://www.kaggle.com/datasets/hosubjeong/bakery-sales/data) 2421 observations, 27 initial columns (Fig 2), with 11% missing values on feature “place” which is the customer’s location. Features with 100% of missing values were not considered and removed before any further preprocessing: “croque monsieur” and “mad garlic”.

In the preprocessing stage, feature engineering was done to enrich the data frame with mainly date features, a feature related to the Purchase Value (low medium or high), and Product Category (food or drink), leaving this dataset with 35 variables.

Regarding the variable types (Fig 2), there are mostly numeric ones as expected, as all products are represented in columns with their respective quantity. The distribution of numeric features was studied through boxplots, where we observed outliers, and through distribution histograms, which most features fitted an exponential distribution. The categoric variables were studied using bar charts.

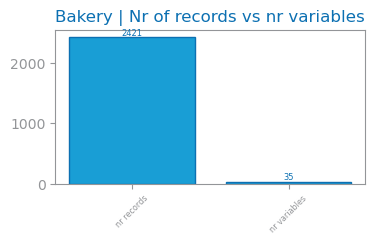


Figure 2: Number of records and variables in dataset 2

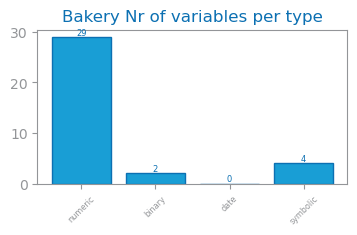


Figure 3: Variable types in dataset 2

Dataset 3: 32.4 million observations, 9 initial columns (Fig 4). 6.4% missing values in variable “days since last order” because it represents users buying for the first time. Additionally, more features were generated around day and time when the order took place. Dataset 3 had as main outliers the “add to cart order” with outlier values ranging between 25 and 145 products per order. “Order number” also had a very high number of outliers with values ranging from 50 to 100 orders.

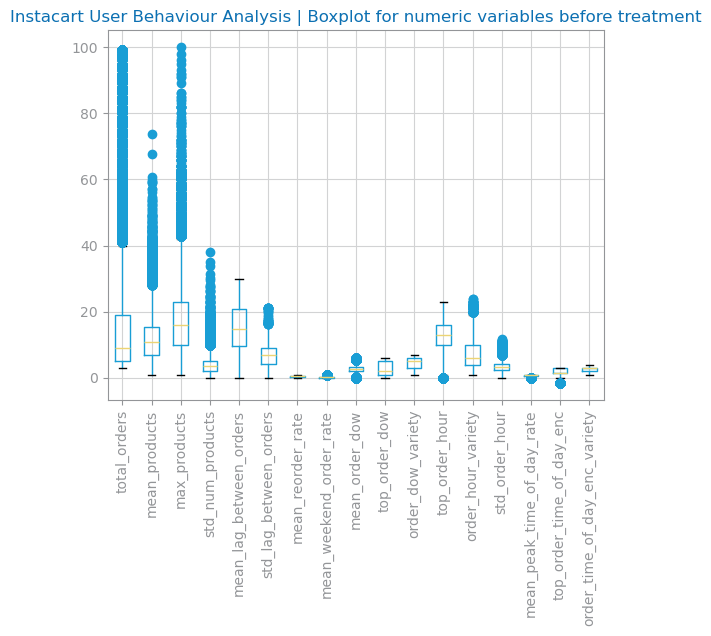


Figure 4: Boxplot for Dataset 2's numeric variables before treatment

From dataset 3 an aggregated version was also generated grouped by user id which features user buying behavior with a total of 206K records and 17 numeric variables. The same outliers from original dataset also had impact on this aggregated version (Figure 4). Looking at frequency, Dataset 3 has orders mostly in the Afternoon and in the Morning on all days of week but especially between 10am and 4pm. (Fig 5).

A blue square with white text

AI-generated content may be incorrect.

A blue rectangular bar graph

AI-generated content may be incorrect.

Figure : Number of records and variables in dataset 3

Most prior orders have been made 6-7 days before or 30 (which may have been capped by dataset provider). 53% of products purchased are reordered by the user. Top purchased goods are fresh fruits and vegetables as well as dairy products. Top purchased products include Bananas, Strawberries, Spinach or Avocados (Fig 5).

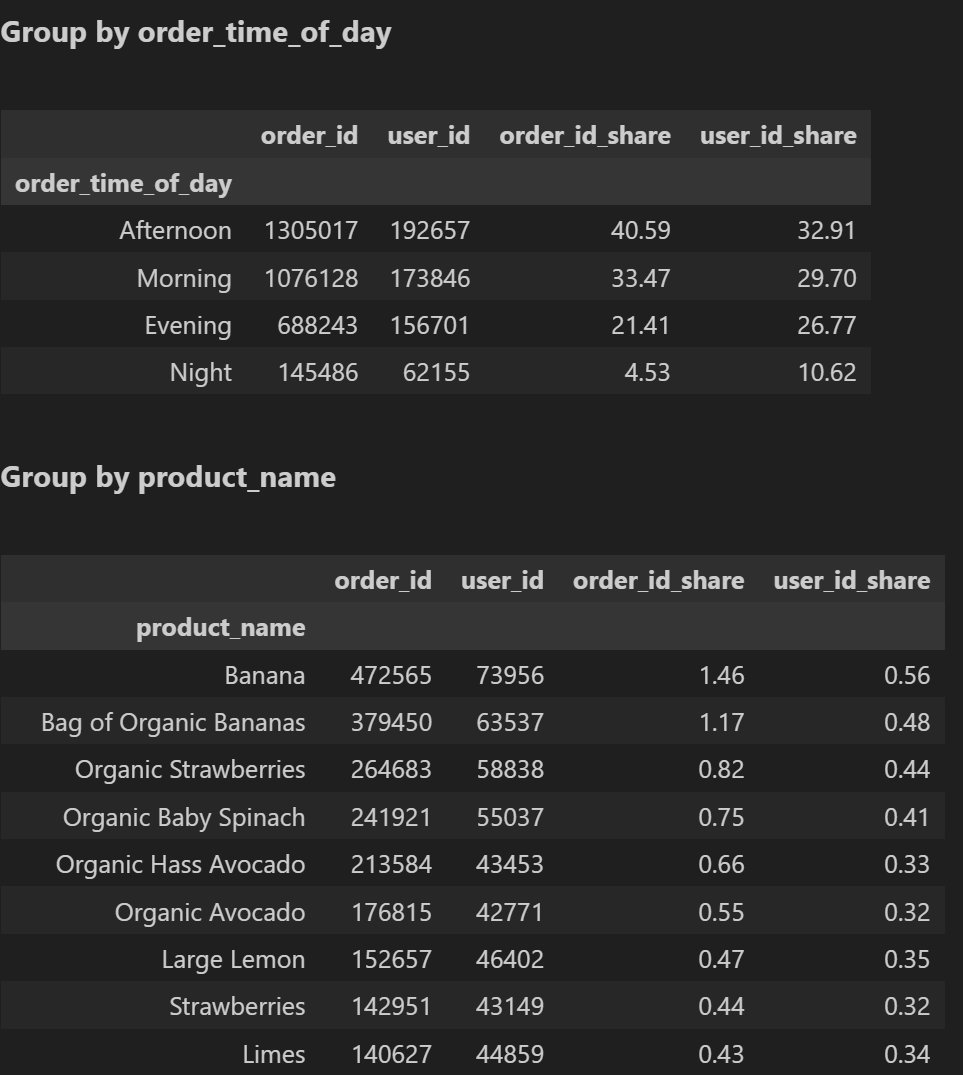


Figure : Top purchases grouped by time of day and product name, in dataset 3

## Feature relevance and dependence

Dataset 1: For Clustering, features with low variance (<0.1), like rain mm, were removed. Redundant features, with a correlation >0.7 were also removed – end borough, hour sin.

Dataset 2: Features with low variance (<0.1) were removed, as they appeared to be irrelevant: “merinque cookies”, “milk tea”, “tiramisu”, “berry ade”, “gateau chocolat”, “hour cos”. Redundant features removed with the same threshold as dataset 2, mainly related to the datetime features.

Feature relevance and dependence was studied for Dataset 3’s User Aggregated Behavior where features with low variance (<0.1) could be removed for Clustering Analysis: ‘mean weekend order rate’, ’mean reorder rate’, ’mean peak time of day rate’. Similar research was applied for redundant features (>0.85 to be more exigent) where these were found: ‘max products’, ’std num products’.

# CLUSTERING

Hard clustering approaches based on hierarchy and partition, namely agglomerative clustering and K-means respectively. We recommend further research to explore soft approaches, for example based on density.

Dataset 1: The goal was to understand the different types of rides in Citi bike according to type of user, bicycle, ride duration and speed, and weather conditions.

Dataset 2: the goal was to group the transactions based on which products were bought, and the time the purchase happened.

Dataset 3: The goal was to understand if there are user groups that buy in different quantities, frequency, if they reorder products or if they buy at different times of day or days of week.

## Reference clustering solutions

For the Agglomerative clustering, a dendrogram was plotted to select the ideal number of clusters according to the highest vertical jump and considering the Silhouette score. Similarly, for K-Means we plotted the elbow method to determine optimal number of clusters to initially use, looking for the point where SSE has the biggest drop, while also considering the Silhouette score.

## Visualization and description

All 3 datasets plotted 2D PCA to visualize, and dataset 2 used t-SNE for further understanding.

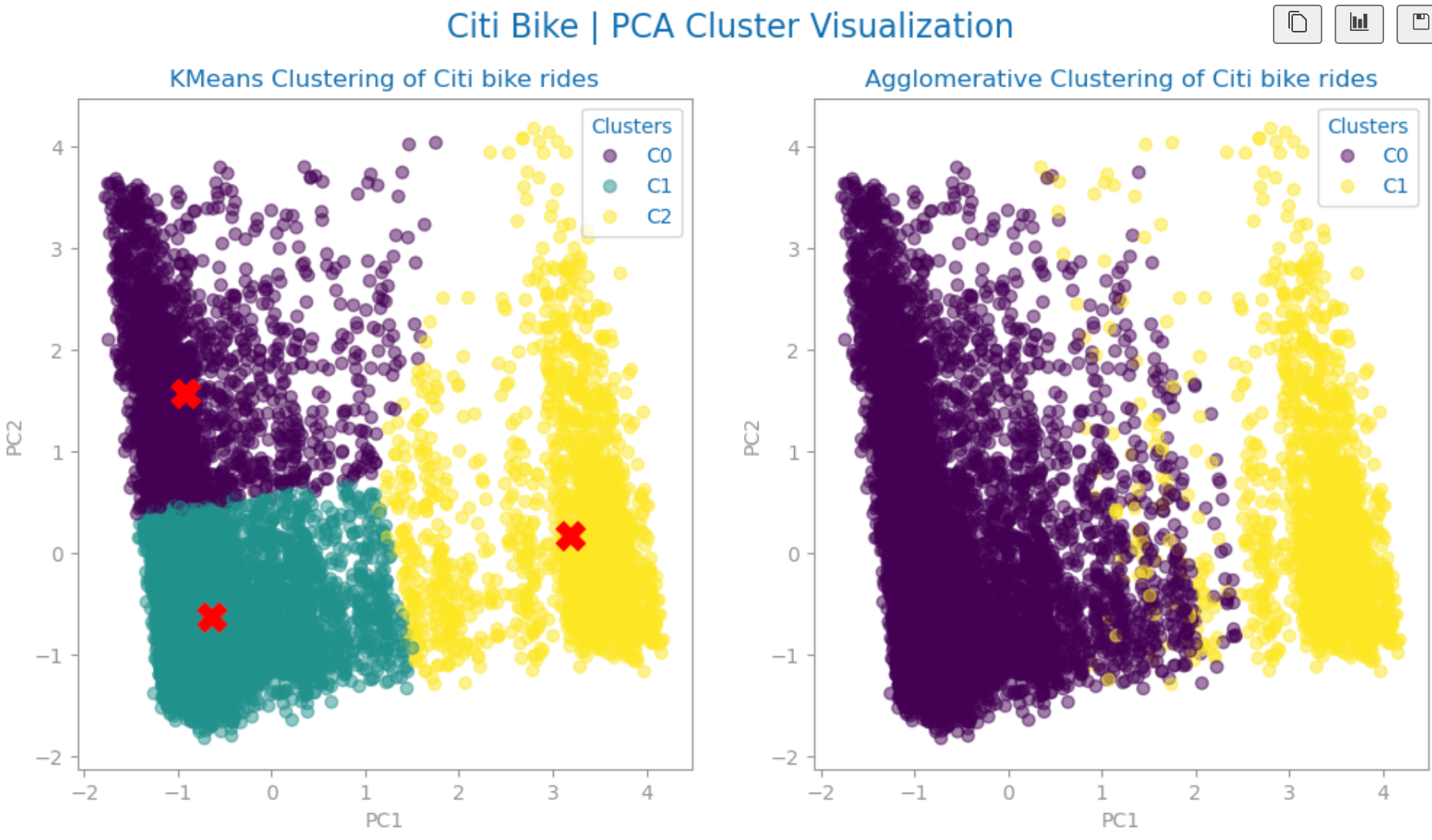
Dataset 1: Using PCA 2D, the clusters were plotted based on the top 2 components (Fig 6). With K-Means, 3 clusters can be observed with moderate cohesion and separation (silhouette of 0,21). Visual inspection showed distinct separation between clusters, highlighting effective grouping after scaling and outlier removal. For Agglomerative Clustering, 2 clusters were defined with a silhouette score of 0.33, reflecting better cohesion. However, some overlap at the cluster boundaries suggested room for improvement in separation. 

Figure : PCA cluster visualization in dataset 1

Dataset 2: Used PCA on the original dataset to plot the 2 principal components (Fig 7) that explain the most variability of the data (around 75% for 5 Principal Components), for both K-Means and Agglomerative clustering solutions. We can observe good cohesion in the purple K-Means cluster, but many observations are still quite separated from each other.

A screenshot of a graph

AI-generated content may be incorrect.

Figure : PCA cluster visualization in dataset 2

To better understand our clusters visually, t-SNE was plotted, revealing an interesting singular major cluster for both clustering solutions, and 2 minor ones (Fig 8).

A screenshot of a graph showing different colored dots

AI-generated content may be incorrect.

Figure : t-SNE cluster visualization in dataset 2

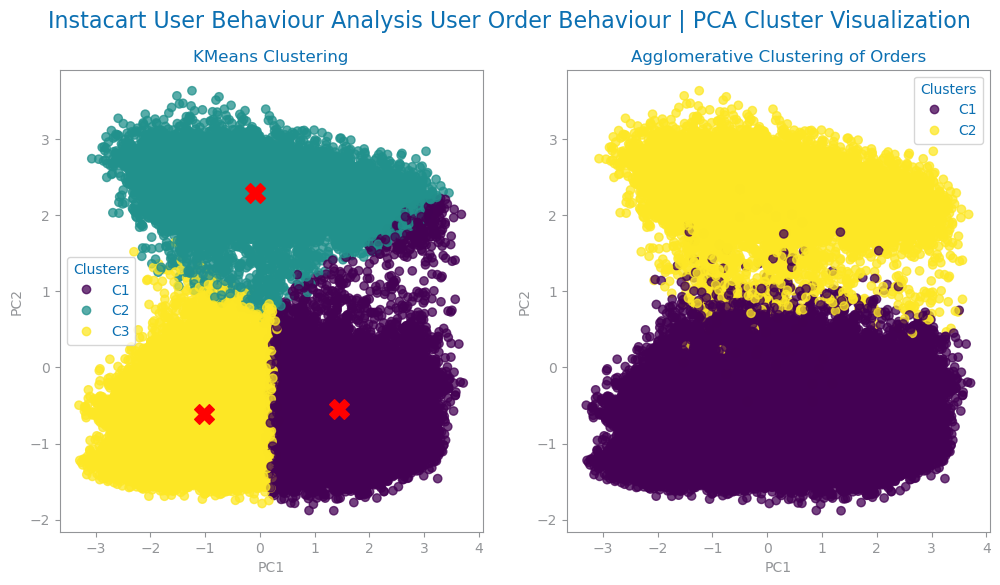
Dataset 3: Used PCA 2D to plot the different clusters (Fig 9). For K-means we have 3 different clusters that are well limited but with no visible separation. Similarly, Agglomerative Clustering showed good cohesion inside the clusters, however some overlapping observed in the boundaries do not show such good separation between the clusters. 

Figure : PCA cluster visualization in dataset 3

## Distances and methods

For all datasets, in the Agglomerative Hierarchical clustering, the impact of using different linkage criteria combined with different distance measures was plotted against Silhouette scores and assessed (Fig 10).

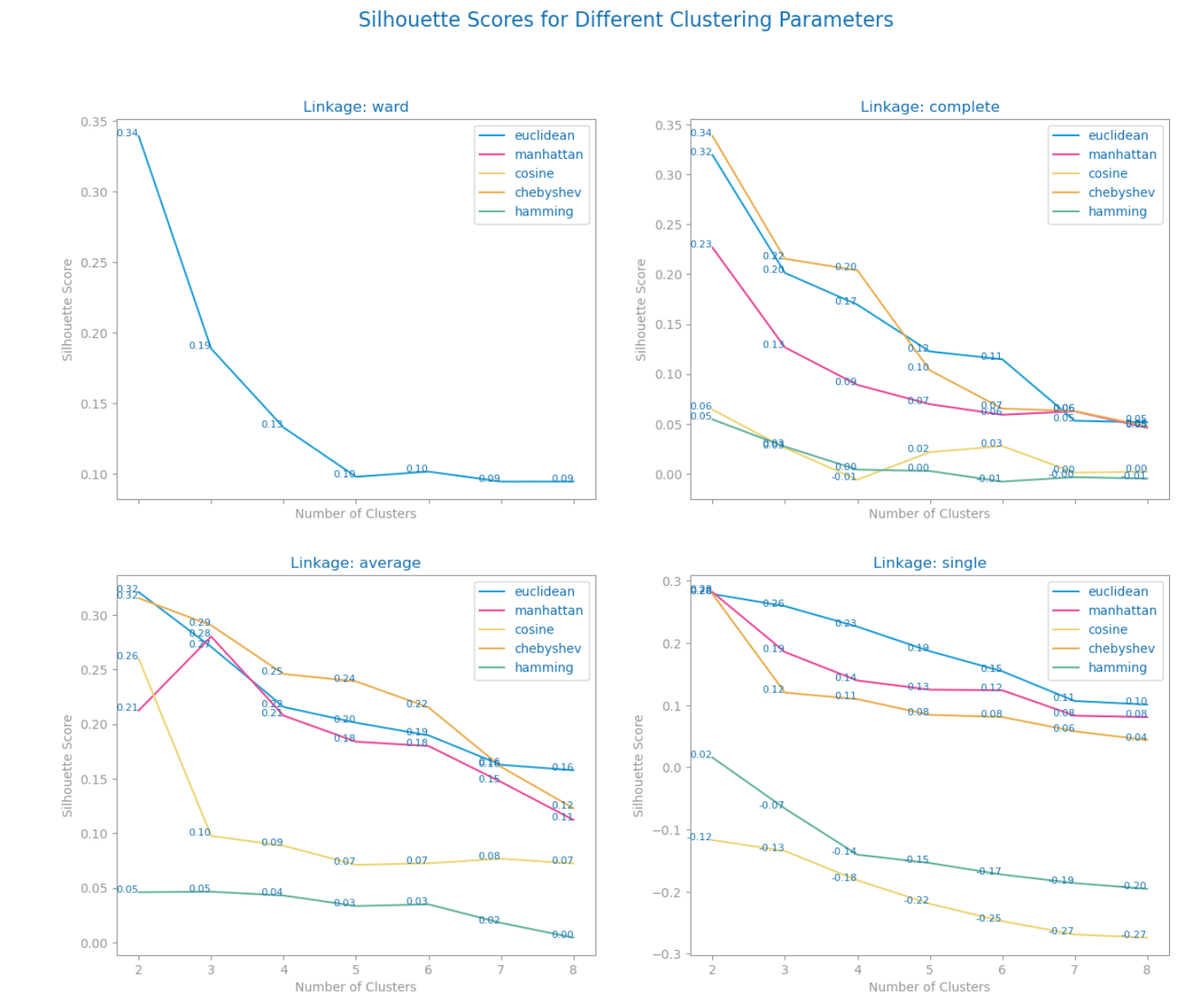


Figure : Silhouette study with different linkage and distance criteria in dataset 2

This investigation resulted in the following solutions for each dataset: dataset 1 and 3 used the same distance Chebyshev, and dataset 2 and 3 used the same linking criteria (Average), which were the solutions with highest Silhouette scores.

Dataset 1: Chebyshev, complete (0,33).

Dataset 2: Manhattan, average (0.89)

Dataset 3: Chebyshev, average (0.24)

## Number of clusters

Dataset 1: In K-Means, according to the elbow method and silhouette score, an optimal number of clusters of 3 clusters resulted in a silhouette of 0,21 (Fig 11, Fig 12). In Agglomerative Hierarchical clustering, several linkage and distance metrics were assessed. The Complete linkage and Chebyshev distance resulted in an optimal number of clusters of 2 with a silhouette of 0,33 (Fig 13).

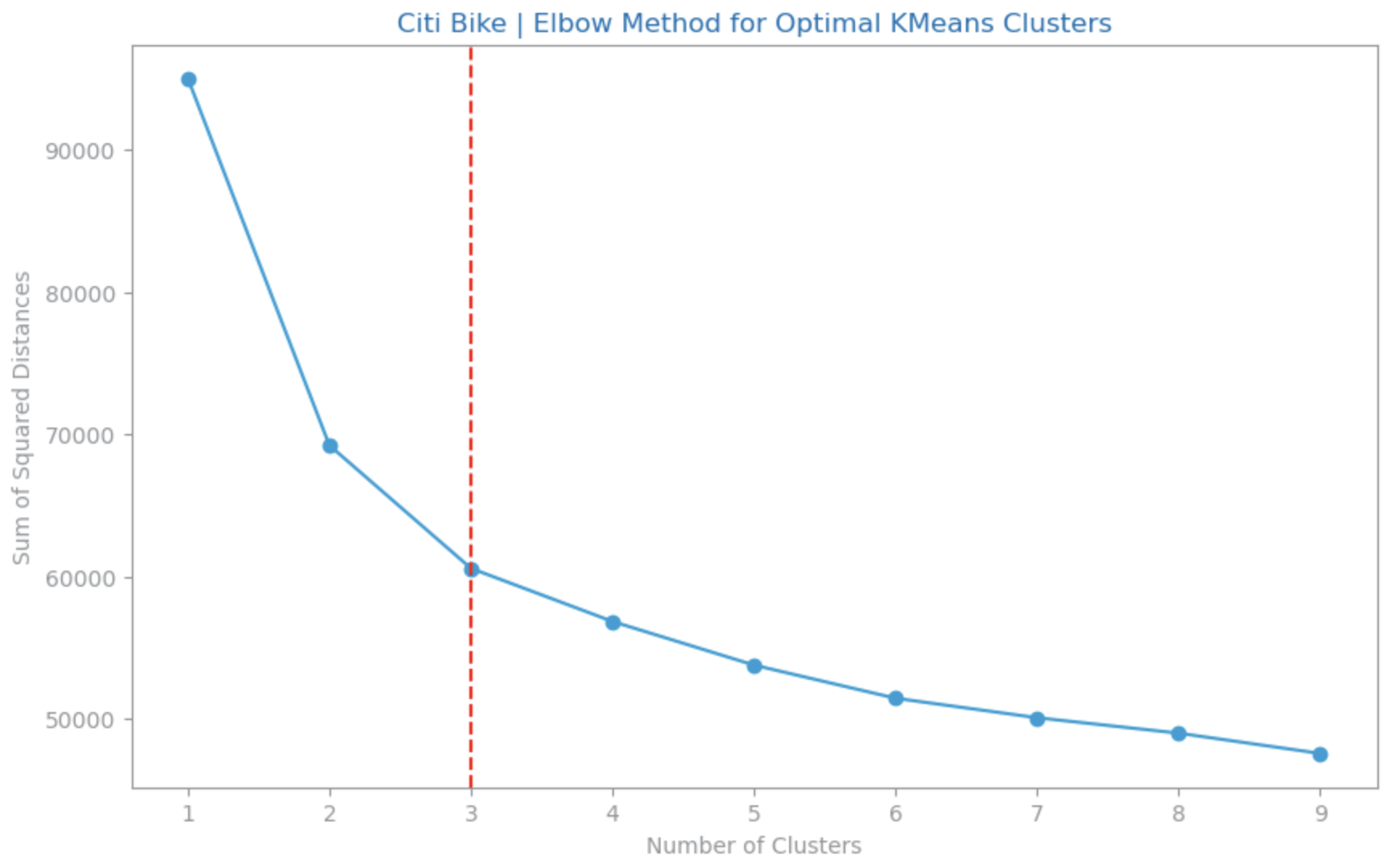


Figure : SSE vs number of clusters for dataset 1

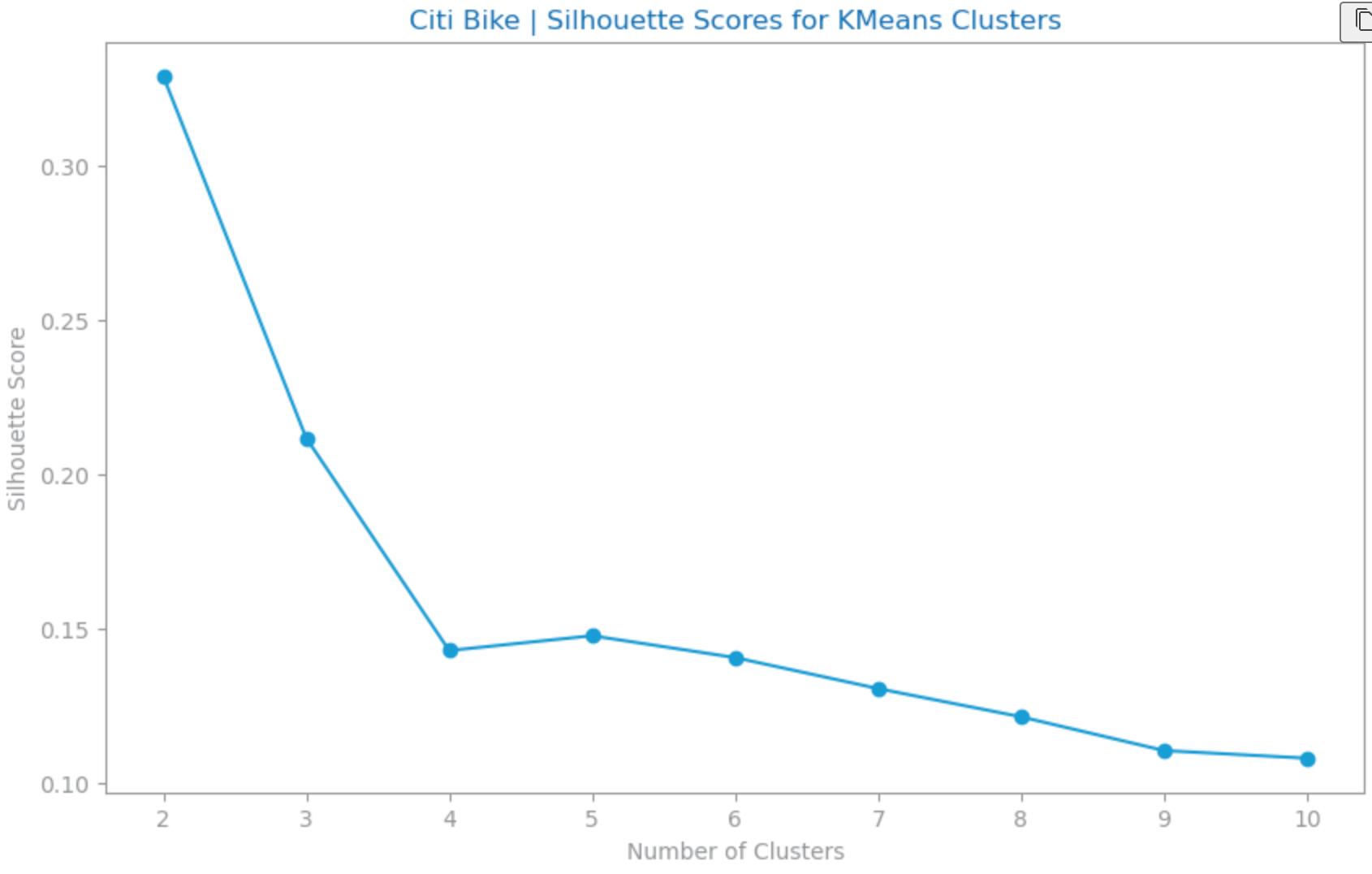


Figure : Silhouette scores for K Means for dataset 1

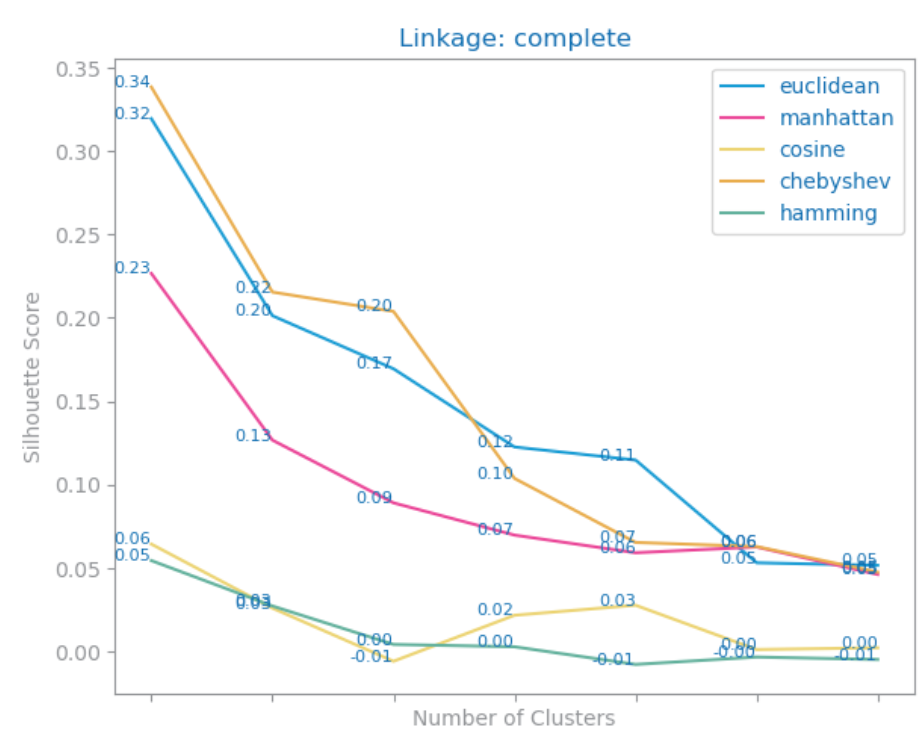


Figure : Silhouette scores for Agglomerative Clustering with different distances for dataset 1

Dataset 2: three Agglomerative Hierarchical clusters using average linkage and Manhattan distance, and three with K-means. By looking at the dendrogram (Fig 14), we would consider 2 clusters, however it was not evident the ability to distinguish between both clusters when we analyzed their characteristics specifically. As an alternative, instead of two clusters, there clusters were considered for the Agglomerative solution.

A graph of a bar chart

AI-generated content may be incorrect.

Figure : Dendrogram for dataset 2

As for the K-Means and considering both the Sum of Square Errors and the Silhouette scores, we defined three final clusters. The elbow method example can be seen below (Fig 15).

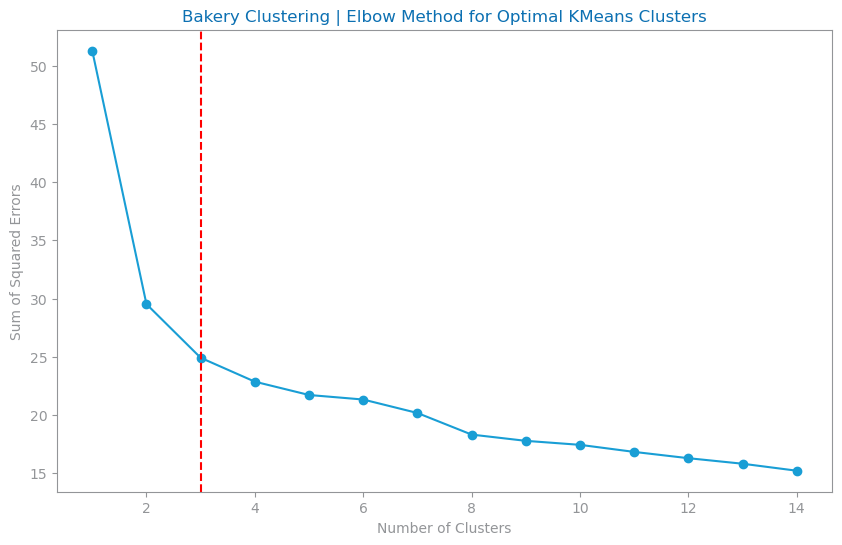


Figure : SSE vs number of clusters for dataset 2

For Dataset 3 clustering through K-Means, the Elbow Method pointed to three clusters with an SSD of ~280k (Figure 16)while the silhouette Study (up to 8 clusters), also suggested the same reaching a silhouette of 0.19 (Fig 17)

As for Agglomerative Clustering, Dataset 3 showed best silhouette performance in Average Linkage with two clusters, in particular the Chebyshev distance which achieved 0.245. The dendrogram also shows a distinctive vertical jump that separates into two clusters. (Figure 18)

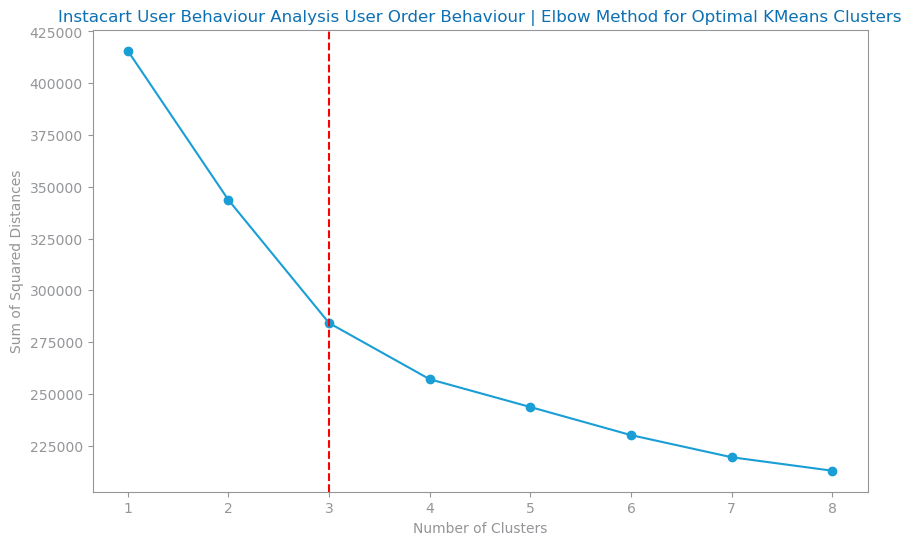


Figure : SSE vs number of clusters for dataset 3



Figure : Silhouette scores for K Means for dataset 3

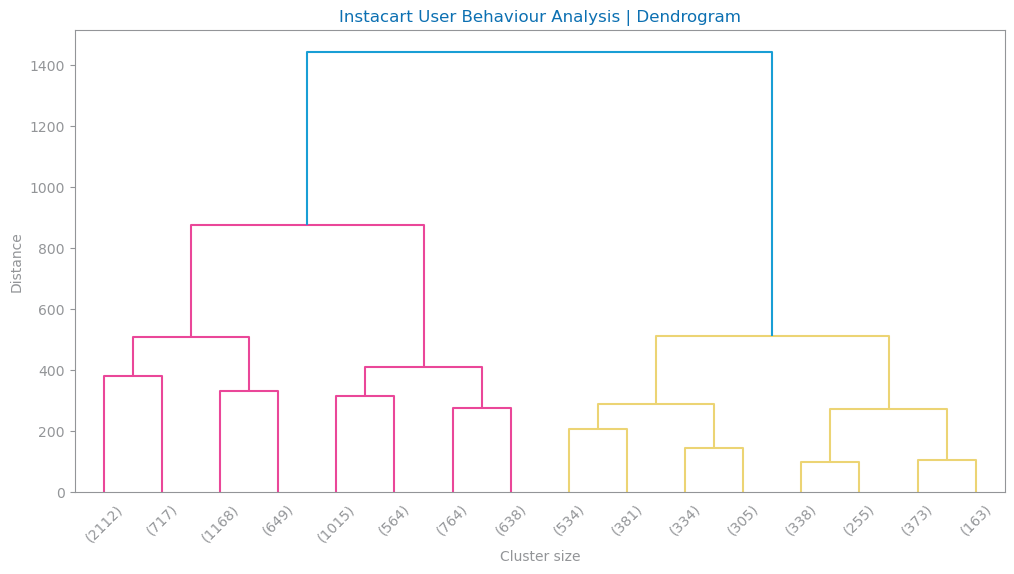


Figure : Dendrogram for dataset 3

## Preprocessing impact

Dataset 1: For Clustering analysis, a sample of 12k records (3,5 of initial dataset) was used to reduce processing time.

To address Variables with many outliers, we truncated extreme values at 2 times the standard deviation to reduce noise in clustering. Robust Scaler was used in scaling, as its less sensitive to outliers than Standard or Min Max, preserving feature relationship and leading to clear cluster formation. Low variance features like rain mm were removed, as they didn't provide useful information. Highly correlated features (Pearson correlation > 0.8), such as end borough and hour sin, were excluded to reduce redundancy. This helped the clustering algorithms focus on the most important features, improving cluster separation and cohesion.

Dataset 2: to address categorical variables, we opted to encode all cyclic variables related to datetime through sin and cosine formulas (example season), and others we encoded using hierarchical logic (Purchase Value). When possible, we encoded using dummification as the library used in this project (sklearn) does not handle categorical features very well.

Moreover, as K-means is sensitive to extreme values, the missing values of “Purchase Value” were filled with the median instead of mean, and other rows that contained missing values were dropped, mainly from “address”.

Additionally, the feature “total” was removed for the clustering approach as it was making cluster visualizations hard to interpret, due to the natural range of the variable.

Regarding scaling, Normalizer was used, as it gave not only the best silhouette scores in comparison with other scaling methods (Fig 26), especially for K-Means, but also contributed to clearer and more concise clusters for this dataset. We opted to not treat any outliers in this dataset nor truncating or removing them as it presented worse Silhouette scores.

Dataset 3: Despite removing outliers entirely, some columns like total orders were negatively impacting the Agglomerative model creation and visualization (). To mitigate this impact in the model, Robust Scaler was used. From this stage, two different approaches got the best results:

- Approach 1: Remove Low variance, remove redundant variables and drop outliers on ‘total orders’, ‘mean products’ and ‘std order hour’

- Approach 2: Keep all variables and truncate 'total orders', 'mean products', 'std order hour', 'max products', 'std num products'.

## Detailed assessment

Dataset 1: In our evaluation of clustering solutions, we tested both K-Means and Agglomerative Clustering using silhouette scores to assess their performance. For K-Means, with 3 clusters, the silhouette score was 0.21, suggesting moderate cohesion and separation between clusters. Despite the lower score, visual inspection of the PCA 2D plot showed distinct and well-separated clusters, indicating that K-Means provided effective grouping after scaling and outlier removal. On the other hand, Agglomerative Clustering, with 2 clusters, had a higher silhouette score of 0.33, reflecting better cohesion. However, there was some overlap at the cluster boundaries, suggesting that the separation could still be improved. Despite the lower silhouette score, the K-Means solution provided clearer cluster separation, which allowed for more meaningful analysis and better segmentation of the data. Therefore, K-Means was selected for further exploration.

Dataset 2: the Silhouette scores together with PCA visualization of the original dataset were examined when looking for the most adequate preprocessing for clustering. For this, we first clustered keeping outliers, secondly with outliers truncated, and another approach by removing outliers. The best approach was to maintain the outliers, and we moved to the second assessment regarding feature relevance.

It was also tested to remove only redundant variables, only irrelevant variables, and both, and we decided to move on with the latter one.

Lastly, a third assessment was followed by clustering with different scaling methods, namely using Standard Scaler, Robust and Normalizer. To note that this dataset did not undergo any outlier handling, so the Robust method would be highly fitting. However, the clustering solutions represented visually were ambiguous. Therefore, Normalizer was preferred in the final solution with a silhouette score of 0.81 and 0.89 ( 20) for K-Means and Agglomerative Clustering, respectively. In the end, final clustering solution chosen was the Agglomerative clustering.

A black background with white text

AI-generated content may be incorrect.

Figure : Silhouette results clustering dataset 2

Reviewing the 2 approaches, we got the following results:

- Approach 1: K-means has a silhouette of 0.185 with 3 clusters and agglomerative 0.275 with 2 clusters. From PCA with 6 components we have 82.7% of total explained variability.

- Approach 2: K-means has a silhouette of 0.19 with 3 clusters and agglomerative 0.235 with 2 clusters. From PCA with 6 components we have 81.6% of total explained variability.

The reason we chose approach 2, despite having slightly lower performance in Agglomerative Clustering, is because it has more records of users with a larger number of orders that were considered as outliers, but these can be very valuable customers. Also having variables such as 'mean weekend order rate', 'mean reorder rate', 'max products', 'std num products' allow us to better target behavior based on weekend and products reordered from users.

Lastly, despite K-means having a lower silhouette score than Agglomerative, this is the preferred Clustering Method for user behavior analysis as it provides higher granularity of cluster’s features.

## Major findings (knowledge acquisition)

Dataset 1: Considering the K-Means Clustering, the feature “ride duration min” and “ride distance km” are the most relevant for Cluster 1, while “cloud cover low pct” is highly relevant for Cluster 3 (). The clusters showed the following characteristics:

* C0: **Long rides** with higher distances (3-5km), occurring **throughout the day**, including both weekdays and weekends, possibly representing leisure or long-distance commuters.
* C1: **Short rides** (<2km), occurring **primarily during morning and evening,** with moderate weather conditions, including both weekdays and weekends, likely representing short-distance commuters and casual users.
* C2: **Short rides** (<2km), on **weekdays**, heavily influenced by **high cloud cover** (Overcast) and specific weather conditions, possibly indicating weather-sensitive usage patterns.

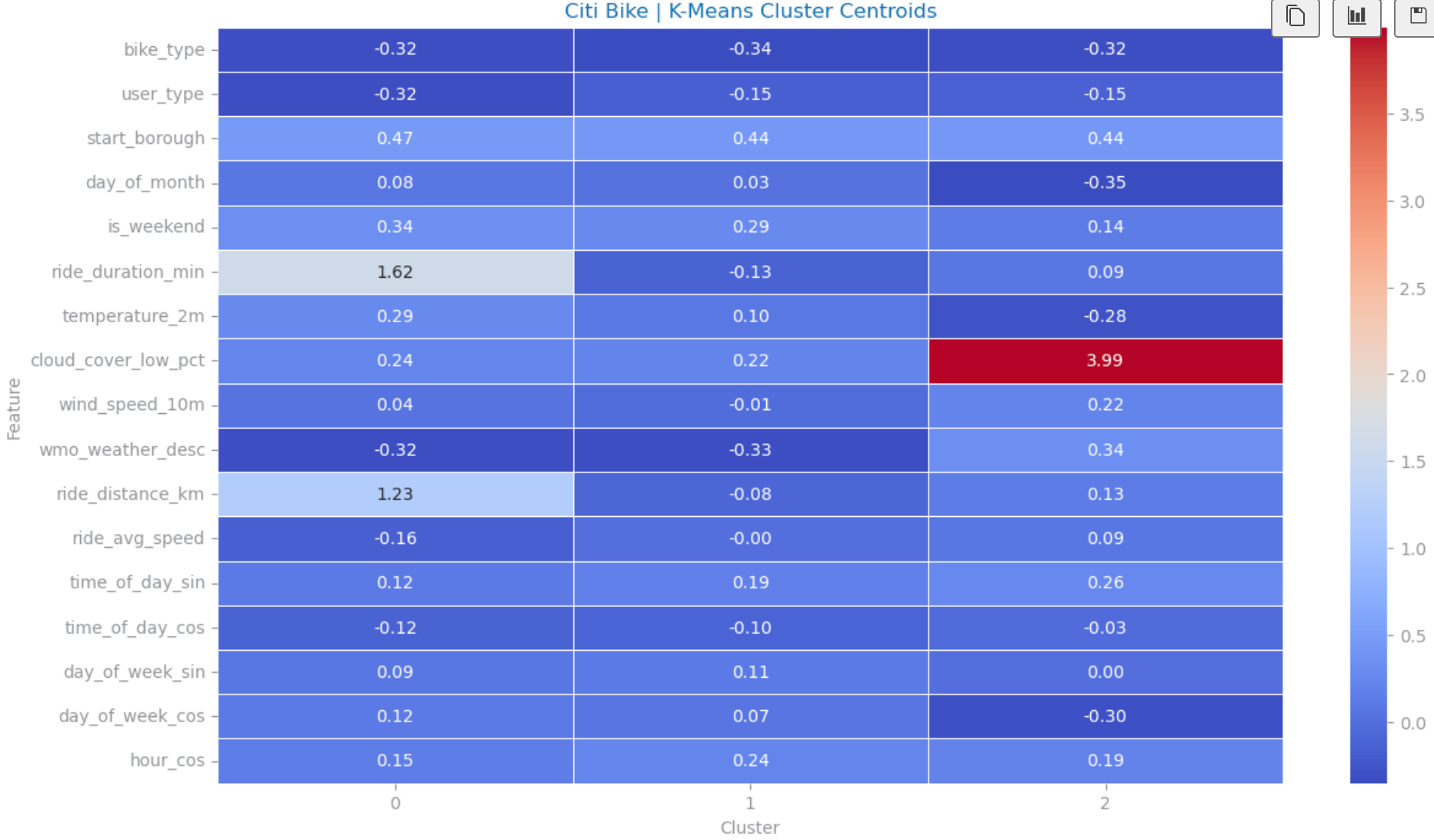


Figure 21: K-Means Cluster Centroids Heatmap dataset 1

Dataset 2: Considering the Agglomerative Clustering solution, which was considered to have clearer cluster definition and characterizations, the following clusters were identified:

* **C1**: Transactions with more angbutter (South Korean pastry), almond croissants, pandoro, done at the beginning of the year, with high purchase value (high transaction value).
* **C2**: Transactions with bread, croissants, and drinks, done on weekends throughout the whole year, and high range of purchase value.
* **C3**: Transactions with orange pound, with low purchase value, done at the end of the month.

Dataset 3:

Looking at the three K-means generated clusters we can see from its centroids some insightful information about the users’ behavior (Figure):

* C1: This cluster contains 33.7% of the user sample, which shows the highest number of orders and products, as well as high variety of days where they purchase. These are most likely the top customers for Instacart that buy frequently and in high quantity. They tend to buy on all days of week and buying specially from 10AM to 3PM. (FIG)
* C2: This cluster contains 20.4% of the user sample, being the smallest cluster. These customers have the special tendency to buy in the evening (FIG).
* C3: This cluster contains 45.84% of the user sample but its probably the lowest value customer group. It shows the lowest number of total orders and variety in purchase days. Although they purchase at peak days, they show low buying frequency. They also have less chance to buy on weekends and reorder items. (FIG)

When joining the Clustering labels with the original dataset we can also have more insightful information:

* More than half of the users from C1 and C2 reorder their products while C3 users do not. (FIG).
* C3 and C1 are the top clusters in total share and have similar behavior for the moment when they buy but the major difference is in the quantity of orders and products, making C1 the top customers that the business should benefit by a rewarding strategy while C3 users might need more incentive to buy more often as per RFM method of segmentation of customers for loyalty programs [1].

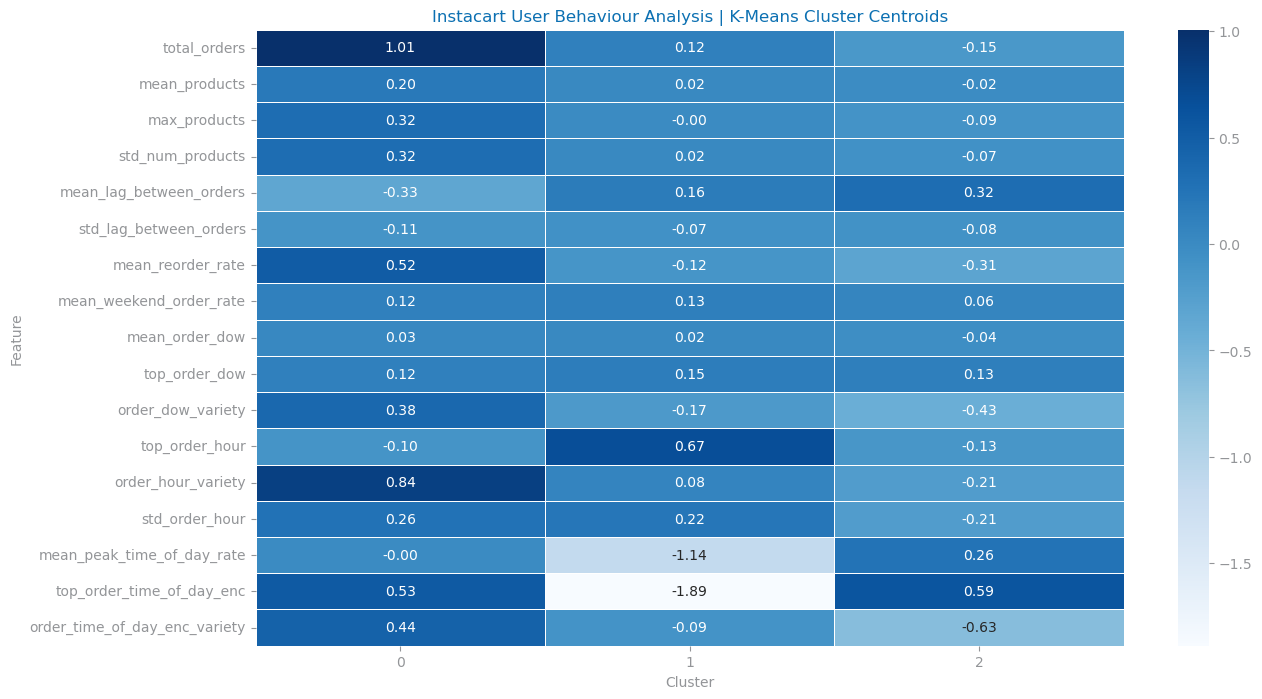


Figure 22: K-Means Cluster Centroids Heatmap dataset 3

# PATTERN ANALYSIS

For Pattern analysis, it is essential to only have binary variables, which implies discretization and one hot-encoding of the variables. These preprocessing steps were done for all three datasets. Overall, we observed unexpectedly frequent patterns (pattern mining) with statistical significance, and unexpectedly discriminative patterns (association rules) by prioritizing support, confidence and lift.

Datasets 2 and 3 have a similar goal

Dataset 1: The goal is to identify and analyze patterns in Citi bike rides between **member** users (annual subscription) and **casual** users (single ride or day pass), focusing on user behavior, ride duration, speed and weather conditions.

Dataset 2: the goal was to understand the association between the bakery products, which ones were frequently bought together, and which ones led to buy others.

Dataset 3: the goal was to understand which products and its respective category levels were most purchased together.

## Reference pattern solutions

Dataset 1: We began by analyzing frequent itemsets using a minimum support of 0.2, up to 100 patterns, and a maximum p-value of <0.01 for both member and casual user rides. For discriminative patterns, we applied a minimum support of 0.2, at least 100 patterns, a minimum confidence of 80%, and a minimum lift of 1.4.

Dataset 2: We looked first into pattern mining, identifying the unexpectedly frequent patterns, and from there we looked at the ones that were discriminative through association rules, filtering for only the features related to the bakery products.

For this dataset it was used min support 0.5, min number of patterns 20, min confidence 0.7. All unexpectedly frequent patterns had a lift higher than 1.4, which indicates that the rules, besides being unexpectedly discriminative, also had a positive correlation between the antecedent and consequent.

Dataset 3: After filtering the dataset for each of the 3 product category levels of the dataset, patterns were found with the following parameters:

1. Department: minimum length of 2, max p value of 0.05, minimum 60% confidence and 1.4 lift. Additional patterns were found at 1.2 lift.
2. Aisle: minimum length of 2, 1.4 lift and 60% confidence with p value of 0.10. Additional patterns were found at 1.2 lift.
3. Product: Over the top 1000 purchased products, 57 patterns were found at a minimum confidence of 30% confidence, 0.10 p-value and 1.4 lift. More patterns were found with length >= 3 and 1.2 lift.

The changes in confidence, p value and lift were done to find unexpected patterns or when no minimum number of patterns were found.

## Preprocessing impact

Dataset 1: For pattern analysis, a sample of 320k records was used to reduce processing time. Outliers in the ‘ride avg speed’ column were removed, and low-variance variables (e.g., ‘wmo weather desc rain’...) were removed during feature selection. Non-binary variables were discretized into three bins (low, medium, high) with equal proportions to balance the data. One-hot encoding was then done on these variables to transform them into binary. Additional variables like ‘start borough’ and ‘end borough’ were removed, as most rides occurred in Manhattan, making patterns involving these boroughs expected and uninteresting for analysis.

Dataset 2: We looked at the features that were related to baked goods, and the remaining ones were dropped. As the products were already one-hot encoded (with 0 and the respective quantity), we transformed all values above 0 into 1, resulting in a data frame with only binary variables.

Dataset 3: The features department, aisle and product name were one-hot encoded over order ids for pattern mining. Since the dataset has 32M records, a sample with 1.6M was created and for product pattern mining we filtered to top 1000 products purchased by users since we do not have computing power to process that much data and prefer to focus on top product associations.

## Major findings (knowledge acquisition)

Dataset 1: For both member and casual users rides, we identified frequent item sets with strong support (over 100 patterns with support >0.3) and high statistical significance (pvalue <0.01), mainly combining ride day and weather conditions **(FIG).** Association rules were also applied to both datasets, resulting in over 100 patterns with support between 0.1 and 0.2, confidence >0.8, and lift >1.4 **(FIG).**. However, none of these patterns provided unexpected or meaningful insights for analyzing or comparing member and casual user rides.

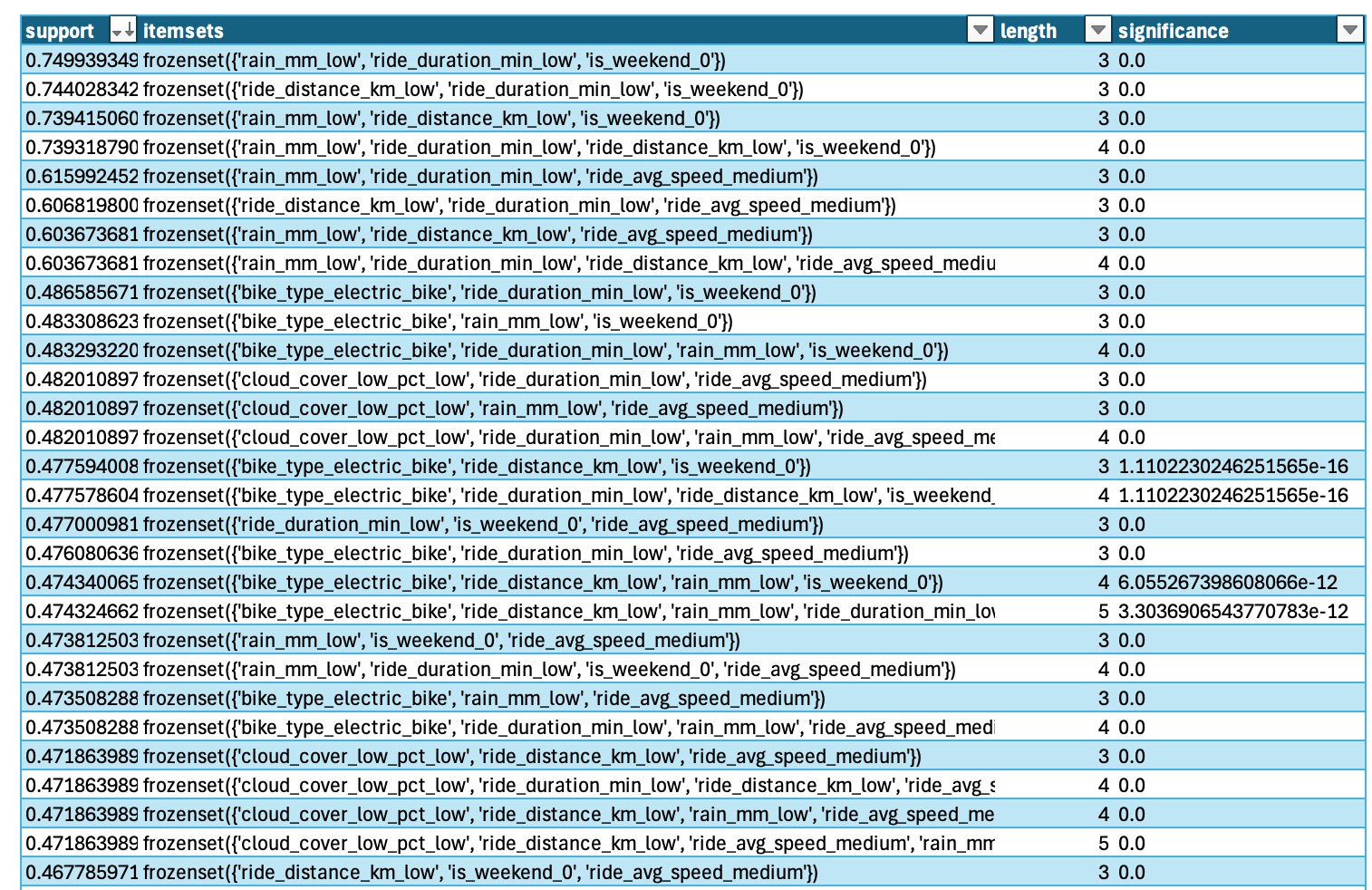


Figure : Top 30 Frequent item set discovered for dataset 1 (member users)



Figure : Top 30 Association Rules discovered for dataset 1 (member users) ordered by confidence.

Dataset 2: Overall, patterns with very low support (around 7% max), and quite high confidence on the rules (70%+) were found. This indicates that although the rules do not happen that often, when they do happen, the consequent also happens in 80% of the cases. We observed that people who buy jam are inclined to also buy bread, and this rule is the most frequent one to happen. Similarly, people who buy jam and ang butter are also inclined to buy bread.

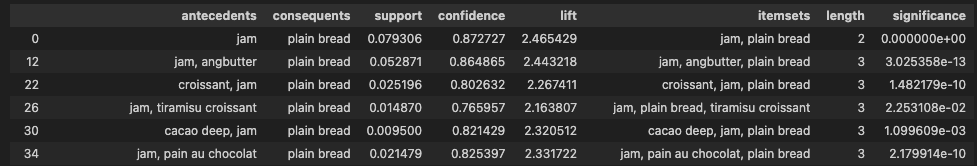


Figure : Main Association Rules discovered for dataset 2

Dataset 3: There is a natural tendency to decreased support in the dataset as data is grouped by the order: department, aisle and finally product name. This happens due to the increased granularity of the items. Here are some pattern insights per product category level:

Department Patterns:

* Produce and Dairy Eggs are the top consequents achieving 12-24% support with great confidence (70-90%) (FIG)
* Support wise, top itemset is (frozen, produce, dairy eggs) at 24% support. (FIG)
* An interesting length=3 pattern with a higher lift at 1.35 we have frozen, produce and dairy eggs. (FIG)

Aisle Patterns:

* Besides frequently seeing fresh vegetables or fresh fruits as top consequents that can achieve a max of 23% support, it is interesting to see dairy products as top antecedents together with these vegetables and fruits such as: yogurt, packaged cheese or milk. These itemsets with length 3 can achieve support as high as 11.7% such as the itemset (yogurt, fresh vegetables, fresh fruits). (FIG)

Product Patterns:

* For length=3 itemsets, only fruit or vegetable products’ patterns have some significant support (0.5-0.6%) like the itemset: (Large Lemon, Organic Baby Spinach, Banana). (FIG)
* In length=2 itemsets, Banana or Bag of Organic Bananas are a frequent consequent with itemsets’ support reaching 1 to 2%, and lift 1.9 to 2.7. Top antecedents include fruits like Organic Avocado, Organic Fuji Apple, Cucumber Kirby or Honeycrisp Apple. However, confidence is not very high for these (~30-40%).(FIG)

# APPENDIX

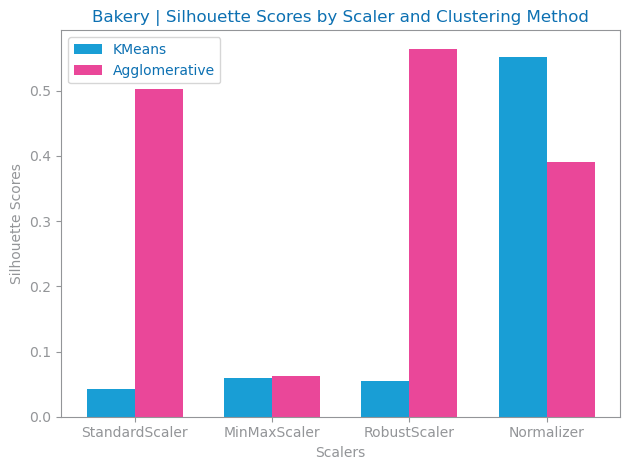


Figure : Silhouette study by Scaler method and Clustering method

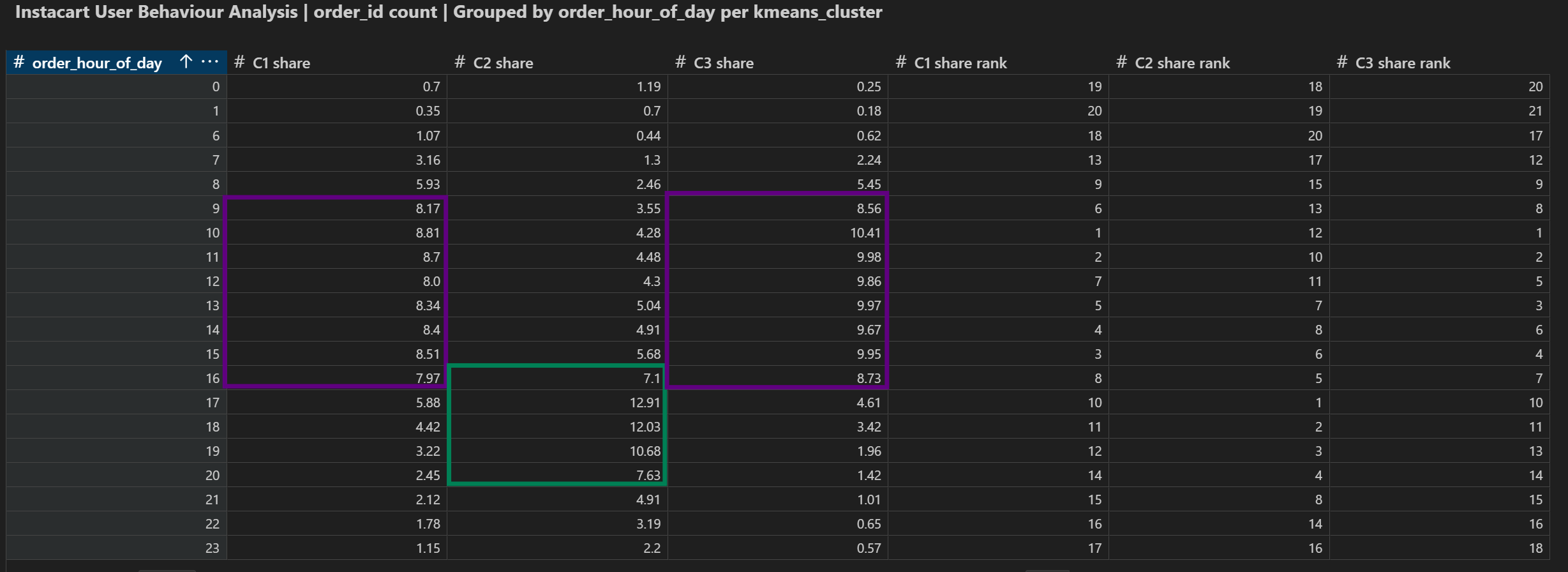


Figure : Order hours of day per Kmeans Cluster Dataset 3

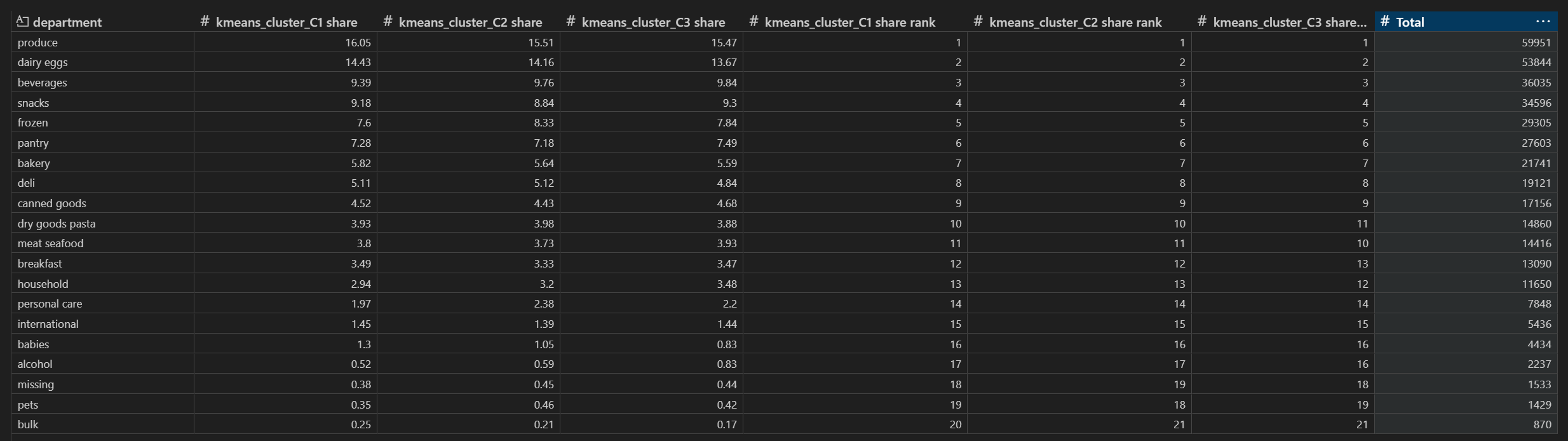


Figure : Department Orders per Kmeans Cluster Dataset 3

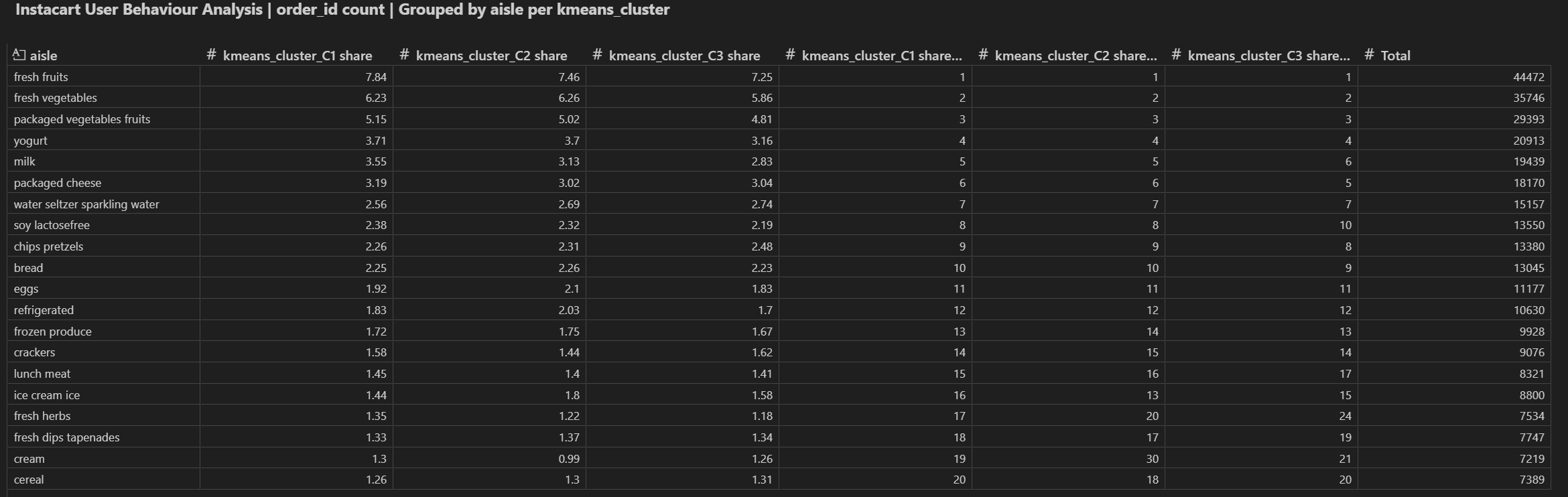


Figure : Aisle Orders per Kmeans Cluster Dataset 3

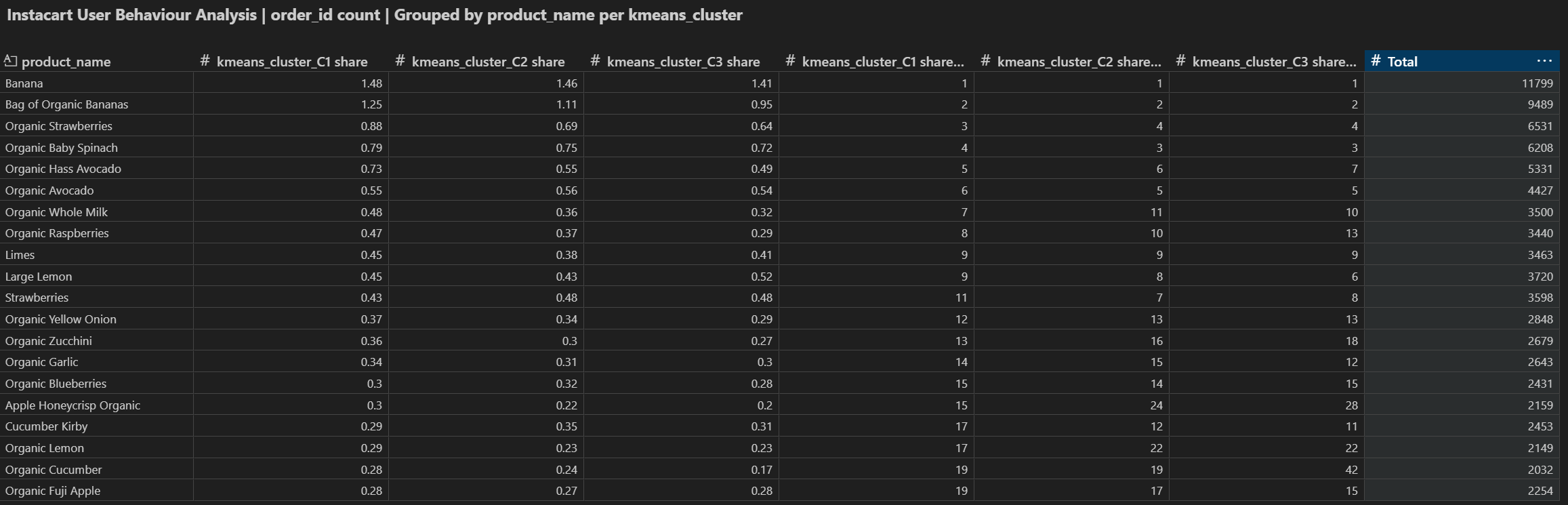


Figure : Product Orders per Kmeans Cluster Dataset 3

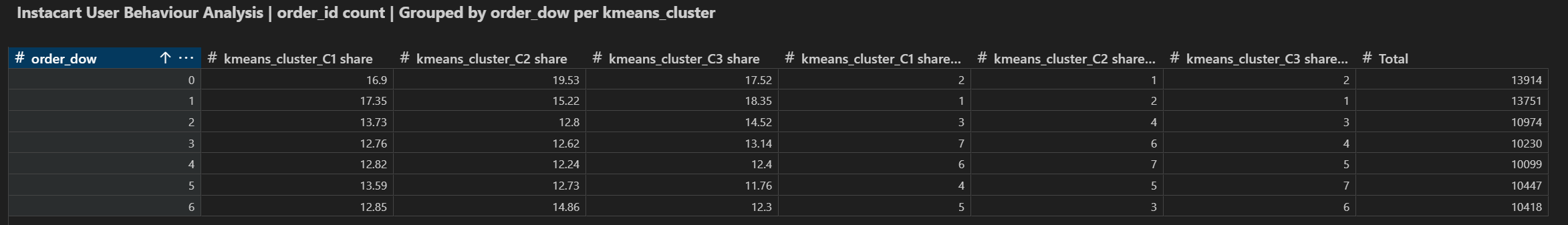
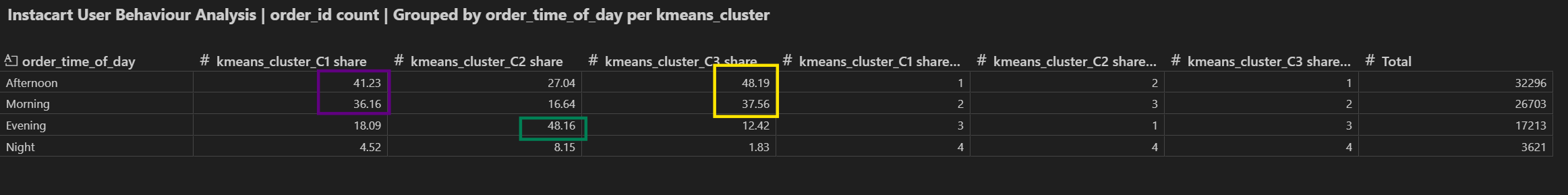


Figure : Day of Week Orders per Kmeans Cluster Dataset 3

Figure : Time of Orders per Kmeans Cluster Dataset 3

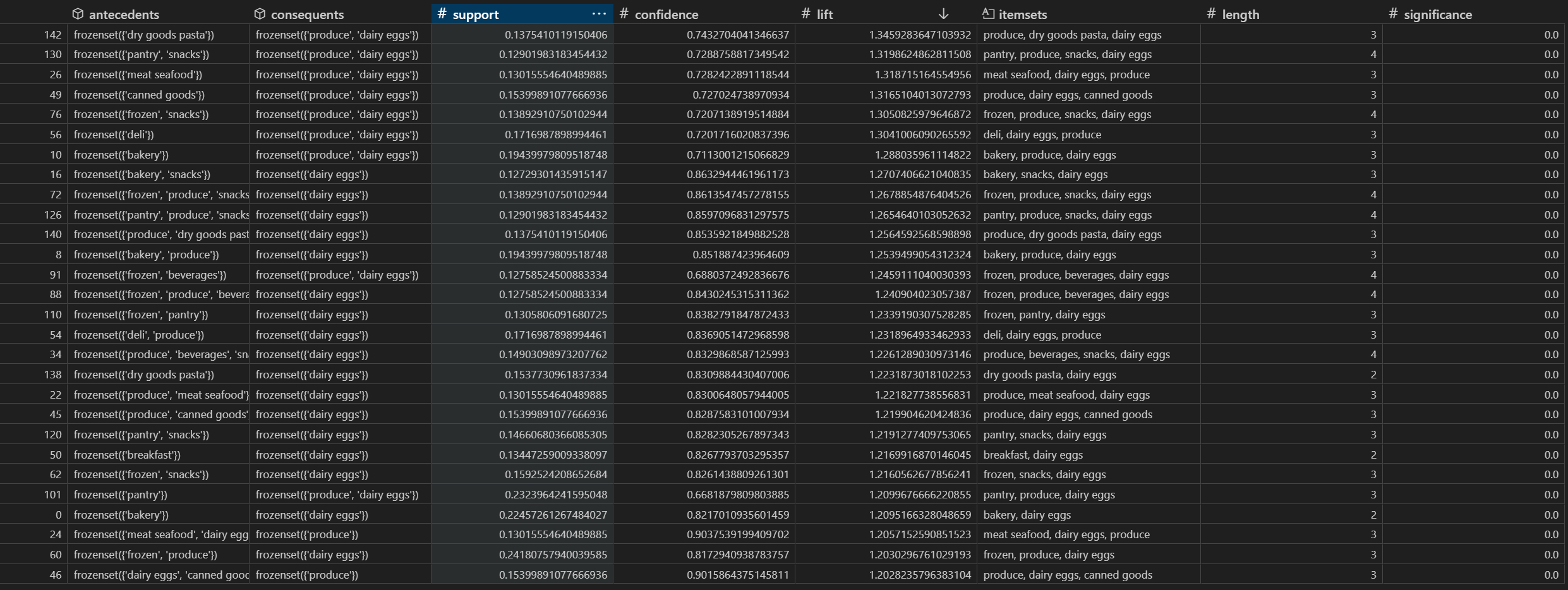
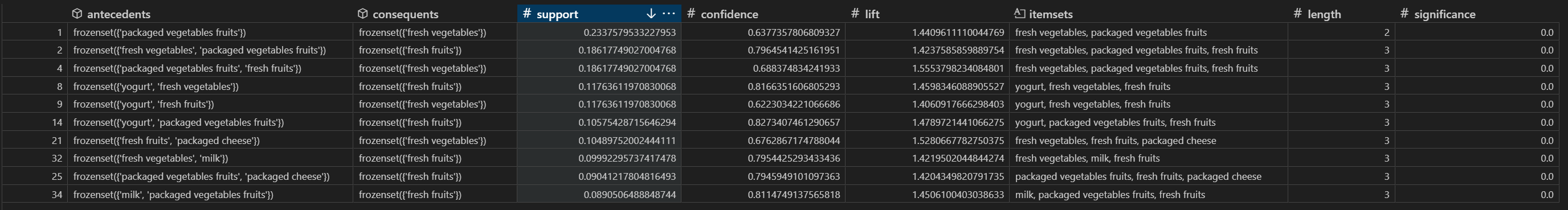
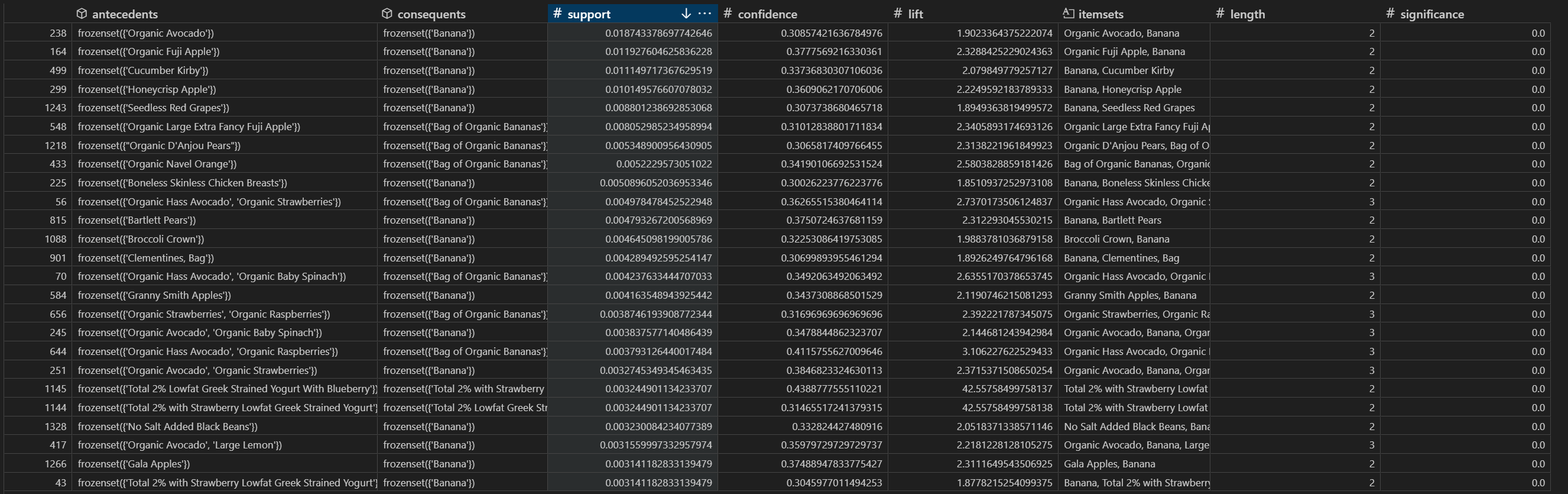


Figure : Department top Patterns in Dataset 3

**Figure** : Aisle top Patterns in Dataset 3

Figure : Product top Patterns in Dataset 3

# REFERENCES

1. Arthur Middleton Hughes. (2010, March 28). Making your database Pay Off Using Recency Frequency and Monetary Analysis. <https://rfm.migmar.com/2010/03/28/making-your-database-pay-off-using-recency-frequency-and-monetary-analysis-by-arthur-middleton-hughes/>