

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**

Dataset 2: New York City bike-sharing system, where each observation represents a bike ride. It contains information on the rides such as bike type (electrical/classic), user type (member/casual) and start/end stations and time.

Dataset 2: South Korean small-sized bakery, each observation is the invoice of a sale made through their 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: Instacart Basket Purchasing Data, 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.

This project has the main purpose of exploring unsupervised learning in 3 distinct datasets from different business domains. 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 dummification, outlier and feature relevance/redundancy study, before implementing non supervised descriptive learning.

In all datasets we clustered using hard clustering approaches, specifically K-Means and Agglomerative. In dataset 1...... In dataset 2, three primary clusters were found, mostly described by the baked goods costumers bought, determined by the time of year, as well as according to how much they’ve spent in the bakery in that purchase. Lastly, in dataset 3 we were able to come up with 3 clusters through Kmeans that were able to explain differences in user buying behavior regarding buying frequency, quantity and datetime preferences.

Lastly, we applied pattern mining with association rules. In dataset 1..... .In dataset 2 the unexpectedly frequent patterns were mostly related to jam and bread.

Lastly, in dataset 3....

# DATA PROFILING

## Statistical analysis

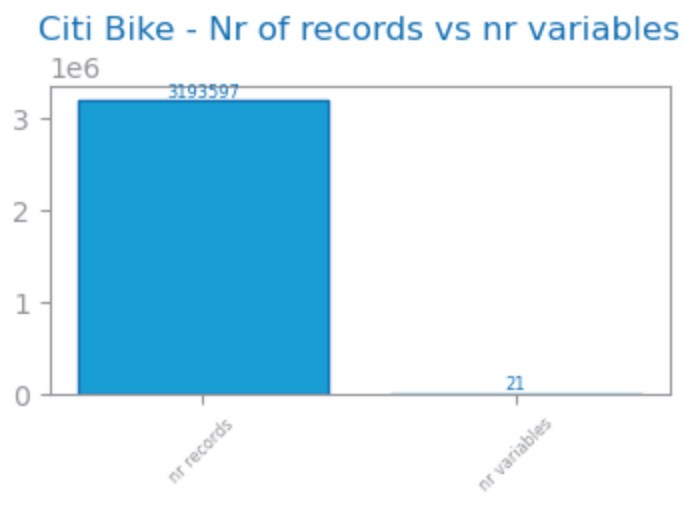


Figure 1: Number of records and variables in dataset 1

Dataset 1: ~3 million observations, 21 initial variables (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 logitude 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 columns had significant outliers with different scaling values, so outlier truncation (using a threshold of 2 standard deviations) was applied to these columns. Due to the presence of continuous variables with widely varying scales, it was necessary to normalize the data using Robust Scaler, ensuring that these features were appropriately scaled.

Dataset 2: extracted from [kaggle](https://www.kaggle.com/datasets/hosubjeong/bakery-sales/data) 2421 observations, 27 initial columns. 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), 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.

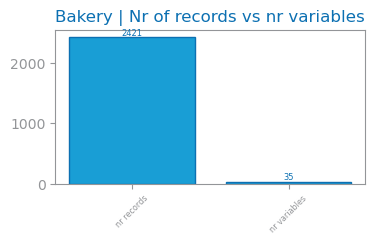


Figure 2: Number of records and variables in dataset 2

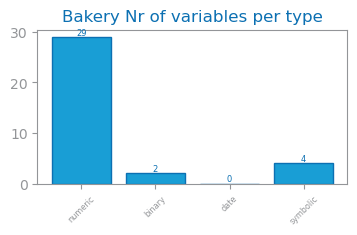
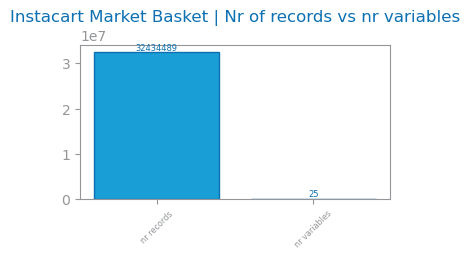


Figure 3: Variable types in dataset 2

Dataset 3: 32.4 million observations, 9 initial columns. 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 it was bought. 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 randgng from 50 to 100 orders. 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. Looking at frequency, Dataset 3 has orders mostly in the Afternoon in the Morning on all days of week but specially between 10 and 16h. 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 user. Top purchased goods are fresh fruits and vegetables as well as dairy products. Top products include Bananas, Strawberries, Spinach or Avocados.



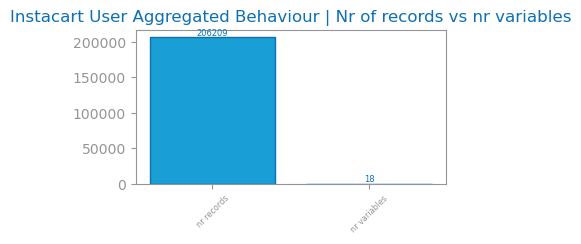


Figure 4: Number of records and variables in dataset 3

## Feature relevance and dependence

Dataset 1: For Clustering, features with low variance (<0.1), like rain mm, were removed. Redundant features, with a correlation >0.8 were also removed – end borough, hour sin.

Dataset 2: dropping irrelevant and redundant variables. 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 (higher than 0.7), mainly related to the datetime features.

Dataset 3: Feature relevance and dependence was studied for Aggregated Behaviour where features with low variance (<0.1) could be removed: ‘mean weekend order rate’,’mean reorder rate’,’mean peak time of day rate’. Similar research was applied for redundant features (>0.85) where these were found: ‘max products’,’std num products’. For clustering, product info was not considered into user aggregated clustering.

# CLUSTERING

Hard clustering approaches based on hierarchy and partition agglomerative clustering and K-means respectively. It would be interesting to explore soft approaches, for example based on density.

Dataset 1: The goal was to understand the different types of rides in city bike system according to type of user, bicycle, ride time, trajectories 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 groups that buy in different quantities, frequency, if they reorder products or if they buy on different times of day or days of week.

## Reference clustering solutions

Agglomerative clustering: dendrogram, select number of clusters according to the highest vertical jump and considering the silhouette score.

K-Means: elbow method to determine optimal number of clusters to initially use, looking for the point where SSE has the biggest drop.

## Visualization and description

Dataset 1: Using PCA 2D, the clusters were plotted based on the top 3 components. For KMeans, with 3 clusters, the silhouette score was 0.21, indicating moderate cohesion and separation. Visual inspection showed distinct separation between clusters, highlighting effective grouping after scaling and outlier removal. For Agglomerative Clustering, with 2 clusters, the silhouette score was 0.33, reflecting better cohesion. However, some overlap at the cluster boundaries suggested room for improvement in separation. Agglomerative Clustering effectively captured the data’s internal structure. Comparison: KMeans provided better separation, while Agglomerative Clustering showed stronger cohesion. More and well separate cluster in k-means provides more insights, so we will describe this solution.

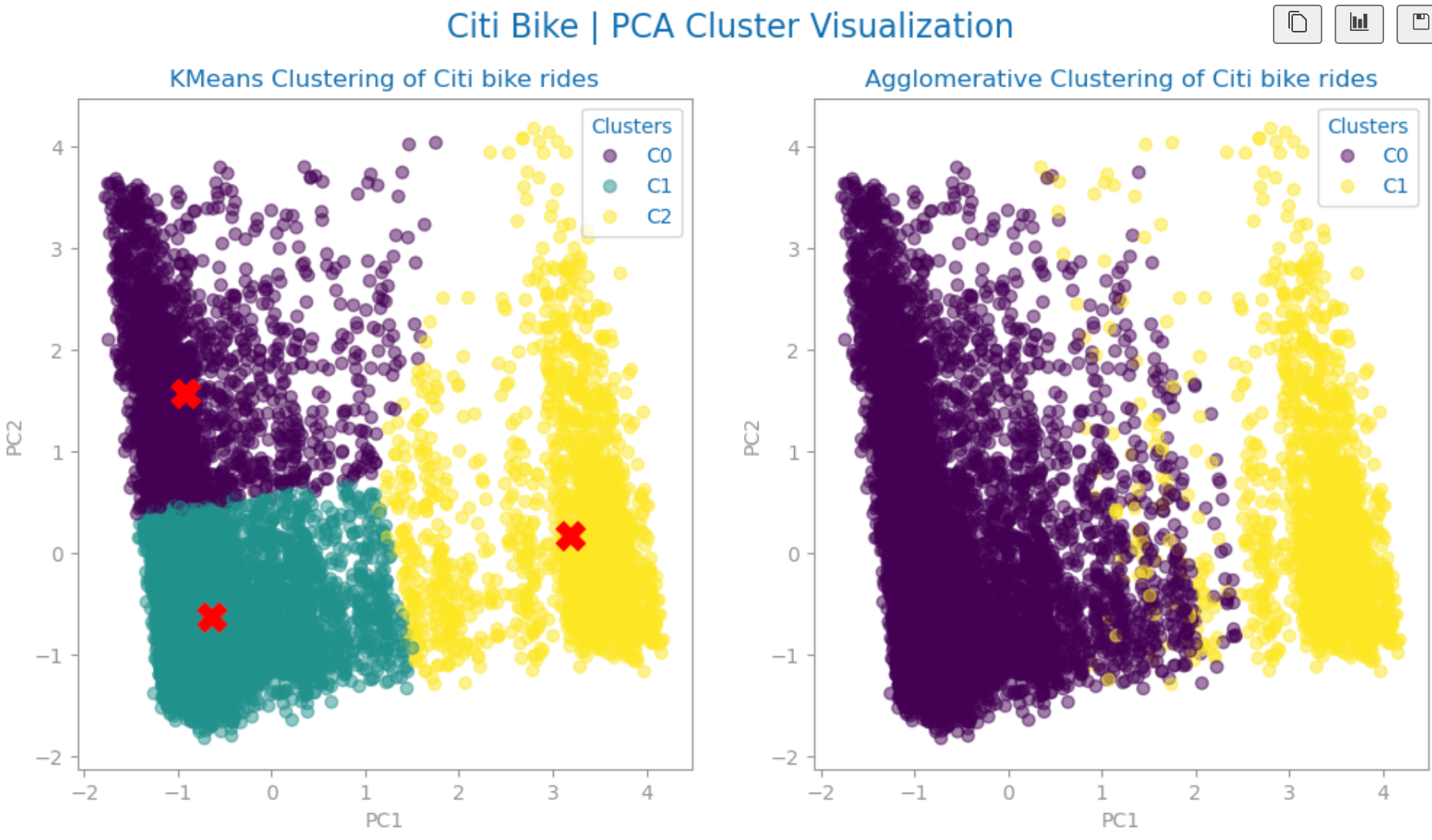


Figure 5: PCA cluster visualization in dataset 1

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

A screenshot of a graph

AI-generated content may be incorrect.

Figure 6: 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:

A screenshot of a graph showing different colored dots

AI-generated content may be incorrect.

Figure 7: t-SNE cluster visualization in dataset 2

Dataset 3: Used PCA 2D to plot the different clusters with the main 3 components. There are two distinct shapes upon which the clusters can be plotted with good separation. For Kmeans we have Cluster 1 and 2 plotted in the larger shape and Cluster 3 in yellow shape. Agglomerative Cluster has 2 clusters visibly separated in purple and yellow.

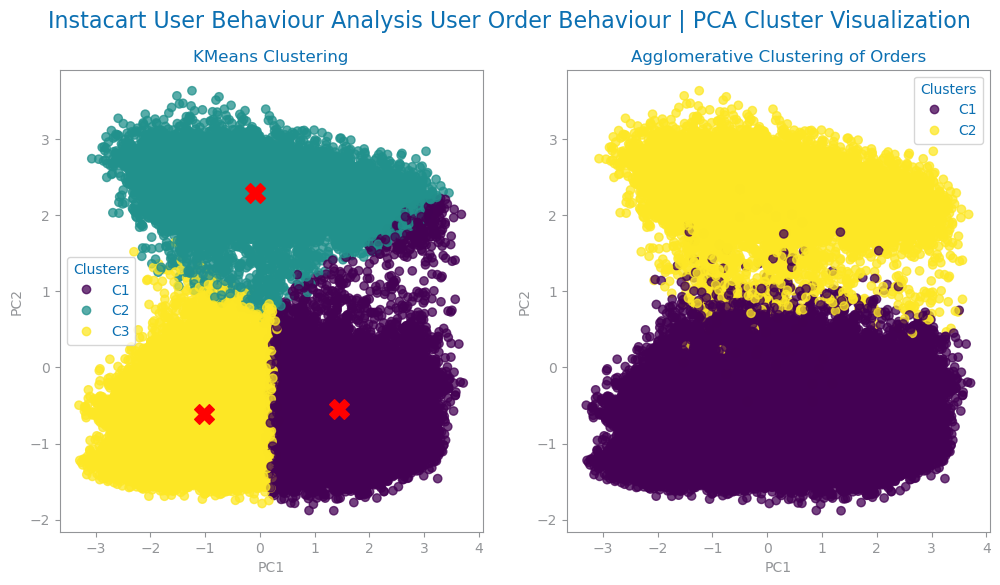
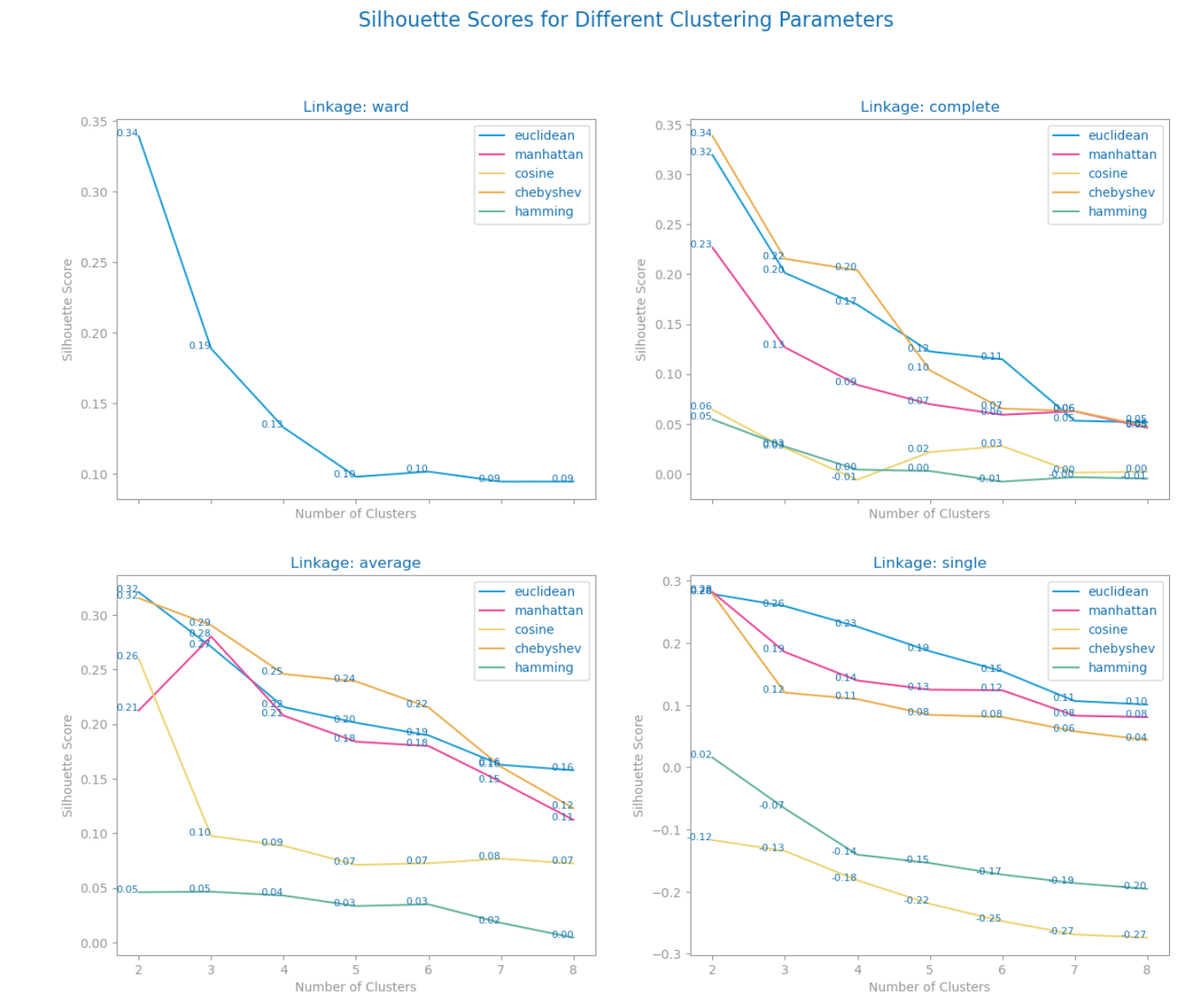


Figure 8: PCA cluster visualization in dataset 3

## Distances and methods

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



A screenshot of a graph

AI-generated content may be incorrect.

Figure 9: Silhouette study with different linkage and distance criteria in dataset 2

This investigation resulted in the following solutions for each dataset:

Dataset 1: Chebyshev, complete.

Dataset 2: Manhattan, average.

Dataset 3: Chebyshev, average

## Number of clusters

Dataset 1: 3 clusters with K-means according to the elbow method and 2 clusters with Agglomerative Hierarchical clustering using Complete linkage and Chebyshev distance.

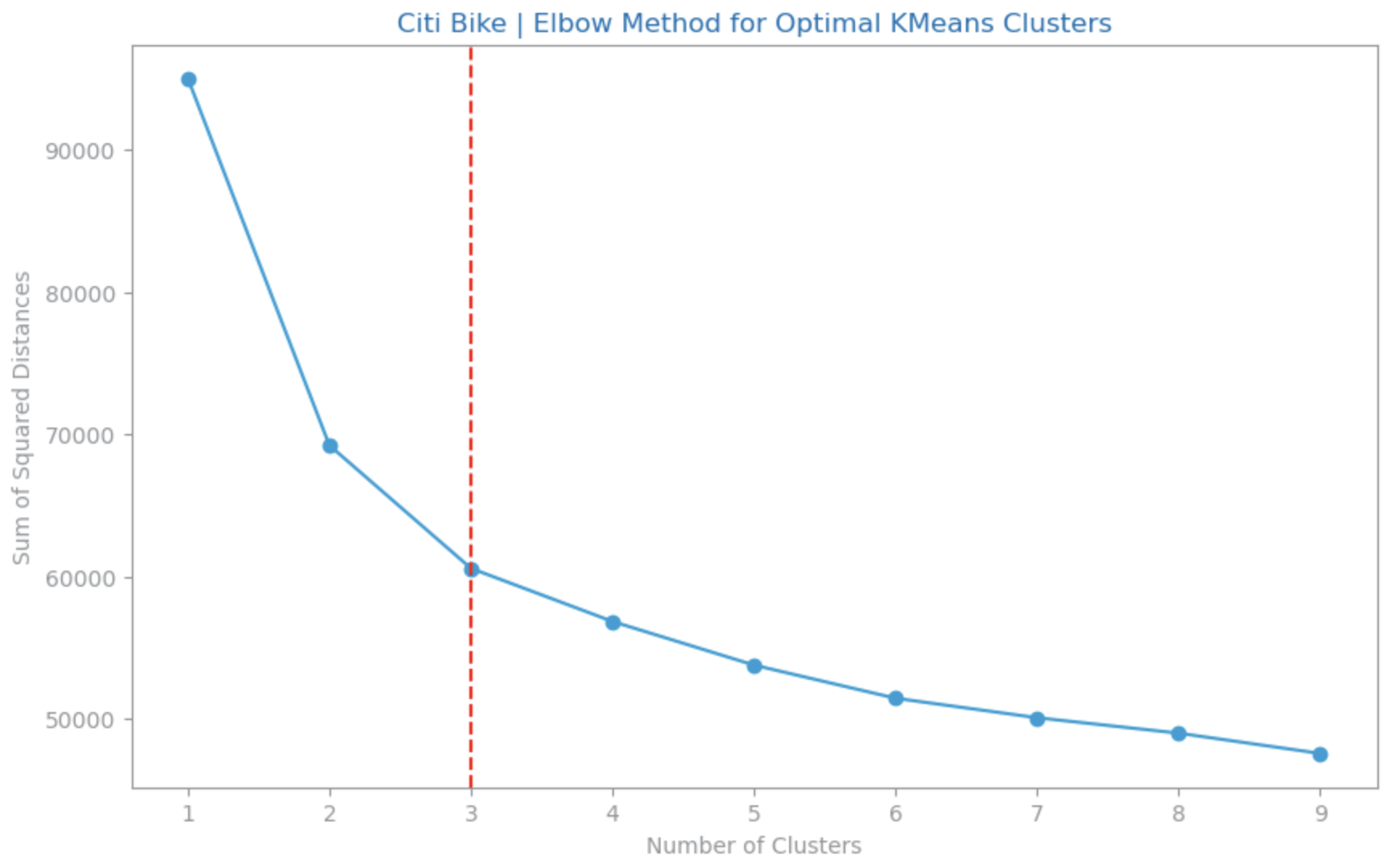


Figure 10: SSE vs number of clusters for dataset 1

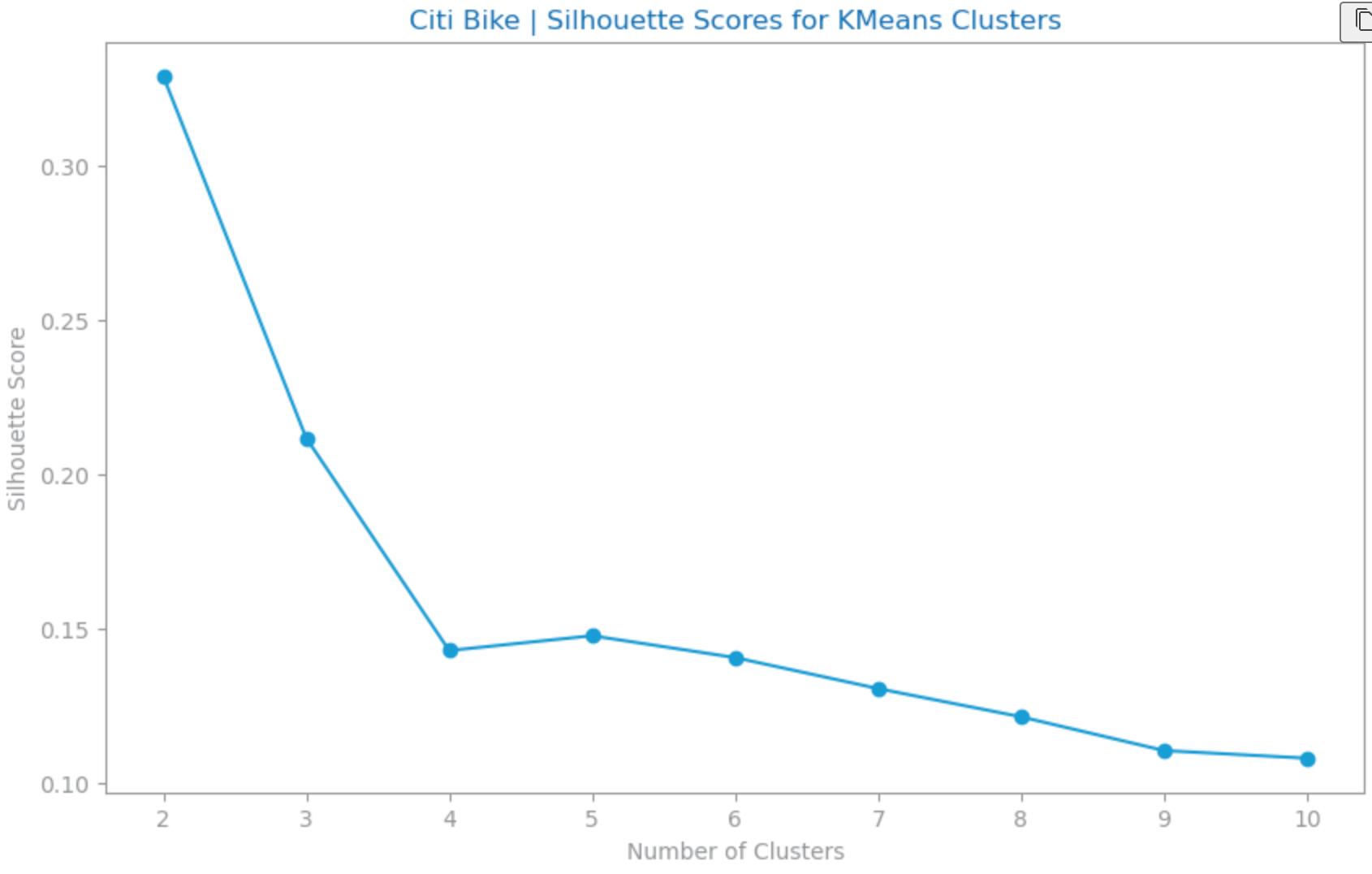
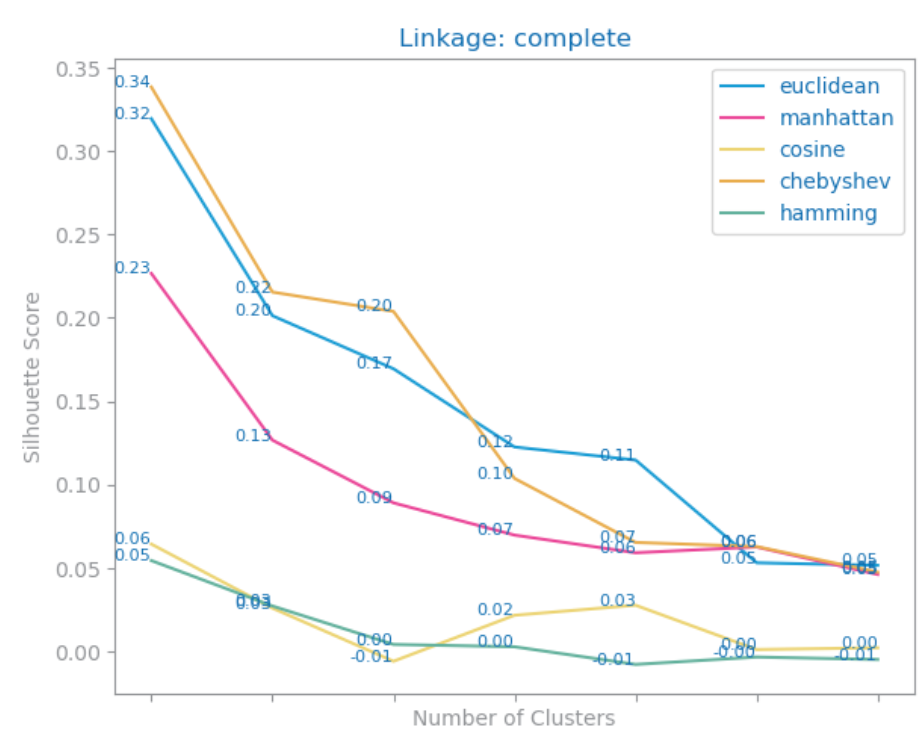
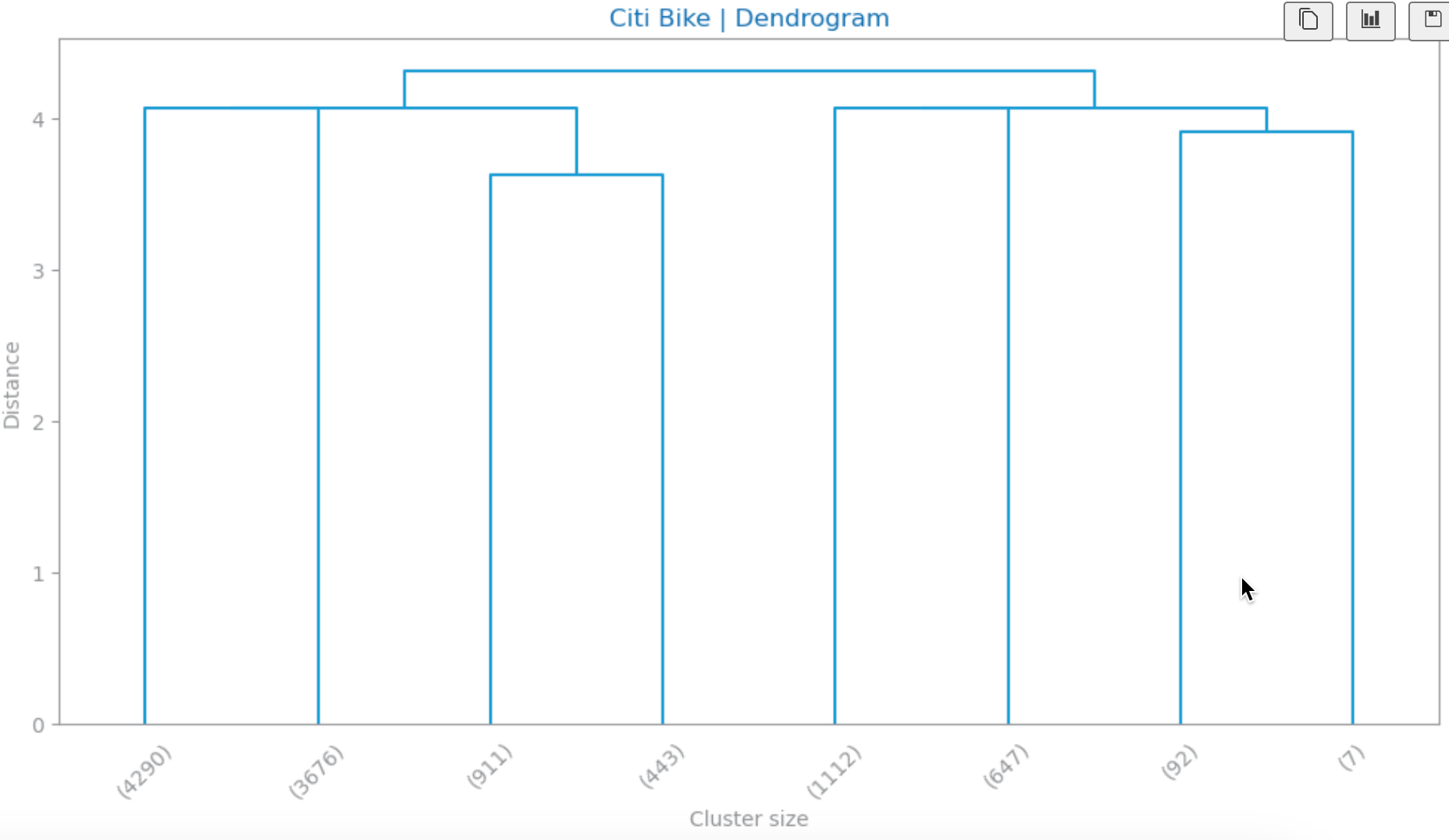
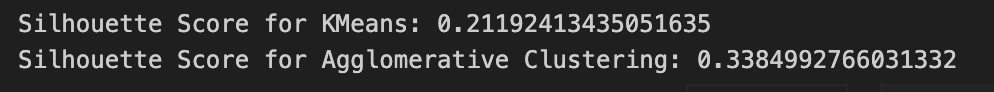


Figure 11: Silhouette scores for K Means for dataset 1







Dataset 2: 3 Agglomerative Hierarchical clusters using average linkage and Manhattan distance, and 3 with K-means. By looking at the dendrogram (Fig), 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 2 clusters, 3 clusters were considered for the Agglomerative solution.

A graph of a bar chart

AI-generated content may be incorrect.

Figure 12: Dendrogram for dataset 2

As for the K-Means and considering both the Sum of Square Errors and the Silhouette scores, we defined 3 final clusters. The elbow method example can be seen below (Fig).

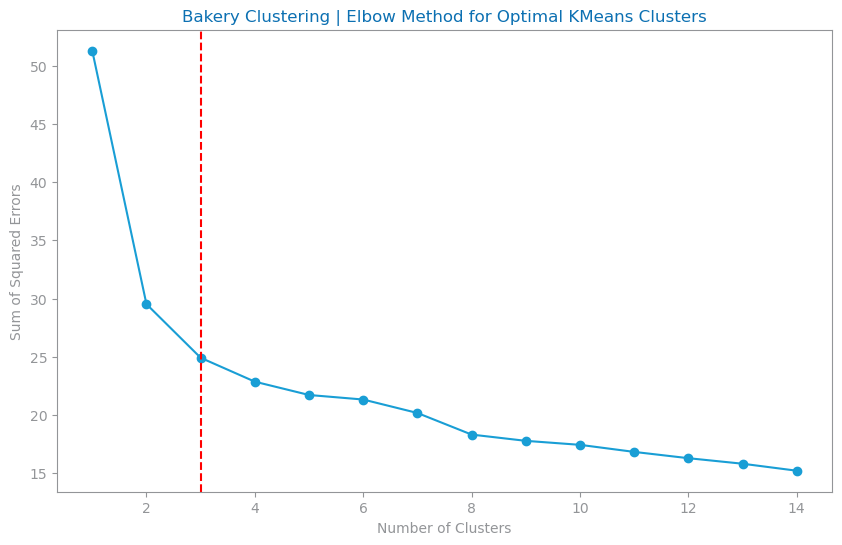


Figure 13: SSE vs number of clusters for dataset 2

Dataset 3: 3 Clusters with K-Means and 2 Clusters with Agglomerative Hierarchical Clustering using Complete Linkage and Euclidean Distance.

## Preprocessing impact

Dataset 1: To address Variables with many outliers, we truncated extreme values at 2 times the standart deviation to reduce noise in clustering. Removing highly correlated variables (pearson>0..8), like ‘end borough’ and ‘hour sin’, improved model stability and cluster separation. Robust Scaler, which handles better outliers than Standart or Min Max, was used, 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, 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), 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. In order to mitigate this impact in the model, RobustScaler 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 Agglomerative Clustering's better cohesion, K-Means offered more well-separated clusters, which provided a greater opportunity for deeper insights into the data. TDespite 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 datasets 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 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 was highly considered. However, the clustering solutions represented visually were ambiguous. Therefore, the final solution included performing scaling using Normalizer, with a silhouette score of 0.81 and 0.89 for K-Means and Agglomerative Clustering, respectively.

A black background with white text

AI-generated content may be incorrect.

Figure 14: Silhouette results clustering dataset 2

Dataset 3: From the two approaches, the approach 2 was chosen:

* Approach 1: Kmeans has a sillouette 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: Kmeans has a sillouette 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 reorder options from users. For both KMeans methods we used Elbow Method and Sillouette studies to find optimal number of clusters. For Agglomerative we used Sillhoette study comparison between distances and metrics as well as dendrogram.

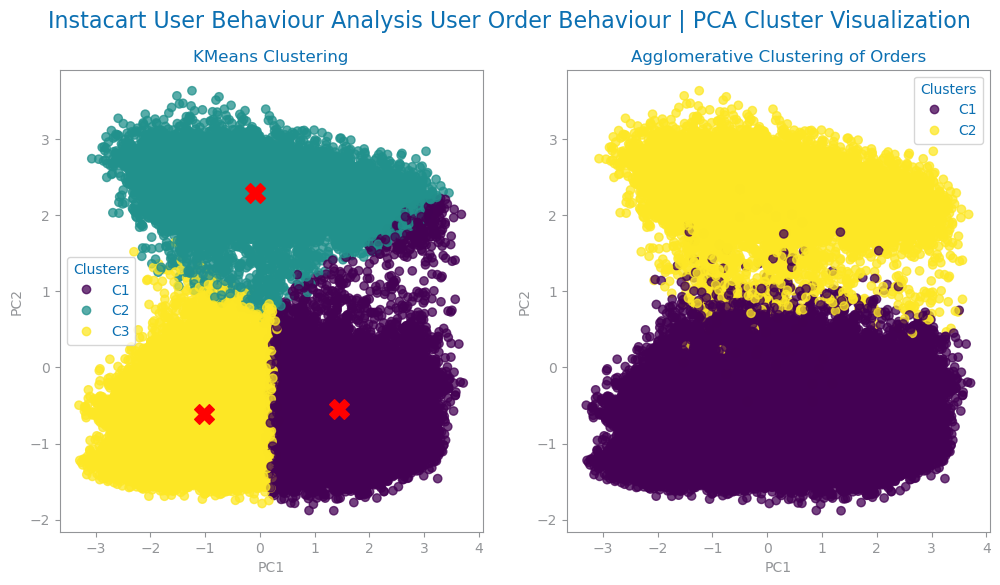


Figure 15: PCA cluster visualization for dataset 3

## Major findings (knowledge acquisition)

Dataset 1: Considering the K-Means solution, 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

* 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.

Considering the Agglomerative solution, we have 2 distinct clusters:

* C0: Moderate-length trips (2-5km), occurring on **clear-sky** days with **moderate temperatures**, primarily made by **members** during weekdays. Peak hours suggest commuting to work or regular activities.
* C1: Moderate-length trips (2-5km), occurring on **cloudy days,** with **lower temperatures**, primarily made by **members** during weekdays. Varied trip times suggest quick, occasional trips.

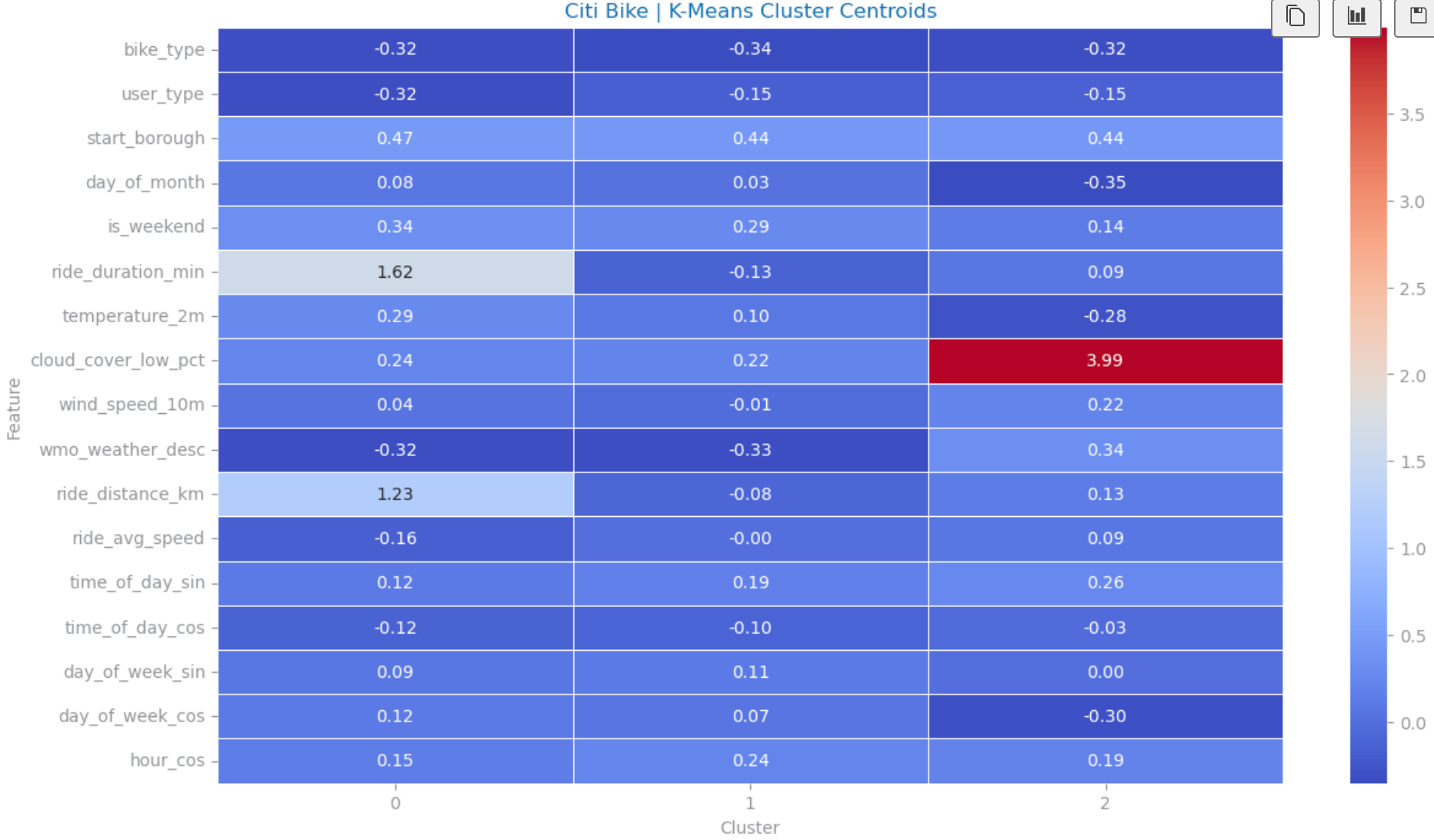


Figure 16: K-Means Cluster Centroids Heatmap dataset 1

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

* **C1**: more angbutter (South Korean pastry), almond croissants, pandoro, buying at the beginning of the year, high purchase value (spending more in their purchase in the bakery),
* **C2**: bread, croissants, and drinks, buying throughout the whole year, and purchase value ranging from low to high, buying on weekends
* **C3**: orange pound, low purchase value, buying at the end of the month

Dataset 3:

Looking at Kmeans Clustering which has 3 clusters we see:

* C1: Cluster with highest value on total orders (15-30+orders), mean products(9-17 products) mean reorder rate(50-70%) and day and time variety. These are most likely the top customers for Instacart which buy frequently and in high quantity. They have high order time and day variety buying on all days of week and buying specially from 10h to 15h times. They represent 33.7% of the user sample.
* C2: Cluster has the highest top order hour and has the lowest mean peak time of day rate and top order time of day enc. This means that these users buy at at a totally different time than the other clusters. They represent 20.4% of the user sample
* C3: Highest top order time of day enc and highest mean lag between orders. Lowest total orders and order variety features. This could mean that these users may tend to buy at the most popular times of day and week but they have a low buying frequency. They also have less chance buy on weekend and reorder items. They represent 45.84% of the user sample
* All clusters have similar top order dow and std lag between orders.
* More than half of the users from C1 and C2 reorder their products while C3 do not. Top order day for C1 is Monday while C2 and C3 have it as Monday
* All 3 Clusters share the same top 9 Departments bought: Produce, dairy eggs, beverages, snacks, frozen, pantry, bakery, deli, canned goods. In Aisle breakdown where all three share same top Aisles: fresh fruits, fresh vegetables and packaged vegetables fruits. This shall be relevant for Pattern Mining task.
* On all clusters there are top 4 products: Banana, Bag of Organic Bananas, Organic Strawberries, Organic Hass Avocado and Organic Baby Spinach. Standouts: C2 with Strawberries in 7th and C3 with Lemons in 6th place.
* C3 and C1 are the top clusters in total share and have similar behaviour 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 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].

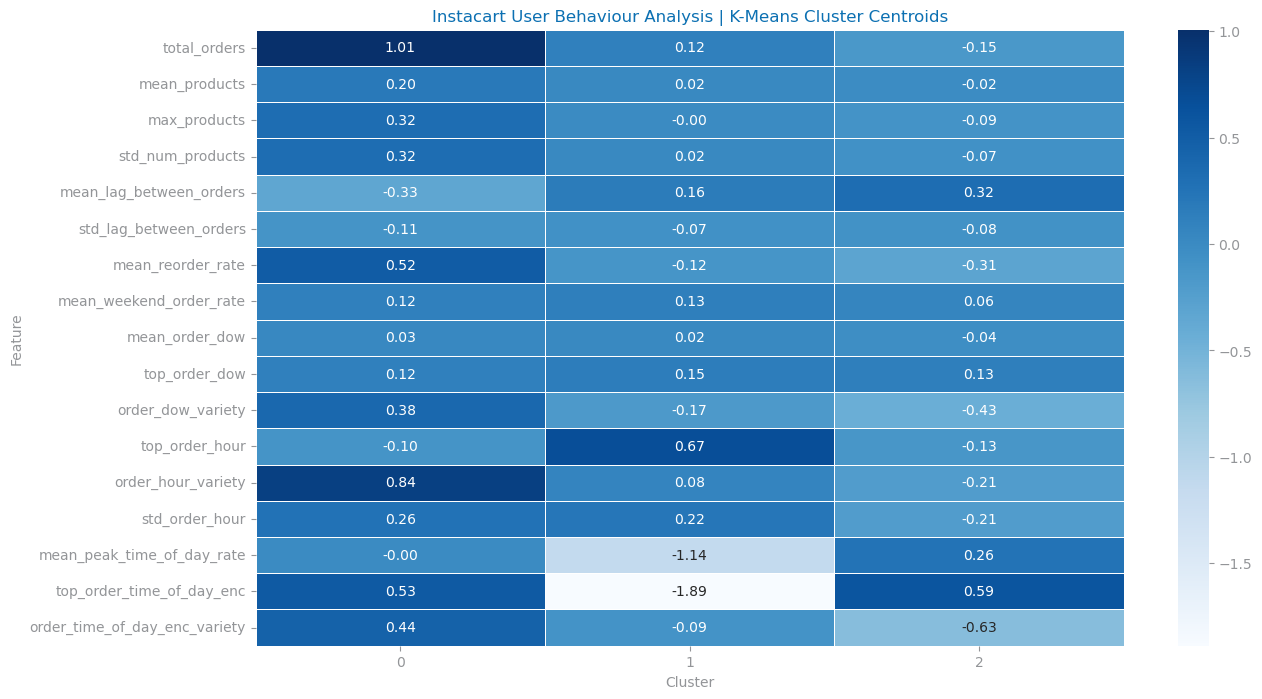


Figure 17: K-Means Cluster Centroids Heatmap dataset 3

Although we do not find as interesting insights from Agglomerative clusters because these have less detail and feature variance between clusters there are some highlights:

* C1: represents 78.8% of user sample vs C2 with 21.2%
* C1 has more total orders (median 10 vs median 7 on C2) with a higher number of orders distribution as well as more products bought in general
* C2 buys more during offpeak times (evenings around 17-19h) vs C1 (peak time at 10h-15h)

# PATTERN ANALYSIS

Observe unexpectedly frequent patterns (pattern mining) with statistical significance (<0.05), and unexpectedly discriminative patterns (association rules) by prioritizing support, confidence and lift.

Dummification + Discretization

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

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

Dataset 3: the goal was to understand which product categories and names 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 datasets. 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, using only the features related to the bakery products.

Dataset 3:

## Preprocessing impact

Dataset 1: For pattern analysis, a sample of 320k records was used to reduce processing time. Outliers in the ‘r'ide avg speed’ column were removed, and low-variance variables (e.g., ‘wmo weather desc rain’...) were removed during feature selection. Non-binary variables were dummyified by discretizing them into three bins (low, medium, high) with equal proportions to balance the data. 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 dummified (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 dataset has department, aisle and product name as features that allow us for this 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

## Detailed assessment

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: At each category level of the dataset patterns were found:

1. Department: 19 patterns were found with minimum lenght of 2, max p value of 0.05, minimum 60% confidence and 1.4 lift. Additional 27 patterns were found at 1.2 lift. These achieved between 9% and 23% support.
2. Aisle: 10 patterns were found at 1.4 lift and 60% confidence with p value of 0.10. Additional 5 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. Additionally 5 patterns were found with length >= 3

A Binomial Distribution was also applied to obtain significance levels for the patterns found.

## Major findings (knowledge acquisition)

Dataset 1: For both member and casual users data, 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. 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. However, none of these patterns provided unexpected or meaningful insights for analyzing or comparing member and casual user rides.

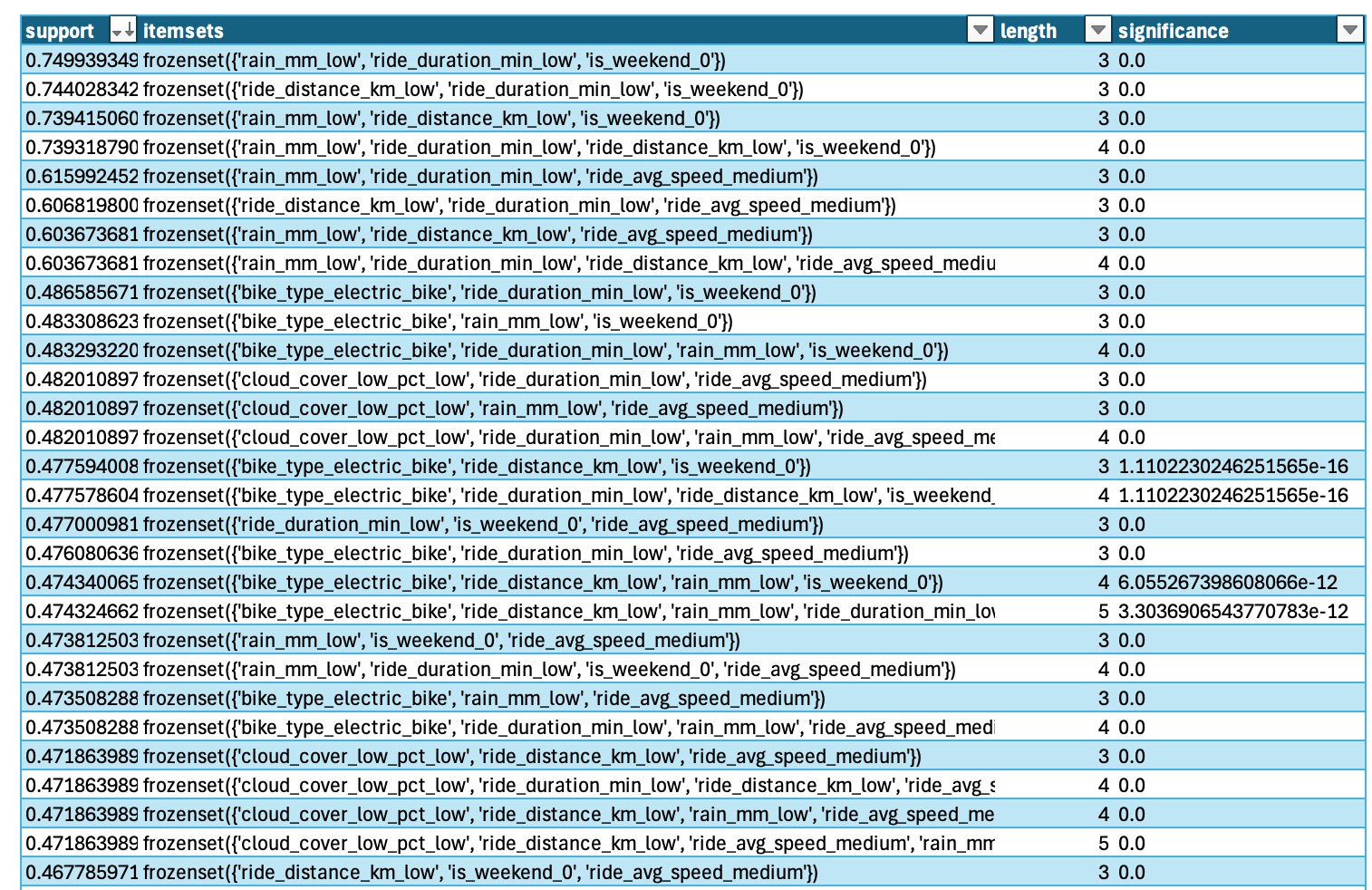


Figure 18: Top 30 Frequent item set discovered for dataset 1 (member users)



Figure 19: Top 30 Association Rules discovered for dataset 1 (member users) ordered by confidence.

Dataset 2: Overall, very low support (around 7% max), and quite high confidence on the rules (70%+). 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 angbutter are also inclined to buy bread.

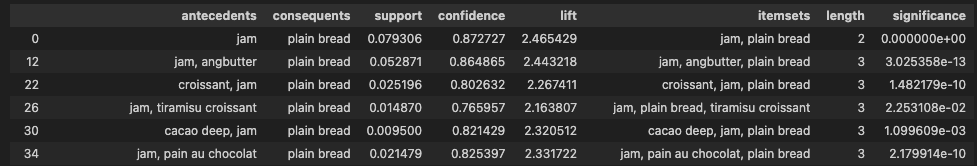


Figure 20: Main Association Rules discovered for dataset 2

Dataset 3: There is a natural tendency to decrease 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.

Department Patterns:

* Produce and Dairy Eggs are top consequents achieving 12-24% support, 70-90% confidence and at least 1,2 lift
* Support wise, top itemset is (frozen, produce with dairy eggs) at 24% support at 81% confidence and 1.2 lift.
* With a slightly higher lift at 1.35 we have frozen, produce and dairy eggs.

Aisle Patterns:

* Besides frequently seeing fresh vegetables or fresh fruits as top consequents that can achieve 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 ttemsets with length 3 can achieve support as high as 11.7%. Eg.: The itemset (yogurt, fresh vegetables, fresh fruits) had a 81% confidence and 1.45 lift.

Product Patterns:

* For length=3 itemsets, only fruit or vegetable products patterns have some significant support (0.5-0.6%). Example: (Large Lemon, Organic Baby Spinach, Banana)
* In length=2 itemsets, Banana or Bag of Organic Bananas are a frequent consequent with itemsets with 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. Confidence is not very high for these though (~30-40%).

# APPENDIX

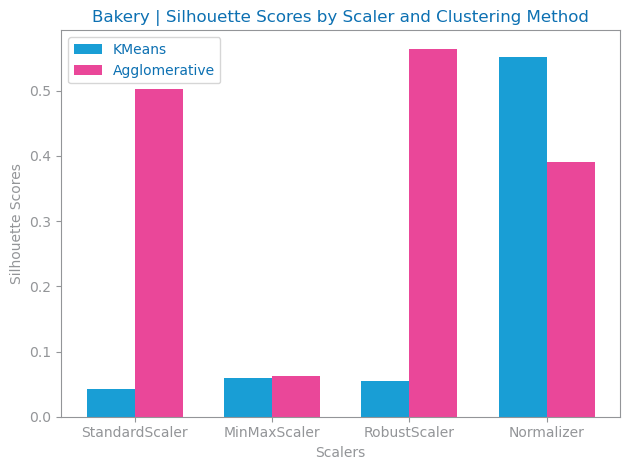


Figure 21: Silhouette study by Scaler method and Clustering method

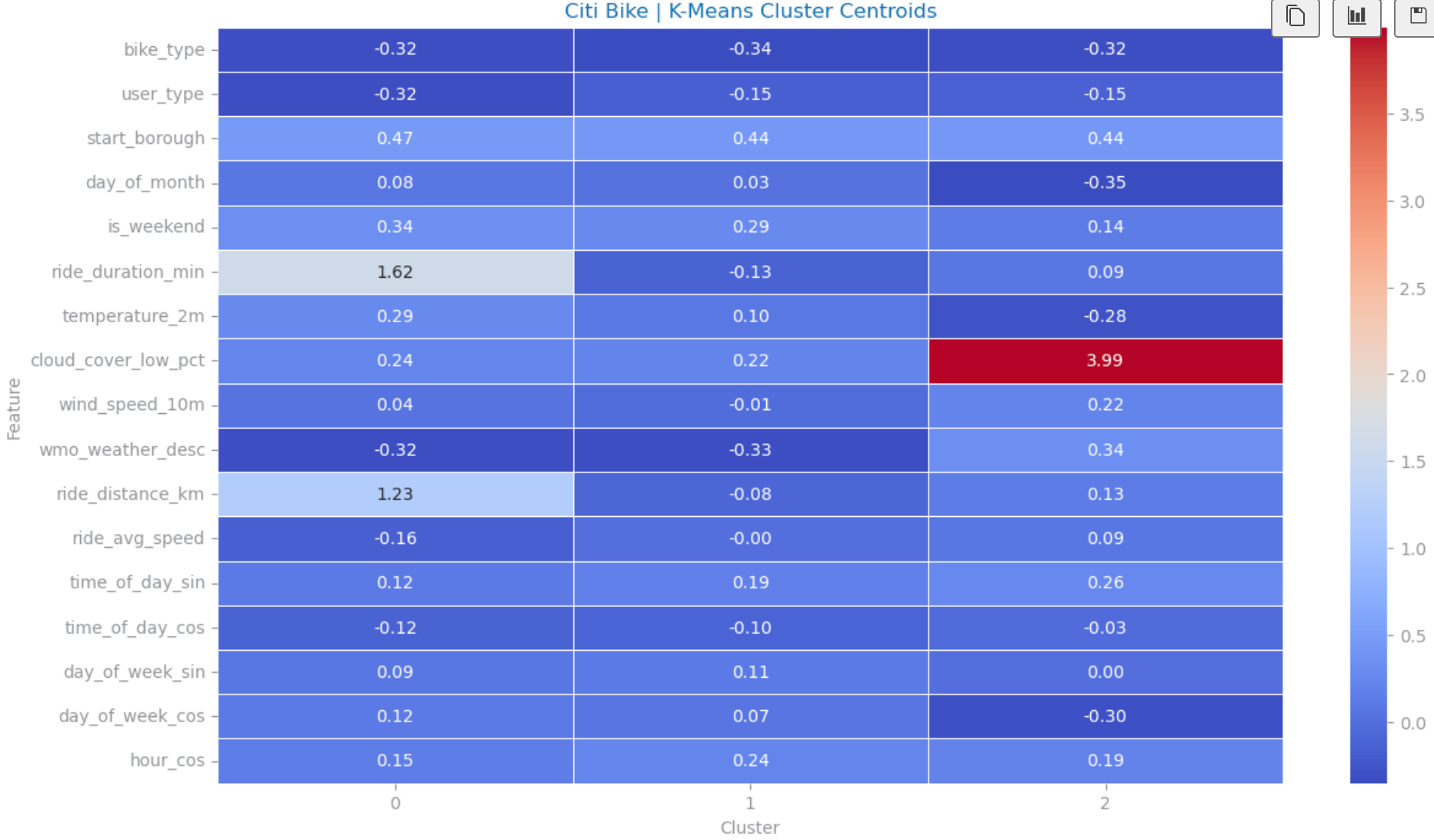


Figure 22: K-Means Cluster Centroids Heatmap dataset 1

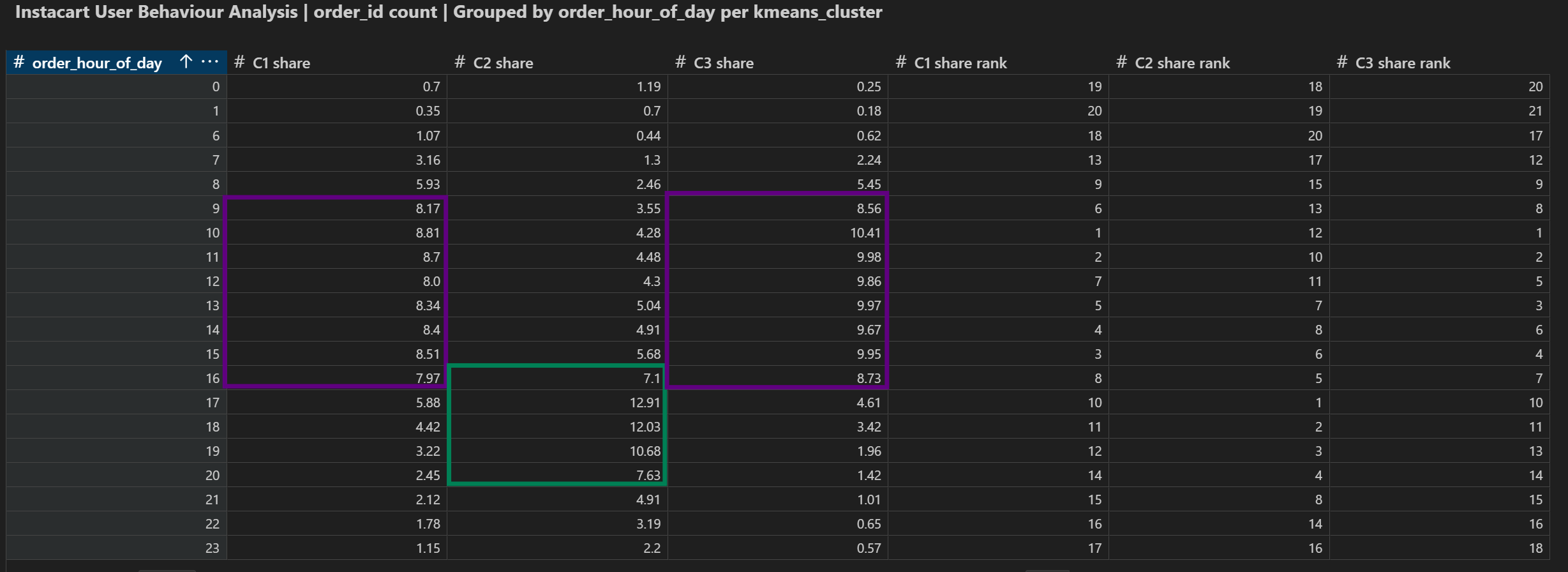


Figure 23: Order hours of day per Kmeans Cluster Dataset 3

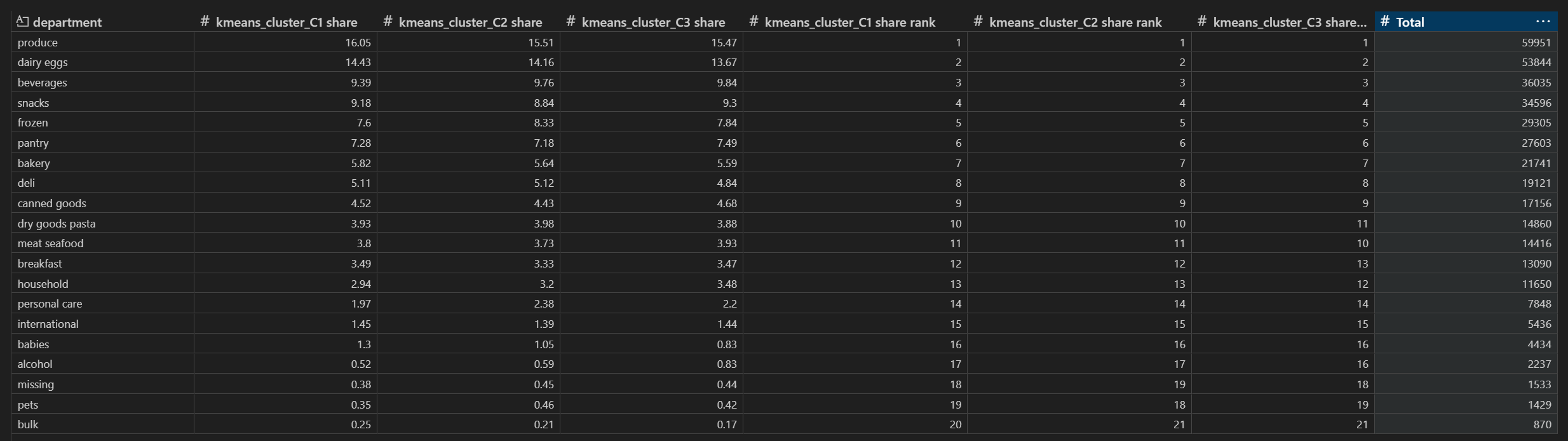


Figure 24: Department Orders per Kmeans Cluster Dataset 3

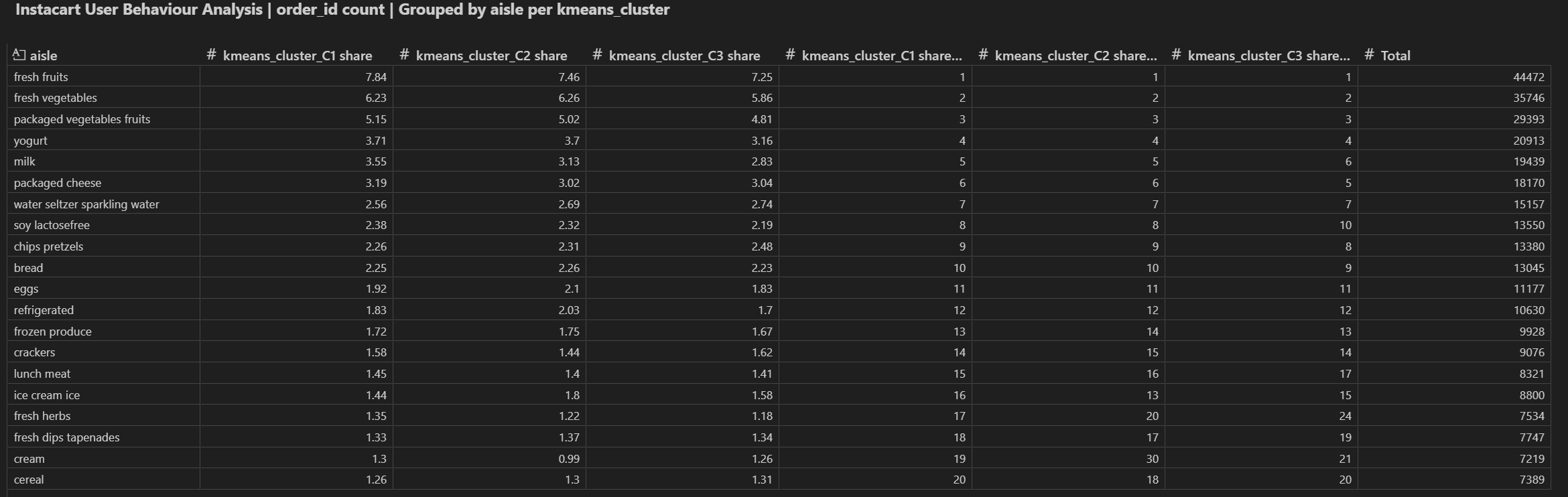


Figure 25: Aisle Orders per Kmeans Cluster Dataset 3

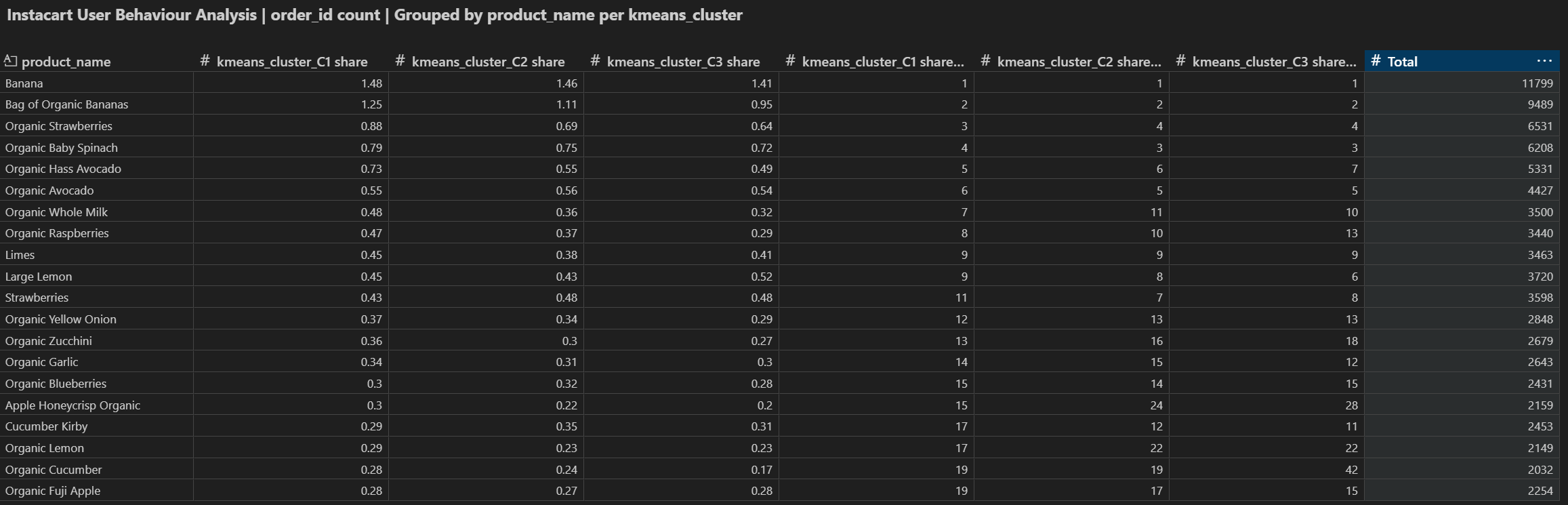


Figure 26: Product Orders per Kmeans Cluster Dataset 3

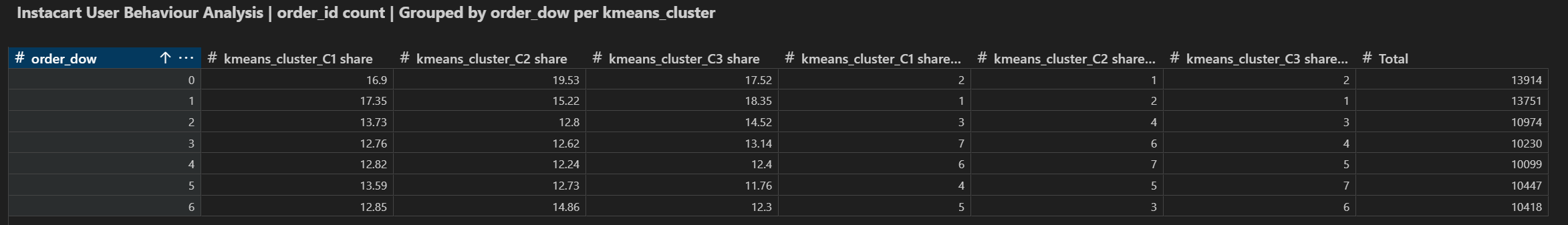
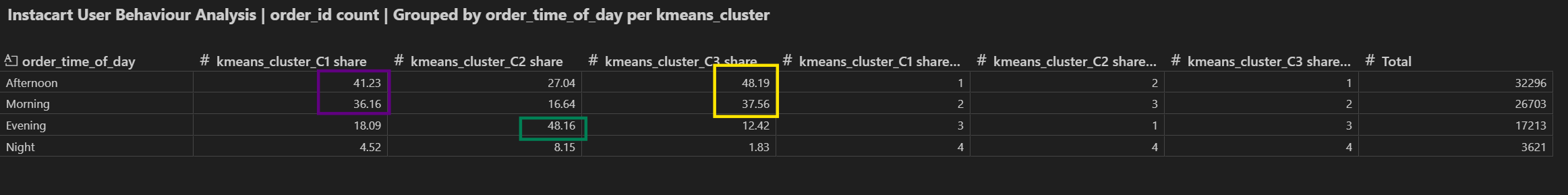


Figure 27: Day of Week Orders per Kmeans Cluster Dataset 3

Figure 28: Time of Orders per Kmeans Cluster Dataset 3

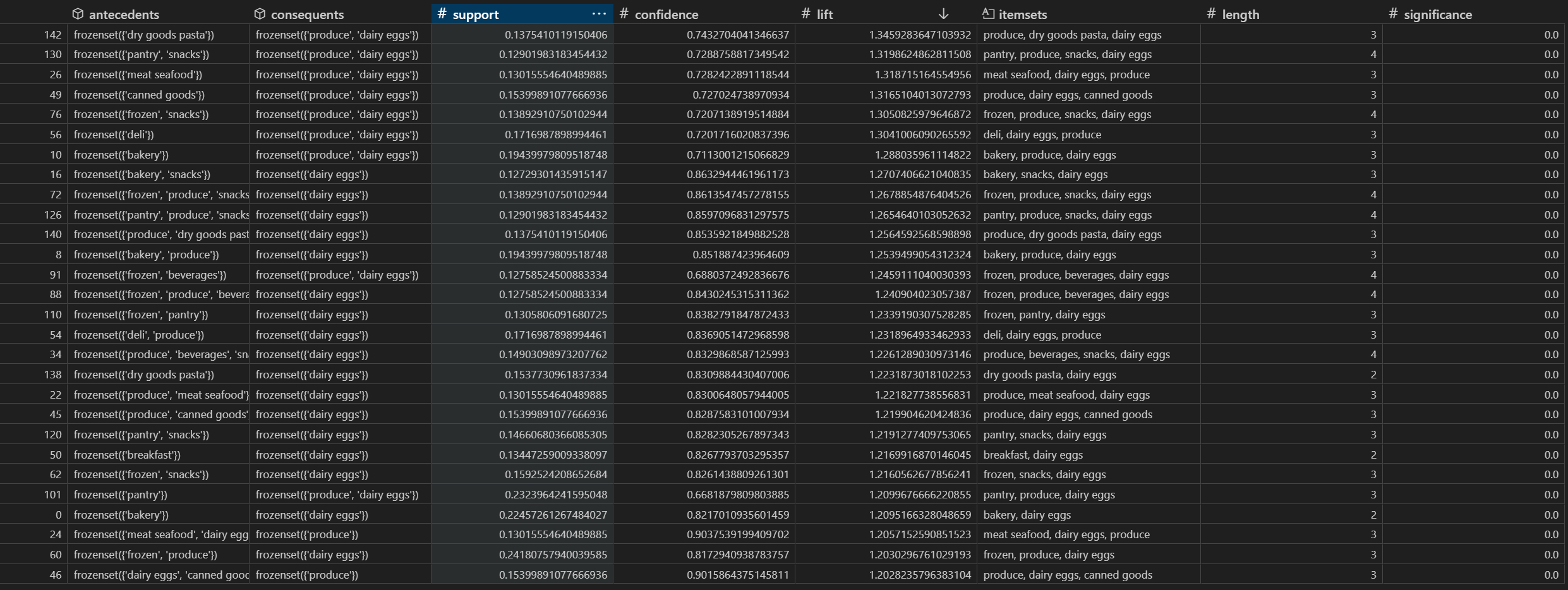
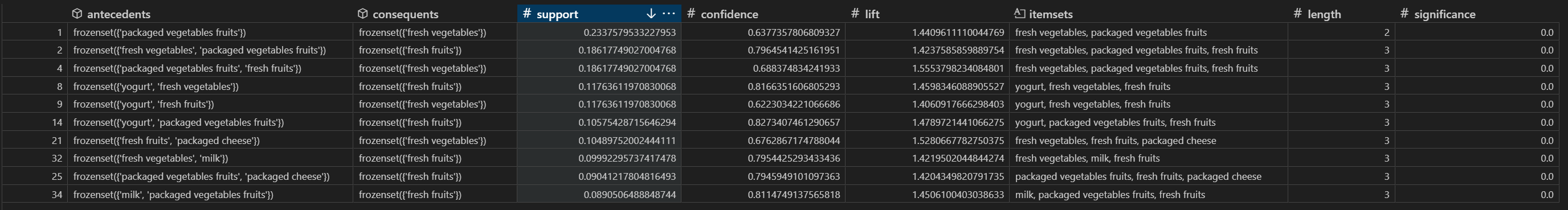
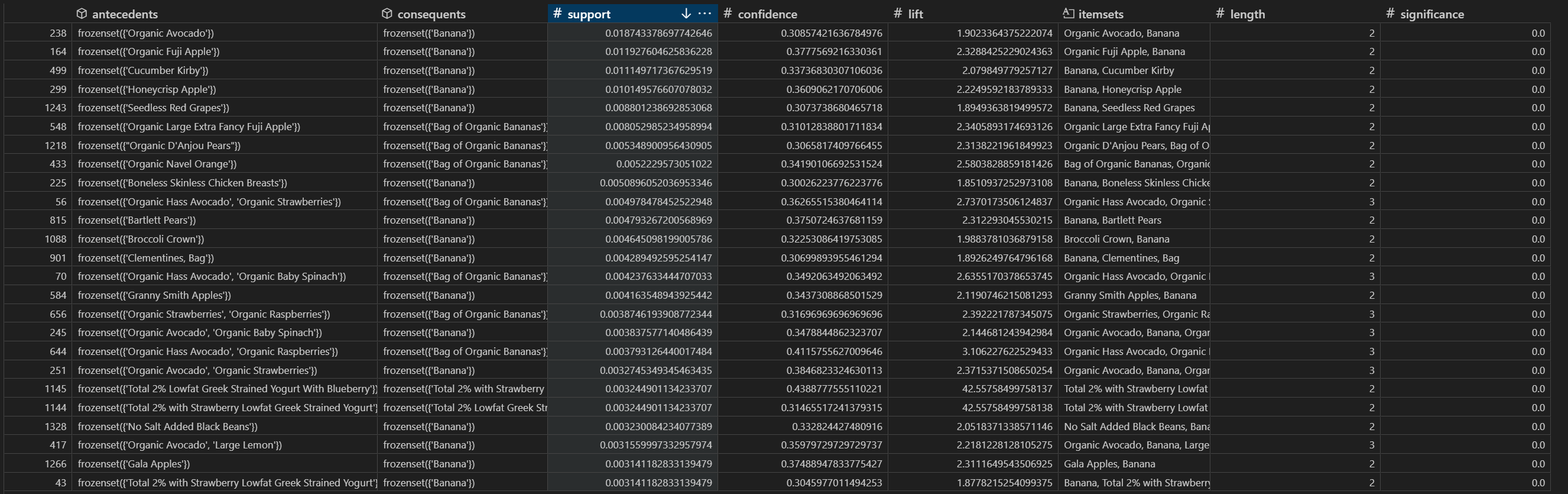


Figure 29: Department top Patterns in Dataset 3

**Figure 30**: Aisle top Patterns in Dataset 3

Figure 31: 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/>