

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**

Major findings.

Dataset 2: New York City bike-sharing system, where each observation represents a bike trip. 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.

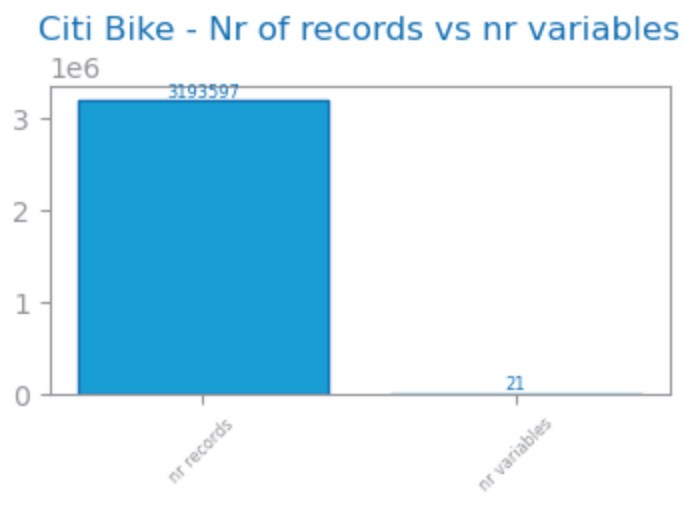
In all datasets we did preprocessing steps like feature engineering, encoding, outlier study and feature relevance/redundancy study.

In all datasets we clustered with hard approaches with the objective of......: in dataset 1..... in dataset 2.... in dataset 3....

In all datasets we did pattern mining with association rules as we wanted to ..... in dataset 1..... in dataset 2.... in dataset 3....

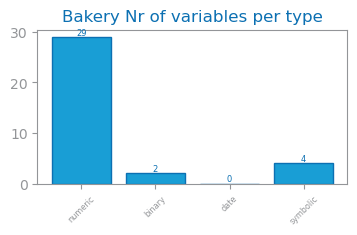
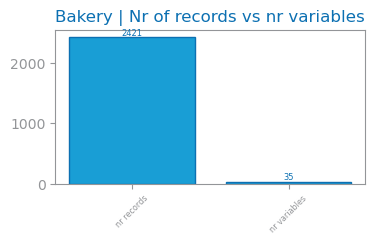
# DATA PROFILING

## Statistical analysis

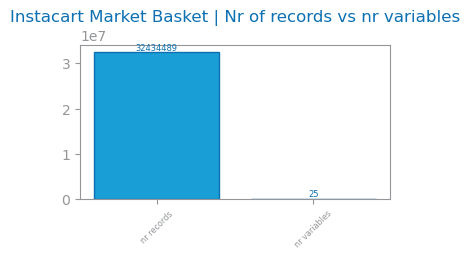


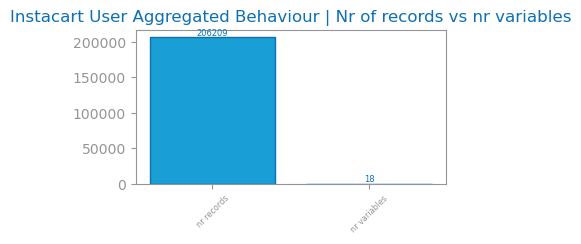
Dataset 1: ~3 million observations, 21 initial variables (11 numeric, 2 binary and 7 symbolic. In preprocessing, new time variables were created. Based on stations names columns, we were able to get also start and end borough (Manathan, Bronx,..). With latitude and logitude of the stations, new columns like ride distance and ride speed were created. Additionally, Meteorological data from New York City was also included to analyze the impact of the weather conditions on ride behavior. As several columns had major outliers with diferente scling values, outleir truncation was applied (nr\_stdev=2) for these columns, ride\_avg\_speed and ride\_duration\_min that were removed from the dataset. As mentioned we have variables with really diferent scaling, in order for that don’t impact the kmeans, based on distance, we also normlaize our dataset with robust Scaler.

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 the categoric variables were studied using bar charts.



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.





## 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 removed with low variance (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) were removed: ‘mean\_weekend\_order\_rate’,’mean\_reorder\_rate’,’mean\_peak\_time\_of\_day\_rate’. It would be interesting to understand the effect of weekend or reordering on clusters but low variance of these features may be explained by orders happening on all days of week and that 50% of all products purchased are reordered. Redundant features removed (>0.85): ‘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.

Goals for each dataset:

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 an at different time or days of week.

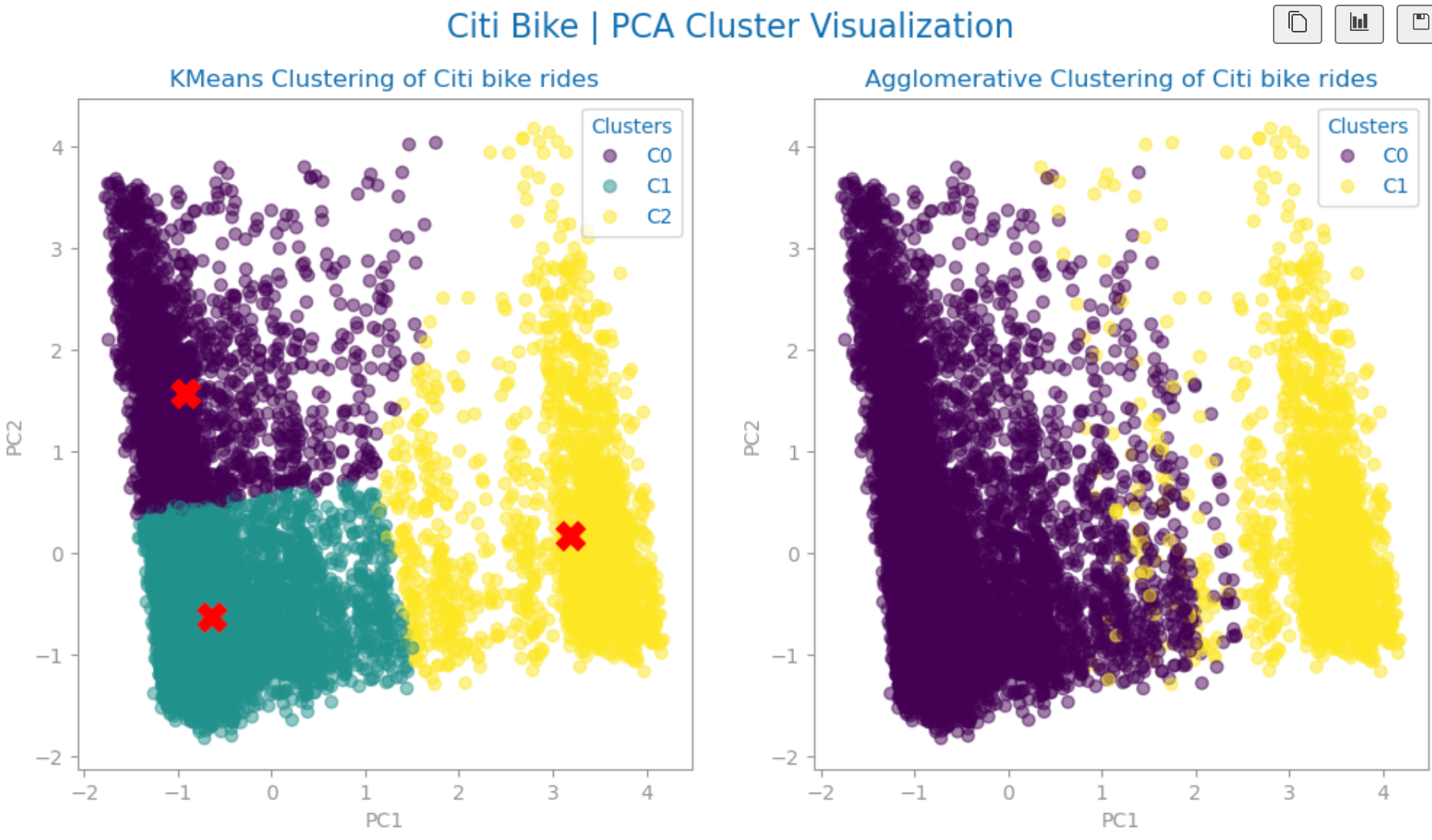
## Reference clustering solutions

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.



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.

## Distances and methods

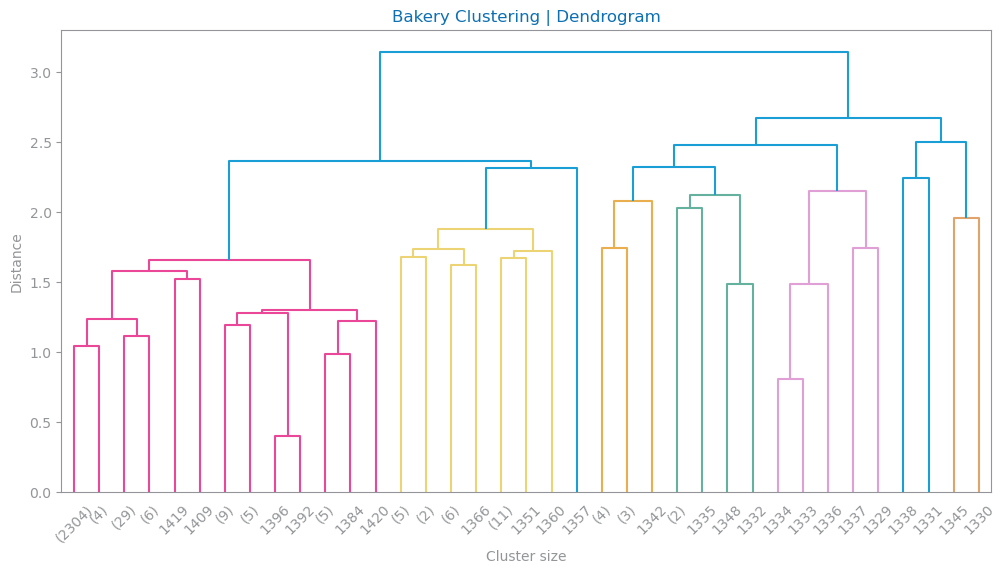
Dataset 1: Chebyshev, complete.

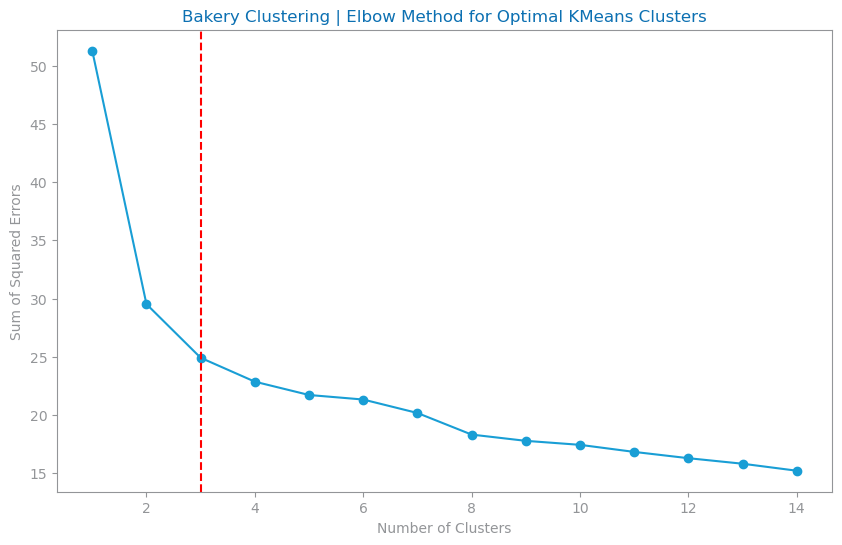
Dataset 2: Manhattan, average.

Dataset 3: Chebyshev, average with a silhouette of 0.275

## Number of clusters

Dataset 1: 3 clusters with K-means. 2 clusters with Hierarchical clustering using Complete linkage and Chebyshev distance.

Dataset 2: 3 clusters with Agglomerative Hierarchical clustering using average linkage and Manhattan distance, and 3 with K-means. 



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: as K-means is sensitive to extreme values, the missing values of “purchase value” was filled with the median instead of mean, and other rows that contained missing values were dropped, mainly from “address”. Removed the feature “total” as it was making cluster visualizations hard to interpret, due to the natural scale of the variable. Scaled using Normalizer 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. We opted to not treat any outliers in this dataset nor truncating or removing them as it presented worse Silhouette scores. ENCODING!!!!

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'.

The major difference is that the second method has more records of users with a larger number of orders that were considered as outliers, but these can be very valuable customers. More variables might also be helpful

## Detailed assessment

Dataset 3: From the two approaches, the approach 2 was chosen despite having a slightly lower silhouette score

* Approach 1:

## 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 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.

Dataset 2: considering the Agglomerative Clustering solution, the following clusters were identified:

* C1: more angbutter, almond croissants, pandoro, buying in the beginning of the year, high purchase value (spending more in their purchase in the bakery), at the end of the month
* C2: bread, croissants, buying throughout the whole year, and purchase value ranging from low to high
* C3: orange pound, low purchase value, no drinks, at the end of the month

Dataset 3:

Looking Agglomerative Cluster we have 2 clusters that have less detail than the Kmeans Model. Main insights include:

* C2 has higher top\_order\_hour with normal values ranging from 16-18h vs C1 that prefer to buy in Morning-Lunchtime
* C2 has a slightly larger std\_order\_hour range (2.75-4.0h) vs C1 (2.5-3.75) which means C1 really tends to concentrate on specific times of day to buy
* C1 has larger number of orders (6-17) vs C2 (5-14)
* Other than that similiraty on other features between the 2 clusters is large

Looking at Kmeans Clustering which has 3 clusters we see:

* Cross Cluster: similar average number of products (8-14); average day of week order is around Midweek (2-4)
* C1: Low number of orders (median ~5 orders). Highest mean\_lag\_between\_orders. C1 represents the users that buy less frequently.
* C2: Buys on almost all different days of week while others have less day variety. Slightly higher mean number of products purchased than other Clusters. Top Cluster in total Orders (ranging 14-13 orders). This group also has the less shorter lag of days between orders (10-16) with higher lower range of st\_lag\_between\_orders probably due to higher number of orders. Also has the highest variety of hours where they buy.
* C3: Top order hour is 17h while others prefer during morning/lunchtime (10-15h)

# 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

Goals for each dataset:

Dataset 1: The goal was to understand the association of user behavior and environmental factors with bike usage patterns in the Citi Bike dataset

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:

Dataset 2: 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

## Preprocessing impact

Dataset 1: We looked at the features that were related to baked goods, and the remaining ones were dropped.

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

## Class-discriminative patterns (optional)

WIP

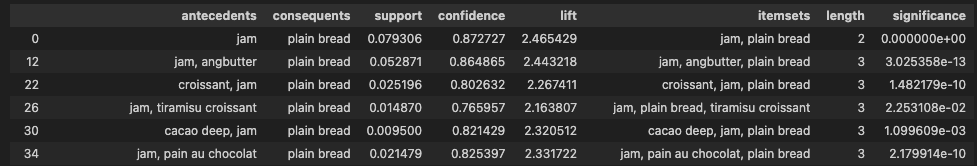
## Detailed assessment

WIP

## Major findings (knowledge acquisition)

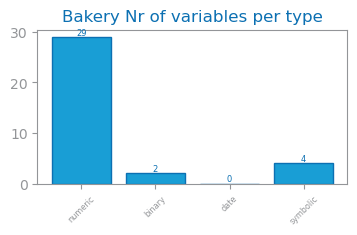
Dataset 1:

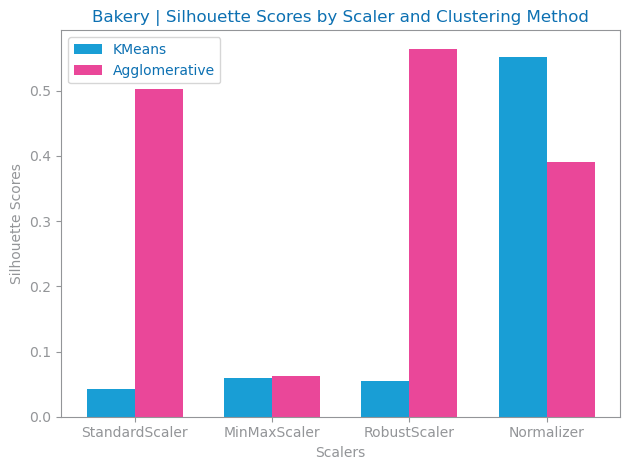
Dataset 2: Overall, very low support (around 7% max), and quite high confidence on the rules (70%+). This indicates that although the rules do not happen that often, when they do happen, the consequent also happens in 80% of the cases. We observed that people who buy jam are inclined to also buy bread,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

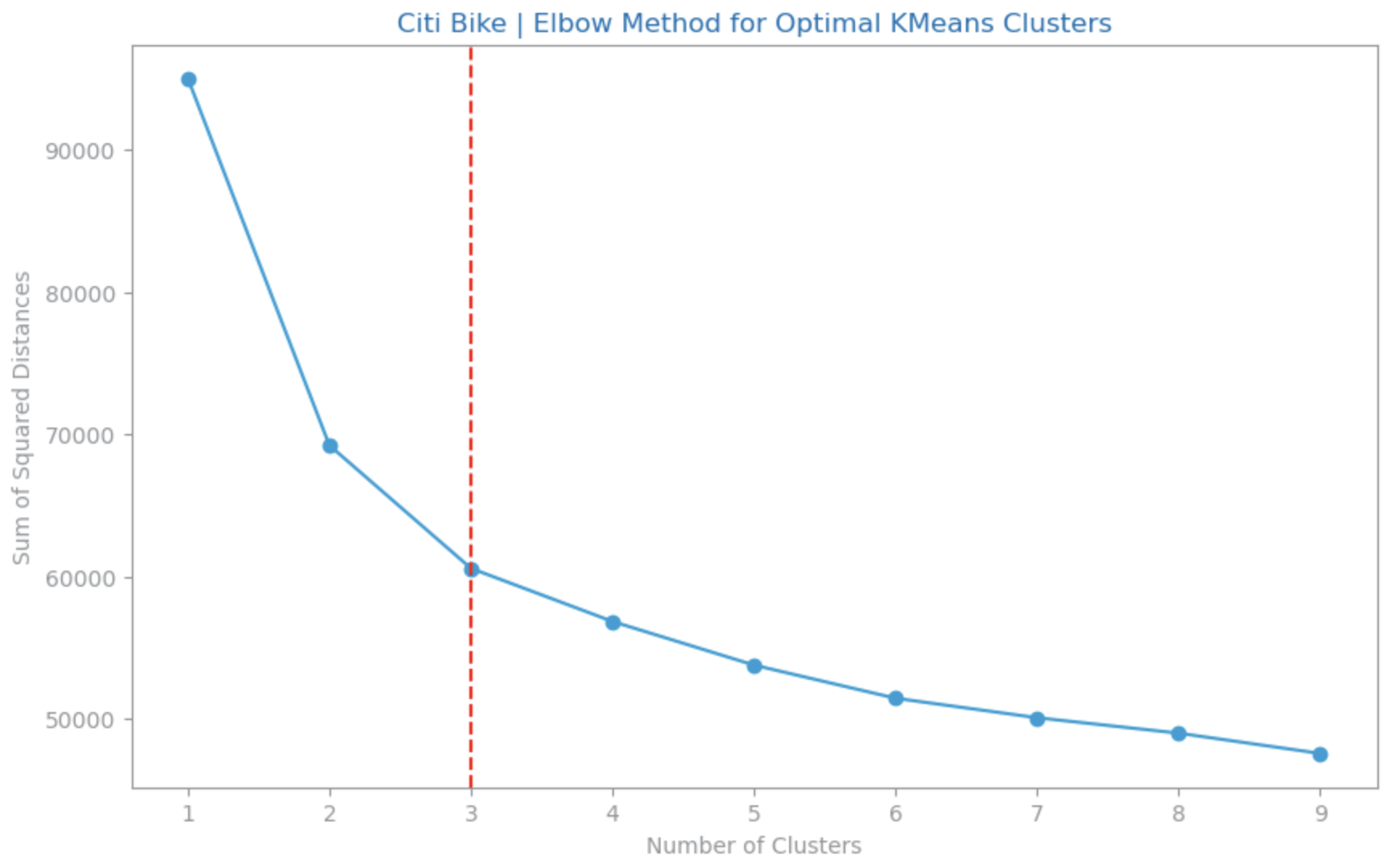


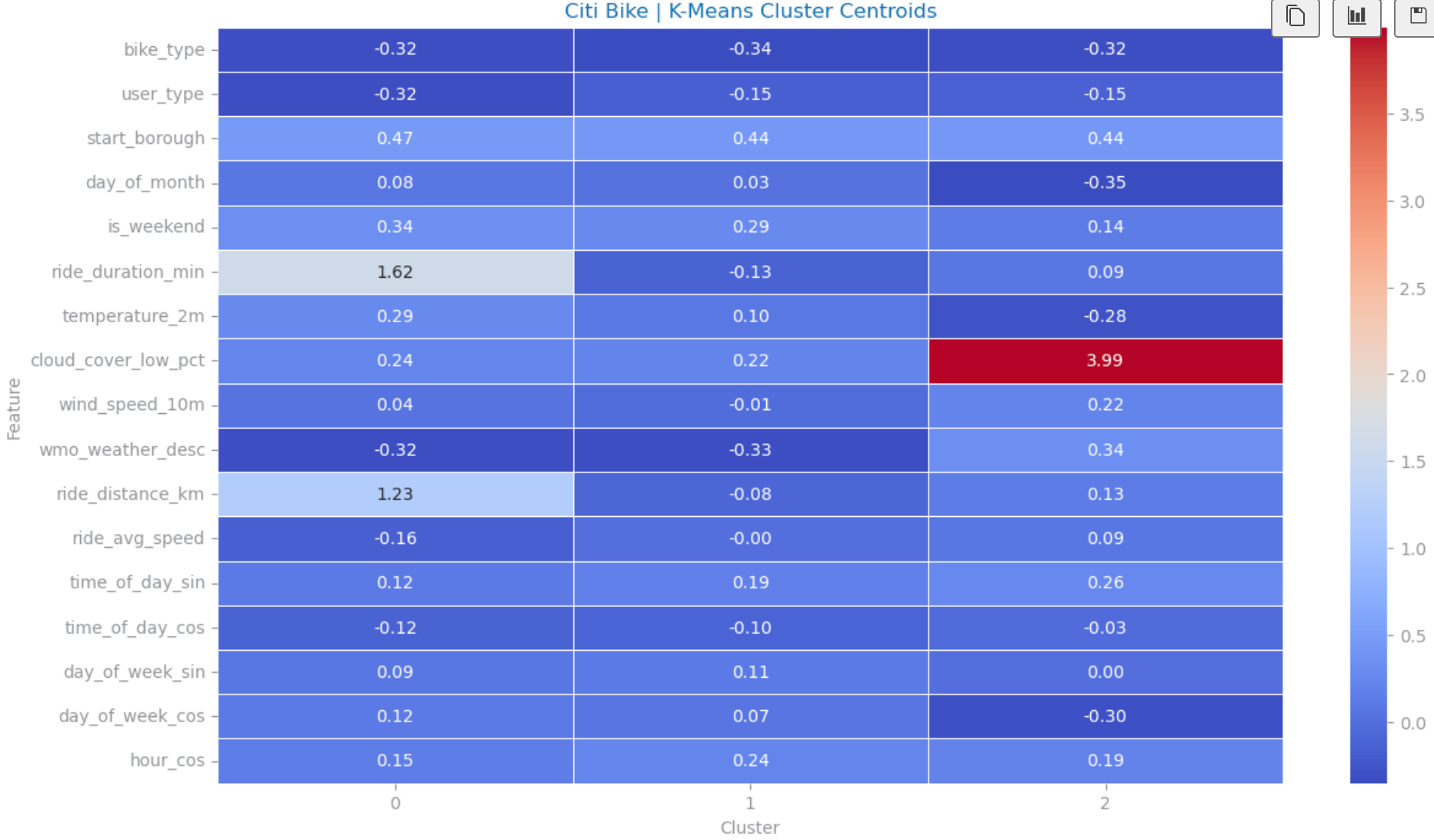
# APPENDIX

Any less focal results or observations can be placed here.









# REFERENCES (*optional*)

1. Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825. DOI= <http://doi.acm.org/10.1145/161468.16147>.
2. Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.