NOVA IMS Information Management School

Mestrado em Data Science and Advanced Analytics, com especialização em Data Science

ABCDEats Inc.

Project of Data Mining

Abstract

This project involved segmenting ABCDEats Inc.'s customers into distinct groups based on shared characteristics or behaviours, aiming to enhance the company's understanding and marketing strategies. The study included detailed data analysis using a dataset with 52 numerical and 4 categorical features. We addressed data cleaning and preprocessing by correcting data types, resolving inconsistencies, handling missing values, and managing outliers. Additionally, we developed ten new features to deepen our insights into customer behaviours, ensuring data integrity throughout. In order to prepare features for clustering, categorical variables were encoded with frequency encoding and one-hot encoding, and numerical features were normalized using Min-Max and Standard Scaling. We explored different clustering configurations using Hierarchical and K-Means clustering, ultimately selecting a six-cluster model that effectively captured customer diversity. Perspective Based Clustering was also tested, dividing features into categories like preference, purchase behaviour and age/time, and applied Hierarchical Clustering within these frameworks. This method confirmed that a combined approach across different perspectives would deliver the best results. The final customer segments from the combined perspective and hierarchical clustering analysis were the following: Group 0 - Regular but high spending customers; Group 1 - Largest group of regular customers; Group 2 - Frequent orders but low loyalty; Group 3 - High frequency, small orders; Group 4 -Best spending customers; Group 5 - Least spending customers. The project concluded with the development of a tailored marketing strategy based on the defined customer segments. Furthermore, a user-friendly interactive interface was created to allow stakeholders to easily access and analyse the clustering results.

Group 19

Jan-Louis Schneider, 20240506
Marta Boavida, 20240519
Matilde Miguel, 20240549
Sofia Gomes, 20240848

Fall Semester 2024-2025

Table of Contents

1 Introduction

In this project, you will assume the role of consultants for ABCDEats Inc. (ABCDE), a fictional food delivery platform collaborating with a variety of restaurants to provide diverse meal options.

In this way, our group went through several phases, namely: data exploration, data cleaning, preprocessing, create new features, transform data, feature selection, clustering and we did an iterative application.

Through this report, you can get a detailed explanation of everything that was done to achieve our goal. The code that made this project possible is available at the following link https://github.com/Gomsofi06/Data_Mining_NOVAIMS

2 Data Exploration

First, we did an initial exploration of the data, such as looking at the first ten rows, print the name of all columns, look at the types of each column and see some statistics.

We conclude that there are 52 numerical and 4 categorical columns. In addition, we analysed each column in detail and discovered that: most of the people who order are between 23 and 31 years old, the person who made the request who is the youngest is 15 years old and the person who made the request who is the greatest is 80 years old (column customer_age); the average number of unique suppliers the customer has ordered from is 3, the min of unique suppliers is 0 and the max of unique suppliers is 41 (column vendor_count); The total of products orders are 31888, most people order between 2 and 7, there are people who ask for nothing and people who ask for 269 products (column product_count); Most people order between 1 and 3 from chain restaurant, but there are people that orders from 83 chain restaurant (column is_chain); the days from the start when the customer first placed an order is between 7 and 45, the max is 90 (column first_order); the days from the start when the customer most recently placed an order is between 49 and 83, the max is 90 (column last_order). Also, we discover that several columns have values 0, such as 'CUI_Cafe', 'CUI_Desserts' and others.

After that, we looked in more detail at the types of each column, and realised that the type of the columns customer_age, first_order and HR_0 was float, so we had to convert value in int. Another detail about incorrect data is between product_count and vendor_count. There is information about products that were not purchased but there is information about sellers who sold them.

Regarding the duplicate values, we calculated the percentage and as the value was very low (0.041%), we decided to remove them.

In order to see the missing values, we first decided to convert in NAN the places that have "" or "-". Then we calculated the percentages of each column that had missing values (last_promo with 52.3%, HR_0 with 3.7%, customer_age with 2.3%, customer_region with 1.4% and first_order with 0.3%). After that we created a dendrogram that would help to understand relationships between columns based on their patterns of missingness, it would help to inform subsequent decisions about imputation or feature engineering 1. The dendrogram reveals one distinct group of features based on their patterns of missing values: the blue part comprises three closely related sub-clusters represented by orange. After that, we created a heatmap that shows correlations between missing values in different columns 2.

The next step was to explore the numerical and categorical features. For a better

understanding of the numerical features, we did some plots for each column 3. The conclusions about plots are: the ages range from 15 to 80, the majority of those who order are aged between 15 and 50 and there are lots of outliers from age 45 onwards (column customer_age); The distribution ranges from 0 to 41, the largest number of single sellers is between 0 and 10, as such, from 10 onwards, you can see quite a few outliers (column vendor_count); They range from 0 to 269, most people order between 2 and 7 products and outliers from 10 onwards are visible, with a large gap after 125 and then a single value at 269 (column product_count); The distribution is between 0 and 83, most of the values are between 1 and 3, from 5 onwards, there are many outliers (column is_chain); The distribution is decreasing, i.e. the number of days from the start of the dataset when the customer first placed an order is greater than afterwards and no outliers visible in this graph (column first_order); The distribution is increasing, i.e. the number of days from the start of the dataset when the customer most recently placed an order is greater and no outliers visible in this graph (column last_order). Also, we create a correlation matrix of numerical features 4. About this graph, we can see: vendor_count and product_count have a correlation of 0.83, suggesting that the number of vendors is strongly related to the number of products, is_chain also has a high correlation (0.83) with product_count, indicating that being part of a network is associated with more products; first_order and vendor_count have a correlation of -0.39, showing that the date of the first order is inversely related to the number of vendors, first_order also has a negative correlation with product_count (-0.36); customer_age has a correlation close to 0 with all the other variables, indicating that customer age is not related to the other characteristics.

For a better understanding of the categorical features, we saw some statistics about 3 variables (Customer_region, last_promo, payment_method) and we conclude that customers are from 8 different regions, most customers are located in region 8670, there are promotions in 3 different categories, most customers use promotions in the delivery category, there are 3 different payment methods used by customers and most customers use Card as their preferred payment method 5.

3 Data Preprocessing

In this part, after exploring our data, we searched for incoherences, handle the missing values that can be imputed with simple techniques that do not tend to leak data, remove outliers and explore the results of our preprocessing.

In order to deal with the data types, we changed the columns 'customer_age', 'first_order', 'HR_0' to int.

The next step was to find inconsistencies. The first one we found was that the average of product_count is 5 and exists one number with value 269. But we decided to treat that value like an outliers. The next was that the column HR_0 only shows values of 0, so we dropped that column. The next incoherence was the column Product_count == 0 and vendor_count >= 1, so we decided to replace product_count = 0 by NaN values. Another strange value that we found was 138 columns where product_count and vendor_count == 0. This is an incoherence because they are not a customer if they have not placed an order yet. We decided to drop because less than 0.5% were affected and these rows have no impact on most features, since there is no product/vendor count and no expenses in cuisines/hours of day.

Dealing with missing values effectively is crucial to ensure our dataset's integrity and the accuracy of our analysis. In case of numerical features, the strategy we used to deal

with missing values is replaced by the median. To treat the missing values in first_order and product_count, we used the technique KNNImputer, which for each point with missing values, finds the K nearest neighbors based on a distance metric and replaces the missing value with the average value of the K nearest neighbors. In case of categorical features, the strategy we used to deal with missing values is replaced by the mode.

Outliers can result from data variability or errors during data collection, entry, or processing. Firstly, to see the outliers in detail, we created graphs 6. To treat the outliers, there are two methods: automatic method and manual method. We conclude that we should see what is common to both methods and remove only that because the automatic method is removing a very high percentage of the data. After doing that, we can see that the date is better distributed 7.

4 Feature Engineering

Creating new features can significantly enhance our analysis by providing additional insights and improving the performance of models. That's why we have created 10 new variables. All the information about them is described in the following table 8.

After that, we processed these new variables, just as we did with the others. We saw the statistics of the numerical and categorical, and concluded that: there are 31098 users, the average time that costumers stay was 34.74 days (column lifetime_days); The average of total amount spend per customer is 35.25 and most users spend around 12.82 to 43.37 (column total_expenses); The average money spend per product is 7.57 (column avg_per_product); The average money spend per order is 10.22 (column avg_per_order); The average number of different cuisines is 0.15 (column culinary_variety); The percentage of order from chain restaurant is 0.63 (column chain_preference); The average of loyalty to vendors is 0.84 (column loyalty_to_vendors); Consumers prefer to order on Saturday (column preferred_order_days); The most commonly hours to order is "12h-18h".

Also, we checked for missing values, but we did not find any and about the outliers we used the same method that we do in the other features. To visualize all the new features created, we did some plots 9, 10 and 11.

5 Data Transformation

After creating new features, we need to do feature encoding, transform some features and scaling them.

Feature encoding is used to transform categorical variables into numerical representations, since most machine learning algorithms cannot work directly with categories. So, first we used frequency encoding to the columns 'last_promo' and 'payment_method'. To the other categorical variables, we used One-hot encoding. The encoded features will not be used for clustering and scaling.

Feature transformation is the process of modifying or transforming variables in a dataset to improve the performance and facilitate analysis. In order to do that, we decided to use a Logarithmic Transformation to all the numerical variables.

After that, we standardized and normalized the numerical variables to improve the performance. First, we used Min-Max Scaling in order to rescale features to [0, 1]. After that, we tried using Standard Scaling, which sets the average to 0 and the standard derivation to 1. However, we noticed that, possibly because there are different scales, the value

of some columns ('first_order', 'last_order', 'is_chain', 'vendor_count', 'product_count') is not correct, so the solution we found was to scale those features together.

6 Clustering

Feature selection is a crucial step to prepare data for clustering analysis, as it allows us to focus on the most relevant variables and exclude those that could skew the results, like nominal variables and one-hot encoded features. For that reason, even though 'customer_region' and one-hot encoded features are numerical features, they will not be considered during clustering analysis.

To start building the clusters, we looked at the ideal number and, as we can see from the graph 12, the best is 4 or 5 or 6 clusters and the best linkage = ward.

After that, we checked the performance with 6, 5 and 4 clusters. We conclude that the separation of the regular customers who avoid chains into an own cluster can be useful and makes sense, especially considering customer-specific advertisements, in which in this case these users could receive different, non-chain-related, advertisements. That is why we prefer the clustering with 6 clusters, seen in 13. For these 6 clusters, we could label the customers in each cluster with the following characteristics. First of all we see that cluster 3 has the biggest amount of data, followed by 1 and 0, whilst the clusters 2, 4 and 5 are the smallest, again seen in 13. Following the average values for each feature per cluster in 13, we decided that customers from cluster 1 are our best spending and most valuable customers. Customers from cluster 2 are the counterpart of these, the customers who spend the least. Customers from clusters 3 and 4 are the regular spending customers without outstanding characteristics, but customers from cluster 4 tend to spend less money whilst customers from cluster 4 are the more frequently spending regular customers. Customers in cluster 5 are a smaller group from group 3, but those who avoid using chains. Customers from cluster 0 do not make a lot of orders, but if they do they order a lot.

After that, we have to choose the k clusters to apply the K-Means algorithm. This algorithm minimizes the sum of squared distances between data points and their corresponding cluster centres. After an analysis as seen in 14 and 15, we conclude that the best performance is with 4 clusters, but we decided to stay with the results from the Hierarchical Clustering since the higher number of clusters leads to a better and more profound distinction between the customers, while the K-means Clustering with only 4 groups of customers joined some customers into groups who can be distinguished reasonably into different groups.

High spenders from hierarchical clustering group 1 mostly moved into k-means group 1. Conversely, the lowest spenders from Hierarchical Clustering group 2, along with those who make infrequent but large orders transitioned into k-means group 3, indicating that this group consists of low spenders and occasional high value customers. Regular customers from hierarchical clustering group 3 and 5, who frequently place orders and particularly those from group 5 who prefer to avoid chain restaurants, largely moved into K-Means Group 0. This group also includes some of the highest spenders from HC Group 1, making it representative of the most frequently "regular" ordering customers. Finally, K-means Group 2 mainly comprises customers from the former hierarchical clustering group 4, characterized by regular but low spending and it also includes a considerable number of customers from HC groups 2 and 3, reflecting a mix of regular ordering patterns and lower expenditures. All this can also be seen in 16 and 17.

Now we divide our features in various perspectives and create clusters for each perspective. We decide for three perspectives, preference based, purchase behaviour based and age/time based 18. We use Hierarchical clustering for this approach since it has proven before that it fits best for our data and specific preferences. After analysing elbow-plots and dendrograms 19, we decide for the best number of clusters being 5. For one perspective, 4 would have been a reasonable number as well, but for the final clustering in the end all perspectives should have the same number of clusters.

To join the different perspective clusterings into one final clustering, we decided to calculate a distance matrix between the three clustering labels for each perspective for each row, and then conduct a final hierarchical clustering based on this distance matrix. We also added different weights to the different perspectives, in order to decrease the impact of the age/time based perspective and increase the impact of the other two perspectives since these seemed more important to us. The silhouette value plot 20 delivered 4 or 6 as best number of clusters, the dendrogram 21 delivered 5 or 6 as best number of clusters, so we decided for 6 clusters. To analyse and label these final clusters, we again looked at the sizes and average values of each cluster, as seen in 22 and 23. We can see that clusters 1 and 3 have the most customers, whilst 4 has the least. Clusters 0, 2 and 5 have similar sizes. With the average values, we can label customers from cluster 4 as the best spending and most valuable customers, whilst customers from cluster 5 are the least spending customers. Customers from cluster 1 are the most regular customers with no specific characteristics. Customers from cluster 0 are a smaller group from group 1, who tend to spend more frequently. Customers from cluster 3 make a lot of orders, but rather small ones. Finally, customers from cluster 2 are a smaller group of cluster 3 who prefer more different places and don't always order from the same places.

Cluster 0: regular but customers who spend more : These customers spend more than average regular users and have higher total expenses. We believe a focus on encouraging continued engagement would be useful in order to help them becoming closer to the best spending customers. A main idea is the offering of premium loyalty rewards, such as free delivery for reaching specific spending thresholds. Also encouraging higher basket sizes by promoting exclusive combos or upsized orders would be a good approach.

Cluster 1: largest group of regular customers: Regular customers represent the core user base with average spending habits and strong loyalty to specific vendors. We recommend to focus on reinforcing their attachment by introducing loyalty programs tied to their favourite vendors, offering exclusive discounts or benefits, since they have reasonably high loyalty to certain vendors. Also encouraging repeat orders by recommending familiar items or combinations they are likely to prefer might help to make them more valuable.

Cluster 2: frequent orders but low loyalty: We recommend to focus on their variety-seeking behaviour by highlighting new or diverse dining options. Personalized dynamic offers can also incentivize them to return to a specific vendor, gradually building loyalty.

Cluster 3: High frequency, small orders: We recommend to encourage larger orders by introducing value bundles or promotional discounts for multi-item purchases. Gamified loyalty programs, such as rewards for hitting a specific number of orders in a month, can also keep them engaged while promoting more consistent activity.

Cluster 4: best spending customers: These customers have high spending, loyalty, and long-term engagement. It is important to keep these customers satisfied. Premium experiences like VIP memberships with perks such as free delivery, priority service, or early access to exclusive menus could reward them for they high spending habits. Also strengthening their loyalty with unique rewards or experiences, such as exclusive restau-

rant partnerships or invitations to food-related events could make them feel valued.

Cluster 5: least spending customers: Simple instruments like discounts and coupons could encourage them to increase their spending habits. After regaining these customers, their behaviour can be observed and they can be put in one of above groups in order to focus on their satisfaction and habits.

7 Interface

We built a user-friendly interactive interface that allows the user to get deeper visual insights into the clusters. When running the interface, the user has four different options to choose 24. He can look into insights in one cluster, he can compare multiple clusters with each other, he can get insights into all clusters or add a new entry and see to which cluster this entry will be added. Additionally he can also select for which clustering approach he wants to get the insights, either the final clustering or one of the perspectives. When choosing Insights in one cluster, the user will see this window 25. Here the user can either view a boxplot, a heatmap or the cluster cohesion for one cluster (which he can choose from the combo box on top). He again can select one clustering approach. In this example image, the user has chosen the boxplot for cluster 1 26. On the bottom left corner the user also has the labels for each group of customers per cluster.

If the user selects "compare clusters" from the main page, he gets to this window 27, where he can select the clusters to be compared and can choose between a radar chart, feature difference bar chart, distance plot and distribution overlap plot. For the feature difference bar chart, the user has to select exactly two clusters to compare, the other plots work with any number of clusters higher than 0. For the distribution overlap plot the user has to choose a feature from the combo box below.

If the user selects "Insights into all clusters" from the main page, he can choose between eight different plots, to get visual insights into the full clustering 28. If the 3D plot was selected, the user can interact with this plot by zooming in and out and moving in all three dimensions around this plot.

Finally, if the user selects "Connect new entry to cluster", he gets to this window 29. Here he is asked to add the necessary values for a new entry of a customer. He can again select between the clustering approach with the combo box on top, only the necessary features will be entered. All entry fields have to be confirmed with the confirm button below. If an entry is wrong there will be an error message, if an entry is has an unusual values (values that don't appear in the original data like this) the confirmation will still be done but the user will receive a warning message since the results now might be unexpected 30. After entering all values, the user can either click on Quick prediction or calculate cluster. The quick prediction calculates the ward distance from this new point to all cluster centroids and adds the point to the nearest cluster, this might not be very accurate 31. If the user chooses calculate cluster, the whole hierarchical clustering will be recalculated with the new point as part of the data, this delivers very precise results but might take some time. If the user selects one of the perspective approaches from the combo box on top this should only take a short amount of time, but if he chooses the final clustering approach, all perspective clusters will be calculated and then the final clustering so this might take a big amount of time and might not work on every machine 32.

8 Conclusion

In this project for ABCDEats Inc., we conducted a comprehensive analysis to enhance decision-making processes. The dataset provided contained 52 numerical and 4 categorical features and it required significant cleaning and preprocessing before we could proceed to clustering analysis.

Initially, we delved into data exploration and preprocessing to ensure data quality and reliability. During this initial phase we identified and corrected data types, addressed missing values and outliers. Additionally, we resolved data inconsistencies by removing irrelevant features and adopted different strategies to correct illogical data points. To manage missing values, we employed three methods: for numerical features, we used median values and the KNN-imputer, while for categorical variables, we opted for mode imputation. We strategically handled outliers using a combination of automatic and manual methods which significantly improved the distribution and integrity of our database while preserving an appropriate amount of data.

We created 10 new features that significantly improved our analysis and model performance. These variables provided new insights about customer behaviour and spending patterns, Afterwards, we conducted an exploratory analysis for the new features, checked the absence of missing values and handled outliers using the same previous approach to ensure data quality and consistency.

To further optimize the dataset for clustering analysis, we transformed categorical variables into numerical formats using frequency and one-hot encoding, and applying logarithmic transformations to numerical variables to normalize data distributions. We further enhanced data quality with Min-Max Scaling to normalize features to a [0, 1] range and adjusted Standard Scaling due to inconsistencies from different scales by scaling features together.

To prepare data for clustering analysis, we selected the most relevant features by excluding those that could skew the results. Specifically, 'customer region' and one-hot encoded features. This step ensured that our clustering was based on the most meaningful data attributes, enhancing the accuracy and reliability of our findings.

For our clustering analysis we tested 4 different approaches to best segment our customer base. We began with Hierarchical clustering, exploring the optimal number of clusters, ultimately selecting 6 clusters for their ability to meaningful differentiate customer behaviours. Following this, we applied K-Means clustering algorithm, which suggested the best performance with four clusters. However, to maintain a more nuanced differentiation among customer groups, we adhered to the six-cluster model from the Hierarchical clustering, which provided a cleared segmentation of customers. We also introduced Perspective Based Clustering, dividing features into categories like preference, purchase behaviour and age/time, and applied Hierarchical Clustering within these frameworks. This method confirmed that a combined approach across different perspectives would deliver the best results, leading us to finalize our clustering with 6 distinct groups. In our final clustering, we calculated a distance matrix across the perspectives to blend the clusters into a cohesive model, weighting the perspectives differently to emphasize most impactful ones. Our final customer segments were the following:

- Group 0 - Regular but high spending customers
- Group 1 - Largest group of regular customers
- Group 2 - Frequent orders but low loyalty

- Group 3 - High frequency, small orders
- Group 4 -Best spending customers
- Group 5 - Least spending customers

This clustering technique revealed to be the most effective to identify diverse customer profiles and provide actionable insights to segment customers and tailor ABCDEats.Inc marketing strategy.

Finally, we developed a user-friendly interactive interface for ABCDEats Inc. which significantly enhances the accessibility and understanding of the customer clustering results. This tool allows for intuitive exploration and analysis of the data, enabling stakeholders to make informed decisions based on clear visual insights into customer behaviours and patterns.

A Appendix

Figure 1: Dendrogram of Missing Values

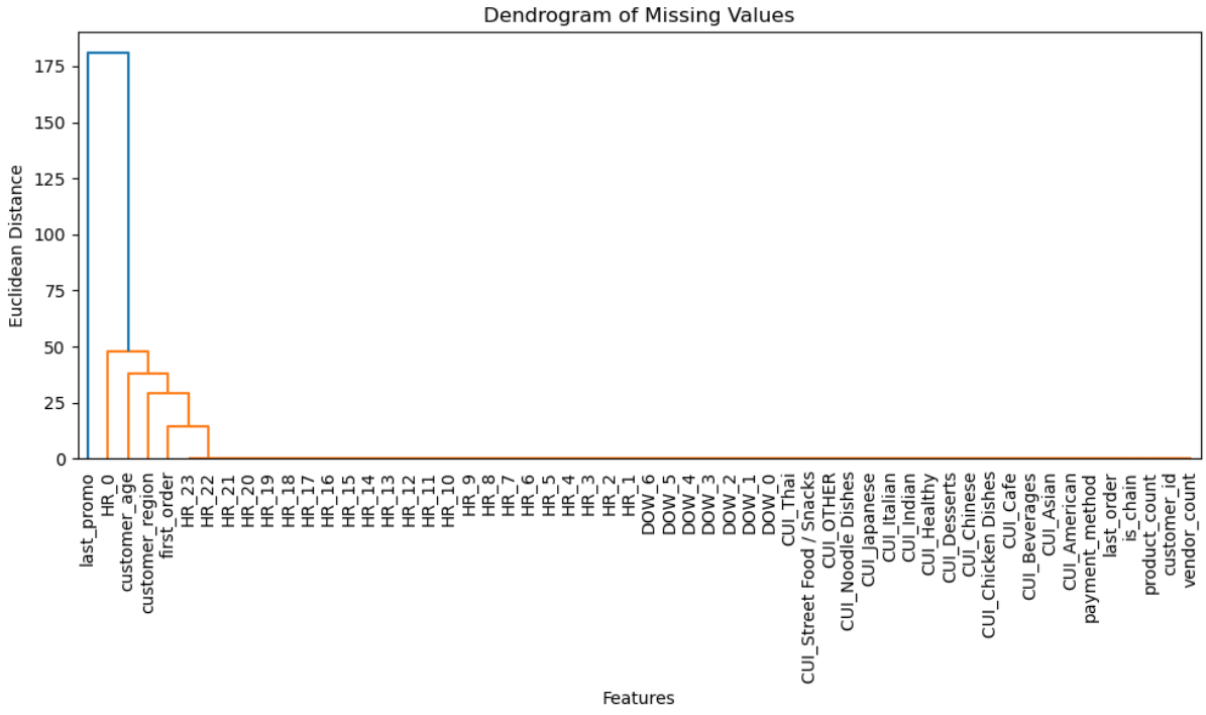

Figure 2: Heatmap of Missing Values

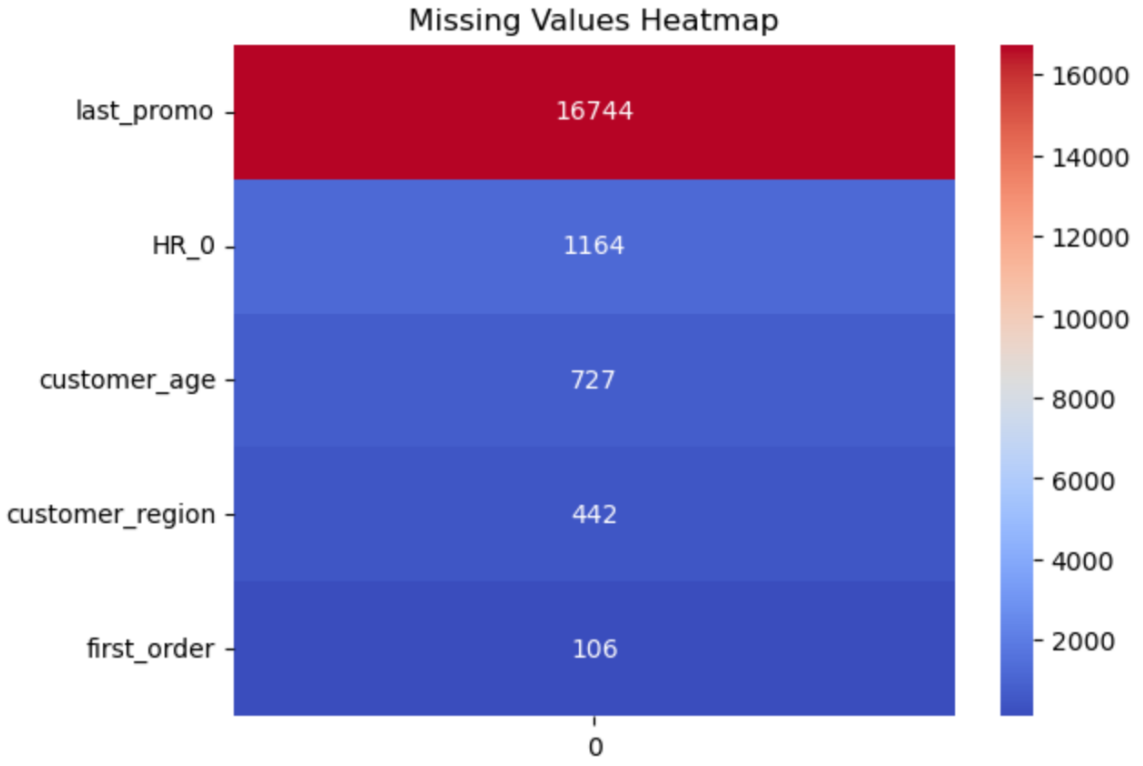

Figure 3: Distributions of numerical features

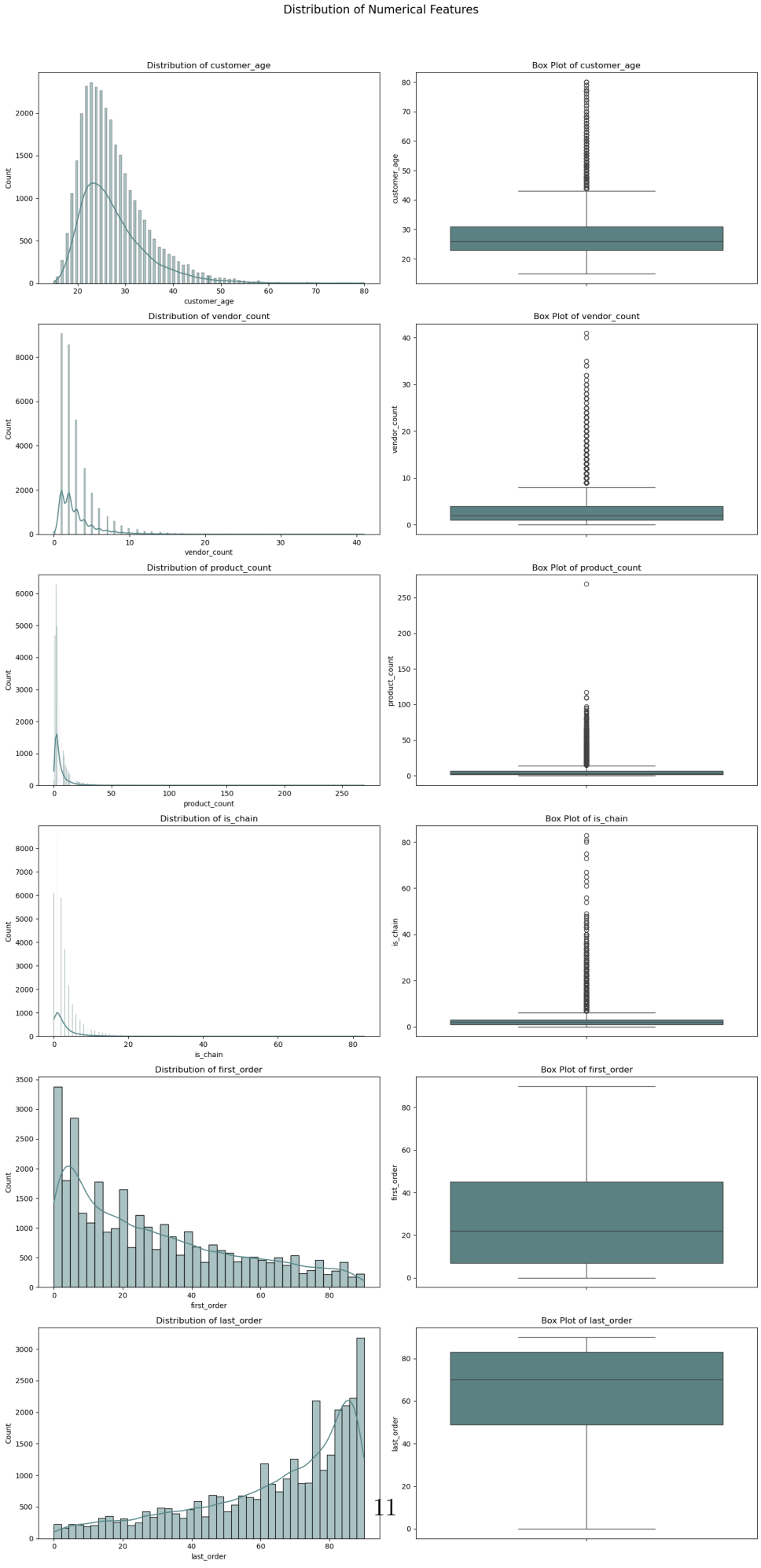

Figure 4: Correlation Heatmap of numerical features

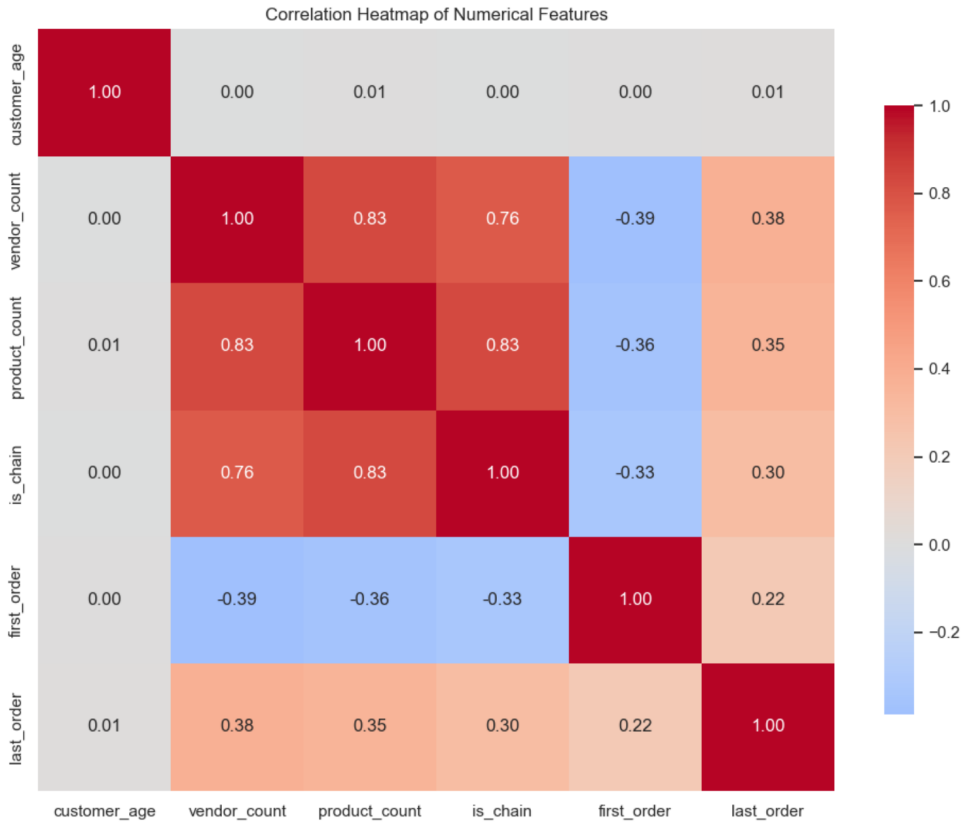

Figure 5: Distribution of categorical features

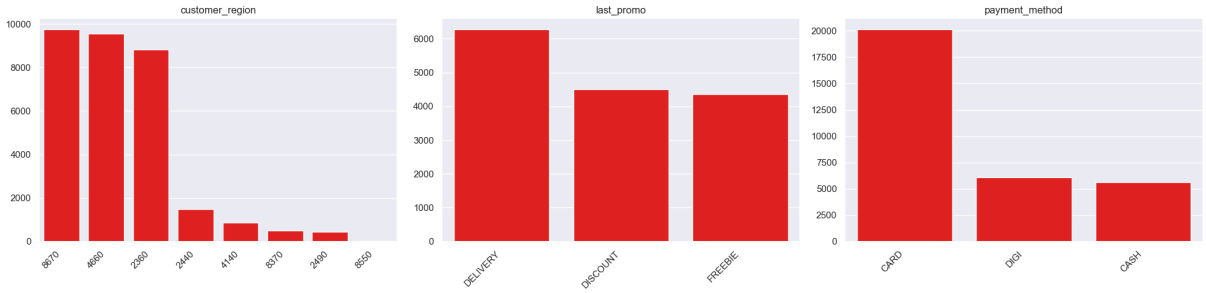

Figure 6: Outliers in dataset

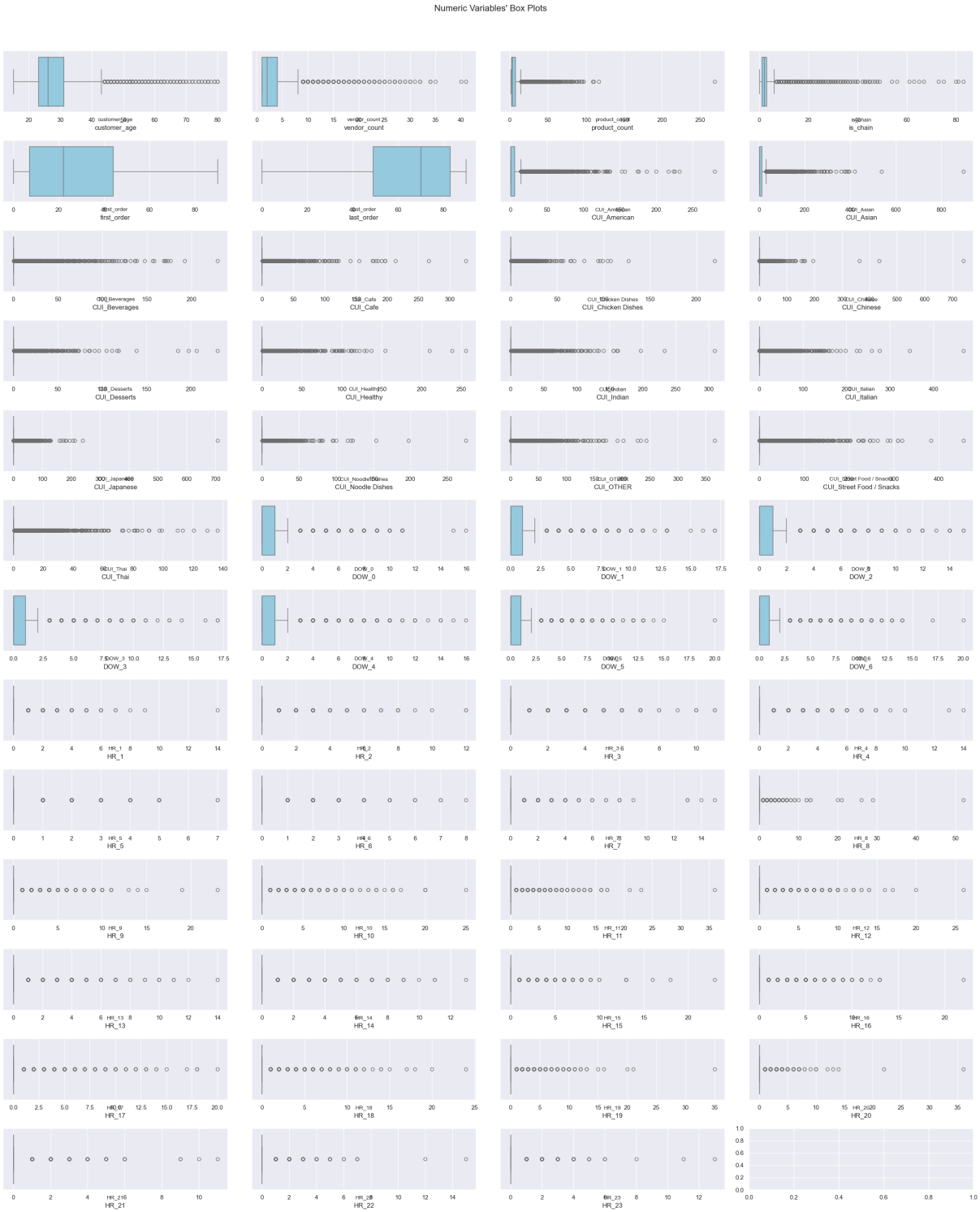

Figure 7: After treat the outliers in dataset

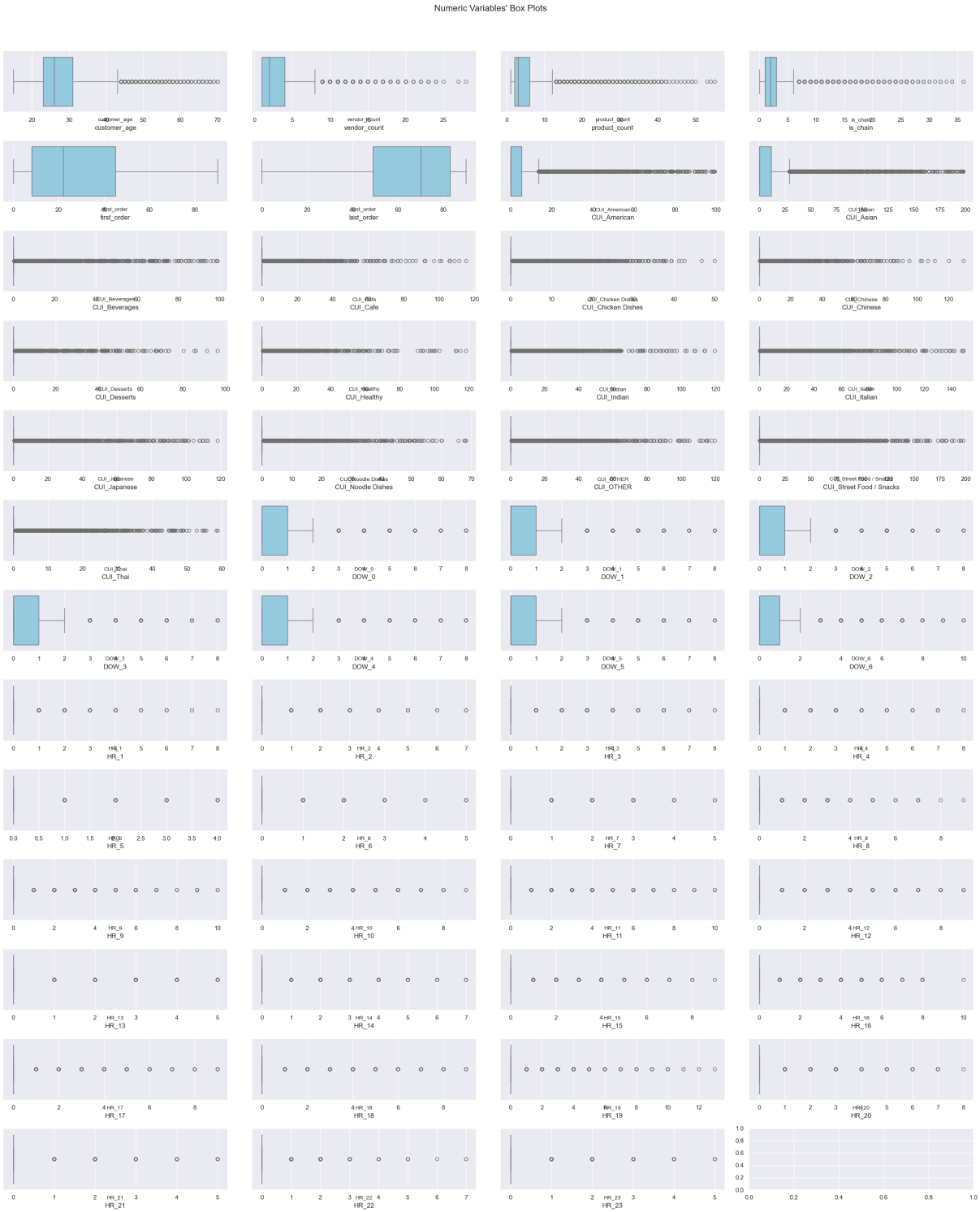

Figure 8: New features

New Feature	how were created	description	why they were created
Customer lifetime	last_order - first_order	time customers stayed with us	possible replacement for first- and last order
most frequent order days	day/days with most orders	shows prefered days for customer	possible replacement for all 7 DOW_ columns
most frequent part of day	divide hours in groups, count orders per group	part of days where most orders where made	possible replacement for all HR_ columns
total expenses	adding all values from CUI_columns	total amount of money spent per customer	important for evaluating customer impact
average per product	total expenses / product count	average money spent per product	shows what kind of products are ordere, more expensive or cheap ones
average per order	total expenses / total number of orders(sum of all DOW_ values)	average money spent per order	shows if customer buys more big orders or only small orders
culinary variety	number of different cuisines ordered from / number of all cuisines	value indicating amount of different cuisines ordered from	useful to differentiate customers in more "open minded" and more conversative
chain preference	is_chain / total orders	percentage of orders made at chain restaurants	useful to differentiate customers in thoose prefering chains or normal restaurants
loyality to vendors	vendor_count / total orders	value indicating loyality of customers in general to restaurants	distinguishes between customers who prefer order from restaurants they know and like and those who order from a lot of restaurants

Figure 9: New numerical features

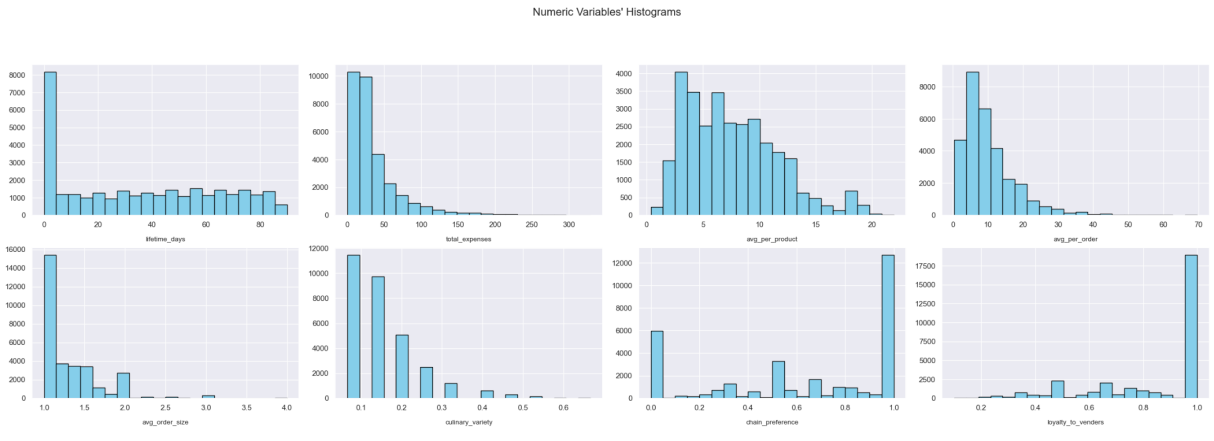

Figure 10: Preference orders day

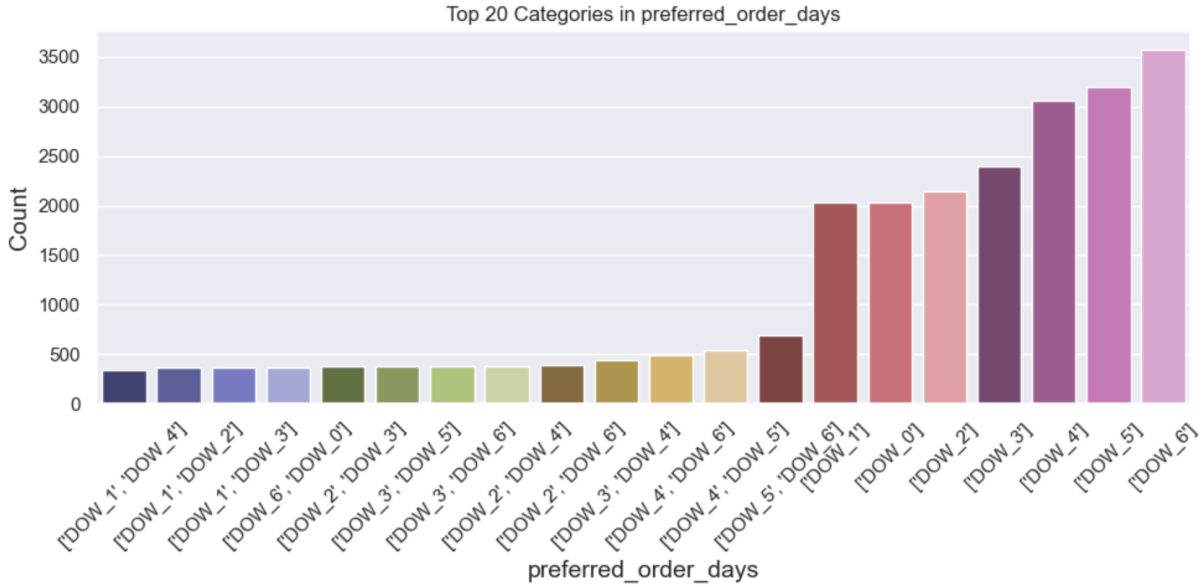

Figure 11: Preference part of day

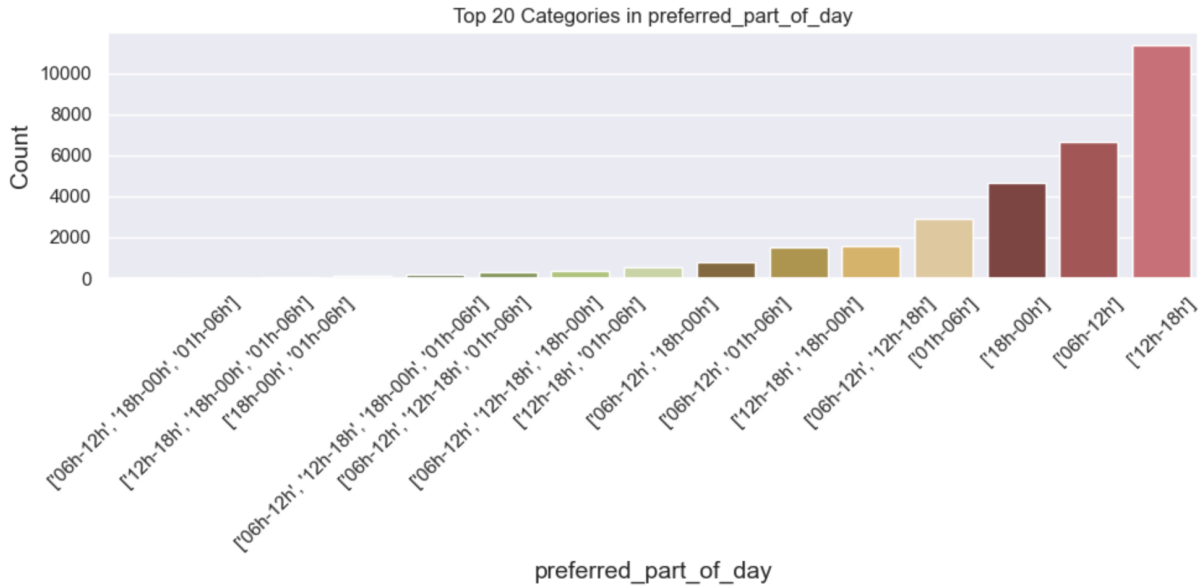

Figure 12: Find the best number of clusters

Figure 13: Analyse the clusters of clustering with 6 clusters

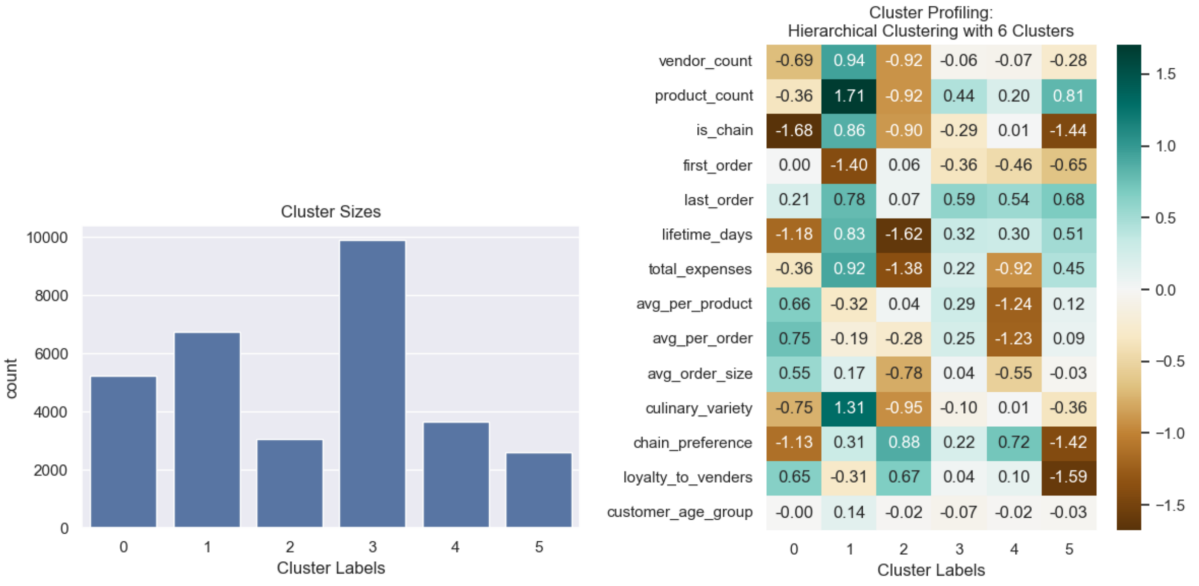

Figure 14: Silhouette for clusters

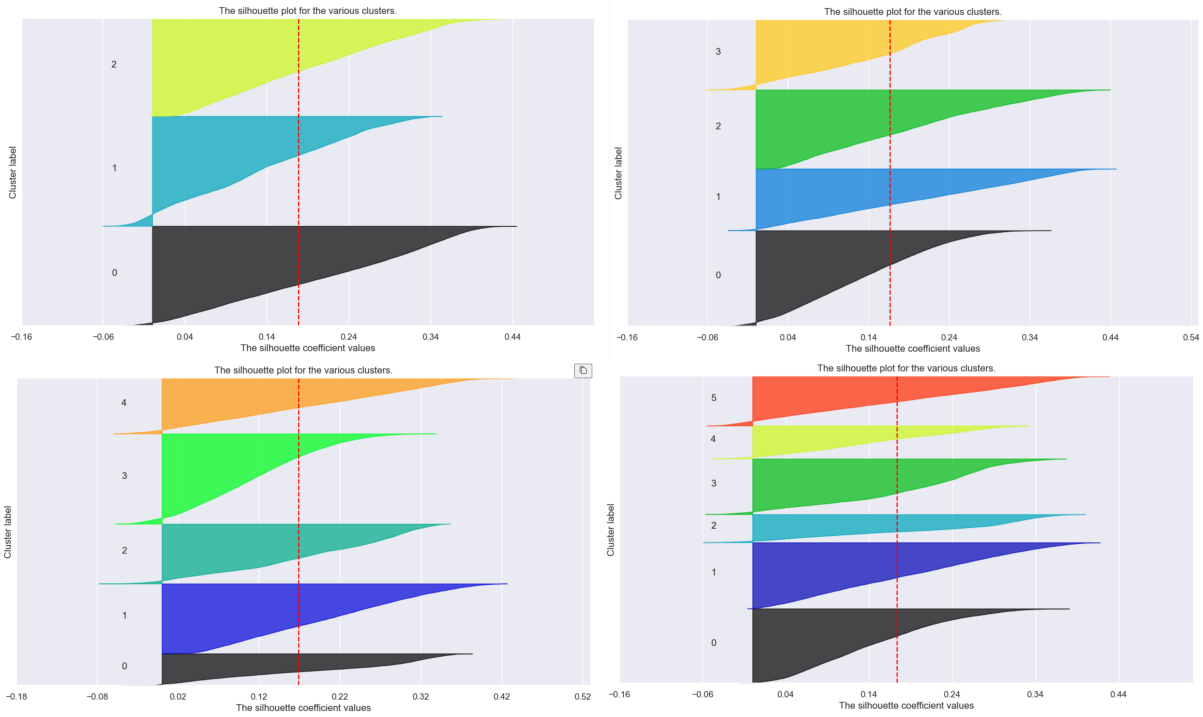

Figure 15: Inertia plot over clusters

Figure 16: Compare Hierarchical clustering and K-means clustering

KM4	0	1	2	3
HC6				
0	870	0	162	4215
1	1188	5170	391	0
2	0	0	976	2103
3	5635	766	2804	686
4	36	138	3488	11
5	2038	222	228	109

Figure 17: Cluster profile

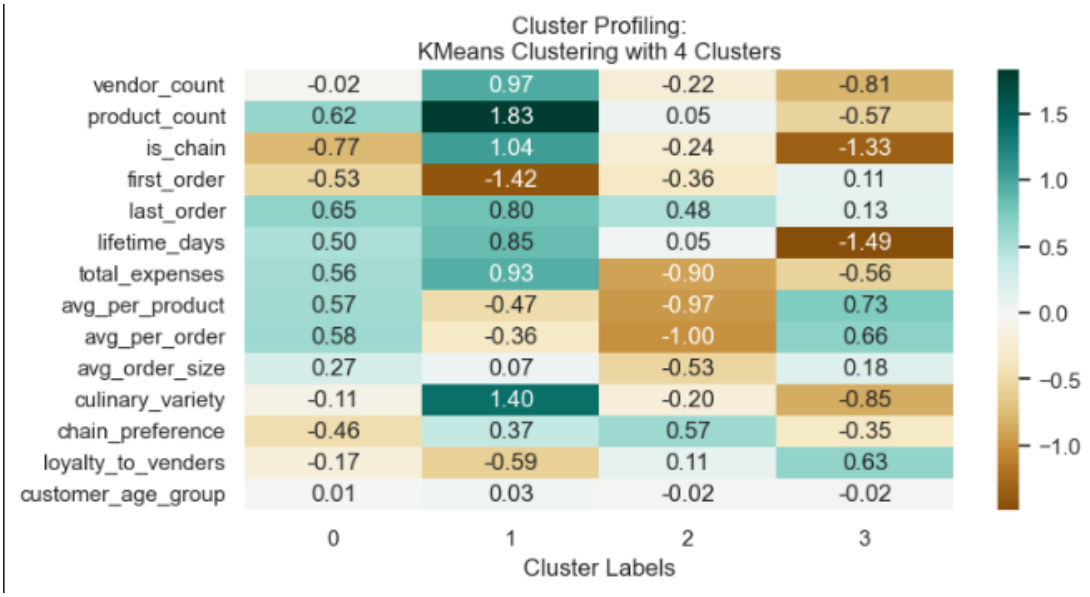

Figure 18: Perspectives Features

```
prefBasedFeatures = ['culinary_variety', 'loyalty_to_venders', 'chain_preference', 'is_chain']
purchaseBasedFeatures = ['vendor_count', 'product_count', 'total_expenses', 'avg_per_product', 'avg_per_order', 'avg_order_size']
ageTimeBasedFeatures = ['first_order', 'last_order', 'lifetime_days', 'customer_age_group']
```

Figure 19: Hierarchicals Plots

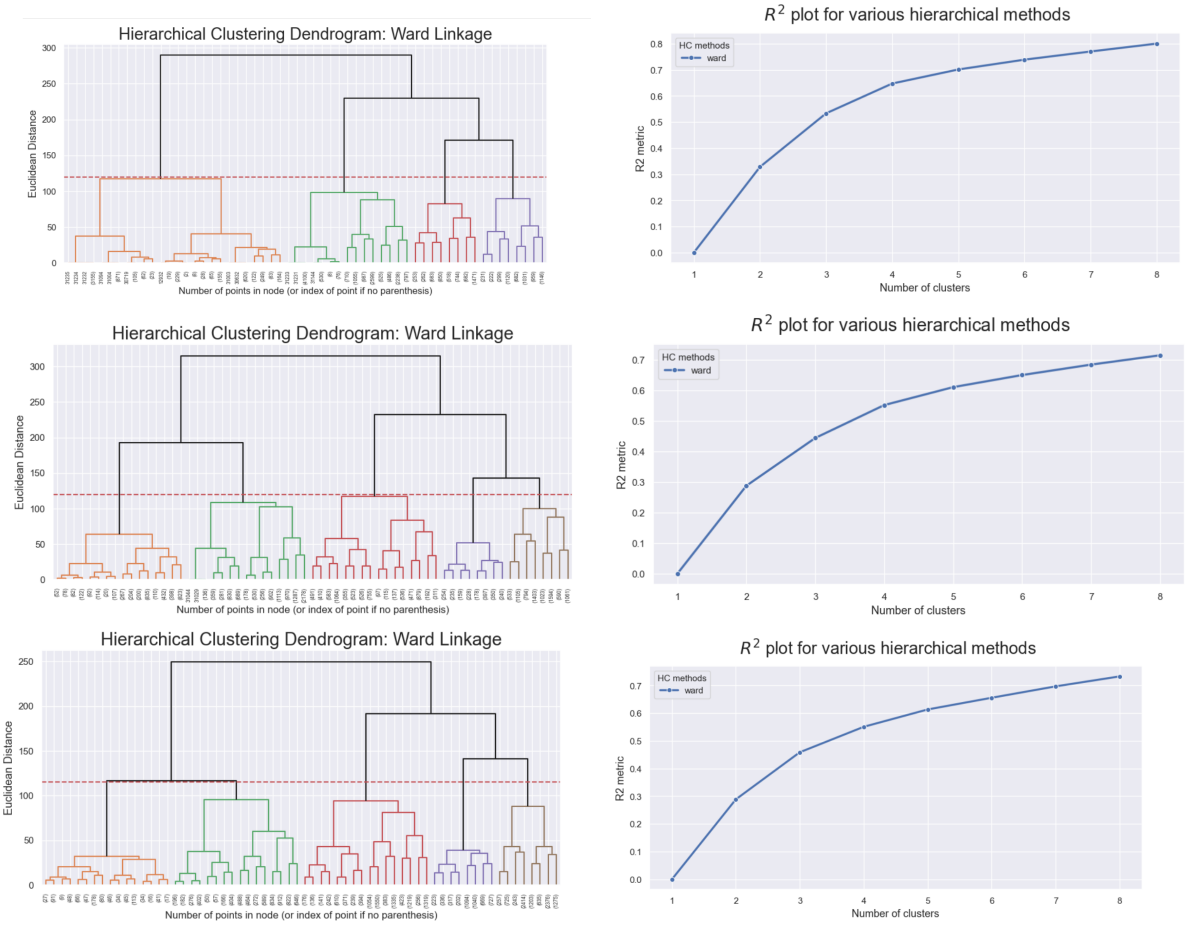

Figure 20: Silhouette of final clustering

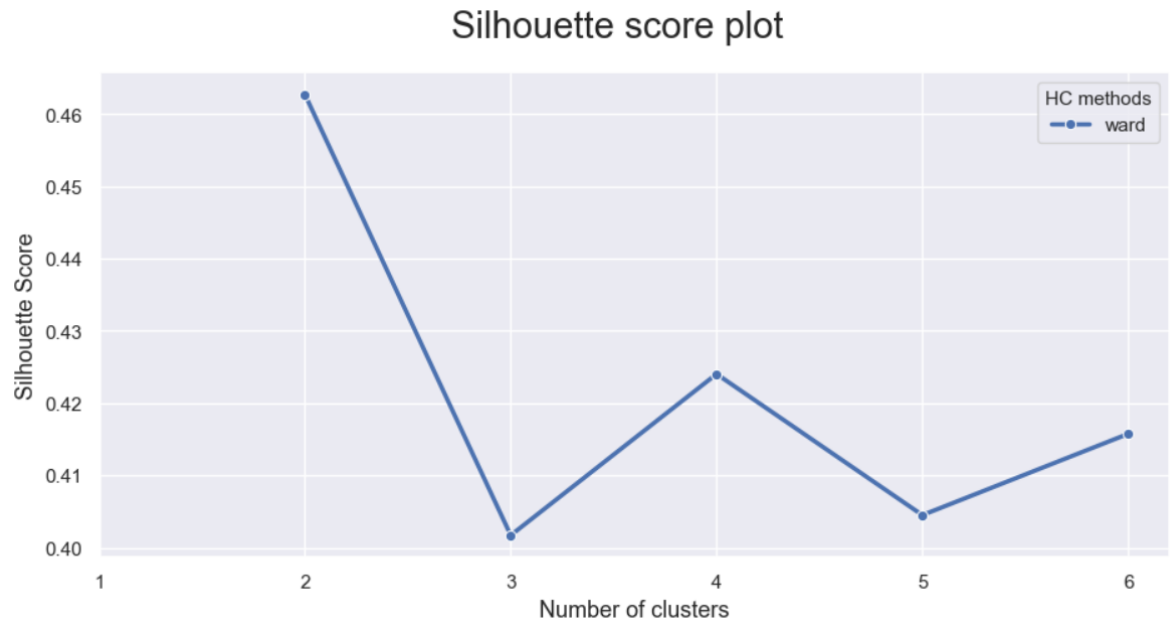

Figure 21: Dendogram of final clustering

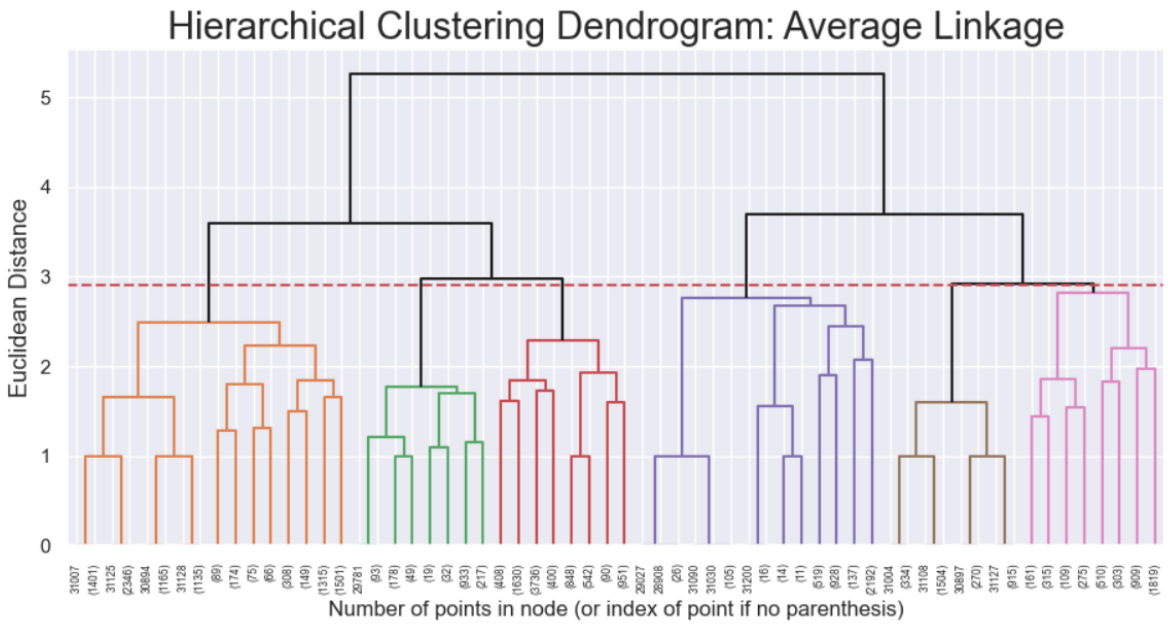

Figure 22: Average of final clustering

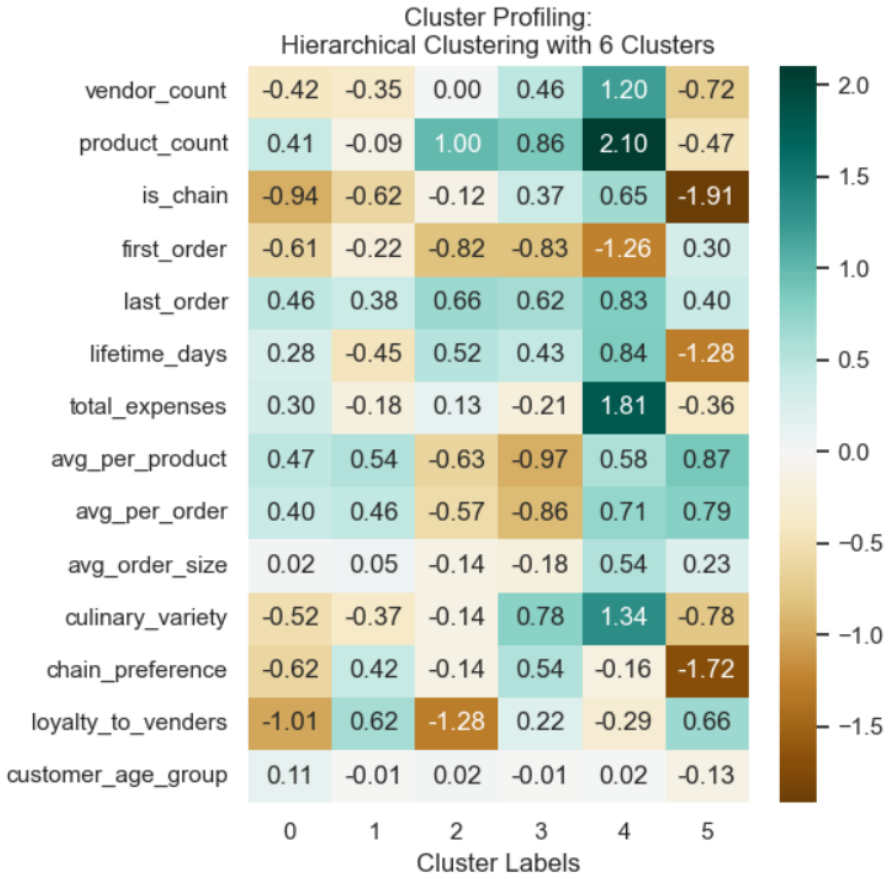

Figure 23: Sizes of final clustering

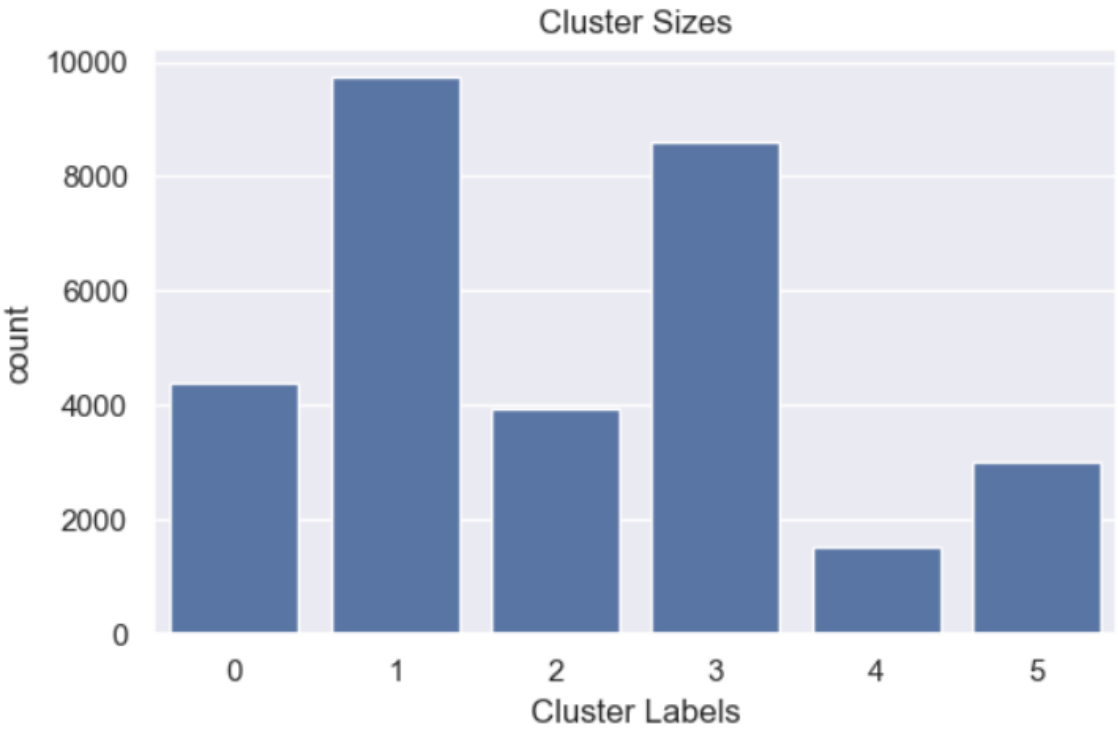

Figure 24: Main page of interface

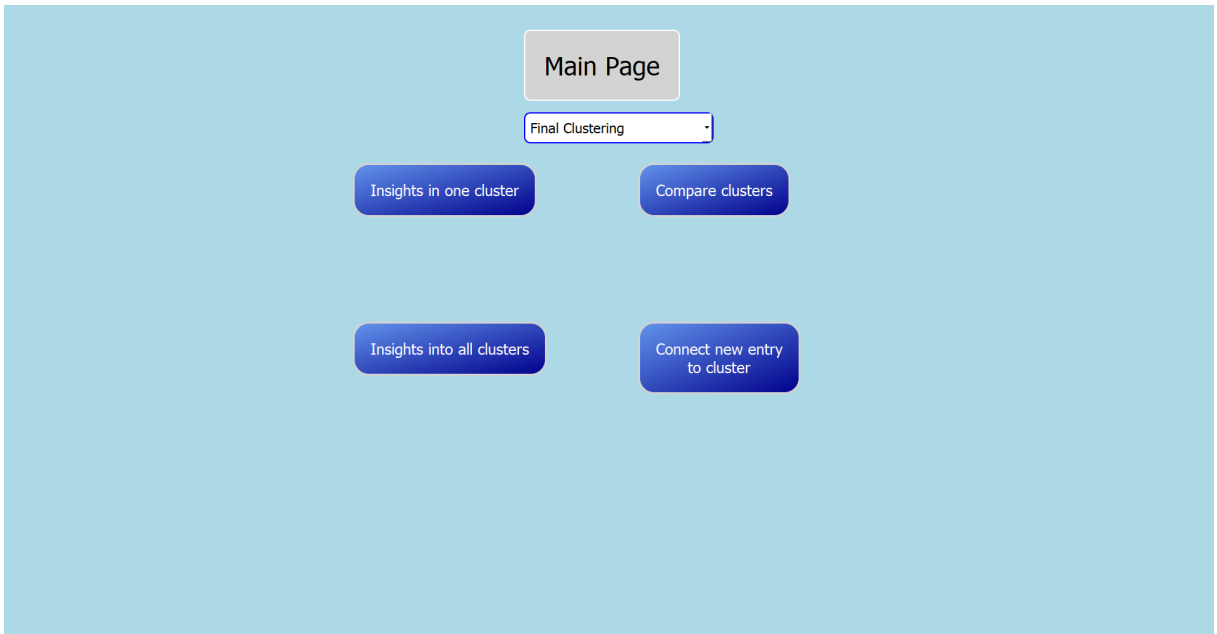

Figure 25: Page of interface 1

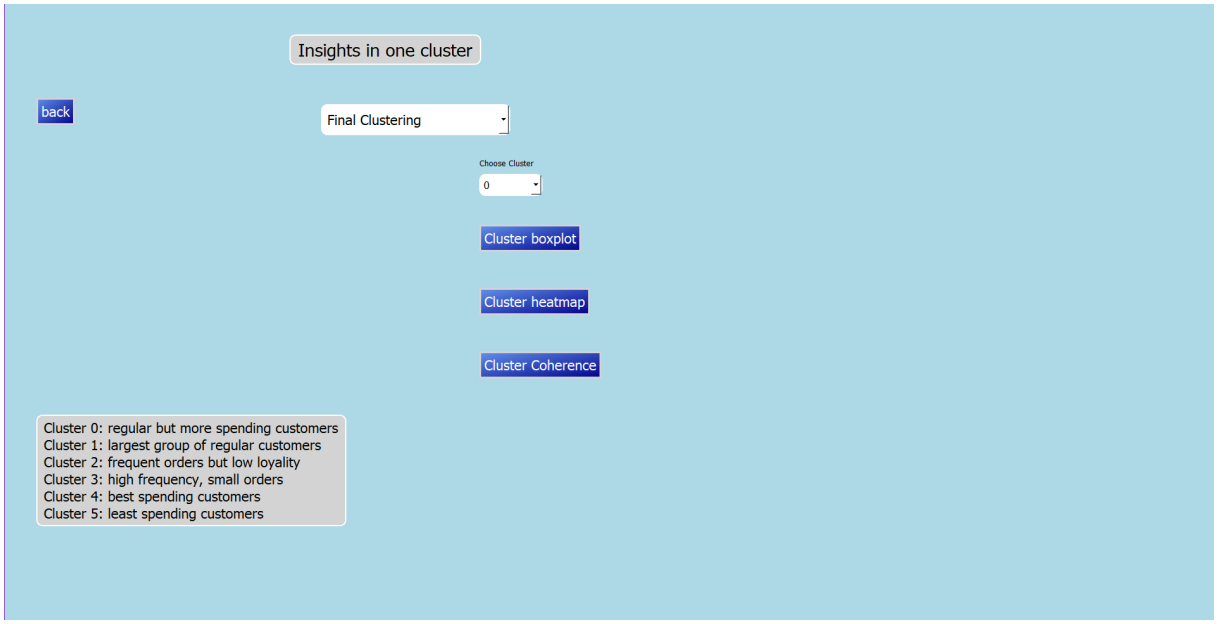

Figure 26: Page of interface level 1

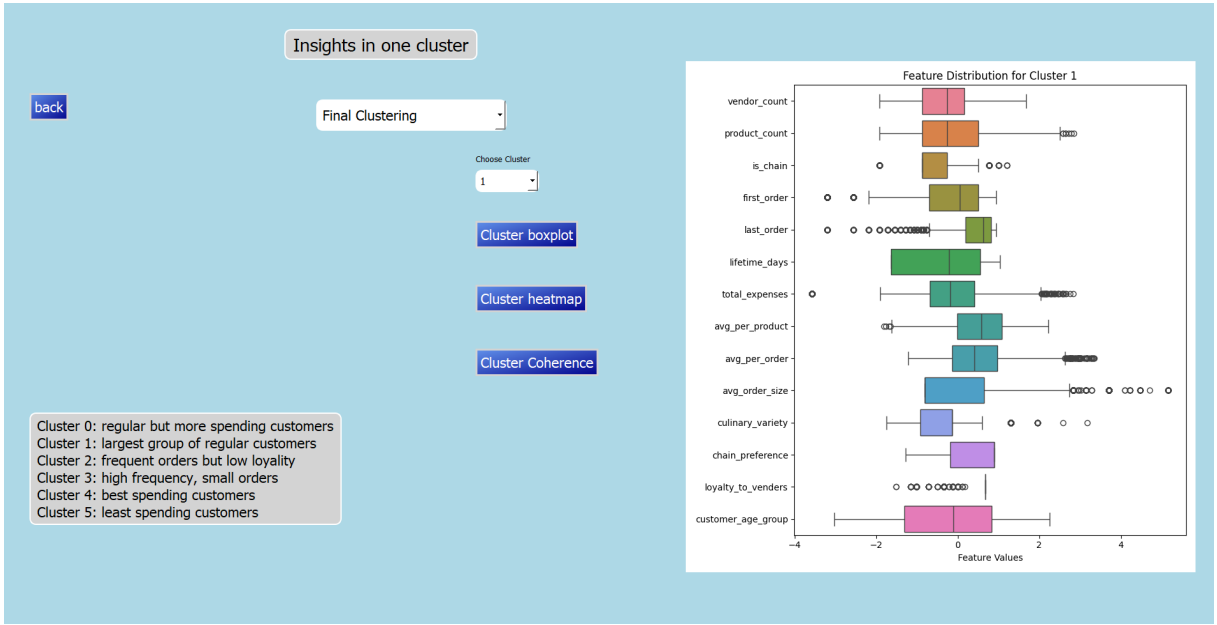

Figure 27: Page of interface 2

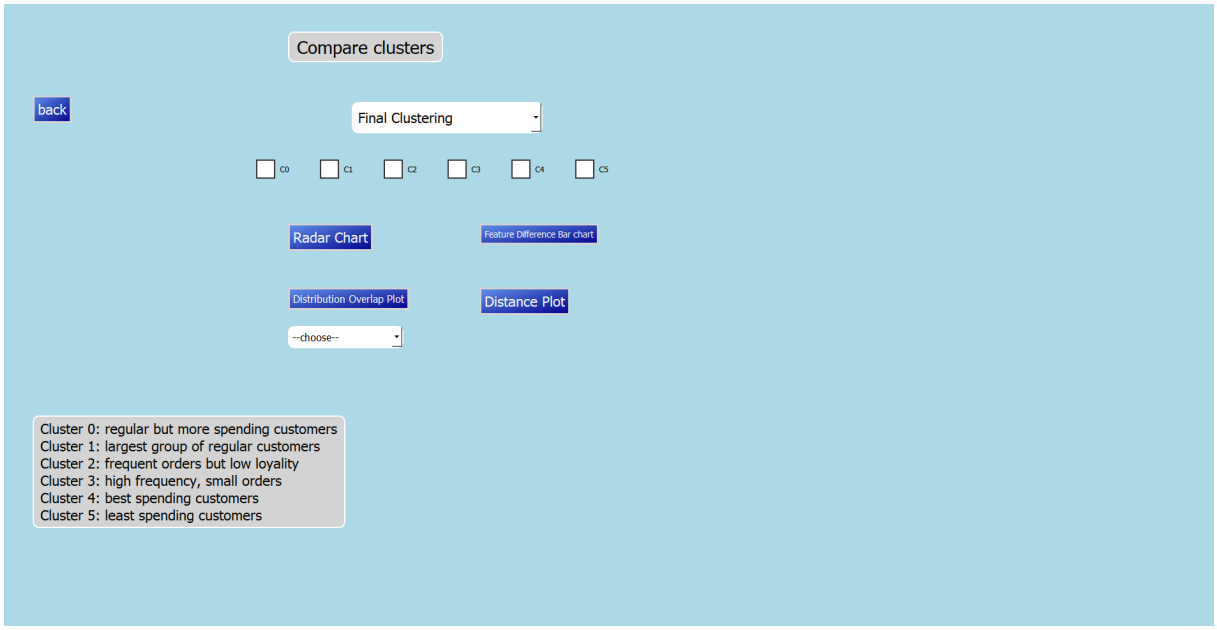

Figure 28: Page of interface 3

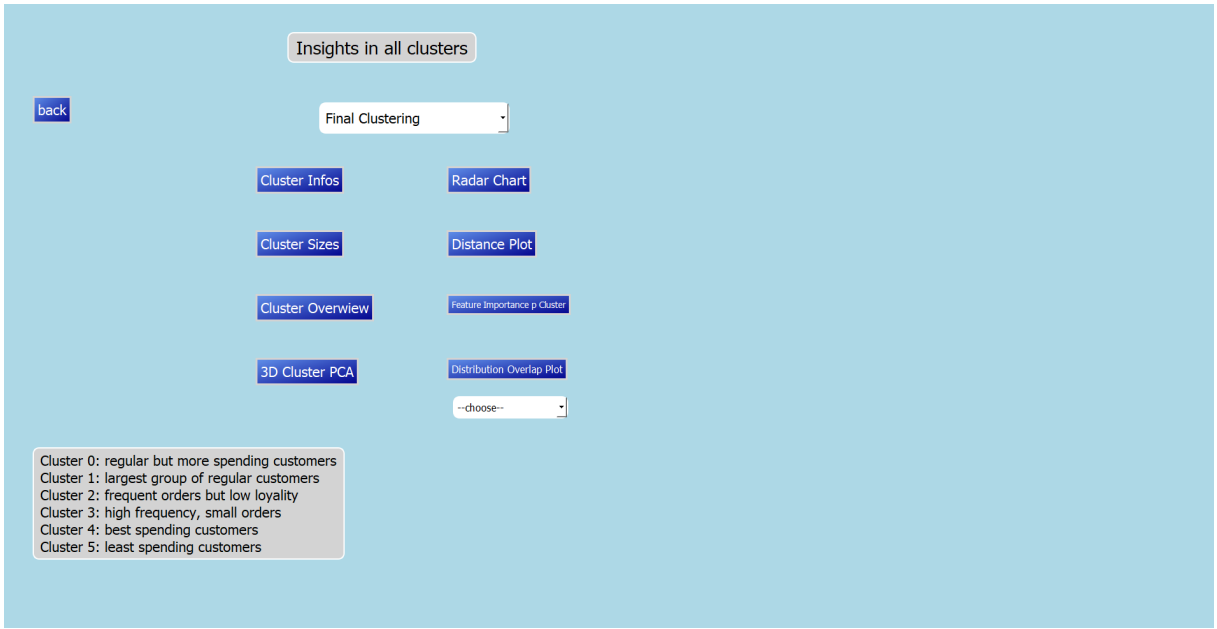

Figure 29: Page of interface 4

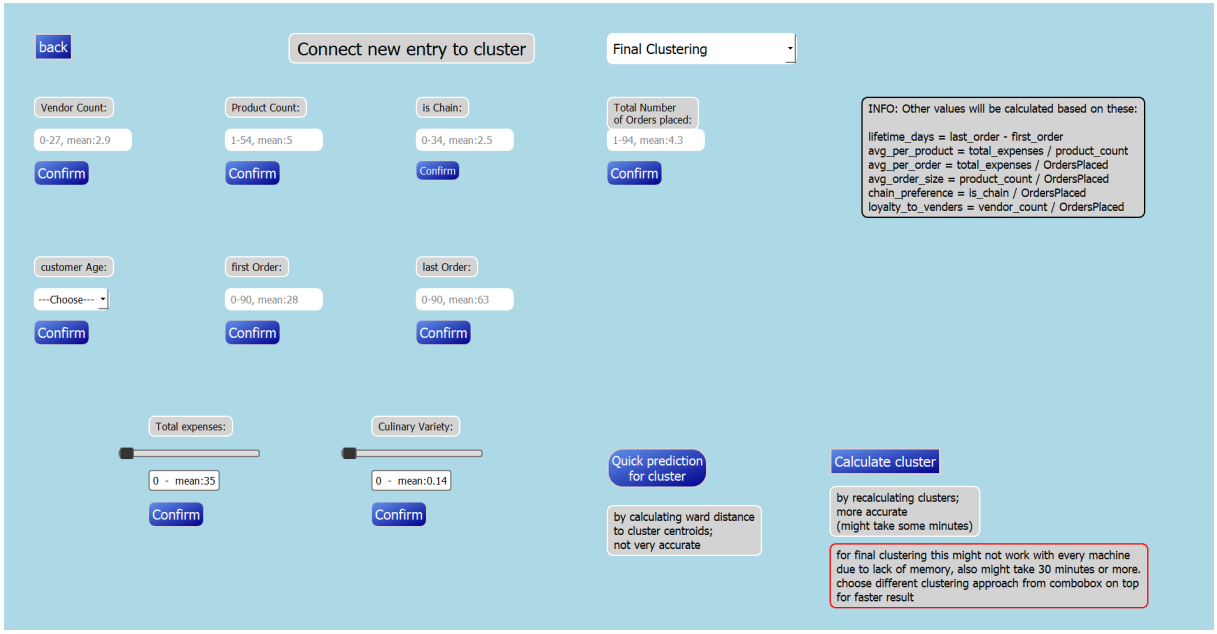

Figure 30: Page of interface error

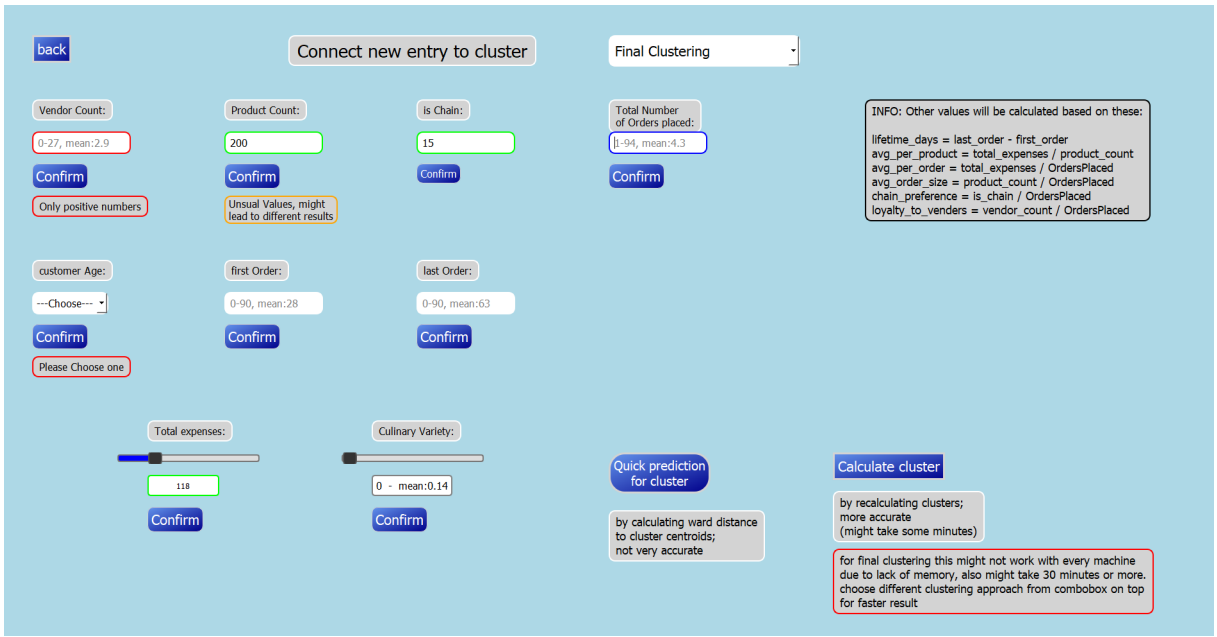

Figure 31: Page of interface output

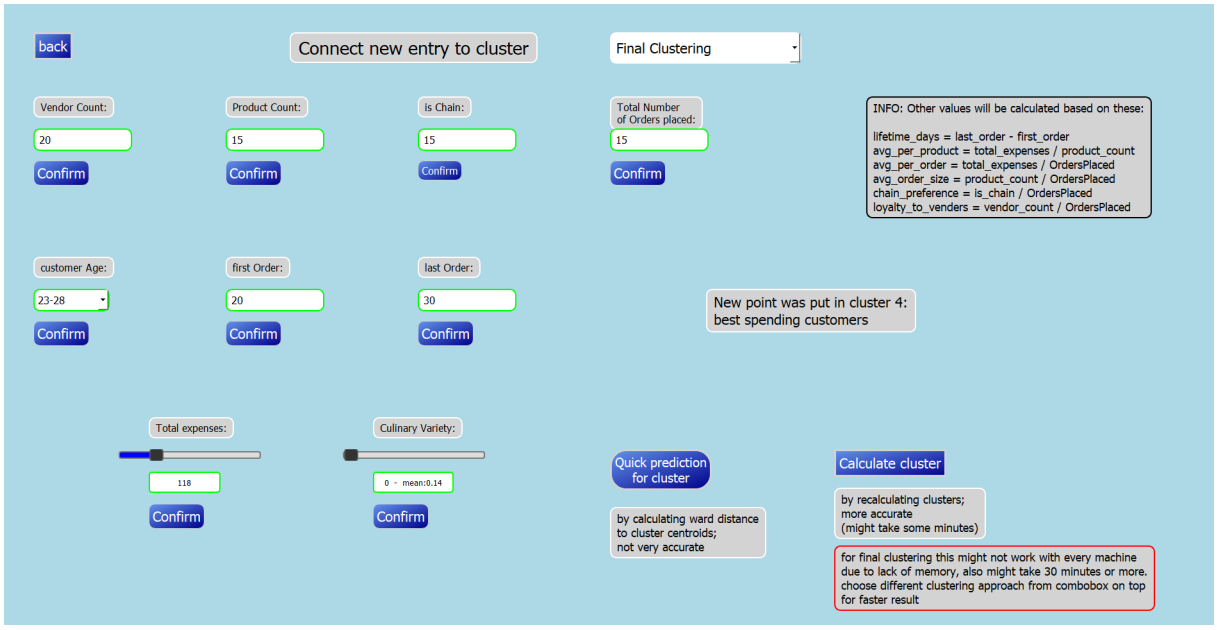

Figure 32: Page of interface output 2