NOVA IMS Information Management School

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

ABCDEats Inc.

Project of Data Mining

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 report we will focus on the user-friendly interactive interface we created. The functionalities and its results will be shown and briefly discussed.

The code for this interface is also availabe at the following link https://github.com/Gomsofi06/Data_Mining_NOVAIMS

2 Main Features

In this part the main features of the interface will be presented briefly. Before running the interface, the notes before running found under run_instructions in the annex should be read carefully.

The features can be seperated in two different groups, features for visual understanding, analysis and comparisation between the clusters and feature for adding new points to the clusters and predicting their cluster belonging. The first group can be seperated in three different parts: Insights into one cluster, compare between multiple clusters and Insights into all clusters. For each of these parts there are multiple visualizations available between which the user can choose.

Now a more detailled insight into the features will be provided. When running the interface, the user has four different options to choose 1. 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 2. 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 can again select one clustering approach. In this example image, the user has chosen the boxplot for cluster 1 3. 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 4, 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. If these conditions are not fullfilled, no plot will be shown and the user will be alerted about the wrong conditions. 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 5. 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 6. 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 7. 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 8. 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 9.

3 Results

In this part, the results and insights gained by this interface will be presented.

The Insights into one cluster section provided simple information about one cluster. It was useful in order to understand the characteristics of each group in the clustering.

The Compare Clusters section had big impact in the comparisation between clusters, especially when deciding for the best number of clusters, two clusters could be specifically analysed in order to decide whether they provide different enough characteristics or not. It was also useful to gain deeper insights into the differences and similiraties for each feature between the clusters.

The Insights into all clusters section combined previous sections into the full overview over the clustering. It helped with final evaluation of the clustering with sizes and average values of all clusters and also provided visual presentation in form of scatter plots between the data points and its clustering group, in 2D (Cluster Overview) and interactively in 3D (3D Cluster PCA).

Finally, the Connect new entry to cluster section provided some experimental possibilities to evaluate if new points also get added to the clusters they should belong to. The Quick prediction function delivers not very precise results which often differ from the expected result. This happens because the Quick prediction function solely bases on the distance to each cluster centroid, but as could be seen in the 2D and 3D scatter plots of the clusters, the clusters are partially overlapping and not totally different, which can lead to not very precise predictions if solely based on distance to cluster centroids. The calculate cluster function was intended to fix this by totally recalculatin the clusters with the new value. Due to perfomance issues, sadly this functionality is not easy to use as it might take a lot of time, especially if the Final Clustering is choosen as in this case four different clusterings have to be recalculated. The results of this functionality however showed promise as the new points were mostly added to the expected clusters.

A Appendix

How to run the programm

Some important notes before running the interface: In the first cell of the interface notebook there can be found some pip install commands which are commented out. It can be tried to run all the imports from the third cell, if this does not work then the pip installs from the first cell should be applied and after the imports should work. If not, then the missing libraries will have to be installed seperately. The main library used for this interface is PyQt5. Furthermore some libraries considering the calculation of distances and usage of clustering algorithms were imported, mostly from sklearn and scipy.

In the second cell the user needs to enter the size of its screen/resolution. The interface was created in 1920x1080. If the first two entries, "current_width" and "current_height" are change to the right values, the interface should look rightly scaled. If something looks odd, probably the entered sizes are not well fitted for the screen and should be changed. The last two entries "base_width" and "base_height" are NOT to be changed as they define the original size in which the interface was created. Changing this would lead to massive problems in the positioning and sizes of widgets on the screen.

After setting the right window size and successfully importing all libraries, the next two cells should be run, the first one to define some global variables and the second one to define the actual code of the interface. Also the second to last cell should be run to define the run() method. Now the interface can be started as often as wished by running the last cell just consisting of the runMain() command.

Figure 1: Main page of interface

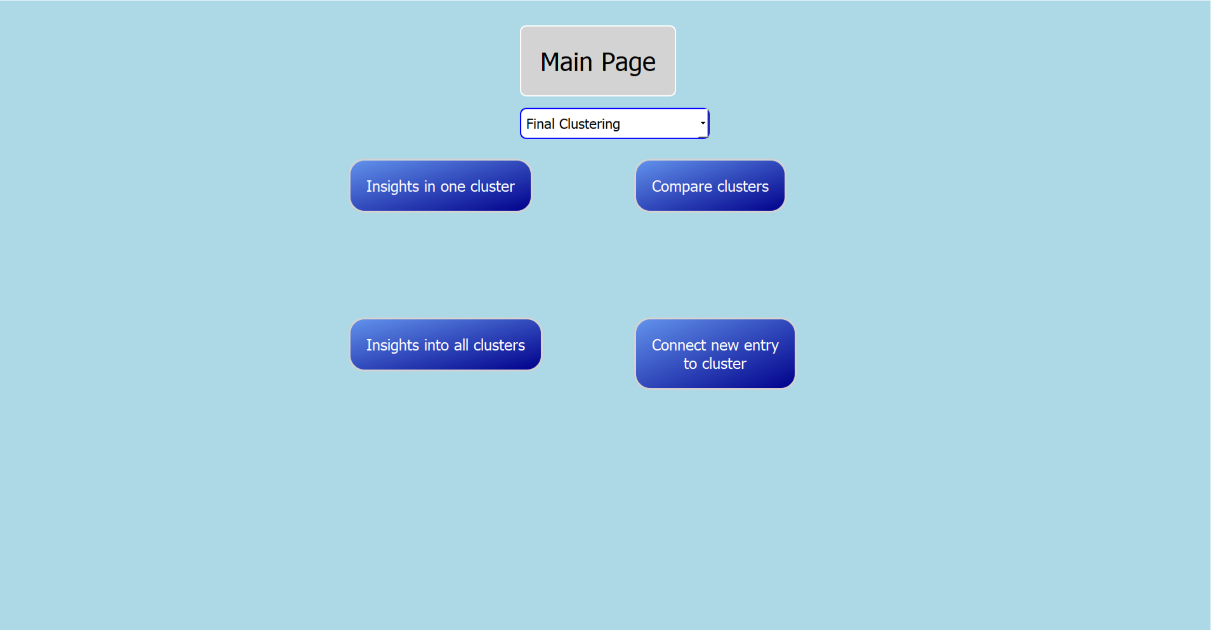
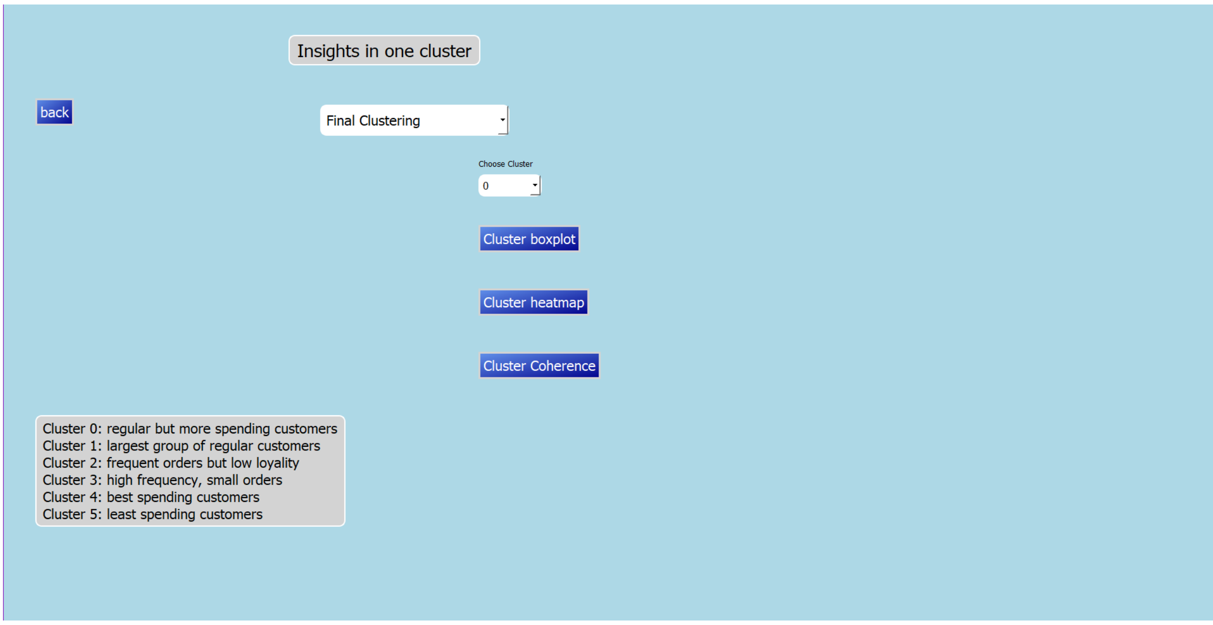
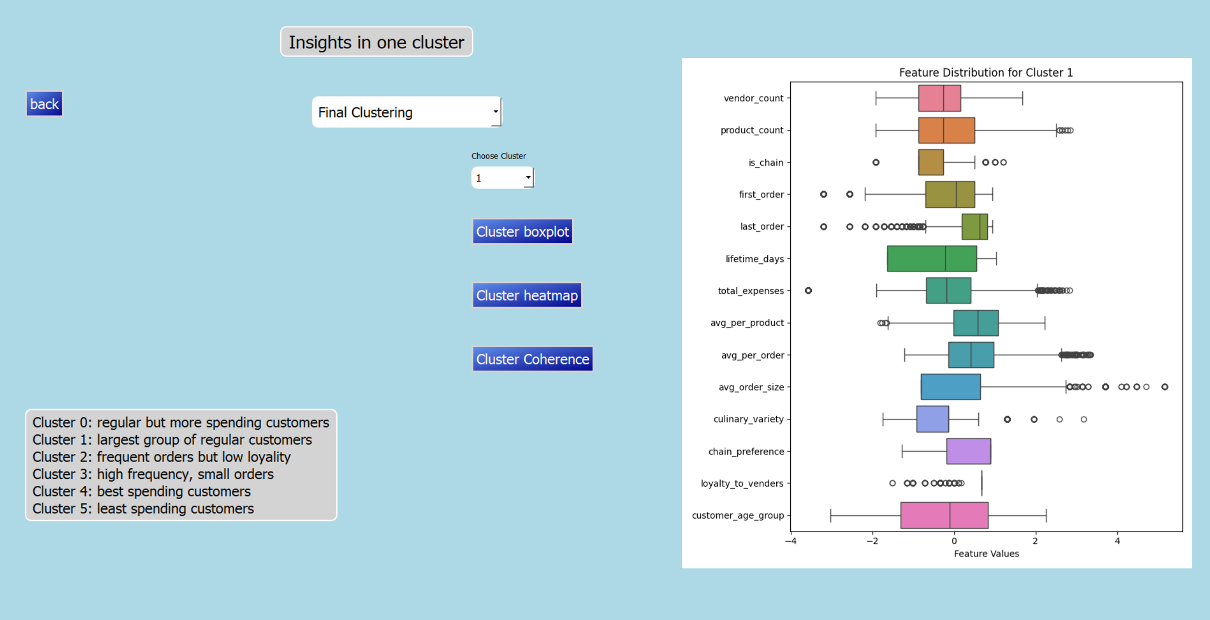


Figure 2: Page of interface 1



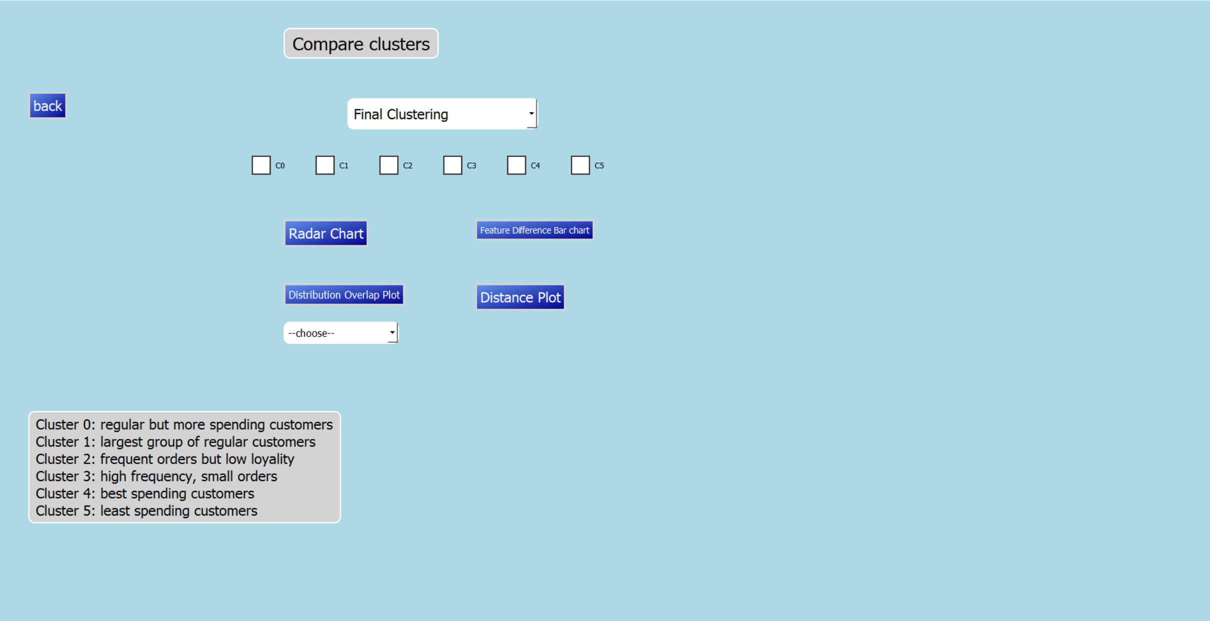
Cluster 0: regular but more spending customers
Cluster 1: largest group of regular customers
Cluster 2: frequent orders but low loyality
Cluster 3: high frequency, small orders
Cluster 4: best spending customers
Cluster 5: least spending customers

Figure 3: Page of interface level 1



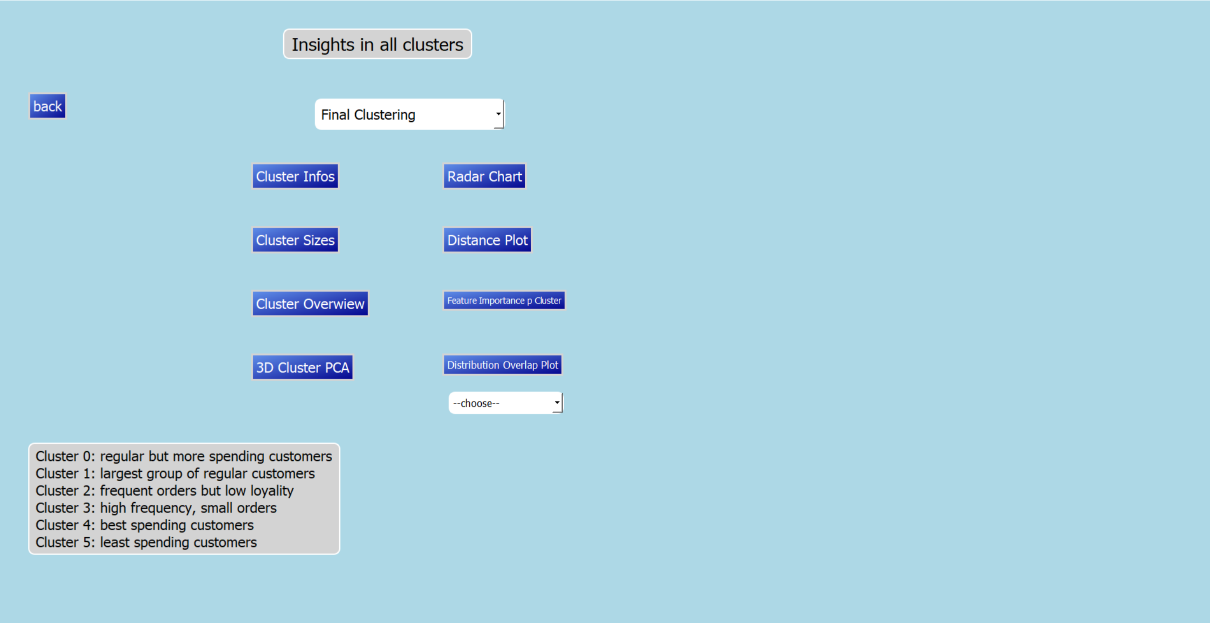
Cluster 0: regular but more spending customers
Cluster 1: largest group of regular customers
Cluster 2: frequent orders but low loyality
Cluster 3: high frequency, small orders
Cluster 4: best spending customers
Cluster 5: least spending customers

Figure 4: Page of interface 2


Compare clusters
back
Final Clustering
□ C0 □ C1 □ C2 □ C3 □ C4 □ C5
Radar Chart
Feature Difference Bar chart
Distribution Overlap Plot
Distance Plot
--choose--

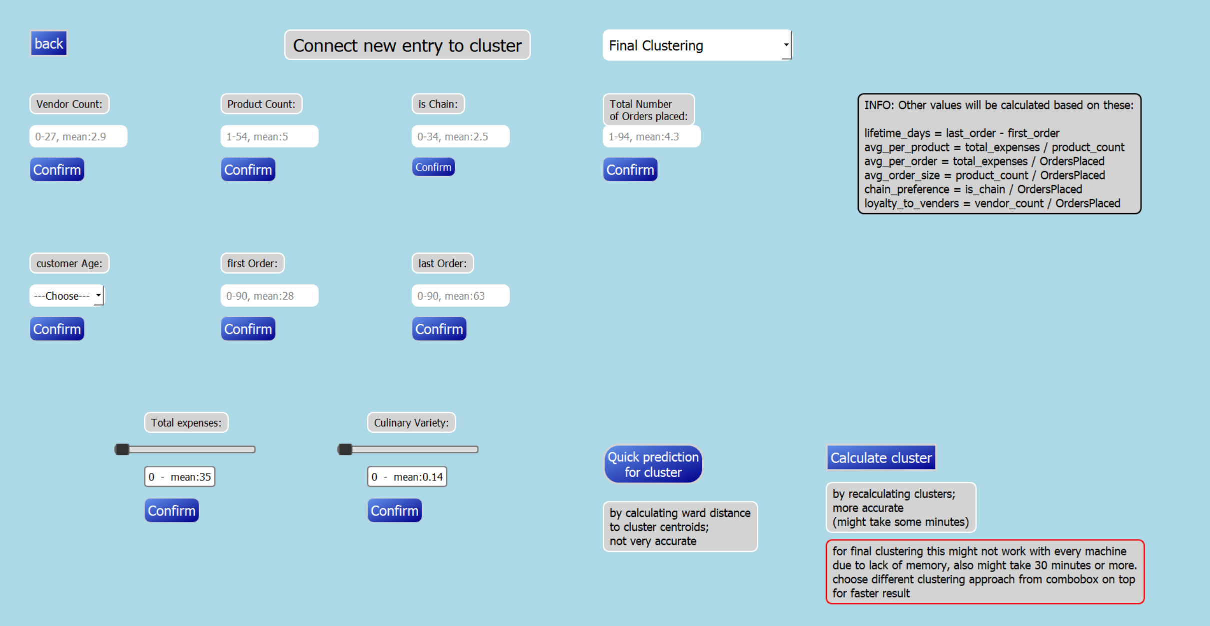
Cluster 0: regular but more spending customers
Cluster 1: largest group of regular customers
Cluster 2: frequent orders but low loyality
Cluster 3: high frequency, small orders
Cluster 4: best spending customers
Cluster 5: least spending customers

Figure 5: Page of interface 3



Cluster 0: regular but more spending customers
Cluster 1: largest group of regular customers
Cluster 2: frequent orders but low loyality
Cluster 3: high frequency, small orders
Cluster 4: best spending customers
Cluster 5: least spending customers

Figure 6: Page of interface 4



INFO: Other values will be calculated based on these:

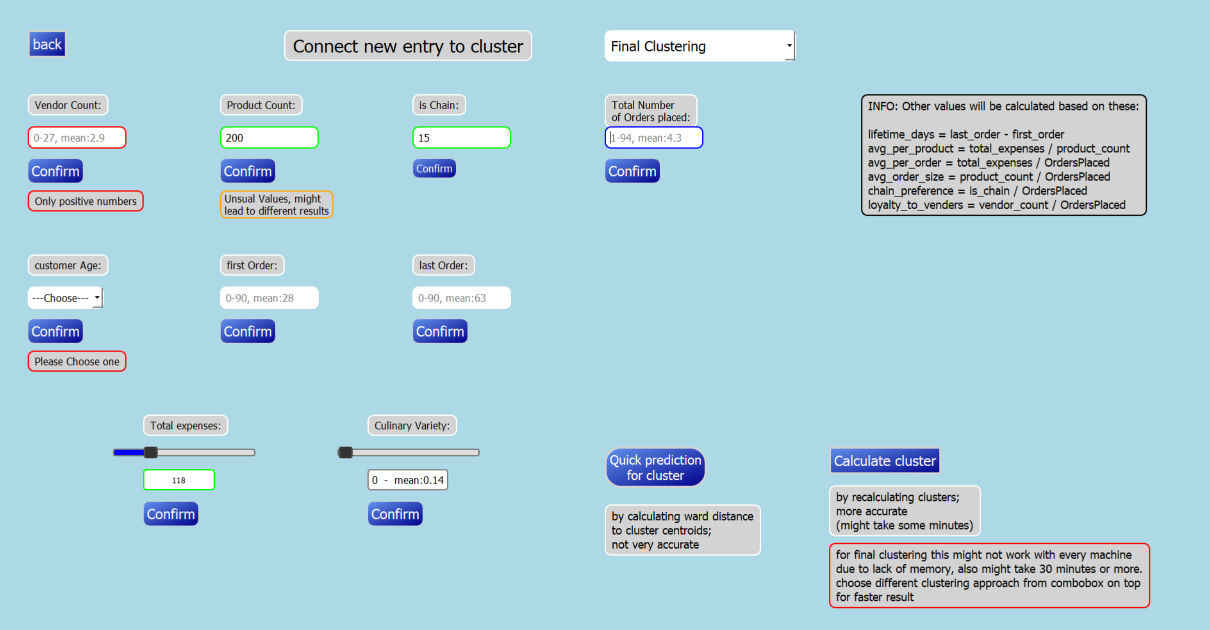
lifetime_days = last_order - first_order
avg_per_product = total_expenses / product_count
avg_per_order = total_expenses / OrdersPlaced
avg_order_size = product_count / OrdersPlaced
chain_preference = is_chain / OrdersPlaced
loyalty_to_venders = vendor_count / OrdersPlaced

by calculating ward distance to cluster centroids; not very accurate

by recalculating clusters; more accurate (might take some minutes)

for final clustering this might not work with every machine due to lack of memory, also might take 30 minutes or more. choose different clustering approach from combobox on top for faster result

Figure 7: Page of interface error



Product Count: 200

is Chain: 15

Total expenses: 118

Only positive numbers

Unsual Values, might lead to different results

Please Choose one

INFO: Other values will be calculated based on these:

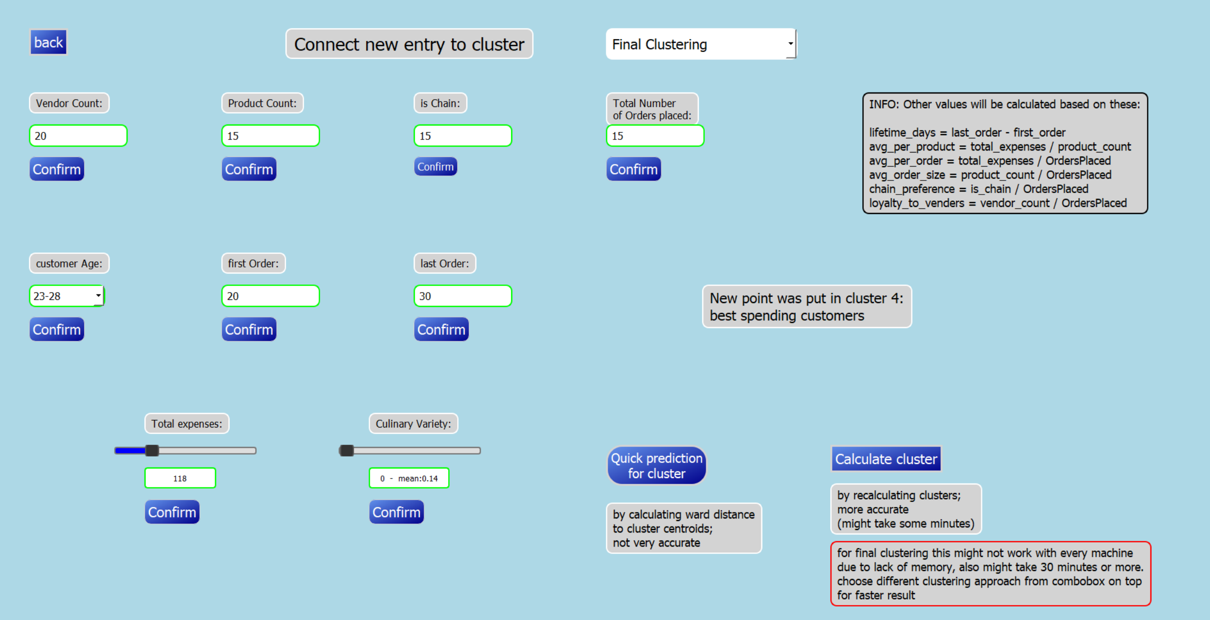
lifetime_days = last_order - first_order
avg_per_product = total_expenses / product_count
avg_per_order = total_expenses / OrdersPlaced
avg_order_size = product_count / OrdersPlaced
chain_preference = is_chain / OrdersPlaced
loyalty_to_venders = vendor_count / OrdersPlaced

by calculating ward distance to cluster centroids; not very accurate

by recalculating clusters; more accurate (might take some minutes)

for final clustering this might not work with every machine due to lack of memory, also might take 30 minutes or more. choose different clustering approach from combobox on top for faster result

Figure 8: Page of interface output



Final Clustering

Vendor Count: 20

Product Count: 15

is Chain: 15

Total Number of Orders placed: 15

customer Age: 23-28

first Order: 20

last Order: 30

Total expenses: 118

Culinary Variety: 0 - mean:0.14

INFO: Other values will be calculated based on these:

lifetime_days = last_order - first_order
avg_per_product = total_expenses / product_count
avg_per_order = total_expenses / OrdersPlaced
avg_order_size = product_count / OrdersPlaced
chain_preference = is_chain / OrdersPlaced
loyalty_to_venders = vendor_count / OrdersPlaced

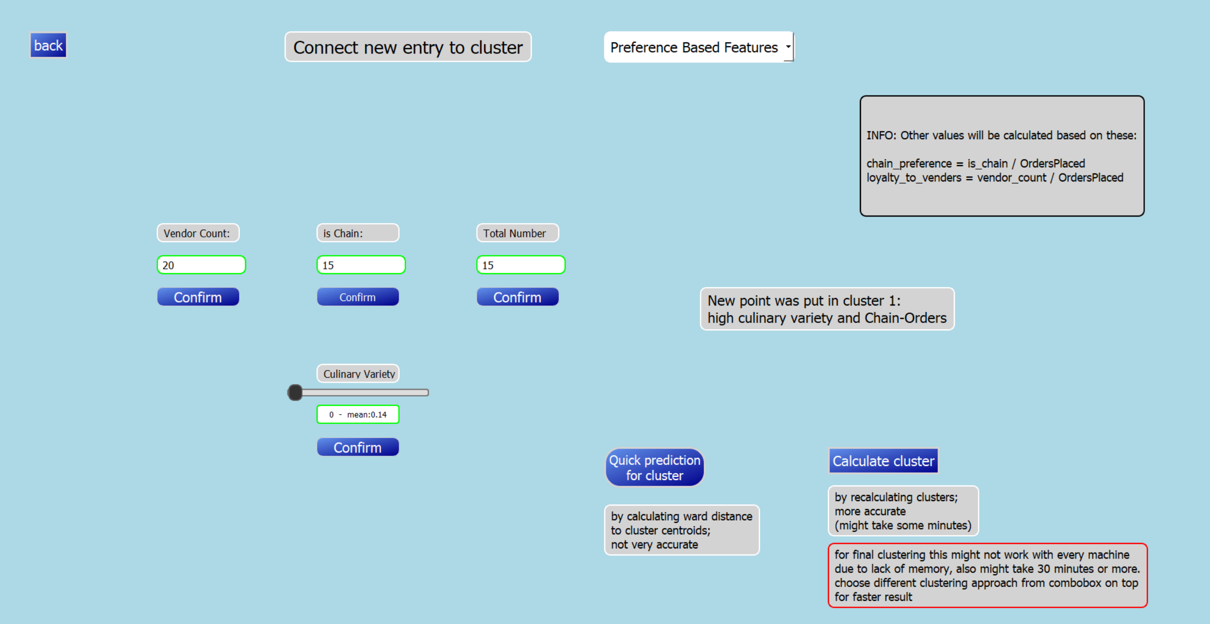
New point was put in cluster 4:
best spending customers

by calculating ward distance
to cluster centroids;
not very accurate

by recalculating clusters;
more accurate
(might take some minutes)

for final clustering this might not work with every machine
due to lack of memory, also might take 30 minutes or more.
choose different clustering approach from combobox on top
for faster result

Figure 9: Page of interface output 2



Preference Based Features

INFO: Other values will be calculated based on these:

chain_preference = is_chain / OrdersPlaced
loyalty_to_venders = vendor_count / OrdersPlaced

Vendor Count: 20

is Chain: 15

Total Number 15

Culinary Variety 0 - mean:0.14

New point was put in cluster 1:
high culinary variety and Chain-Orders

by calculating ward distance
to cluster centroids;
not very accurate

by recalculating clusters;
more accurate
(might take some minutes)

for final clustering this might not work with every machine
due to lack of memory, also might take 30 minutes or more.
choose different clustering approach from combobox on top
for faster result