DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc.
Clustering Analysis

Group 69

Paulo Martins, r2015469

Beatriz Fonseca, r20201599

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

In this report we present our findings regarding clustering analysis. We started by defining our customer base as the customers that recorded more than one purchase in the business and then we further filtered them based on their recency, frequency and monetary patterns and order diversity.

To these customers were then applied clustering techniques, such as Hierarchical Clustering, KMeans, Gaussian Mixture Models, Self-Organizing Maps and Spectral Clustering.

As for the perspectives, we assessed the customer behavior based on Spending, Geography, Cuisines and Time.

2. PREPROCESSING

Before starting the preprocessing efforts, an auxiliary dataframe was created excluding all customers which were flagged as non-regulars, i.e., customers that made only one purchase.

Then, using the regular customers, we proceeded to compute the outlier bounds and purged the outliers on the original dataset.

Following was the missing values analysis, which were detected in *cust_age*. To impute these values, we tested 3 alternatives on sample data: mean, median and KNN-imputer. The mean and KNN-imputer yielded similar results, so we went ahead and imputed the missing values using the mean, due to the lower computational complexity.

Once the data had no more missing values, we proceeded with the encoding of the categorical features. This is where we part ways with the non-regular customers and proceed only with the regular ones. We performed data standardization on the regular customers data and removed the multidimensional outliers using Local Outlier Factor.

Lastly, we performed PCA on *avg_amt_per_day*, *avg_product_per_day*, *avg_order_per_day*, *n_product* and *n_order*. We attempted to create 3 components, but only 2 yielded good loadings on these variables. **PC0** was represented by *avg_amt_per_day*, *avg_product_per_day* and *avg_order_per_day* - to which we called *transaction_volume*; **PC1** was related with *n_product* and *n_order* - which we named *interaction_rate*.

3. SPENDING AND ORDER DIVERSITY

3.1. RFM Clustering

To better understand the customer base, we implemented a kind of RFM analysis. We computed the *recency*, frequency (*n_order*) and monetary (*total_amt*) for each customer (regular or not) and assigned them a category between 1 and 3 for each of these values. This data had not yet been scaled and so we proceeded to do so.

We were now ready to start the clustering efforts using AgglomerativeClustering and KMeans. We used the first to compute the centroids and used them to initialize the KMeans algorithm, with 4 clusters - this returned a silhouette score of 0.42.

3.2. Spending and Order Diversity Clustering

Continuing our efforts to better define the customer based, we implemented the same algorithm workflow. However, this time we used the customers previously defined as regulars and the features *total_amt*, *n_cuisines*, *n_vendor* and *n_product*. This clustering yielded a silhouette score of 0.36.

Having information about the customers' placement in both clustering attempts, we compared both results. It was noted that 37% of the customers belonged to cluster 0 of the RFM and cluster 3 of the Spending clustering. Both clusters were represented by the lowest spending and interaction possible. Therefore, we disregard these customers as part of the customer base.

Using only the customer base from now on, we attempt again to cluster the customers under this Spending and Order Diversity perspective. We recompute the centroids and clusters, using 4 clusters. This returned a silhouette score of 0.37 and an R^2 of 0.64. As for the cluster profiling (See Annex 1, 2):

Cluster 0 – Adventurous High-Spenders: These customers are explorers, frequently trying new cuisines and vendors. They also buy many products and spend a significant amount of money. They value variety and have high purchasing power, likely enjoying discovering new options and experiences.

Cluster 1 – Loyal High-Spenders: These customers spend a lot of money and buy many products but prefer sticking to familiar cuisines and vendors. They exhibit loyalty to a select range of offerings while demonstrating significant spending capacity.

Cluster 2 – Low-Spending Minimalists: These customers have the lowest spending, try the fewest cuisines and vendors, and purchase the least products. They are cost-conscious and not very exploratory, possibly focusing on essentials or sticking to a routine.

Cluster 3 – **Exploratory Budget-Conscious**: These customers enjoy trying different cuisines and vendors but do not purchase many products or spend much money. They prioritize variety and experiences but are budget-conscious or limit their purchases.

4. GEOGRAPHY

During our initial exploration of the data, we found clear signs that spending patterns were separatable by city - see Annex 4. We hypothesized that this provides a natural clustering solution, and so we applied Spectral Clustering to a mix of metric and categorical features. The selected features were: *per_chain_order*, *log_total_amt*, *avg_amt_per_product*, *n_cuisines*, *cust_city_2.0*, *cust_city_4.0* and *cust_city_8.0* (the hot encoding of city).

Naturally, an appropriate metric needed to be selected, that could handle hybrid data, so the "Gower Distance" was chosen. The distance matrix of the data points was calculated, and a rotation kernel was applied. The Laplacian Matrix and respective eigenpair solutions were obtained, for a pre-defined number of clusters (3).

The resulting solution's eigenvectors were used to transform the original features, with the resulting dataframe being normalized, and then clustered using KMeans.

The final solution achieved an R-squared of 0.43, but silhouette was not calculated as, spectral clusters are not necessarily spheroids. Pair plots for features demonstrated that each city was separated into its own cluster - confirming our initial suspicions. And in fact, all variables demonstrated a good degree of separation of their clusters, both in pairwise comparisons, as well as within their own distributions.

Given this, Cluster 0 can then be named "City8", Cluster 1 can be named "City4" and "Cluster 2" can be named "City2".

City2's customers show a preference for chained restaurant food, having a preference for spending less in aggregate than their peers, and less on each product, while being the ones that tend to purchase from the highest number of cuisines.

City4's customers display a moderate propensity towards spending, as well as a moderate interest in experimenting with new cuisines.

City8's customers have a below average propensity to consume from chained restaurants, while having a high propensity to spend more both in aggregate as well as per product.

In absolute terms, **City8** customers spend on average $76.52 dollars, **City4** customers $58.69 and **City2** customers $32.04.

5. CUISINES

To cluster the cuisines, we used sparse Principal Component Analysis (sPCA), a technique inspired by PCA that relaxes the restriction on orthogonality of component eigenvectors, allowing for linear dependence, which is especially designed for sparse data; as a result, the components are not loadings (which project data onto new axes) but sparse coefficients that encode the contribution of the selected features. Using this method, we were able to represent the cuisines variables using only two components.

Component 1 – Preference for Casual and Street-Style Dining: High contribution from *Asian*, *Beverages*, *Desserts* and *Street Food and Snacks*; and negative contribution from *Cafe*, *Indian*, *Italian*, *Other* and *Thai*. This may indicate a preference for casual, street-style dining over more formal dining options.

Component 2 – Preference for Comfort/Chinese-style Meals: High contribution from *Chinese*, *Noodle Dishes*, *Chicken Dishes* and *Other*; and negative contribution from *Asian*, *Street Food and Snacks*, *American*, *Cafe* and *Italian*. This may indicate a preference for comfort/chinese-style meals.

Recalling that customer cuisine preferences were separable using the marginal propensity towards consumption i.e. the logs of *total_amt*, and *avg_amt_per_product*, into two distinct distributions - a Gaussian Mixture Model was used to attempt to cluster the cuisine components with these logged variables.

Best results were obtained for three clusters, with an R^2 of 0.53, and a Silhouette Score: 0.34. Given the inherent sparsity of the data, this was deemed significant.

Cluster 0 was associated with negative values for components 1 and 2 of the sPCA and marginally positive values for the log variables, implying a small propensity to cuisines that have negative coefficients in both components 1 and 2.

Cluster 1 was associated with a large propensity towards consumption, with high logged feature values, and a high coefficient for component 1, and negative coefficient for component 2, meaning these customers display a strong preferences for positive coefficient cuisines in component 1 and negative coefficient cuisines in component 2.

Cluster 2 shows a strong preference for component 2 positive coefficients and is only slightly positively associated with component 1 negative coefficient cuisines, while exhibiting a trend to expend as little as possible and opt for cheap products.

Given this analysis, we label the clusters "Grab to Go", "Comfort", "Italian American".

Regarding categorical variable profiling, "Grab to Go" customers display the largest probability of not having used any discounts in their previous purchases. While customers in the "Italian American" cluster had the lowest. Conversely, digital payment was most frequent among cluster 2, while card was preferred by cluster 0.

Lastly, reverting our scaling we get the mean values for our clusters, we can see that "Comfort" customers tend to experiment the least with cuisines, but spend the most $76.73. "Italian American" customers purchase the highest number of products - 8.85 on average - but spending the least with an average of $30.08 spent. With "Grab to Go" customers displaying the median behavior across these criteria.

Lastly, with respect to city/region, there is a clear distinction: "Grab to Go" customers reside in city 4 (approx. 84%), "Comfort" customer in city 8 (approx. 97%) and "Italian American" customers in city 2 (approx. 92%).

This lets us conclude that there is a very high association between cuisine preferences, amount spent and city/region; as both our results from spectral clustering, perfectly align with our sPCA, GMM methodology. And that any marketing strategy must be preferably based on geography, and spending power.

6. TIME

As for this perspective, we attempt to cluster the data based on temporal features - the hours (HR_* features) and the days of week (DOW_* features).

For starters, Non-negative Matrix Factorization (NMF) was applied, and 4 factors were created. Factor 1 is related with hours 1 to 6 and all days of week; Factor 2 is related with hours 9 to 12 and all days of week; Factor 3 is related with hours 13, 14 and 19 to 21 and all days of week; and Factor 4 is related with hours 15 to 18 and all days of week. To our working dataset*, avg_amt_per_product*, *n_chain* and *n_cuisines* were also added.

After, a Gaussian Mixture Model with 4 components was applied and returned good visual results. The computed R² was 0.32.

We believed we could do better, so we used Self-Organizing Maps (SOM). As per visual inspection, we concluded Factor 1 was related with *total_amt* and Factor 3 was related with *n_chain* and *n_cuisines*.

Then, we proceeded to apply KMeans clustering, which returned a silhouette score of 0.33 and an R² of 0.40. The profiling of these clusters tells us that:

Cluster 0 – Premiums: High loading on *avg_amt_per_product*. These customers buy expensive products having no time preference. Almost no customers are from city 2.

Cluster 1 – Nighttime Gourmet: High loadings on Factor 1 and *avg_amt_per_product*. These customers buy expensive products during the night. These customers are almost exclusively from city 8.

Cluster 2 – Afternoon Moderate Spenders: Moderate loading on *avg_amt_per_product*. These customers tend to buy in the mid afternoon and average-priced products. Most customers are from region 2360.

Cluster 3 – Adventurous Workforce: High loadings on Factor 3 and *n_cuisines*. These customers like to eat from different cuisines during lunch or dinner times and from chain restaurants. Most customers are from region 2360.

7. FINAL RECOMMENDATIONS

This report finds strong evidence that customers can be segmented effectively by their region, by the hours at which they choose to buy, and the amount that they are willing to spend. As such regarding the time dimension, we suggest time-sensitive advertisement at times that closely correlate with the found customer segments. We further recommend that promotional campaigns that align cuisines and regions be devised based on these clusters.

The information gathered further suggests the need for A/B testing, to confirm these early findings, and a marketing strategy be devised around them.

REFERENCES

Local Outlier Factor

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (pp. 93–104). ACM. https://doi.org/10.1145/342009.335388

Spectral Clustering

Luxburg, U. V. (2007). A tutorial on spectral clustering. Statistics and Computing, 17(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z

Gower Distance

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. Biometrics, 27(4), 857–871. https://doi.org/10.2307/2528823

Sparse Principal Component Analysis

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15(2), 265–286. https://doi.org/10.1198/106186006X113430

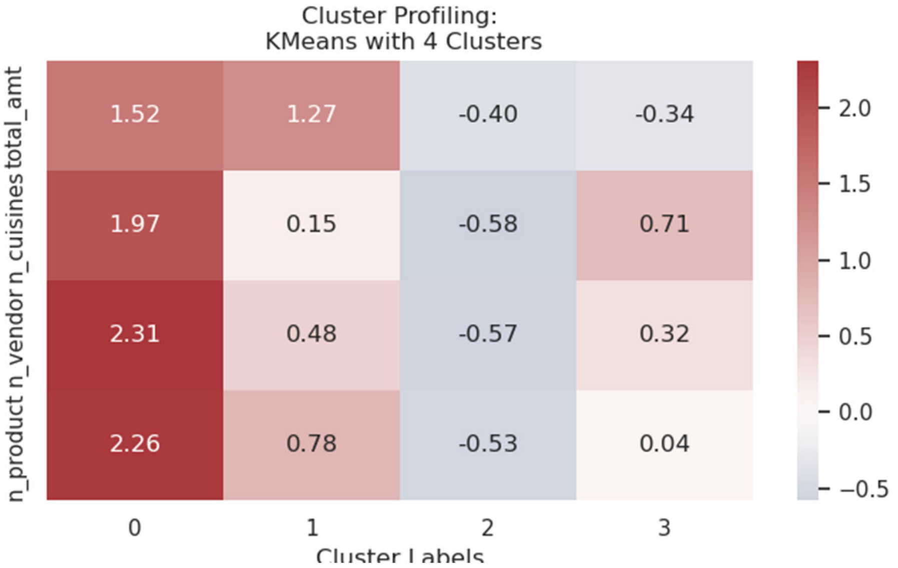
Non Negative Matrix Factorization

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788–791. https://doi.org/10.1038/44565

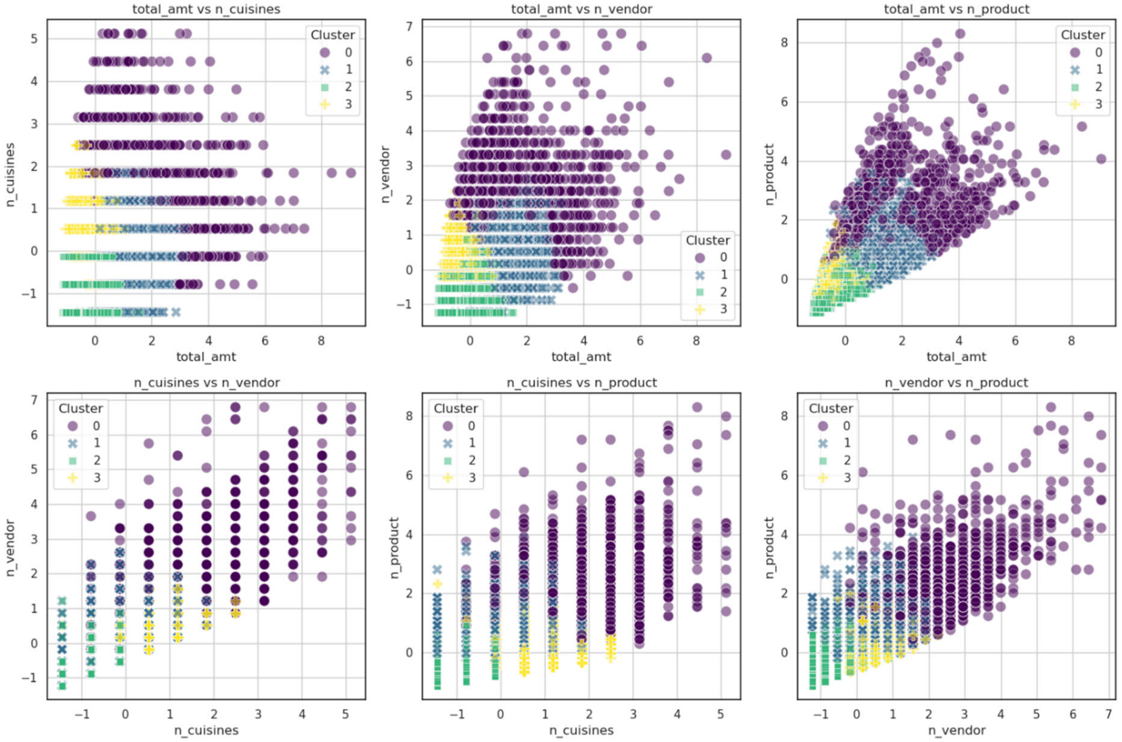
Gaussian Mixture Models

Reynolds, D. A. (2009). Gaussian mixture models. In S. Z. Li & A. Jain (Eds.), Encyclopedia of biometrics (pp. 659–663). Springer. https://doi.org/10.1007/978-0-387-73003-5_196

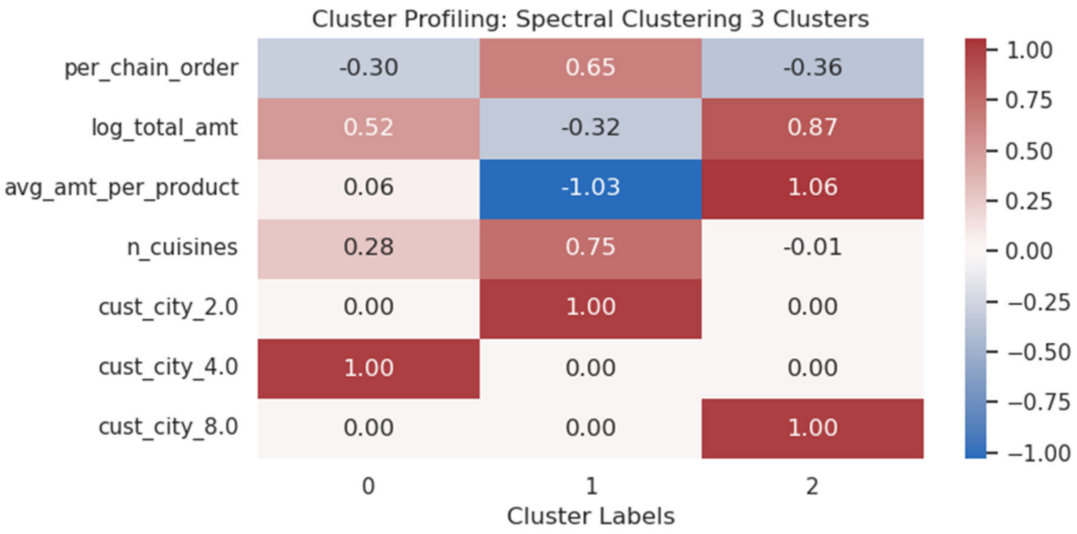
ANNEXES



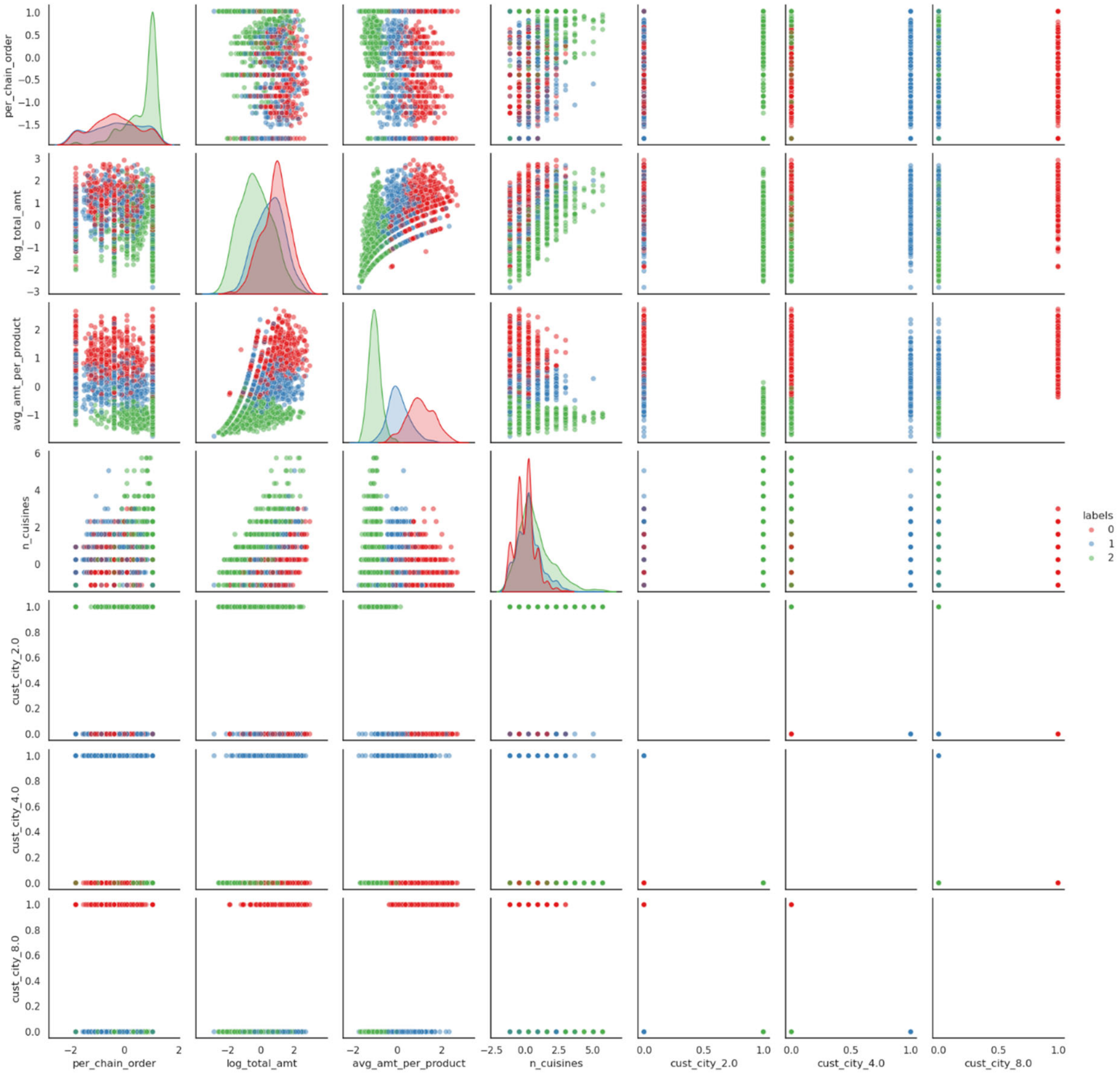
Annex 1



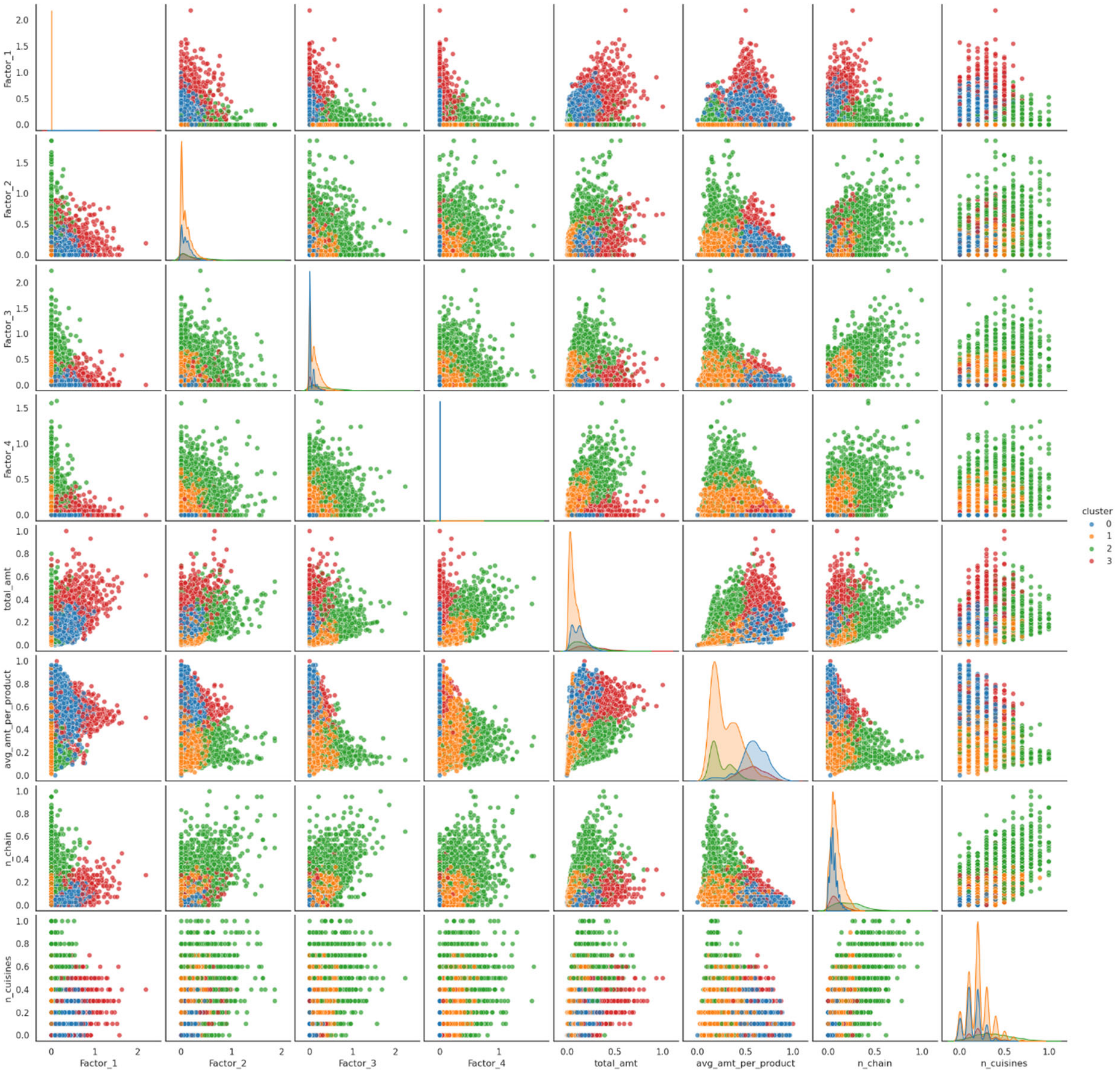
Annex 2


Cluster Profiling: Spectral Clustering 3 Clusters

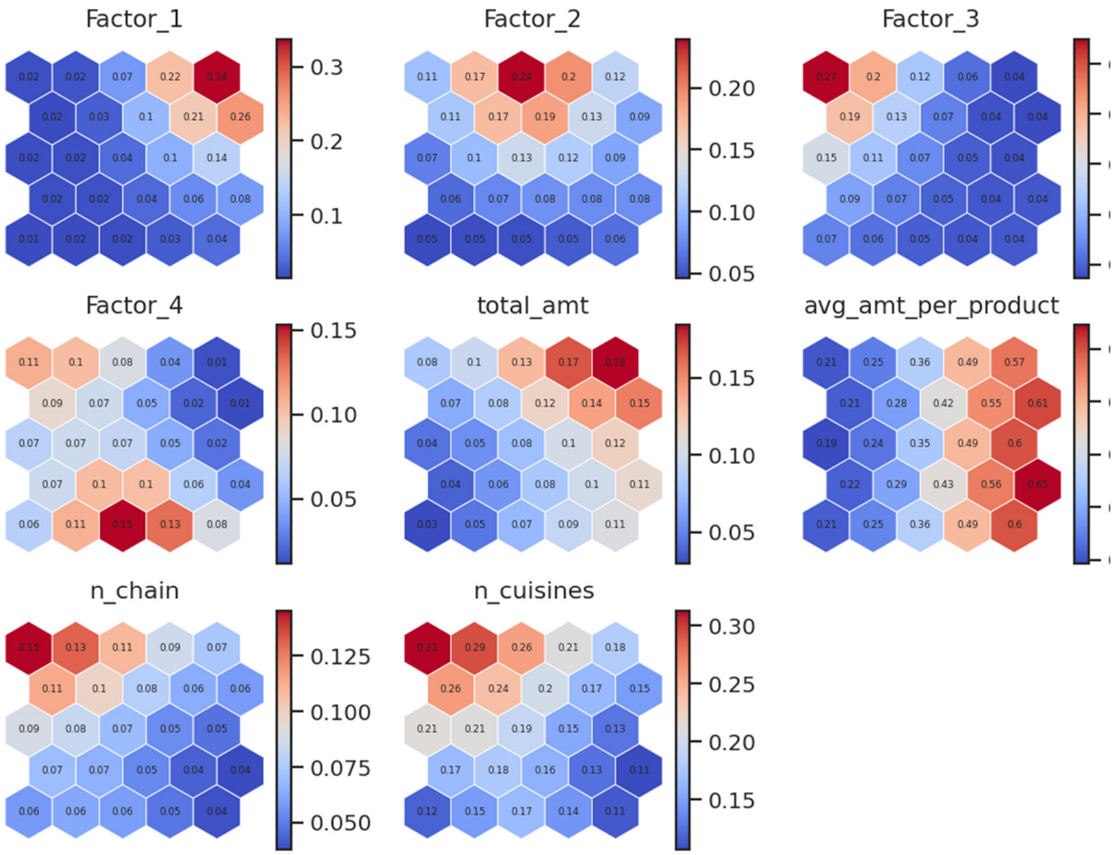
Annex 3


per_chain_order
log_total_amt
avg_amt_per_product
n_cuisines
cust_city_2.0
cust_city_4.0
cust_city_8.0
1.0
0.5
0.0
−0.5
−1.0
−1.5
3
2
1
0
−1
−2
−3
6
4
0.8
0.6
0.4
0.2
−2
−2.5
2.5
5.0
labels
0
1
2

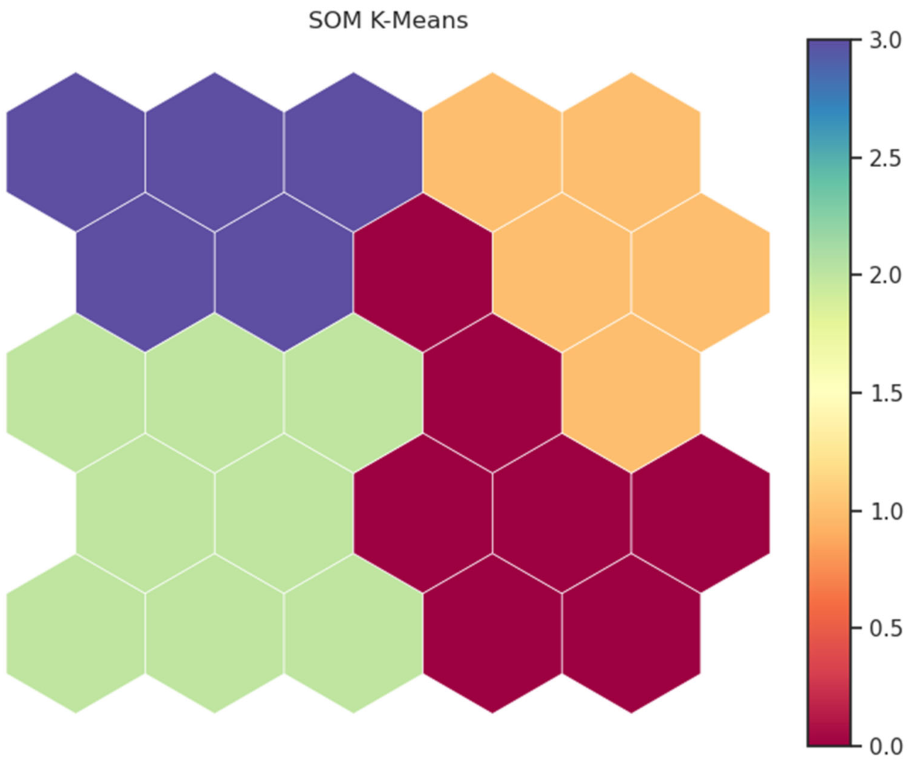
Annex 4


Factor_1
Factor_2
Factor_3
Factor_4
total_amt
avg_amt_per_product
n_chain
n_cuisines
cluster
0
1
2
3

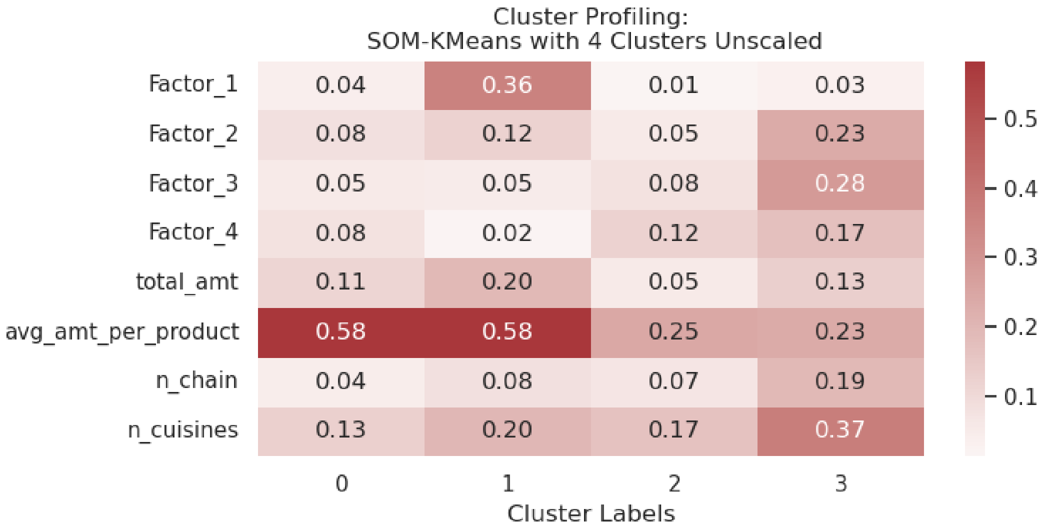
Annex 5


Factor_1
0.02 0.02 0.07 0.22 0.34
0.02 0.03 0.1 0.21 0.26
0.02 0.02 0.04 0.1 0.14
0.02 0.02 0.04 0.06 0.08
0.01 0.02 0.02 0.03 0.04
0.3
0.2
0.1
Factor_2
0.11 0.17 0.24 0.2 0.12
0.11 0.17 0.19 0.13 0.09
0.07 0.1 0.13 0.12 0.09
0.06 0.07 0.08 0.08 0.08
0.05 0.05 0.05 0.05 0.06
0.20
0.15
0.10
0.05
Factor_3
0.27 0.2 0.12 0.06 0.04
0.19 0.13 0.07 0.04 0.04
0.15 0.11 0.07 0.05 0.04
0.09 0.07 0.05 0.04 0.04
0.07 0.06 0.05 0.04 0.04
Factor_4
0.11 0.1 0.08 0.04 0.01
0.09 0.07 0.05 0.02 0.01
0.07 0.07 0.07 0.05 0.02
0.07 0.1 0.1 0.06 0.04
0.06 0.11 0.15 0.13 0.08
0.15
0.10
0.05
total_amt
0.08 0.1 0.13 0.17 0.18
0.07 0.08 0.12 0.14 0.15
0.04 0.05 0.08 0.1 0.12
0.04 0.06 0.08 0.1 0.11
0.03 0.05 0.07 0.09 0.11
0.15
0.10
0.05
avg_amt_per_product
0.21 0.25 0.36 0.49 0.57
0.21 0.28 0.42 0.55 0.61
0.19 0.24 0.35 0.49 0.6
0.22 0.29 0.43 0.56 0.65
0.21 0.25 0.36 0.49 0.6
n_chain
0.15 0.13 0.11 0.09 0.07
0.11 0.1 0.08 0.06 0.06
0.09 0.08 0.07 0.05 0.05
0.07 0.07 0.05 0.04 0.04
0.06 0.06 0.06 0.05 0.04
0.125
0.100
0.075
0.050
n_cuisines
0.31 0.29 0.26 0.21 0.18
0.26 0.24 0.2 0.17 0.15
0.21 0.21 0.19 0.15 0.13
0.17 0.18 0.16 0.13 0.11
0.12 0.15 0.17 0.14 0.11
0.30
0.25
0.20
0.15

Annex 6

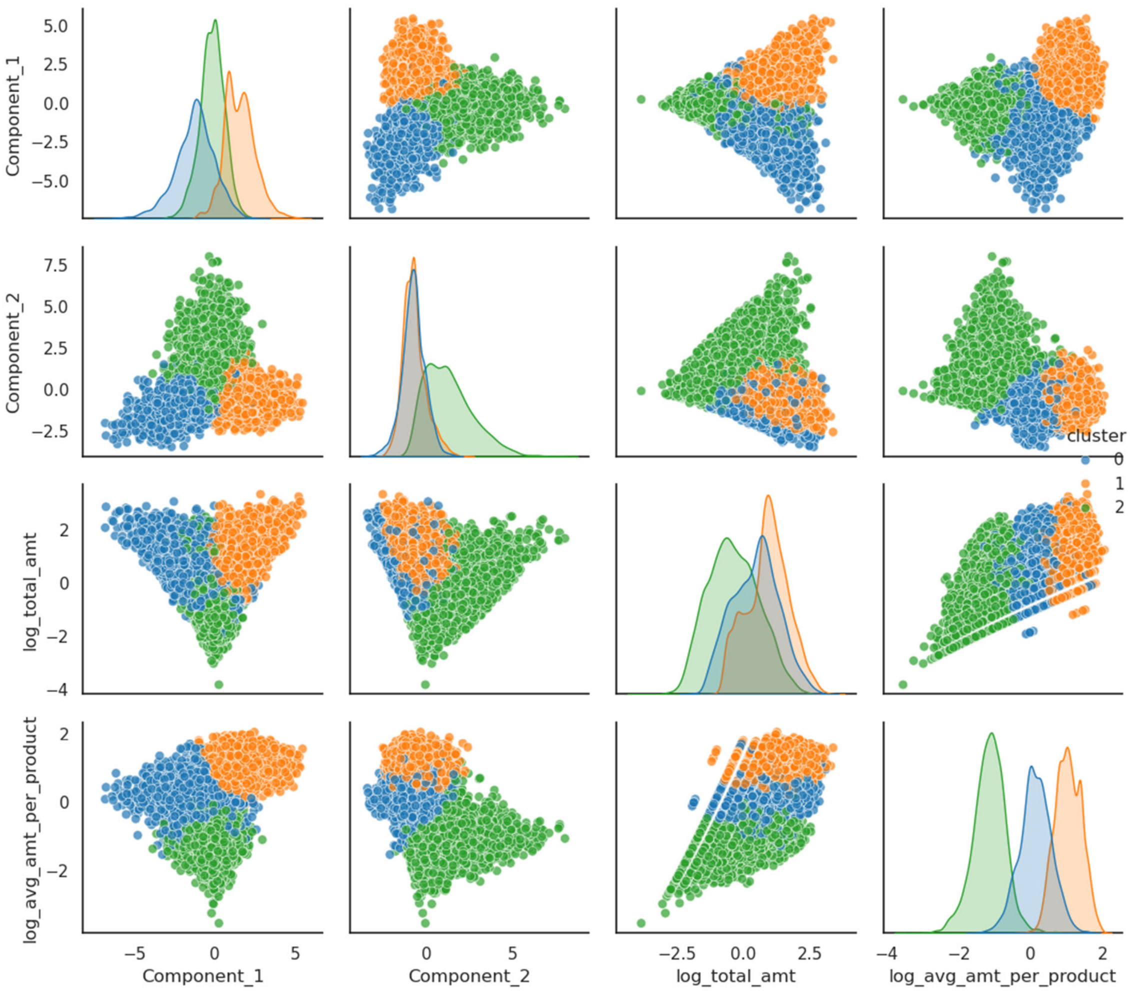


Annex 7

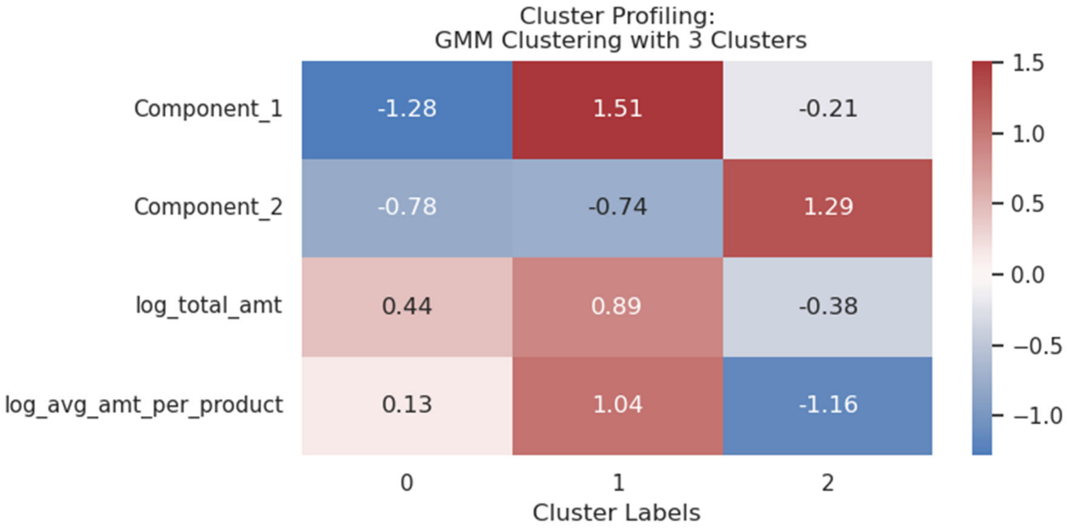


	0	1	2	3
Factor_1	0.04	0.36	0.01	0.03
Factor_2	0.08	0.12	0.05	0.23
Factor_3	0.05	0.05	0.08	0.28
Factor_4	0.08	0.02	0.12	0.17
total_amt	0.11	0.20	0.05	0.13
avg_amt_per_product	0.58	0.58	0.25	0.23
n_chain	0.04	0.08	0.07	0.19
n_cuisines	0.13	0.20	0.17	0.37

Annex 8


Component_1
Component_2
log_total_amt
log_avg_amt_per_product
5.0
2.5
0.0
−2.5
−5.0
7.5
5.0
2.5
0.0
−2.5
2
0
−2
−4
2
0
−2
−5
0
5
0
5
−2.5
0.0
2.5
−4
−2
0
2
cluster
0
1
2

Annex 9


Cluster Profiling:
GMM Clustering with 3 Clusters
Component_1
Component_2
log_total_amt
log_avg_amt_per_product
-1.28
1.51
-0.21
-0.78
-0.74
1.29
0.44
0.89
-0.38
0.13
1.04
-1.16
0
1
2
Cluster Labels
1.5
1.0
0.5
0.0
−0.5
−1.0

Annex 10