DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc.
Customer Segmentation

Group 05

Daniel Caridade, 20211588

Gonçalo Teles, 20211684

Gonçalo Peres, 20211625

João Venichand, 20211644

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

The food delivery market has evolved from being perceived as a luxury service to becoming an integral part of everyday life. This transformation is underscored by its revenue of €531.70 million in Portugal in 2024, with further growth anticipated to €627.98 million in 2025 [1]. The market experienced significant acceleration during the COVID-19 pandemic as consumers, confined to their homes, increasingly relied on delivery services for essential and non-essential goods [2].

In this context, ABCDEats Inc. hired our team of NOVA IMS students to develop a data-driven customer segmentation strategy leveraging domain expertise in data mining, to develop a clustering solution to segment their customers. This initiative began with an exploratory data analysis, which revealed key insights, such as the presence of 138 inactive customers during the dataset's time span, followed by preprocessing addressing data inconsistencies, such as inactivity, feature engineering by joining some cuisines regionally to reduce dimensionally, and imputed missing values using K-Nearest Neighbors ending with extreme outliers' removal to enhance clustering performance. Feature selection techniques identified *product_count*, *last_order*, and *first_order* as redundant, leading to their exclusion.

Clustering algorithms were then applied from two perspectives: customer activity and cuisine preferences. Our results demonstrated that the K-Means algorithm outperformed others in customer activity segmentation, while hierarchical clustering proved more effective for cuisine-based segmentation. Finally, the clusters were combined and profiled using hierarchical clustering to provide actionable recommendations.

2. EXPLORATORY DATA ANALYSIS

As part of our analysis of the dataset provided by ABCDEats Inc., a detailed exploratory data analysis (EDA) report was already delivered mid-semester. However, subsequent investigations revealed an important new insight: 156 customers had a recorded value of 0 in the *product_count* feature, indicating that they did not purchase any products from the company. Upon further analysis, it was clear that only 138 of these 156 customers had no recorded interactions with the company, which aligns with their *product_count* values. Conversely, 18 customers exhibited interactions with the company, suggesting that their *product_count* values of 0 are erroneous.

In addition to identifying this anomaly, our team engineered several new features based on the original dataset. These new features, which were not included in the initial EDA report, offer valuable insights and are discussed in detail in the following section.

3. ADDITIONAL FEATURE EXPLORATION

As previously mentioned, new features were created to provide additional insights into ABCDEats Inc. customers, whose descriptions can be found in Appendix 1. Regarding the distributions of these features, most exhibit a right-skewed pattern, except *dayswus*, which follows a left-skewed distribution. Additionally, all features, aside from *recency* and *dayswus*, contain outliers.

The analysis of the *recency* feature revealed that **0.1%** of customers made their last purchase on the final day of the dataset, **0.33%** on the first day, and **6736** customers made their last purchase 45 days ago, which represents half of the maximum days in the dataset. The features *D_Orders* and *H_Orders* demonstrated discrepancies in their totals which shouldn't be the case and this is due to the missing data in the feature *HR_0*. The *dayswus* feature revealed that many customers first interacted with the company long ago, with all making at least two purchases.

An analysis of the features *Orders_Weekdays* and *Orders_Weekend* revealed that the majority of orders, **71.15%**, are placed on weekdays. However, when analysed on a per-day basis, weekends account for a slightly higher share of orders at **14.42%** of the total orders per day, compared to **14.23%** for weekdays. The additional time-of-day features (*Early Morning*, *Morning*, *Afternoon*, *Evening*, and *Night*) revealed that the afternoon accounts for the largest share of orders at **32.4%**, which aligns with lunch hours. This is followed by the evening, which represents **25.8%** of orders and corresponds to dinner time, while the morning accounts for **21.8%** of the total orders.

Revenue analysis by cuisine type showed that *CUI_Asian* contributes the highest share (**26%)**, followed by *CUI_American* (**12.74%**) and *CUI_Street Food/Snacks* (**10.21%**). The lowest contributors are *CUI_Noodle Dishes* (**1.86%**) and *CUI_Chicken Dishes* (**2.01%**).

Lastly, a noticeable imbalance in revenue distribution among cuisines was observed, highlighted by the Lorenz curve (Appendix 2), that indicates the top 20% of cuisines account for nearly **48.95%** of the total revenue.

4. DATA PREPROCESSING

4.1. Treat Incoherences

In this stage, the incoherences identified during EDA were treated, starting with the removal of 13 duplicated customer entries from the dataset. Then the values equal to "-" in column *customer_region* were changed into "Unknown", and in column *last_promo* to "NONE", since we considered those values in this last feature as not having used a promotion in their last purchase.

The next step was the removal of 138 instances who had no interactions with the company, since those did not make any purchases during the dataset's time span, so they do not qualify as active customers. Following this, all customers under the age of 16 were removed, in compliance with GDPR regulations, which prohibit the processing of minors' data without parental consent, which we do not have [3]. Lastly, values in *product_count* equal to 0 were set to missing, as we observed that those customers made a purchased in the company.

4.2. Feature Engineering

The feature engineering process began by reducing the dimensionality of the cuisine-related columns. This was achieved by aggregating features based on shared relationships (e.g., consistently positive or negative correlations) and logical or regional similarities. For example, the feature *CUI_Chicken Dishes* was aggregated into *CUI_Chinese,* since this food type originated in China. Similarly, *CUI_Italian* and

CUI_American were merged into feature *CUI_Western*, as both represent cuisines from the Western world. *CUI_Thai* and *CUI_Indian* were consolidated into *CUI_IndianOceanic* to reflect their shared demographic proximity to the Indian Ocean, while *CUI_Street Food / Snacks* and *CUI_Beverages* were combined into *CUI_Street_Food_Beverages*. Then the categorical features were frequency encoded, since their numerical representation is required for the imputation algorithms applied in the next section. The final step was to remove the feature *is_chain* due to metadata inconsistencies and limited analytical value because **71%** of its values ranged from 0-3, yet the feature exhibited high cardinality, making it unsuitable for meaningful cluster profiling.

4.3. Missing value imputation

The process of imputing missing values was a very straightforward one, starting by imputing the missing values in *HR_0* as the difference between *D_Orders* and *H_Orders* and missing values in *first_order* were imputed as 0 if the value in *last_order* was 0. For the remaining features with missing values (*customer_age*, *first_order* and *product_count*) KNNImputer with 5 neighbours was applied with data being scaled previously using Z-score and with unnecessary features being removed to avoid computing unnecessary distances as their information is already captured by aggregated features. The removed features included those that have an aggregated version (e.g. *DOW_* columns have *Orders_Weekday* and *Orders_Weekend*), as well as *D_Orders* and *H_Orders*, which were solely created to impute missing values in *HR_0.* After the imputation process, features that depended on the values of imputed features were recalculated to ensure consistency in the data and the categorical features were transformed back into their original scale.

4.4. Outlier Treatment

The final step in our pre-processing pipeline was outlier treatment, a critical stage to ensure the integrity and reliability of the clustering analysis. Outliers can disproportionately influence clustering algorithms, particularly the ones presented afterwards in this report, that rely on Euclidean distances. Their presence can distort results by creating isolated clusters of outliers or assigning outliers to their own clusters, thereby undermining the representativeness and effectiveness of the segmentation.

Given the large prevalence of outliers observed during the EDA, we focused exclusively on treating the features intended for clustering. Rather than removing all outliers, we targeted only the most extreme cases, as detailed in Appendix 3 and 4. This approach balanced the need to minimize distortion in the clustering results while retaining sufficient data for customer segmentation. As a result of the entire pre-processing pipeline, **3%** of the dataset was removed, which is within the **5% threshold** widely regarded as acceptable.

5. FEATURE SELECTION

For feature selection, our approach focused on identifying redundant features to be removed from the analysis starting by examining the variance of all numerical features and finding that none were univariate, indicating that all held meaningful variability and should be considered for analysis. Following this, Spearman correlations between numerical features were evaluated, as shown in Appendix 5 revealing perfect correlations between *dayswus* and *first_order*, as well as between *recency* and *last_order* and high correlations of feature *product_count* with both *vendor_count* and

Orders_Weekday. Based on these observations, *first_order* and *last_order* were removed, since *dayswus* and *recency* provide more direct and interpretable measures of customer behaviour. Specifically, *dayswus* captures customer tenure and loyalty without requiring the combined analysis of *first_order* and *last_order* and *recency* offers a clear measure of recent customer engagement, which is easier to interpret and more actionable than the raw value provided by *last_order*, that requires knowing the range of days in the dataset a priori for a proper interpretation. Furthermore, removing this features allows us to retain the information from the 2 features it is correlated with while reducing the redundancy in the analysis.

6. CLUSTERING

Our clustering initiative focused on selecting key features from the dataset to create two different perspectives for customer segmentation. The first perspective, called **customer activity**, analyses customer interactions with the company to better understand their engagement patterns and the second one **cuisine preferences** perspective examines different customer groups to identify the types of cuisines they are most likely to purchase.

6.1. Customer Activity Perspective

The customer activity perspective includes the features *Orders_Weekday, Orders_Weekend, Early_Morning, Morning, Afternoon, Evening, Night, dayswus and recency* since these provide us insights into the frequency, time and day of ordering that defines our customers.

Two clustering algorithms were experimented (K-Means and hierarchical clustering) to find the best clustering solution. Concerning K-Means, we utilized two criteria to pick the optimal amount of clusters for the algorithm, Inertia (also known as the elbow method) and Silhouette score. Initially these provided an optimal solution of 2 clusters, which was unsatisfactory since they only took into account the clients that made a lot of purchases and those that did less (dividing them into an imbalanced 90%/10% split of the customers). This solution provided a huge cluster imbalance and very generic clusters, not allowing us to provide ABCDEats Inc. with an effective marketing approach to boost customer engagement. To face this issue the Johnson SU transformation was applied to our data, a technique that transforms non-normalized, skewed data into a normal distribution and that has been found to be particularly effective for features with significant skewness and heavy tails [4–5], as is the case with our data. Applying this transformation provided better results, which is coherent with the fact that K-Means performs better with normally distributed data [6], with 4 clusters being used as inertia implied an optimum amount of clusters between 2 and 4, and the silhouette score showing 4 as the solution with the best score, this solution offered a more detailed view of customer engagement patterns compared to the previous 2 cluster solution.

Hierarchical clustering was also performed for both normalized and denormalized data and for both data structures, the Ward linkage method proved to be the most effective, providing higher or equal R^2 values when compared to other linkage methods for any number of clusters up to 10. Then the dendrograms for both data structures was analysed, indicated solutions with 4 and 6 clusters respectively. The solution with 4 clusters was considered better (the one with features normalized) by analysing the contingency table in Appendix 6, since the other solution provided a higher class imbalance in the clusters with some clusters having around a third of the dataset, while others have

only 171 observations (less than 1% of the data). This presents a problem for our goal of making a generalizable marketing campaign to appeal to each of our different customer segments, since it might not make sense to make a custom campaign for such a small percentage of our customers.

By comparing the best hierarchical and non-hierarchical solution we decided to choose K-Means with 4 clusters as our final customer segmentation for this perspective because even though they presented very similar final results in terms of profiling, K-Means achieves it much faster and is able to reassign observations to different clusters unlike hierarchical methods. By analysing our clusters, whose centroids are detailed in Appendix 7, we can identify distinct customer behaviour patterns:

- **Cluster 0: Late Birds -** Customers in this cluster rarely place orders in the morning, focusing their activity primarily in the afternoon and evening.
- **Cluster 1: Early Birds -** This cluster includes customers who predominantly order in the morning, with minimal activity at other times. However, occasional orders are placed during the night, possibly representing early-morning activity around 4 a.m.
- **Cluster 2: Best Customers -** These customers exhibit consistent ordering behaviour throughout the day and across all times. They have been with the company the longest and place orders far more frequently than others, making them the most valuable segment.
- **Cluster 3: Worst Customers -** This group orders infrequently, has the shortest tenure and typically places orders late in the day and on weekends, suggesting occasional app use for parties or late-night activities, weekend-focused activities.

For a visual representation of these clusters, Figure 1 provides a two-dimensional visualization created using t-SNE that shows a clear separation between the clusters.

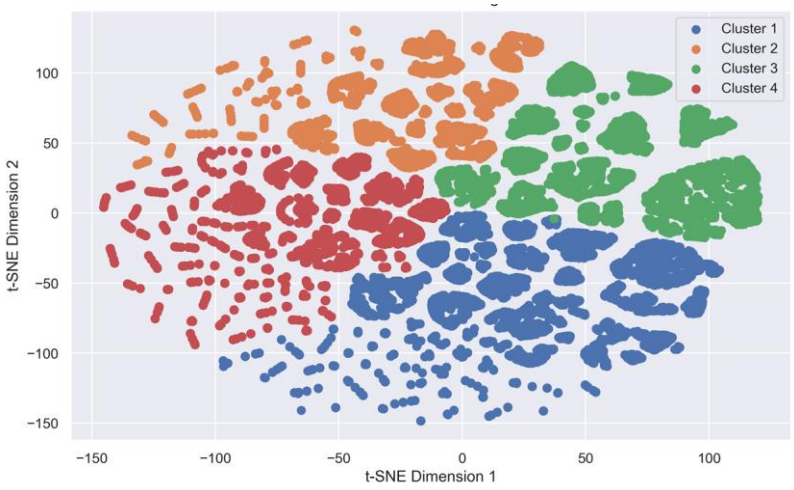


Figure 1: t-SNE Visualization of Customer Activity Segmentation

6.2. Cuisine Preferences Perspective

The cuisine preferences perspective encompassed the features *CUI_Asian*, *CUI_Cafe*, *CUI_Chinese*, *CUI_Desserts*, *CUI_Healthy*, *CUI_Japanese*, *CUI_Noodle_Dishes*, *CUI_OTHER*, *CUI_SoutheastAsia*, *CUI_Western* and *CUI_Street_Food_Beverages*, as these provide insights into customer preferences across the diverse food categories that the company offers being ideal for segmentation.

The same approach of the customer activity perspective was followed with both K-Means and Hierarchical clustering being tested, with and without feature normalization, and the conclusions were fairly the same with the clustering algorithms performing better when features are normalized, since for hierarchical clustering they provided and interpretable solution with fewer clusters and in K-Means a solution with more clusters that does not split observation by those that buy a lot and those that don't.

In our analysis, the hierarchical clustering algorithm provided a superior solution compared to K-Means. This conclusion, supported by the contingency table in Appendix 8, is based in two major factors: a) the hierarchical approach generated 5 clusters, compared to 10 from K-Means, which significantly simplify profiling by reducing the analysis from 40 to 20 clusters and b) the 10-cluster solution from K-Means resulted in minimal differentiation between some clusters, leading to overlapping customer profiles. For ABCDEats Inc., this could result in similar but slightly varied marketing campaigns targeting niche segments, increasing marketing costs and reducing overall effectiveness, which by adopting the hierarchical clustering solution allows achieving more distinct customer segments and streamline marketing efforts, optimizing both cost and impact.

Analysing our 5 clusters centroids (Appendix 9), we can identify the following distinct cuisine preferences:

- **Cluster 1: Western Cuisine Lovers -** Customers that primarily prefer Western food and tend to avoid some specific Eastern cuisine, as they rarely order Japanese or Noodle Dishes.
- **Cluster 2: Asian Street Food Enthusiasts -** Customers that favour Asian food and street food, as well as beverages.
- **Cluster 3: Japan-Western Fusion Fans -** Customers that enjoy both Japanese and Western food, and also purchase a significant amount of Asian dishes.
- **Cluster 4: Diverse Food Explorers -** Customers that mainly buy other types of food and enjoy exploring different cuisine types.
- **Cluster 5: Sushi & Beyond Lovers -** Customers that predominantly buy Japanese food, with a moderate interest in Asian cuisines.

Lastly, a 2 dimension representation using t-SNE was performed, presented in Figure 3.

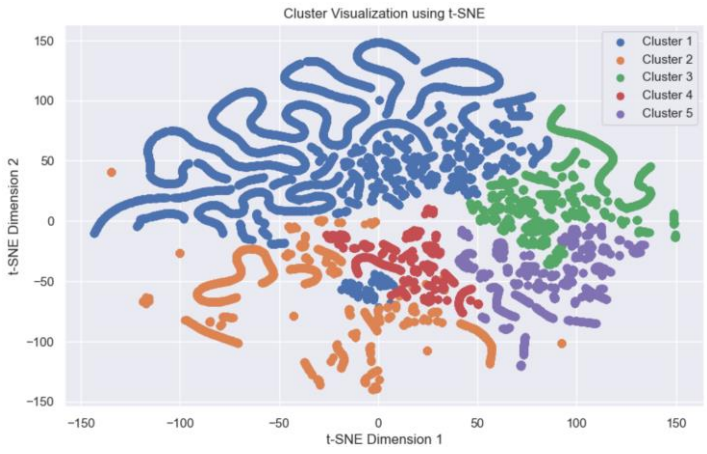


Figure 2: t-SNE Visualisation of Cuisine Preference Segmentation

6.3. Joining both perspectives

The next step in our customer segmentation initiative involved integrating both perspectives, resulting in a total of 20 clusters. However, this number of clusters poses significant challenges: a) the time-intensity of analysing such a high volume of clusters and b) the significant costs of 20 customer segments for ABCDEats Inc.'s Marketing department campaigns. To address these challenges, a hierarchical clustering algorithm to the centroids of the 20 clusters was applied to consolidate them into a more manageable and cost-effective number.

Based on the analysis of the dendrogram in Appendix 10, a solution with 6 clusters would be optimal for the final cluster merging. While a 4-cluster configuration was also considered, we opted for 6 clusters to minimize the size imbalance between them, since a smaller number of clusters would lead to disproportionately large clusters, overshadowing smaller clusters in visualisations and analyses. By selecting 6 clusters, we achieved a more balanced and practical solution that aligns with both analytical needs and marketing constraints of ABCDEats Inc. The hierarchical clustering results, detailing the final cluster assignments for each pair of customer activity and cuisine preference clusters, are summarized in Appendix 11.

7. CLUSTER ANALYSIS

With the final cluster solution that aggregates the two perspectives finally defined, it is important to answer some key questions like: In what do the clusters differ? What type of customers do they represent? And why should we care?

By using the parallel coordinates graph for the 6 clusters shown below on figure 3 we can observe visually the main differences in our clusters (in the numeric features) and interpret them. With a simple glance it is easy to determine that the most divisive features between our clusters are *vendor_count, CUI_Japanese, Order_Weekday, Morning, Afternoon and Evening*.

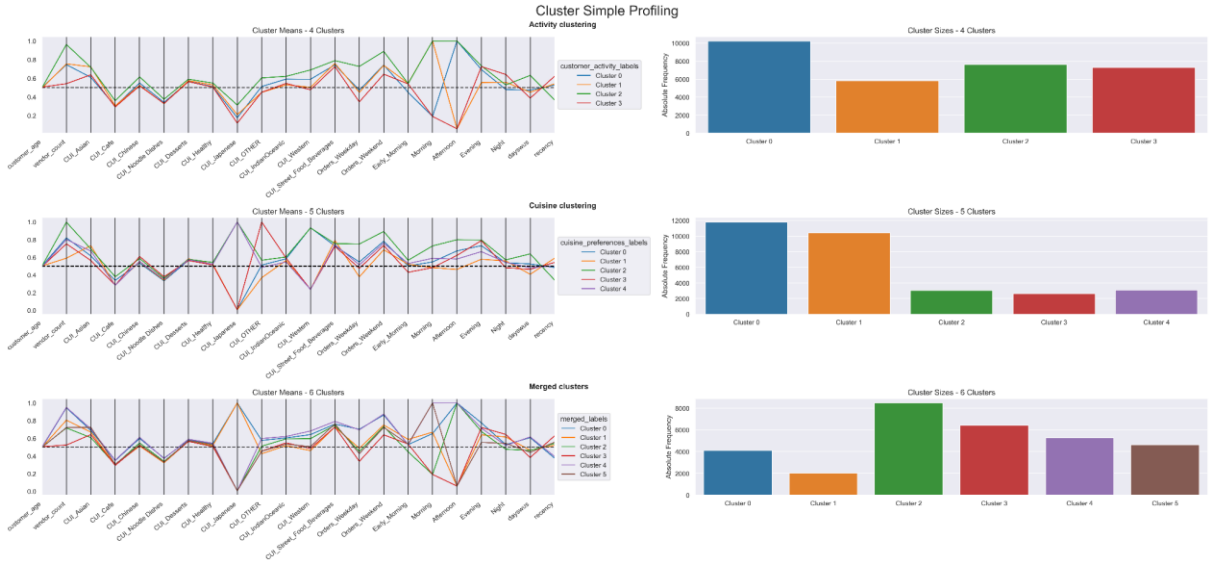


Figure 3: Comparative Cluster Profiling Across Different Clustering Perspectives

Through a more careful analysis we can observe that the best customers are located in both cluster 0 and cluster 4, being constantly above in all features, except *recency* where they prove to be our most frequent customers (by having a low value), in feature *Morning* where Cluster 0 is behind and in *CUI_Japanese* where cluster 4 is behind. This is important because these are the clients we want to foster most and incentivize to keep ordering a lot at all days, like they already do. **Cluster 4 represents the "Best Customers"** because even though they dislike Japanese food, they still order any day at any time, while **cluster 0 was labelled as "Japanese Food lovers"** since they prefer Japanese cuisine more than most other clusters of customers.

The remaining clusters all order less during the weekdays than the above-mentioned clusters. Clients from **cluster 1** are one of the few that order Japanese food but unlike cluster 0 they rarely order in the afternoon, specifically **"Japanese breakfast lovers"**. Cluster 2 and 5 clients can be seen as polar opposites in their ordering schedule behaviour, while cluster 2 customers differ from most other clusters tendencies by never ordering in the morning (probably due to being asleep or not having a habit of ordering breakfast at all) and ordering mostly in the afternoon, cluster 5 only tends to order in the morning and very rarely in the afternoon, earning **cluster 2 the name of "Lunch Enthusiasts"** and **cluster 5 the "Breakfast enthusiasts"** title.

Lastly, **cluster 3 represents our worst clients**! The ones we want to change and incentivize to order more, since they seem the closest to being lost clients. These clients order infrequently, have been with us the least, and tend to order the least of any other cluster independent of day or time.

In terms of differences in the categorical features, the clusters only seem to be significantly different in the *customer_region* feature, showing that certain behaviours are more common in certain areas than others. Region 8670 is mostly represented in cluster 1, 3 and 5 while the second and third-biggest regions, 4660 and 2360 respectively, are mostly represented in clusters 0, 2 and 4.

Figure 4 presents a 2-dimensional UMAP visualization of our final clusters. UMAP was chosen over t-SNE in this visualization for its computational efficiency and slightly better data representation, as supported by the literature [7]. The visualization shows a good cluster separation, though some imperfections remain, which could be addressed in future work.

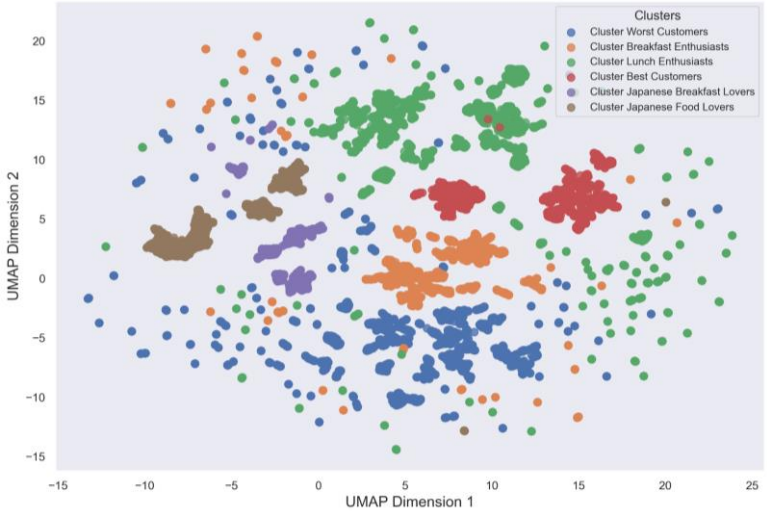


Figure 4: UMAP Representation of Final Clusters

8. MARKETING APPROACHES

The final part of this work is to propose actionable recommendations to ABCDEats Inc., derived from the customer segmentation analysis, aimed at boosting customer engagement and increasing the company's revenue streams. Based on these insights, the following recommendations are put forward:

- **Cluster 0 - Japanese Food Lovers:** Create curated Japanese meal combos with popular dishes like sushi platters, ramen bowls, bento boxes, and miso soup. Host tasting events where customers can try new or premium Japanese products, such as speciality sushi rolls, matcha desserts, or different types of sake. When a new dish reaches the company, they should promote on the website and app to keep customers updated on the latest offerings, helping them discover new favourites. For maximum impact, run promotions in the afternoon, as this is when customers are most likely to engage. However, running promotions at other times of the day can also encourage customers to make purchases at different hours.
- **Cluster 1 - Japanese Breakfast Lovers:** For these customers, marketing strategies from the previous cluster can be used, enhanced by limited-time breakfast specials, by offering dishes that are only available during breakfast hours to encourage customers to try new meals, which will make customers more willing to explore new options and increase the frequency of their orders.
- **Cluster 2 - Lunch Enthusiasts:** Create a Cuisine Roadmap Campaign where customers are encouraged to try different types of cuisine each day. Every day, they'll be prompted to order a specific cuisine type, such as Asian, Western, Japanese. When customers reach a milestone, like trying 4 different cuisine types, they'll earn discounts that can be used for lunch purchases on the app. If customers complete the entire journey and try all the cuisine types, they'll unlock larger discounts. Additionally, offer targeted lunchtime discounts to further boost order frequency, encouraging customers to make purchases during peak hours.
- **Cluster 3 - Worst Customers:** Run social media campaigns with influencers to showcase the latest updates and new offerings from the business. At the same time, send targeted notifications to customers offering significant discounts, encouraging them to make at least one more purchase. By tracking customer orders, the company can gather data to create personalized marketing campaigns, specifically targeting customers at risk of churning, which requires extra effort and a deeper understanding of customer preferences.
- **Cluster 4 - Best Customers:** Offer a loyalty card program that rewards customers with points for every purchase, that can be redeemed for cash-back or discounts. For instance, clients earns 1 point per euro spent and redeem 100 points for a 10€ discount on the next purchase. Run daily promotions at different times to incentive their continuos purchase behaviour and create cross-cuisine discounts to promote repeat and variety purchases, for instance if a customer buys Western food they get 10% off in their next purchase of Asian food.
- **Cluster 5: Breakfast Enthusiasts:** Offer discounts and breakfast combos during breakfast hours. For example, customers can get a coffee, croissant, and orange juice for just €2.50, or enjoy 15% off their breakfast purchase. Run a reward campaign where customers earn 1 ticket for every breakfast purchase, after collecting 12 tickets their next breakfast is free. This encourages more frequent visits during the morning and rewards loyal customers, increasing both satisfaction and revenue.

9. CONCLUSION

In conclusion, this study utilized advanced data science techniques to segment ABCDEats Inc.'s customer base into six distinct clusters based on activity and cuisine preferences. Through a structured pre-processing pipeline and exploratory data analysis (EDA), key insights were uncovered, such as metadata inconsistencies and customer behaviour patterns, like the preference for afternoon orders and varying affinities for Japanese cuisine. Both k-means and hierarchical clustering showed strengths in different perspectives, with no clear winner, but feature normalization significantly improved cluster interpretability, aligning with literature suggesting its importance in clustering performance.

Although there were some minor imperfections in cluster separation during profiling, so future work could explore techniques like SOM or DBSCAN for potentially clearer separation. Ultimately, this study met its objective of proposing actionable recommendations, such as targeted cuisine campaigns for lunch enthusiasts and reward campaigns offering free breakfasts for breakfast enthusiasts, that are aimed at boosting customer engagement and revenue streams of ABCDEats Inc.

10. REFERENCES

[1] *Online Food Delivery—Portugal | Statista Market Forecast*. (n.d.). Statista. Retrieved 2 January 2025, from https://www.statista.com/outlook/emo/online-food-delivery/portugal

[2] Lin, R. (2021). *The Effects of Covid-19 on the Online Food Delivery Industry*. 203–207. https://doi.org/10.2991/assehr.k.211209.033

[3] Art. 8 GDPR – Conditions applicable to child's consent in relation to information society services. (n.d.). *General Data Protection Regulation (GDPR)*. Retrieved 30 December 2024, from https://gdpr-info.eu/art-8-gdpr/

[4] Bean, A. T. (2017). Transformations and Bayesian Estimation of Skewed and Heavy-Tailed Densities [The Ohio State University]. https://etd.ohiolink.edu/acprod/odb_etd/etd/r/1501/10?clear=10&p10_accession_num=osu1503015935192212

[5] Hayashi, H., Kurita, Y., & Tsuji, T. (2015). A non-Gaussian approach for biosignal classification based on the Johnson SU translation system. 2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA), 115–120. https://doi.org/10.1109/IWCIA.2015.7449473

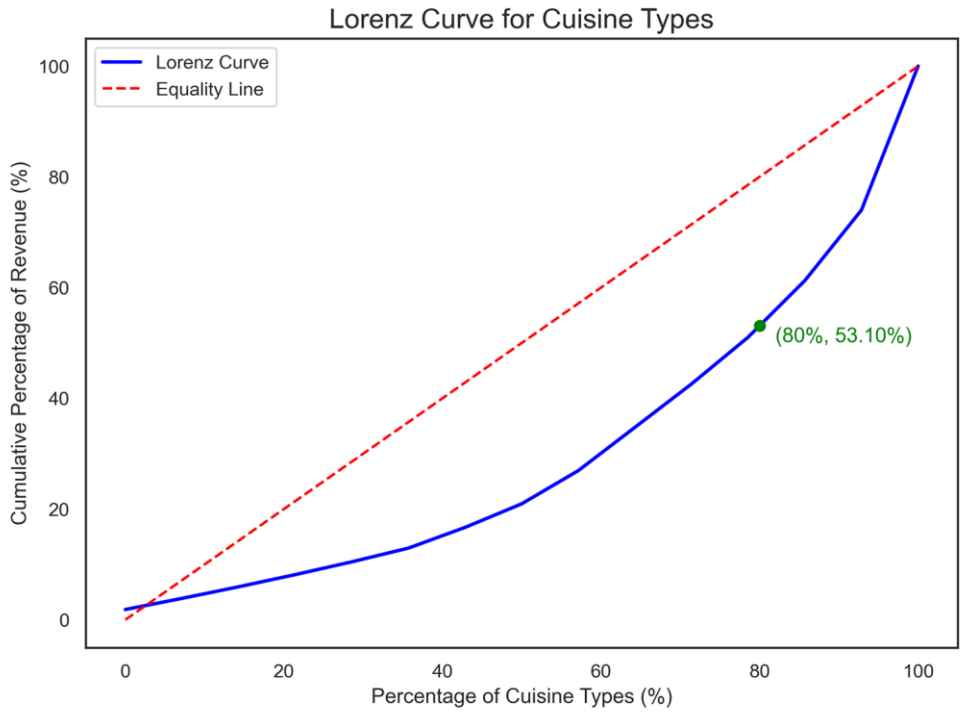
[6] Wu, J., Xiong, H., & Chen, J. (2009). Adapting the right measures for K-means clustering. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 877–886. https://doi.org/10.1145/1557019.1557115

[7] Pal, K., & Sharma, M. (2020). Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space. 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 1106–1110. https://doi.org/10.1109/I-SMAC49090.2020.9243502

11. APPENDIX

Appendix 1: Additional Features Description

Feature	Description
recency	Number of days since the customer's last purchase.
dayswus	Number of days since the customer's first purchase.
D_Orders	Total number of orders across all day-of-week (DOW_) columns.
H_Orders	Total number of orders across all hour-based (HR_) columns.
Orders_Weekday	Total number of orders placed on weekdays (Monday through Friday).
Orders_Weekend	Total number of orders placed on weekends (Saturday and Sunday).
Early Morning	Total number of orders placed between 5:00 a.m. and 8:59 a.m.
Morning	Total number of orders placed between 9:00 a.m. and 11:59 a.m.
Afternoon	Total number of orders placed between 12:00 p.m. and 4:59 p.m.
Evening	Total number of orders placed between 5:00 p.m. and 8:59 p.m.
Night	Total number of orders placed between 9:00 p.m. and 4:59 a.m.



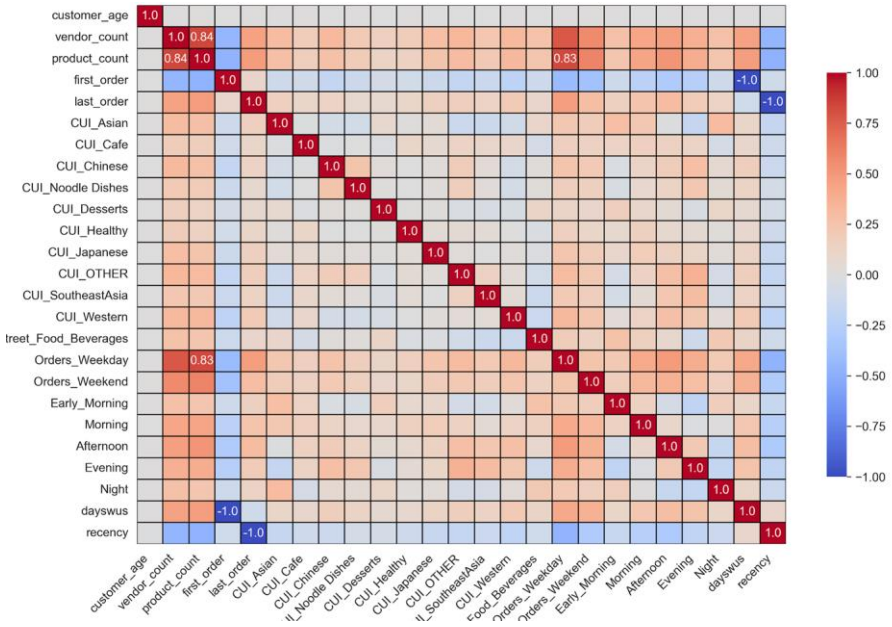
Appendix 2: Lorenz Curve of Revenue Distribution by Cuisine Type

Appendix 3: Outlier Removal in the cuisine columns

Feature	Rule to removal
CUI_Asian	Remove all instances where values exceed 35
CUI_Cafe	Remove all instances where values exceed 50
CUI_Chinese	Remove all instances where values exceed 80
CUI_Desserts	Remove all instances where values exceed 40
CUI_Healthly	Remove all instances where values exceed 40
CUI_Japanese	Remove all instances where values exceed 70
CUI_Noodle Dishes	Remove all instances where values exceed 60
CUI_IndianOceanic	Remove all instances where values exceed 35
CUI_Western	Remove all instances where values exceed 25
CUI_Street_Food_Beverages	Remove all instances where values exceed 25

Appendix 4: Outlier Removal of the remaining columns

Feature	Rule to removal
Early_Morning	Remove all instances where values exceed 60
Morning	Remove all instances where values exceed 20
Evening	Remove all instances where values exceed 22
Night	Remove all instances where values exceed 20



Appendix 5: Correlation Matrix of Numerical Features

Appendix 6: Contingency Table Comparing Hierarchical Clustering with and without Feature Normalization

	0	*1*	*2*	*3*
0	*190*	*206*	*501*	*237*
1	*2565*	*4846*	*4158*	*2213*
2	*1*	*5*	*163*	*2*
3	*2949*	*2837*	*747*	*1981*
4	*14*	*226*	*1235*	*108*
5	*1549*	*2105*	*802*	*1291*

Appendix 7: Customer Activity Segmentation Cluster Centroids.

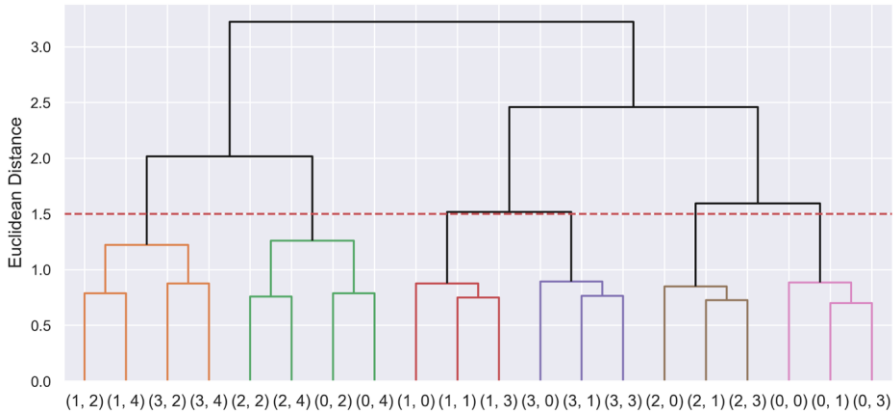
	0	1	2	3
Orders_Weekday	0.47	0.45	0.73	0.35
Orders_Weekend	0.74	0.74	0.89	0.64
Early_Morning	0.45	0.55	0.54	0.54
Morning	0.19	1	1	0.19
Afternoon	1	0.06	1	0.06
Evening	0.69	0.56	0.73	0.73
Night	0.48	0.56	0.53	0.64
dayswus	0.47	0.45	0.63	0.39
recency	0.53	0.54	0.37	0.62

Appendix 8: Contingency Table: Best Hierarchical vs. K-Means Solution

	0	1	2	3	4
0	42	5201	0	0	0
1	6318	0	0	0	0
2	2535	0	0	0	0
3	0	0	55	0	2736
4	0	0	2029	0	0
5	154	3913	0	0	0
6	38	1257	0	0	0
7	0	0	961	0	364
8	123	17	0	2615	0
9	2573	0	0	0	0

Appendix 9: Cuisine Preferences Segmentation Cluster Centroids

	0	1	2	3	4
CUI_Asian	0.6165	0.7322	0.6941	0.5641	0.6703
CUI_Cafe	0.3390	0.2836	0.3796	0.2836	0.2836
CUI_Chinese	0.5353	0.5509	0.5867	0.6094	0.5535
CUI_Desserts	0.5659	0.5823	0.5777	0.5635	0.5727
CUI_Healthy	0.5190	0.5222	0.5420	0.5133	0.5230
CUI_Japanese	0.0055	0.0055	1	0.0055	1
CUI_Noodle Dishes	0.3333	0.3404	0.3632	0.3797	0.3506
CUI_Other	0.5124	0.3688	0.5673	1	0.4859
CUI_IndianOceanic	0.5832	0.5644	0.6033	0.5916	0.5476
CUI_Western	0.9350	0.2361	0.9331	0.2361	0.2361
CUI_Street_Food_Beverages	0.7288	0.7918	0.7553	0.7241	0.7438



Appendix 10: Dendrogram of the Hierarchical Merging of Cluster Centroids

Appendix 11: Cluster Mapping Across Customer Activity, Preferences, and Merged Labels

Customer Activity Label	*Customer Preferences Label*	*Merged Clusters Label*
0	*0, 1, 3*	*2*
0	*2, 4*	*0*
1	*0, 1, 3*	*5*
1	*2, 4*	*1*
2	*0, 1, 3*	*4*
2	*2, 4*	*2*
3	*0, 1, 3*	*3*
3	*2, 4*	*1*