DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc.
Bonus Project

Group 05

Daniel Caridade, 20211588

Gonçalo Teles, 20211684

Gonçalo Peres, 20211625

João Venichand, 20211644

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

Customer segmentation is a critical component of modern business strategies, enabling companies to tailor their services, optimize marketing efforts, and enhance customer satisfaction. For ABCDEats Inc., a customer segmentation initiative was previously developed to group customers based on shared behaviours and characteristics. However, while the clustering solution provided valuable insights, there remained a significant gap in delivering a tool that allowed stakeholders to fully leverage these insights through an interactive and exploratory platform.

To address this gap, our team developed an interactive application designed to enhance the exploration and visualization of customer segments. Built using Python, the application integrates advanced libraries such as **Streamlit** and **Plotly** to create an intuitive and user-friendly interface. The platform incorporates dimensionality reduction techniques such as UMAP and t-SNE, enabling users to visualize complex clustering results in a simplified and interpretable format.

The application offers a range of functionalities, including the ability to explore, visualize, and compare clusters dynamically. Users can apply filters, such as customer region, payment method, and promotions used in the last purchase to draw insights that were not achieved in previous work and that facilitate data-driven decision-making. Finally, all this work (including EDA, customer segmentation and interactive visualization platform) are publicly available at GitHub[1], so users around the globe can take inspiration in this work and develop their own customer segmentation initiatives and visualization applications.

2. APPLICATION FEATURES

2.1. HOME PAGE

The Home Page (Visual 1) serves as an introductory interface of the application, giving us a general view of the dataset and providing the first contact with the Navigation Menu. It includes:

- **Dataset Overview:** Users are presented with the 5 first rows of the dataset to give an idea about the customer data on which the cluster analysis was performed.
- **Navigation Menu:** A side menu available on all pages of the application that allows access to other sections, such as *Cluster Comparison*, *Cluster Comparison* or *Filters*.

2.2. CLUSTER VISUALIZATION

The **Cluster Visualization** tab (Visuals 2 and 3) is the platform's main feature, displaying clustering results for the **Customer Activity** and **Cuisine Preferences** perspectives. Using **UMAP** and **t-SNE**, users can explore how customer segments are distributed in space, with the ability to zoom in, zoom out, and filter clusters. It is also important to highlight the distinct user interactions for the two types of visualizations in this feature. In the **UMAP** visualization, clusters are filtered by deselecting them from

[1] https://github.com/DanieLLL5/Data-Mining-Project-Group5.git

a dropdown menu located above the graph. In contrast, in the **t-SNE** visualization, clusters are selected by clicking their corresponding labels in the legend on the right-hand side of the visual.

2.3. CLUSTER COMPARISON

The Cluster Comparison page (Visuals 4, 5 and 6) allows for a detailed side-by-side cluster analysis, aiding the user in getting a better understanding of the distinct characteristics of each customer segment. Key features include:

- **Box Plot by feature:** Includes a dropdown menu allowing users to choose a feature for analysis. Once selected, a visual comparison of the clusters using box plots, highlighting the distribution, median, and outliers within each cluster is generated enabling users to easily observe and compare how different clusters behave based on the selected feature.
- **Cluster Centroids:** A table that displays the centroids of each cluster, presenting the mean values of all attributes within each cluster that allows users to identify central tendencies of various customer segments and relevant differences in mean values.
- **Clustering Profiling:** Includes bar charts and parallel coordinate plots that highlight the main differences between clusters for the two perspectives analysed, as well as for the profiled clustering solution. These visualizations emphasize the key characteristics and frequencies of each cluster, enabling a concise and straightforward comparison across multiple features. Unlike the other interactive elements of the platform, this visual is static due to conflicts with other functionalities within this feature.

2.4. FILTERS

This section (Visual 7) allows users to apply different filters based on customer attributes. The functionalities available are:

- **Filters:** Users can filter the data based on customers' regions, promotional history and payment method for a detailed cluster examination.
- **Real-Time Data Update:** When filters are applied, both the descriptive statistics table and the filtered data table contained in this feature are instantly updated, providing a new perspective into the different clusters based on the filtering done by the user.

2.5. ABOUT US

The last building block of this platform is an About Us section (Visual 8), where our team is introduced with our names and student numbers. There is also a brief remark highlighting our enthusiasm for leveraging cutting-edge technology and data exploration.

3. TECHNOLOGY STACK

The application was built in Python, leveraging various powerful libraries to facilitate the interactive exploration of customer segments. These libraries include:

- **Streamlit:** This library simplifies the creation of interactive web interfaces, allowing real-time updates based on user inputs. It enables the deployment of our interactive platform on a web

page with just a single command. Streamlit was chosen over Bokeh due to its ease of use, as it allows the app to be developed directly as a Python script without the need to manage any APIs. While Streamlit offers fewer customization options compared to Bokeh, its straightforward scripting approach made it a more practical choice for our needs.

- **Plotly**: A library that generates interactive visualizations for exploring clusters, making use of the UMAP and t-SNE visuals to enhance the analytical capabilities of the platform. The library was chosen for its ability to create dynamic visualizations that make it easier to explore data relationships and cluster patterns, providing deeper insights and enhanced user experience.

- **Data Manipulation and Filtering :** Libraries such as Pandas and NumPy were utilized to ensure the dataset is properly formatted for analysis and visualization.

- **Visualization libraries:** Libraries such as Matplotlib, Seaborn, UMAP, and t-SNE were utilized to effectively visualize the clustering solutions, providing insights from two different perspectives and showcasing the final clustering results.

4. INSIGHTS AND RESULTS

By virtue of this application, it is easier to draw some insights that were previously not addressed in customer segmentation. For instance, using the *Cluster Comparison* feature, we can identify that th **Best Customer** and the **Japanese Food Lovers** clusters interact with significantly more vendors. This suggests that these customers prefer a diverse range of vendors rather than staying exclusive to one, which can be used in future marketing campaigns. Additionally, the analysis reveals that our **Worst Customers** tend to be older on average, which aligns with their observed behavior patterns. The *Cluster Visualization* feature allows us a more side to side analysis of the clusters in all the perspectives and final profiling which allows us to combine how different features and cluster algorithms used provided us with very different shapes of the clusters. Overall, the application provided more granular insights, improving our understanding of the segmentation and helping us refine the marketing approach.

5. CONCLUSION

This interactive application empowers ABCDEats Inc. to explore and engage with customer segmentation analysis more effectively. By leveraging dimensionality reduction techniques such as UMAP and t-SNE, combined with powerful Python libraries like Streamlit and Plotly, the app provides clear, insightful visualizations of customer clusters. These capabilities enhance the company's ability to make data-driven decisions, refine marketing strategies, and improve customer interactions.

Beyond its current functionality, this app sets the foundation for more advanced analytical capabilities in ABCDEats Inc. For example, integrating it with a continuously updated database would allow new customers to be automatically assigned to clusters and enable dynamic re-computation of clusters as needed. Such developments would greatly enhance analytical capabilities, making the application not just a tool for immediate insights but also a foundation for future innovations in data-driven decision-making.

6. ANNEX

6.1. ACCESS GUIDE

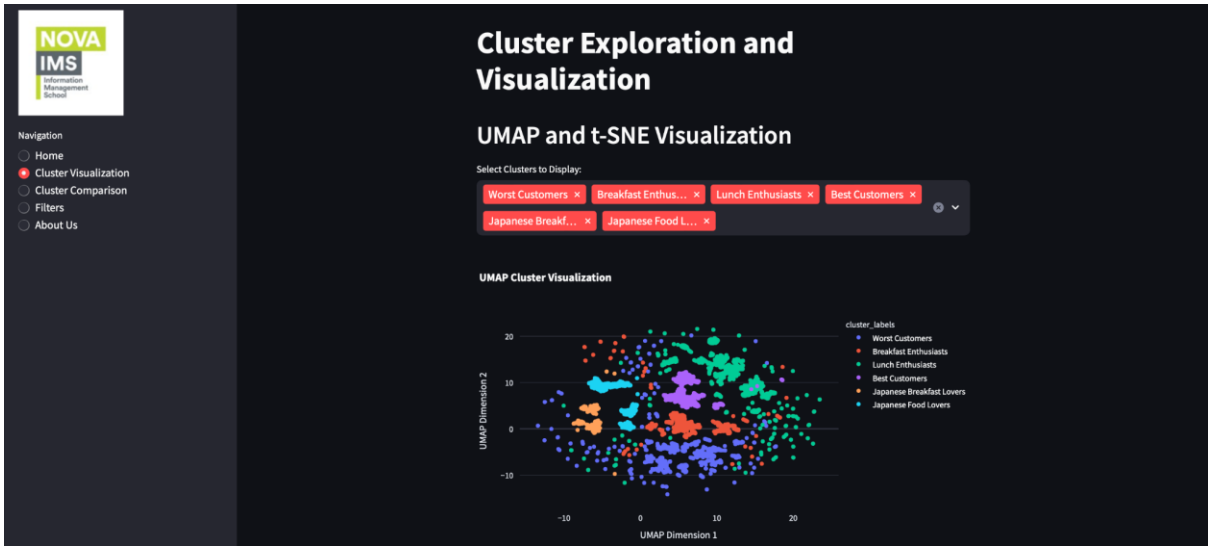
Accessing the application is simple and involves just a few steps, more specifically:

1. Install the Streamlit library by running the following command: (*pip install streamlit)*
2. Open the Command Prompt and run:
   cd path\to\where\app\is\stored
   streamlit run DM2425_BonusProject_05.py
3. If you are using a Mac, Open the Terminal and run:
   cd path\to\where\app\is\stored
   streamlit run DM2425_BonusProject_05.py

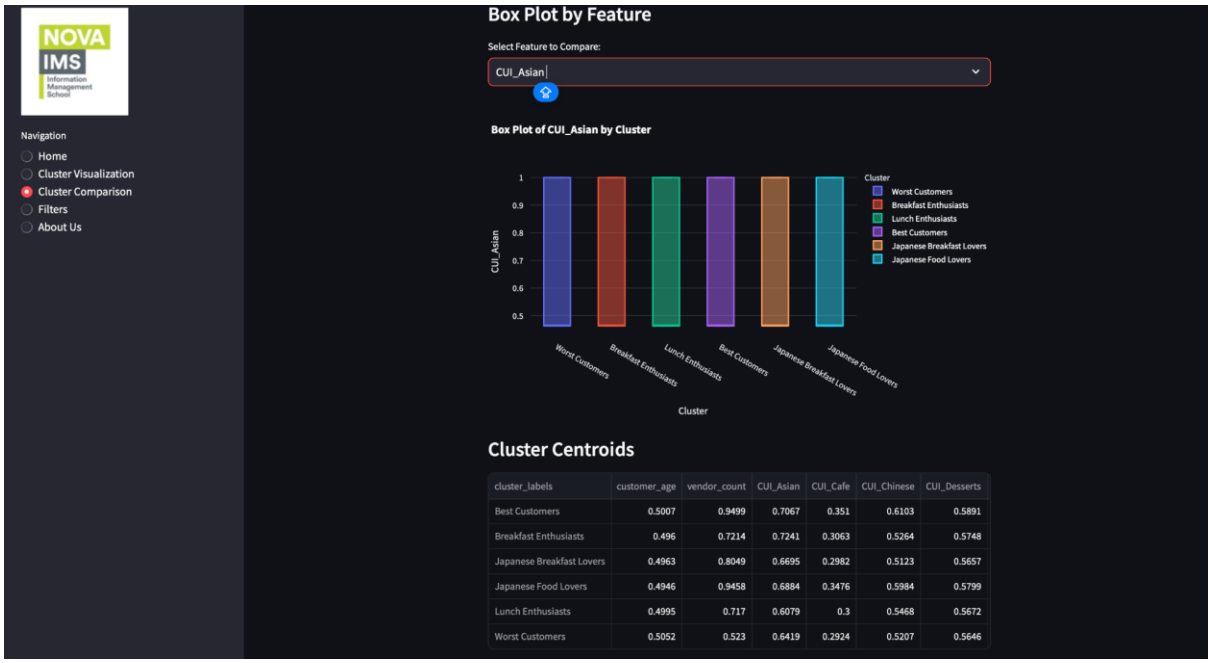
Once these commands are executed, a new browser tab will automatically open, displaying the interface of the application, regardless of whether you're using Windows or Mac.

6.2. VISUALS

Visual 1: Application Main Page

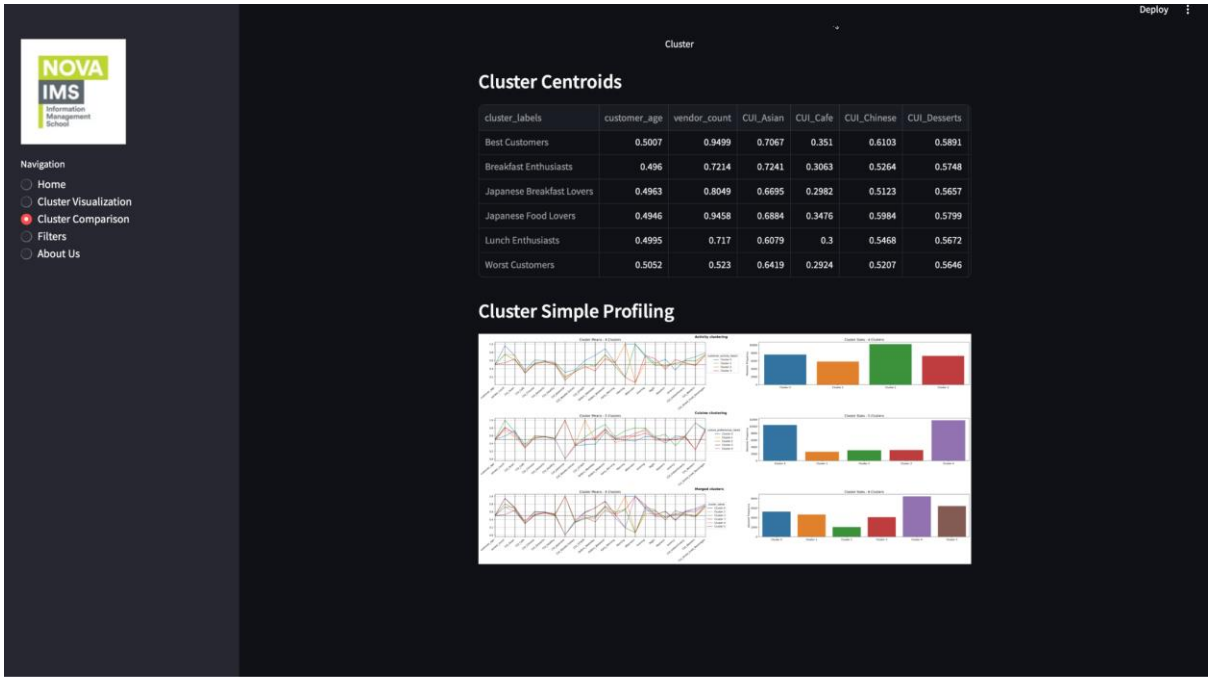


Visuals 2 and 3: Cluster Visualization Tab

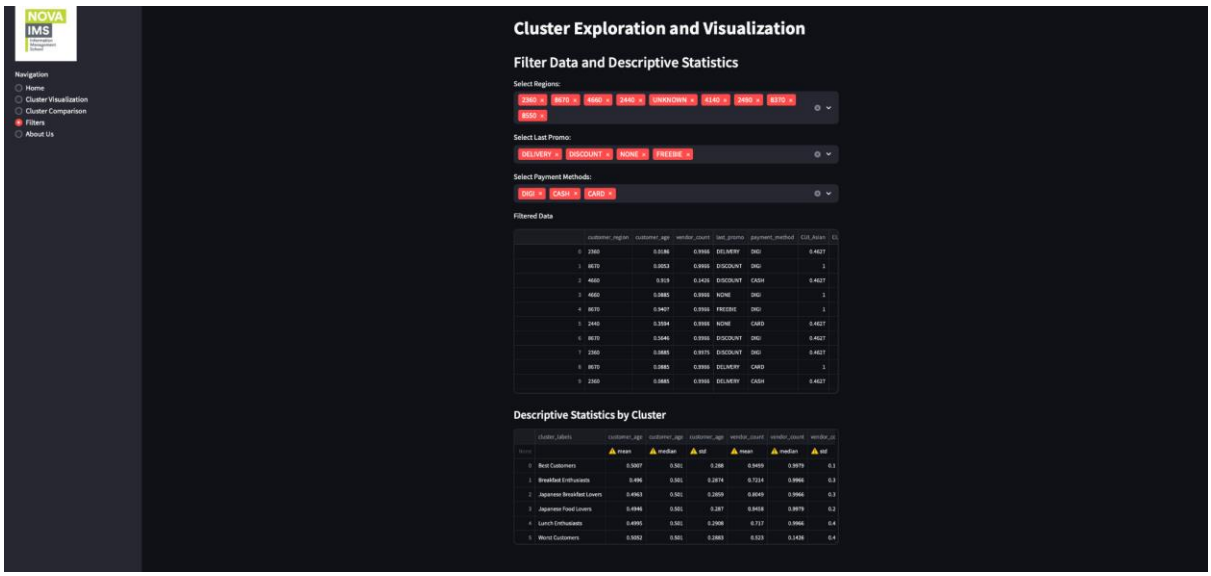


Cluster Centroids

cluster_labels	customer_age	vendor_count	CUI_Asian	CUI_Cafe	CUI_Chinese	CUI_Desserts
Best Customers	0.5007	0.9499	0.7067	0.351	0.6103	0.5891
Breakfast Enthusiasts	0.496	0.7214	0.7241	0.3063	0.5264	0.5748
Japanese Breakfast Lovers	0.4963	0.8049	0.6695	0.2982	0.5123	0.5657
Japanese Food Lovers	0.4946	0.9458	0.6884	0.3476	0.5984	0.5799
Lunch Enthusiasts	0.4995	0.717	0.6079	0.3	0.5468	0.5672
Worst Customers	0.5052	0.523	0.6419	0.2924	0.5207	0.5646



Visuals 4, 5 and 6: Cluster Comparison Tab



Visual 7: Filters Tab

Visual 8: About Us Tab