DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc. (Optional Part)

Group 37

André Silvestre, 20240502

Filipa Pereira, 20240509

Umeima Mahomed, 20240543

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

This report delineates the **deployment** phase of the **CRISP-DM methodology** (**Annex A.**) that has been followed, describing the application and its suite of interactive data exploration as well as customer segmentation tools for *ABCDEats*. Its main objective is to offer an interactive and easy-to-use interface for stakeholders to investigate a synthetic customer database, over a 3-month period, to analyse customer behaviours, identifying useful patterns for strategic decisions. Through the visualization of such complex data, this tool provides information that helps to uncover the different customer segments, which in turn can aid to the development of a targeted marketing and increase of business.

The goal of the application is to facilitate data-driven decision-making, by providing an affordance to the users to actively dive into EDA results and to evaluate custom segmentation efficacy. Through doing so, it presents the use of data mining techniques to address practical business problems.

2. IMPLEMENTATION

This proposed *front-end* interactive dashboard was built using Python programming language and incorporating the *Streamlit* library [5] which enables easy development of web apps based on data science workflows.

Among the challenges faced was the requirement to transform standardized data back to its original scale to make exploration and analysis more comprehensible and user-friendly. The data had been standardized for clustering purposes, which changed the values of the metrics, e.g., the age variable (like others) started taking negative and even decimal values, which would severely impede the goals we had with the app. As a result, the scaler object that was used to preprocess the data was later used to revert all the standardized metrics back to their original scale. In addition to this, to align with the original metadata, several variables were converted to integers. This ensured the dashboard presented data in its simplest and most interpretable form, making variable values easier to understand.

The application allows users to **filter** their target customers by their desired characteristics (demographics, orders, purchasing patterns, etc.). This input is controlled to ensure the resulting selection satisfies logical conditions. For instance, if users select a first order date that is later than the last order date, an error will be issued. In addition, all parameters available for adjustment are offered as pre-defined values, preventing users from choosing unwanted or illogical values. These filters are, however, only applied once the user chooses to and presses the button, allowing the user to decide when to apply it.

This interactive dashboard was deployed for public use on *Streamlit's* cloud platform (**Annex A**). Although the free tier of *Streamlit* offers a simple and quick deployment option, resource limitations pose a potential risk of application unavailability or performance degradation with heavy traffic. It is important to note, however, that during our testing, the application consistently performed well and did not experience any instances of failure or reduced response time. However, the positive testing does not dismiss that such problems may occur under other load conditions. Regardless, the app remains fully functional when run locally.

3. MAIN FEATURES

The application is structured into 2 main sections, accessible via a tabbed interface: "**EDA"** (Exploratory Data Analysis) and **"Final Customer Segments"** (**Figure B1**). The key functionalities include:

- **Interactive Data Filtering:** Users can apply multiple filters to the dataset via a sidebar, allowing for targeted analysis based on cuisine preferences, day and time of order, region, age group, payment method, vendor count, product count, chain usage, and last promo.
- **Exploratory Data Analysis (EDA):** The "EDA" tab focuses on providing insights from the dataset after the application of pre-processing techniques and before clustering. The analysis leverages a variety of visual aids, including line graphs, histograms, box graphs, bar graphs, scatter graphs, sunburst diagrams and assorted tables. As a result, users can visualize the entire dataset and more easily explore the relationships between all the components.
- **Customer Segmentation Analysis:** The application helps in understanding and analysing the options of customer segmentation available in a dataset by allowing users to visualize plots of several acquaintance categories. According to the "Final Customer Segments", each of these groups has unique characteristics and behaviours that can be pictured and help elaborate on the concept of segmentation. In addition, the application allows an individual to restrict the types of customer segments displayed to them.
- **Interactive Visualizations:** A wide array of visualizations using *Plotly* [6] is implemented to effectively present complex data in a digestible format, including interactive chart controls and more information about a data point by moving their mouse cursor over it, to facilitate assessment.

4. RESULTS

In the "**EDA**" tab, among various plots presented, we can immediately see on the top the time series displaying the daily trends of first and last orders, and histograms of several numerical variables (*customer age*, *vendor count*, *product count*, *total amount spent*), consistent with the analysis we presented in **Part 1** of the project.[1] Following this, we have an analysis of several variables by region using box plots, bar charts, and scatter plots. From the scatter plot (**Figure B2**), we can observe that consumers in *Region 8* tend to spend more than those in *Region 2*, although those in *Region 2* make more orders[2].

Further along, we find an analysis of the last promotion and payment method used, which we also analysed in **Part 1**. The chart (**Figure B3**) reveals that *Region 2* has the highest percentage of consumers who prefer to eat at chain restaurants. This preference is measured by the proportion of people within this region that are above 80% when calculating *chain_count/order_count*. The remaining regions exhibiting similar behaviour in this regard. Next, we find a scatter plot (**Figure B4**) assessing each consumer's most spent cuisine, along with their average amount spent and number of orders. Generally, it's noticeable that consumers who make fewer orders tend to have a higher average spend, regardless of their preferred cuisine. Also, hovering over the data points shows that the more orders a customer makes, the higher their total number of food types ordered is, as would be expected. For lower order counts, individuals who spend proportionally more on *Street Food* and *Snacks* have a

[1] Some additional preprocessing was added compared to **Part 1**, but the key insights remain the same.

[2] Due to the overlaid nature of some scatter plots, visual interpretations of categories may be slightly skewed, particularly given the dataset's distribution. For these plots, we recommend focusing on specific groups rather than making overall comparisons.

higher average amount spent. Conversely, for higher order counts (10 or more orders), those who spend proportionally more on *Asian* cuisine tend to have the highest average spend.

Finally, we arrive at the sunburst chart (**Figure B5**), which illustrates the most frequent days and times customers place orders. For instance, we can filter for *Asian* cuisine, meaning only consumers who spent some money on *Asian* cuisine are considered. We can then observe that both for *Monday* and *Sunday*, the most common order hour shifts from *11 am* to *3 am* – a valuable piece of information, particularly for *ABCDEats* and *Asian* cuisine restaurants. More of these essential insights can be discovered by interacting with the filters.

Moving to the "**Final Customer Segments**" tab, we can see some analyses like those we've seen so far, but this time by cluster, which were created in **Part 2** of the project, and other entirely new analyses. In this tab, the user can select and analyse specific clusters of interest using a filter. Right at the beginning, it's possible to compare the number of consumers in each cluster and observe a scatter plot between total amount spent and number of orders.

Moving a bit further, we can explore the segment profiles with the help of bar charts showing the average *number of orders*, *chain count*, *product count*, among others. Here, for example, to analyse the clusters for *Region 2*, we can apply a filter (**Figure B6**) and it's possible to see that the percentage of consumers in clusters 3 and 4 increased significantly (which was expected given the cluster analysis done in **Part 2**) and increased slightly in cluster 1, decreasing in the remaining ones. Even so, the average number of orders made by consumers in cluster 5 is significantly higher after filtering for *Region 2*. The few consumers who remain in this cluster like to place many orders and, consequently, the *average value of chain count* is also higher. On the other hand, regarding the *average amount spent per order*, it decreased significantly in cluster 5 and 2, causing clusters 3 and 4 not to have average spending values on *ABCDEats* lower than the others as happened without the filter. And, in fact, in terms of *absolute values spent*, in *Region 2* it is cluster 3 that spends the most.

When scrolling further, the sunburst chart (**Figure B7**) provides an in-depth view of the composition of each segment in terms of customer *region*, *last promo type* and *payment method*, which helps visualize the influence of these aspects on customer behaviour within each group. We can see that regardless of the cluster they belong to, consumers who had a promo in their last purchase made proportionally far more payments with card. Finally, at the bottom of this section, we have a funnel chart to visualize the distribution of most spent cuisines within each segment, where it stands out that although *Asian* cuisine is where most consumers spend their money, if we look by cluster (**Figure B8**), we can see that consumers in clusters 3 and 4 actually spend more of their money on *Other* cuisine. [3]

5. CONCLUSION

This application presents itself as a platform that allows for effective and individually tailored decision-making, by enabling stakeholders to analyse customer data, discover relevant trends and learn about the segment-specific behaviours. It includes the ability to perform interactive filtering and answering concrete business queries such as*: "Which customer segments prefer specific cuisines?", "How do regional preferences impact ordering patterns?", "What are the most common payment methods across different customer groups?"*, and *"Which days and times are most popular for orders in each cluster?".* Therefore, enabling *ABCDEats* to boost customer satisfaction, loyalty and retention.

[3] Note that cluster numbering in this application (starting from 1) differs from that in **Part 2** of the project (starting from 0).

BIBLIOGRAPHICAL REFERENCES

[1] Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly.

[2] Waqar, M. (2023, October 26). *Unlocking CRISP-DM: Your Path to Data Science Success*. Medium. https://medium.com/@mwaqarbatalvi/mastering-the-crisp-dm-framework-your-path-to-successful-data-science-projects-56f15d6f4c54

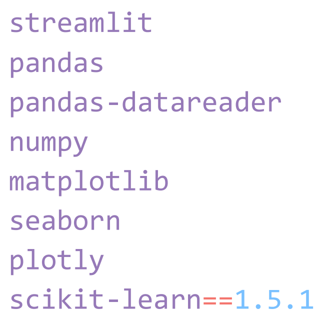
[3] Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly.

[4] Waqar, M. (2023, October 26). *Unlocking CRISP-DM: Your Path to Data Science Success*. Medium. https://medium.com/@mwaqarbatalvi/mastering-the-crisp-dm-framework-your-path-to-successful-data-science-projects-56f15d6f4c54

[5] Streamlit. (2024). *Streamlit — The fastest way to create data apps*. Www.streamlit.io. https://www.streamlit.io/

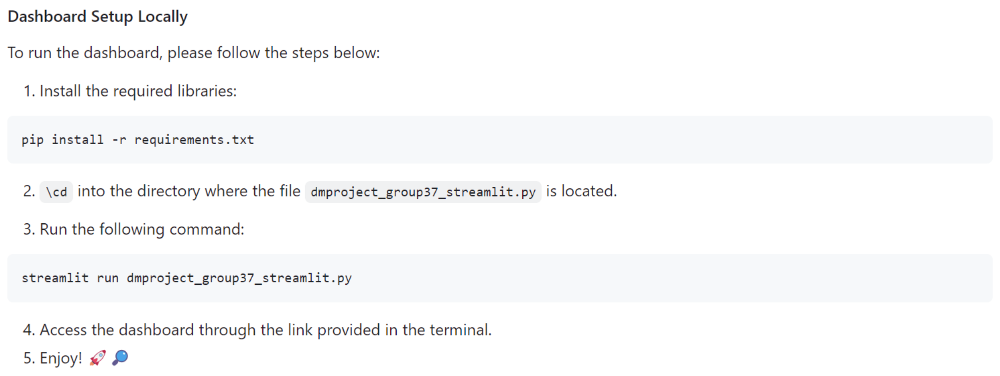
[6] Plotly. (2023). *Plotly Python Graphing Library*. Plotly.com. https://plotly.com/python/

APPENDIX A. IMPLEMENTATION DETAILS [4] [5]



```
streamlit
pandas
pandas-datareader
numpy
matplotlib
seaborn
plotly
scikit-learn==1.5.1
```

Figure A1 - Python libraries used (same as *requirements.txt*).



Dashboard Setup Locally

To run the dashboard, please follow the steps below:

1. Install the required libraries:

```
pip install -r requirements.txt
```

2. `\cd` into the directory where the file `dmproject_group37_streamlit.py` is located.
3. Run the following command:

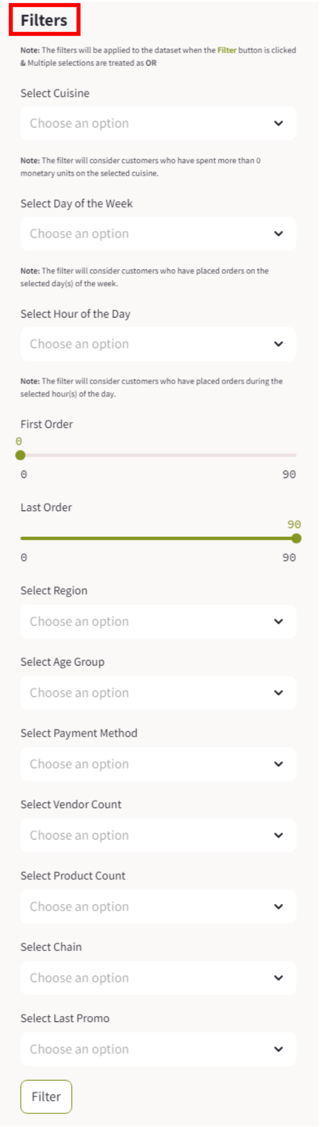
```
streamlit run dmproject_group37_streamlit.py
```

4. Access the dashboard through the link provided in the terminal.
5. Enjoy! 🚀 🔎

Figure A2 - Instructions for setup app locally *(README.md)*.

[4] **GitHub Web Application Repository:** https://github.com/Silvestre17/DM_Dashboard_Group37
[5] **Streamlit Application Link:** https://dm-project-abcdeats-group37.streamlit.app/

APPENDIX B. RESULTS



Filters

Note: The filters will be applied to the dataset when the **Filter** button is clicked & Multiple selections are treated as **OR**

Select Cuisine

Choose an option

Note: The filter will consider customers who have spent more than 0 monetary units on the selected cuisine.

Select Day of the Week

Choose an option

Note: The filter will consider customers who have placed orders on the selected day(s) of the week.

Select Hour of the Day

Choose an option

Note: The filter will consider customers who have placed orders during the selected hour(s) of the day.

First Order

Last Order

Select Region

Choose an option

Select Age Group

Choose an option

Select Payment Method

Choose an option

Select Vendor Count

Choose an option

Select Product Count

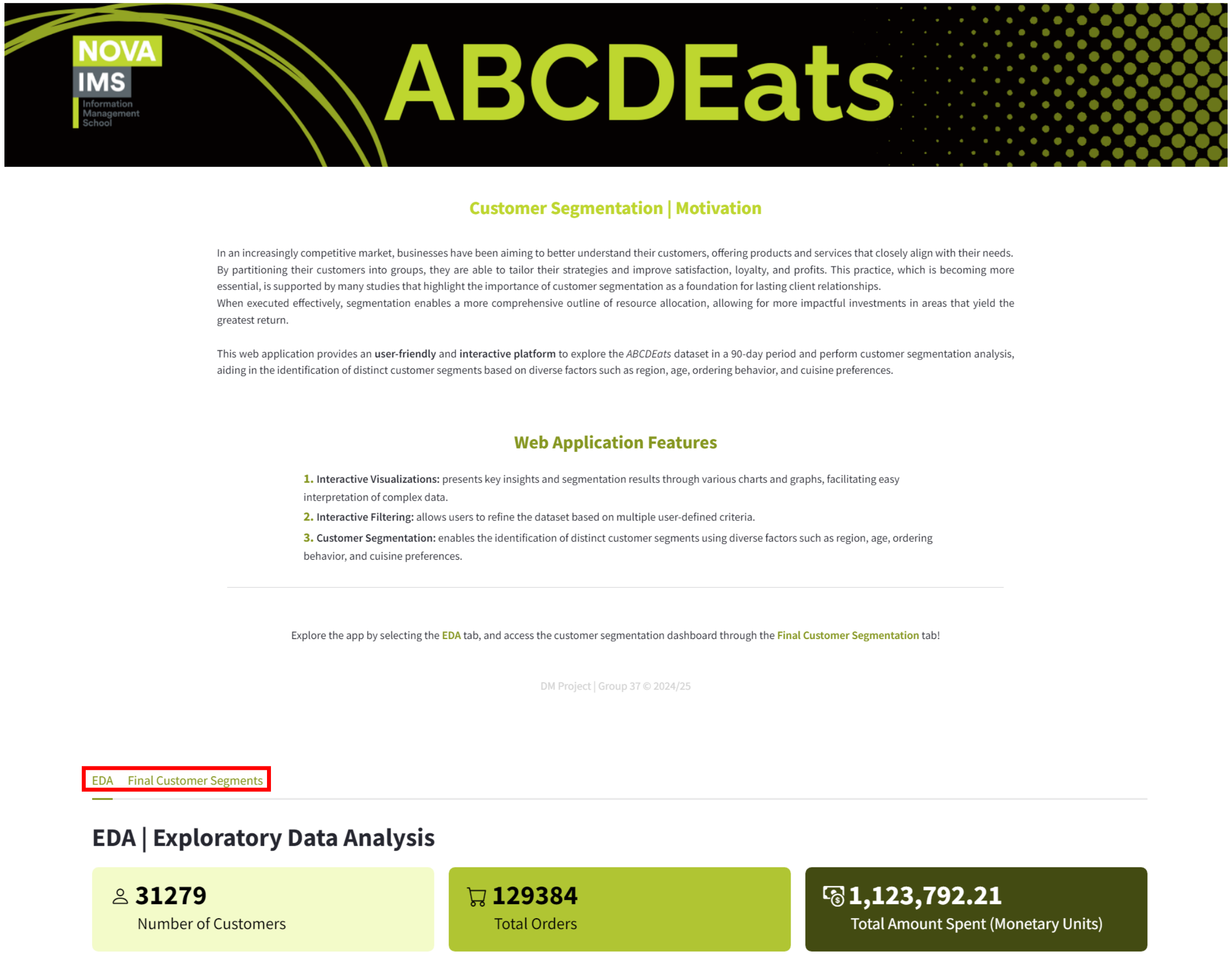
Choose an option

Select Chain

Choose an option

Select Last Promo

Choose an option



Customer Segmentation | Motivation

In an increasingly competitive market, businesses have been aiming to better understand their customers, offering products and services that closely align with their needs. By partitioning their customers into groups, they are able to tailor their strategies and improve satisfaction, loyalty, and profits. This practice, which is becoming more essential, is supported by many studies that highlight the importance of customer segmentation as a foundation for lasting client relationships.
When executed effectively, segmentation enables a more comprehensive outline of resource allocation, allowing for more impactful investments in areas that yield the greatest return.

This web application provides an **user-friendly** and **interactive platform** to explore the *ABCDEats* dataset in a 90-day period and perform customer segmentation analysis, aiding in the identification of distinct customer segments based on diverse factors such as region, age, ordering behavior, and cuisine preferences.

Web Application Features

1. **Interactive Visualizations:** presents key insights and segmentation results through various charts and graphs, facilitating easy interpretation of complex data.
2. **Interactive Filtering:** allows users to refine the dataset based on multiple user-defined criteria.
3. **Customer Segmentation:** enables the identification of distinct customer segments using diverse factors such as region, age, ordering behavior, and cuisine preferences.

Explore the app by selecting the EDA tab, and access the customer segmentation dashboard through the Final Customer Segmentation tab!

DM Project | Group 37 © 2024/25

EDA | Exploratory Data Analysis

31279 Number of Customers

129384 Total Orders

1,123,792.21 Total Amount Spent (Monetary Units)

Figure B1 – Main Features of the Web Application.

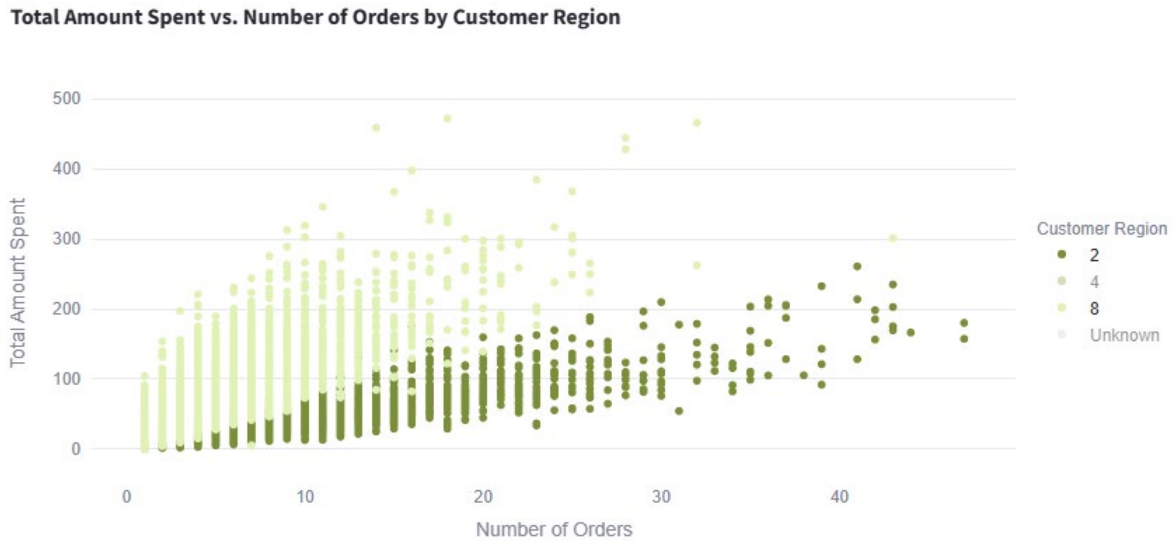


Figure B2 – Relationship between Total Amount Spent and Number of Orders for customers in *Regions 2* and *Region 8*.

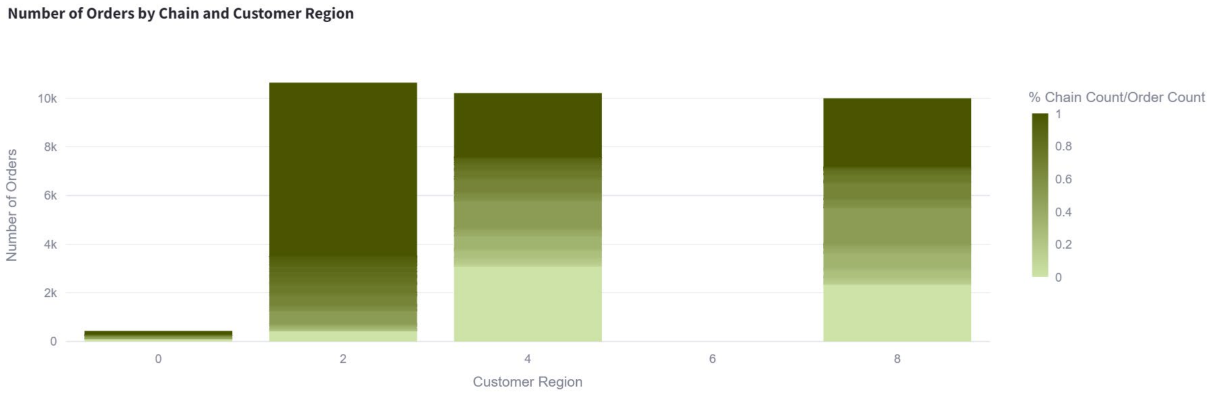


Figure B3 – Proportion of Orders placed at Chain restaurants by Region.

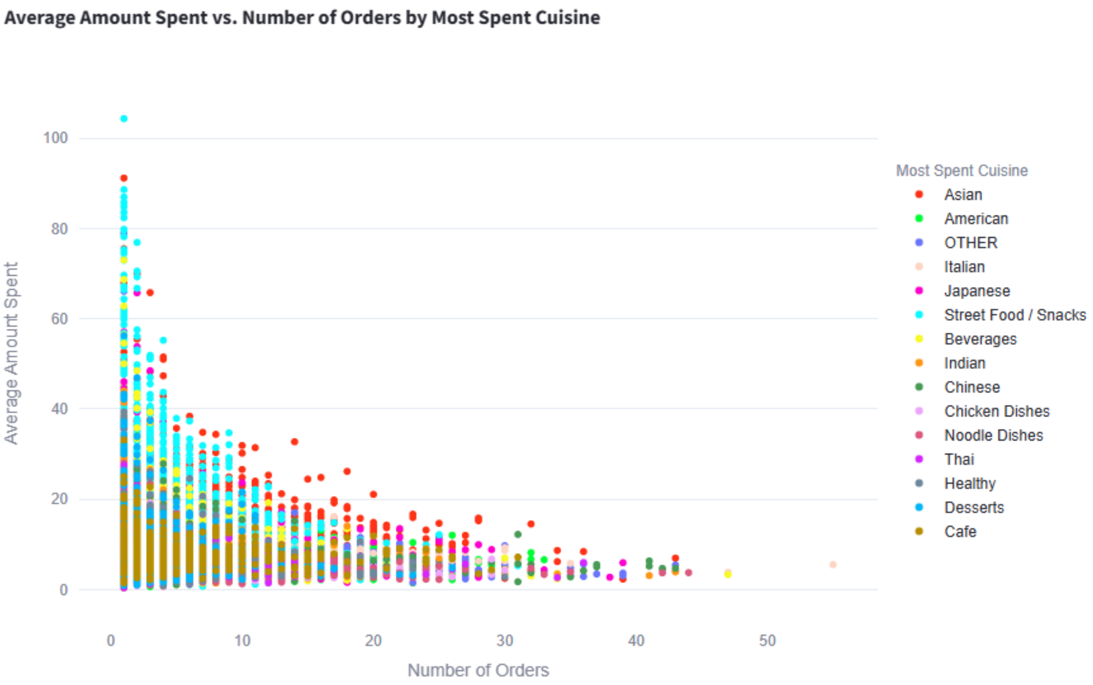
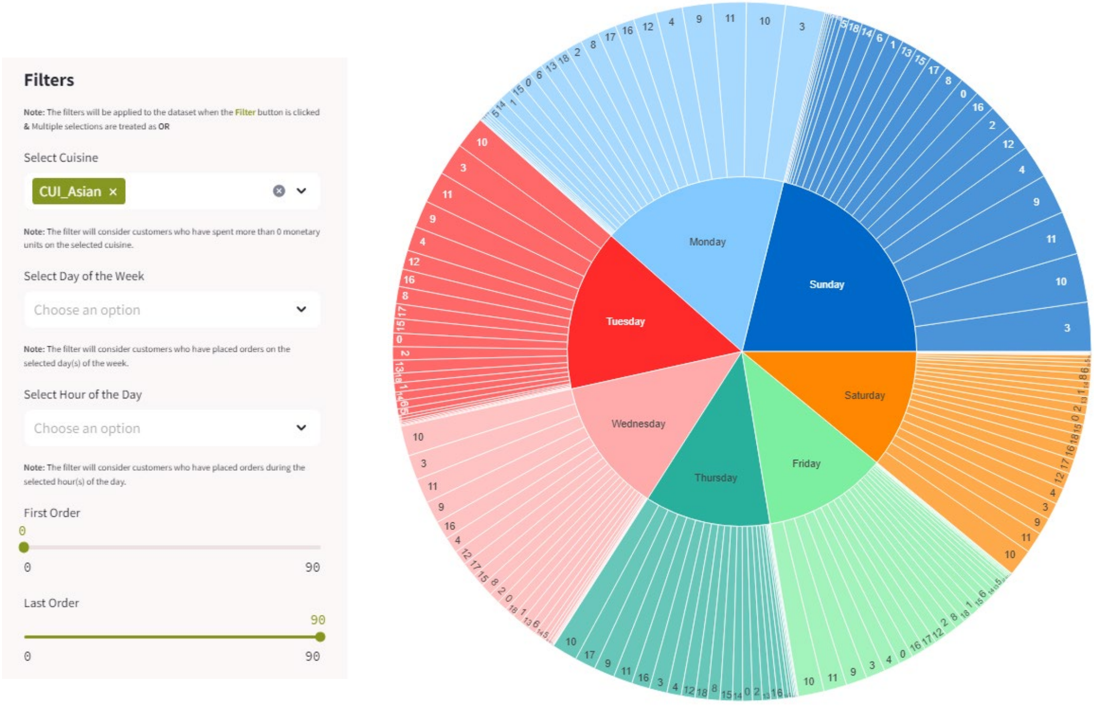


Figure B4 – Average Amount Spent vs. Number of Orders by Most Spent Cuisine.



Filters

Note: The filters will be applied to the dataset when the Filter button is clicked & Multiple selections are treated as **OR**

Select Cuisine

CUI_Asian ×

Note: The filter will consider customers who have spent more than 0 monetary units on the selected cuisine.

Select Day of the Week

Choose an option

Note: The filter will consider customers who have placed orders on the selected day(s) of the week.

Select Hour of the Day

Choose an option

Note: The filter will consider customers who have placed orders during the selected hour(s) of the day.

First Order

0

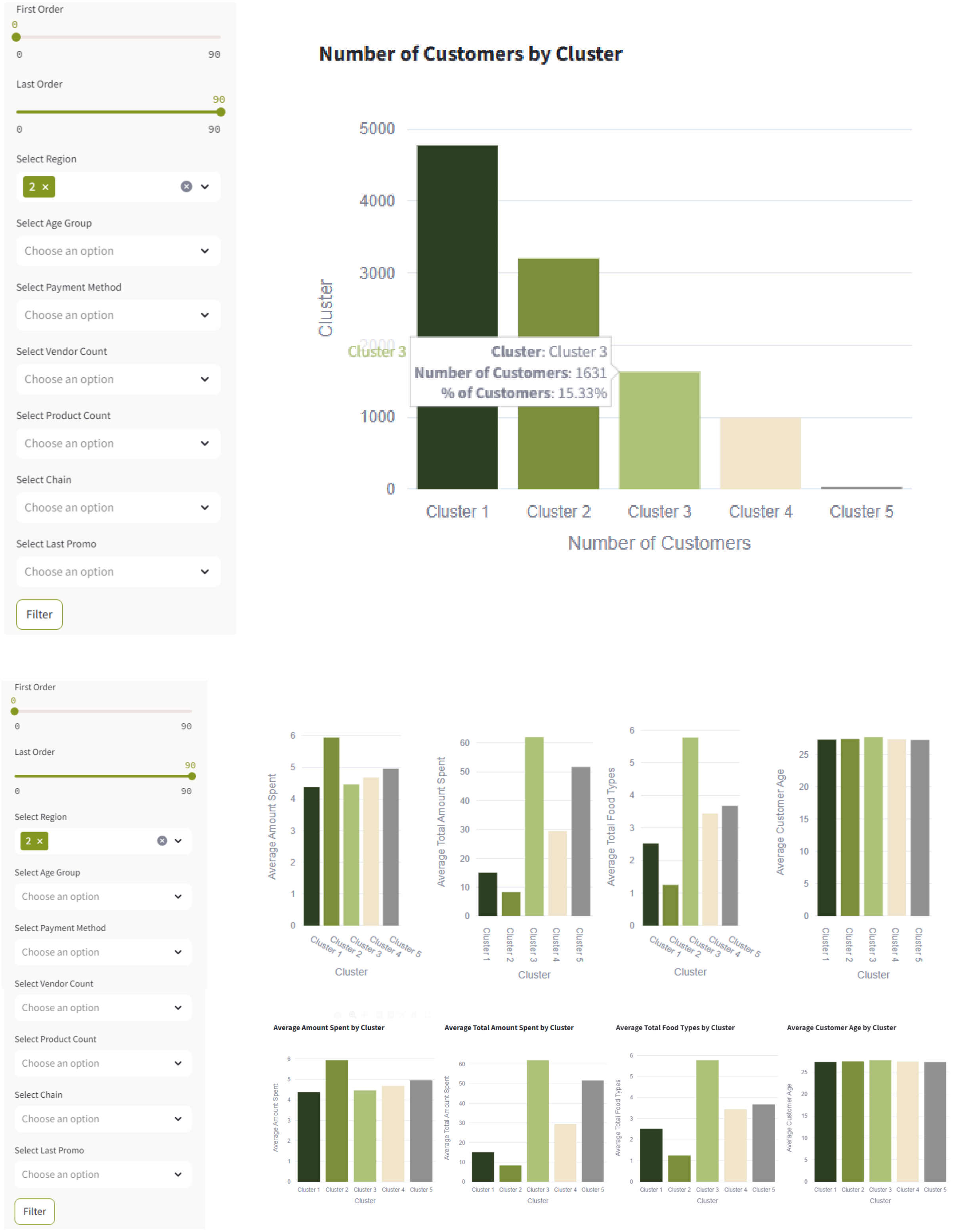
0 90

Last Order

90

0 90

Figure B5 – Most Common Day of the Week and Hour of the Day.



First Order

0

0 90

Last Order

90

0 90

Select Region

2 ×

Select Age Group

Choose an option

Select Payment Method

Choose an option

Select Vendor Count

Choose an option

Select Product Count

Choose an option

Select Chain

Choose an option

Select Last Promo

Choose an option

Filter

First Order

0

0 90

Last Order

90

0 90

Select Region

2 ×

Select Age Group

Choose an option

Select Payment Method

Choose an option

Select Vendor Count

Choose an option

Select Product Count

Choose an option

Select Chain

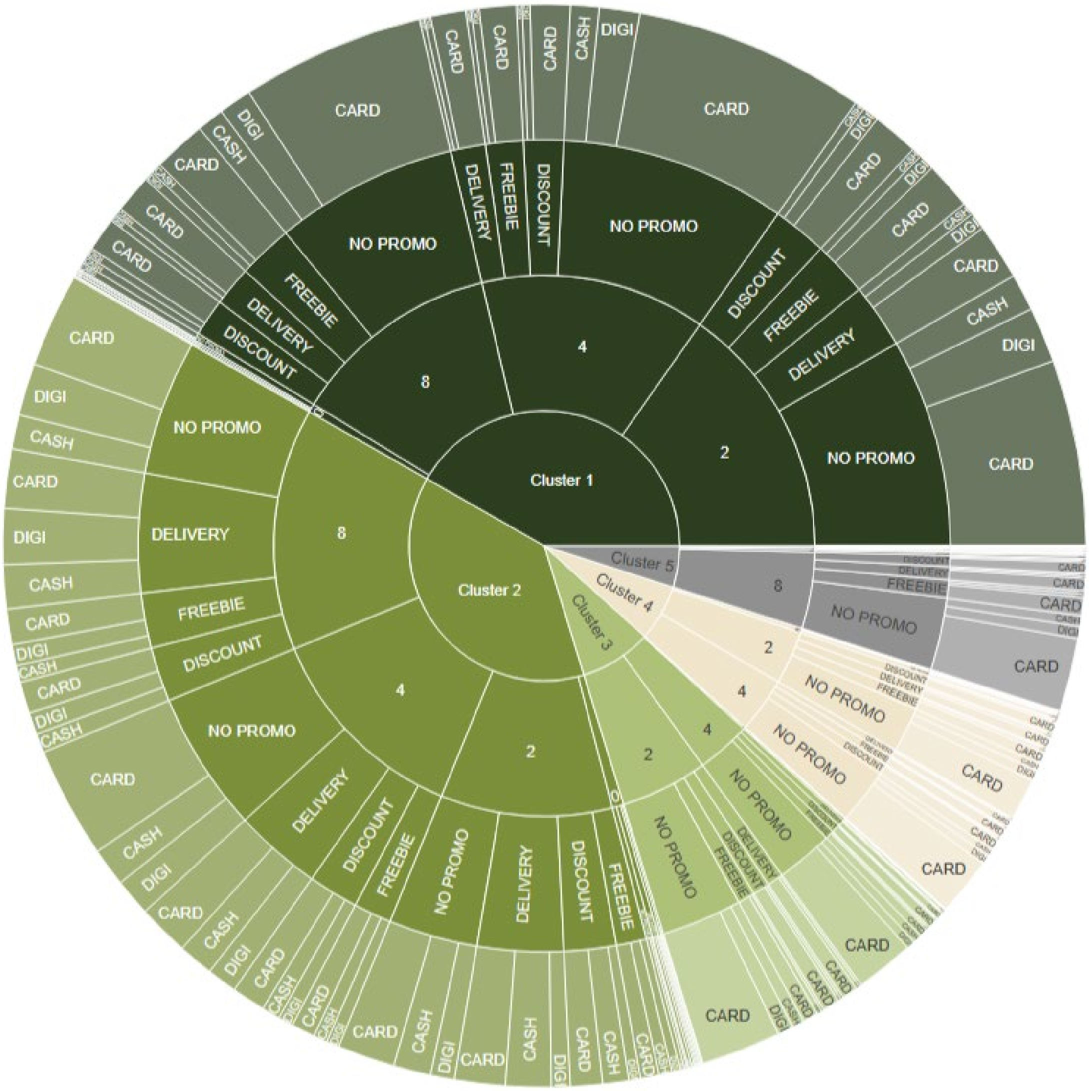
Choose an option

Select Last Promo

Choose an option

Filter

Figure B6 – Analysis of Clusters in *Regions 2*.



Sunburst Plot of Customer Region, Last Promo, and Payment Method by Cluster

Figure B7 – Customer Region, Last Promo, and Payment Method by Cluster.

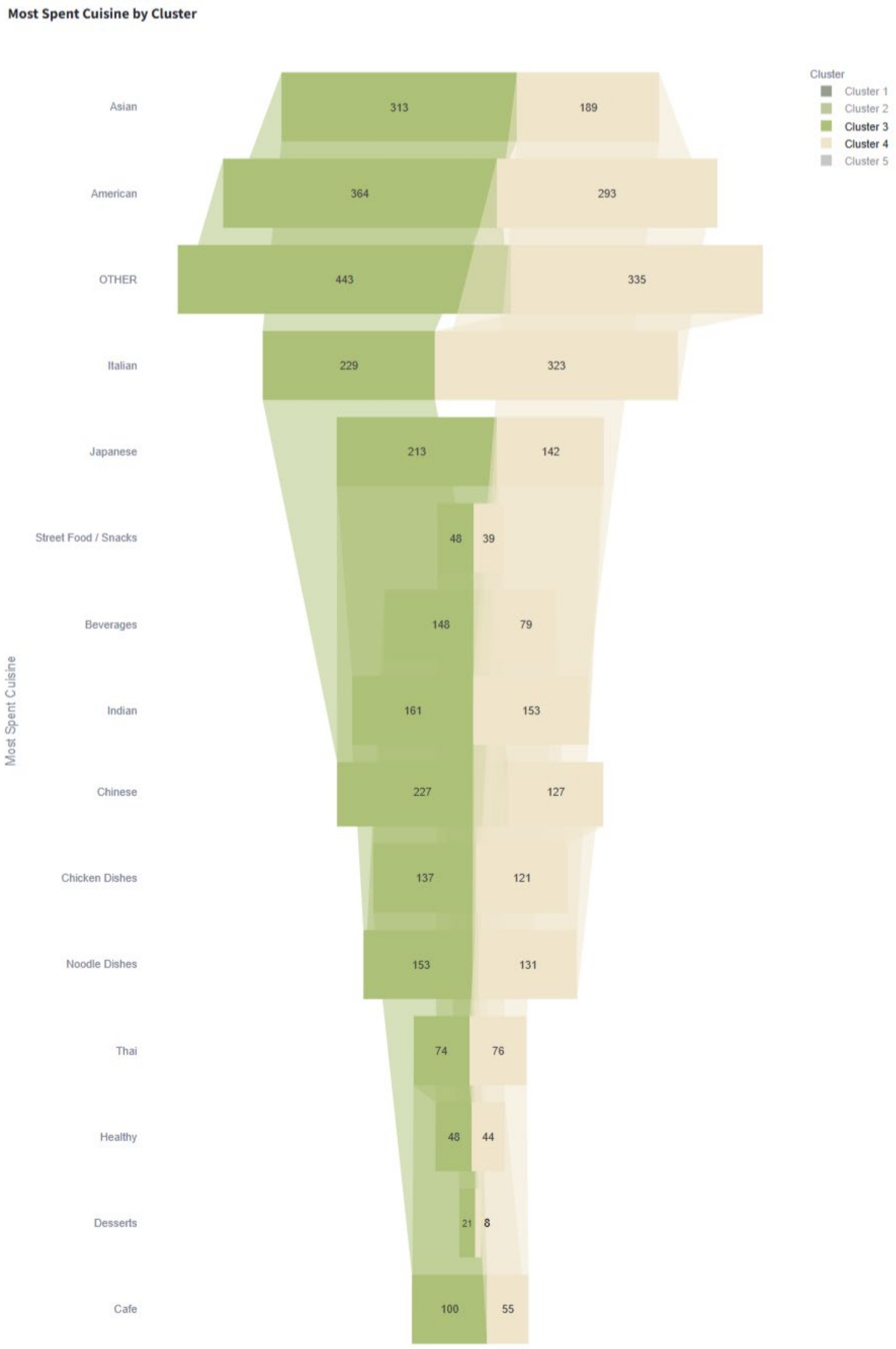


Figure B8 – Distribution of Top Cuisines by Cluster.

ANNEX A. CRISP-DM

To address the defined objectives, the **CRISP-DM** methodology was employed. This methodology is a dynamic and fluid iterative approach (**Figure A1**) and relies on the exploration and refinement of its different stages to achieve better results, thereby supporting more effective and meaningful decision-making. By definition, this methodology comprises six phases: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, and *Deployment*, with various tasks within each phase that will be explored in this project. [1]

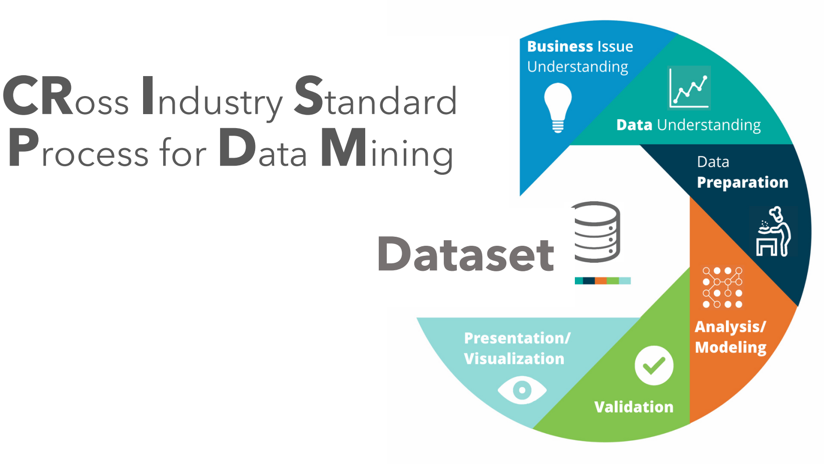


Figure A1 – CRISP-DM Methodology Cycle.
[Adapted from [2]]

Business Understanding

This initial phase, which was started during the initial phase of this project, centred on defining the project's objectives and requirements from the perspective of *ABCDEats Inc.* The aim was to understand the need for effective customer segmentation, recognizing that today's consumers are increasingly selective and demand personalized experiences. Therefore, our goal was to develop a data-driven customer segmentation strategy, enabling to tailor its services and marketing more effectively.

Data Understanding

The second phase, also detailed in our first report, involved collecting and exploring the *ABCDEats* customer dataset, which included data from 3 cities over 3 months, with focus on customer behaviours and purchases. During this phase, we focused in understanding the dataset's quality and suitability for customer segmentation, by calculating descriptive statistics and visualizing various aspects of the data, allowing for a first understanding of the data's characteristics and identifying specific patterns.

Data Preparation

We then moved to data preparation, where we selected, cleaned, and transformed the data. Key tasks included handling inconsistencies, imputing missing values, using deterministic approaches, or imputation, while creating new features. This stage also involved addressing outlier values, to ensure that the data was reliable and well-structured for modelling. We used Principal Component Analysis (PCA) to reduce dimensionality and create new features.

Modelling

During this phase, we applied a range of clustering algorithms, such as Hierarchical Clustering (HC), K-Means, Self-Organizing Maps (SOM), Mean Shift, DBSCAN, and Gaussian Mixture Models (GMM). We also chose to use value-based and behaviour-based perspectives in the clustering process. Each method was assessed based on different metrics such as the inertia, silhouette score, and R^2. In the end, a combination of methods was used to construct the final solution, which was then further profiled and explained.

Evaluation

After the modelling phase, the focus was on evaluating the clustering results, using metrics and visualizations to assess the quality of each model, and by comparing different approaches. The evaluation phase also considered which approach best described the underlying structures of the data.

Deployment

The deployment phase, beyond the typical presentation of results, involved the development of an interactive web application using *Streamlit*. This application, provided as an optional yet valuable extension to the project, empowers *ABCDEats Inc.* to explore and interact with the EDA and customer segmentation analysis in a user-friendly and visually intuitive manner.