Data Mining Project

Master in Data Science and Advanced Analytics

**NOVA Information Management School**

Universidade Nova de Lisboa

ABCDEats Inc.

**Group 26**

Diogo Miguel Calisto Rodrigues, 20240512

Daniel Rodrigues Rainho, 20240607

Duarte Queiróz Miguel, 20240608

Fall Semester 2024-2025

TABLE OF CONTENTS

[1. Introduction 1](#_Toc180343678)

[2. Data Description 1](#_Toc180343679)

[2.1.1. Level 3 title 2](#_Toc180343680)

[2.1.1.1. Level 4 title 2](#_Toc180343681)

[3. Data cleaning and pre-processing 3](#_Toc180343682)

[Bibliographical References (Optional, Not included in page limit) 4](#_Toc180343683)

[Appendix A (Optional, Not included in page limit) 5](#_Toc180343684)

[Annexes (Optional, Not included in page limit) 6](#_Toc180343685)

# Introduction

This project aims to act as consultants for a fictional food delivery service called ABCDEats Inc. In this case, our goal is to analyze all the customers data collected over three months from three different cities and assist the service in developing a data-driven strategy for various customer segments.

We are free to try and analyze various approaches and perspectives in this project with the intention of giving the company a final segmentation to help them develop a marketing strategy.

# Data Description

The sample we received contains **31885** observations and 56 variables that we will need to manage for an easier understanding of the problem. In the following table there’s a description of them.

|  |  |  |
| --- | --- | --- |
| **VARIABLE** | **TYPE** | **DESCRIPTION** |
| customer\_id | object | Customer ID |
| customer\_region | object | Geographic region where the customer is located |
| customer\_age | float64 | Age of the Customer |
| vendor\_count | int64 | Number of unique vendors the customer has ordered from |
| product\_count | int64 | Total number of products the customer has ordered |
| is\_chain | int64 | Number of times the costumer ordered from a chain restaurant |
| first\_order | float64 | Number of days from the start of the dataset when the customer  first placed an order. |
| last\_order | int64 | Number of days from the start of the dataset when the customer  most recently placed an order. |
| last\_promo | object | The category of the promotion or discount most recently used  by the customer. |
| payment\_method | object | Method most recently used by the customer to pay for their orders |
| CUI\_American, CUI\_Asian, CUI\_Chinese, CUI\_Italian,etc. | float64 | The amount in monetary units spent by the customer from the  indicated type of cuisine. |
| DOW\_0 to Dow\_6 | int64 | Number of orders placed on each day of the week  (0 = Sunday, 6 = Saturday). |
| HR\_0 | float64 | Number of orders placed during each hour of the day (0 = midnight, 23 = 11 PM). |
| HR\_1 to HR\_23 | int64 |

To obtain trustworthy result, we must check our data and clean it. The first thing we notice is that our sample has **13 duplicate observations** that we will delete from our database. Secondly, we had to check the missing values. We have **727 missing values in customer\_age (around 2.28%)** and 106 missing values in **first\_order (around 0.33%)**. Further forward, these missing values will be processed.

# Incoherence Checking

After the data description, we had to analyze the coherence of our data and whether it made sense in our context. The first case that was analyzed was whether the sum of the number of orders placed on each day of the week is equal to the number of orders placed on each hour of the day. We saw that **30711** observations complied with this rule and **1164** did not.

The second case was that the total number of orders cannot be smaller than the vendor count. All of the observations comply with the rule. We have no problem here.

NAO PERCEBI O IF PRODUCT COUNT IS ZERO, IF PAYMENT METHOD, IF LAST PROMO

The last case is the last order cannot be before the first order. None of them reject this rule.

# Analysis of Variables

To see what distribution the variables take and possible outliers, we decided to do an analysis of each variable we got. Let’s analyze the **customer\_id** first. Our idea is to set this one as the index of our DataFrame, it is beneficial because identifies each customer, making it easier to locate and manage individual data efficiently. There are no duplicated rows, so we set this variable as the index.

Next, we have the **customer\_age**. In our analysis of the data, we observed that the age variable had 2.28% of its values missing. Since our clustering model cannot accommodate missing values, we will address this issue during the preprocessing stage. As we can see in figure 1 by examining this histogram and boxplot, we can see that the age distribution is asymmetric. This is likely due to outliers, which we can visualize in the boxplot.

Uma imagem com texto, diagrama, captura de ecrã, Gráfico

Descrição gerada automaticamente

Figure 1

For a better visualization of our data, we will exclude the outliers for now:

QUEREM COLOCAR SEM OS OUTLIERS??

For the **customer\_region** we just wanted to see the distribution of clients for each region using a pie chart plot like the one in the next figure.

Uma imagem com texto, captura de ecrã, diagrama, Gráficos

Descrição gerada automaticamente

The three main clients are from the regions **8670**, **4660** and **2360**. Further, our idea is to create a relationship with those regions and a marketing approach since they are the main targets of our customers.

VENDOR COUNT É LITERALMENTE A MESMA CENA QUE O CUSTOMER\_AGE

# Relationships between Features

Logo

Description automatically generated

Figure 2.1 – Illustrative figure

Note that figure labels should be included after the figure. Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.

Table 2.1 – Illustrative table

|  |  |
| --- | --- |
| **Title** | **Title** |
| Text | Number |
| Text | Number |
| Text | Number |

The student can freely choose the table design, as long as it remains consistent throughout the document. Note that table labels should always be included before the table. Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.

### Level 3 title

Example of an unnumbered list:

* Item 1
* Item 2
* Item 3

#### Level 4 title

Example of a numbered list:

1. Item 1
2. Item 2
3. Item 3

# Bibliographical References (Optional, Not included in page limit)

Use APA Style for the entire document

We suggest that students use a reference manager system (Zotero, Mendeley, EndNote),

Please review the style guide at: <https://apastyle.apa.org/style-grammar-guidelines/references/examples>:

Author, A. A., Author, B. B., & Author, C. C. (Year). Title of article. *Title of Periodical, volume number* (issue number), pages.

# Appendix A (Optional, Not included in page limit)

[Appendixes are for materials, tables, or more explanation material only done by the student]

# Annexes (Optional, Not included in page limit)

[Annexes are optional, since they have material and sources not developed by the students, so in most cases referencing them is enough]