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

[Introduction 1](#_Toc181459311)

[Data Description 1](#_Toc181459312)

[Incoherence Checking 2](#_Toc181459313)

[Analysis of Variables 2](#_Toc181459314)

[Creation of New Variables 4](#_Toc181459315)

[Relations between Variables 5](#_Toc181459316)

[1.1.1. Level 3 title 6](#_Toc181459317)

[1.1.1.1. Level 4 title 6](#_Toc181459318)

[Bibliographical References (Optional, Not included in page limit) 8](#_Toc181459319)

[Appendix A (Optional, Not included in page limit) 9](#_Toc181459320)

[Annexes (Optional, Not included in page limit) 10](#_Toc181459321)

# 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 restaurante (\*) |
| 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 |

(\*) Here we decided to change the description of the variable. The description given by the problem did not correspond well to the data and in our interpretation, this description better suits the type of the problem, and the date received.

To obtain trustworthy results, 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%)**, 106 missing values in **first\_order (around 0.33%)** and 1165 missing values in **HR\_0** **(around 3.65%)**. 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 the observations comply with the rule. We have no problem here.

With the data we have, to make sense in the context of this work, we decided to check if there are values in the product\_count and in the vendor\_count with value 0. For the first one we confirmed **156** rows with value 0 and for the second one **138** rows with value 0. Those 138 rows coincide with the lines whose product\_count is 0. Right now, we will interpret them as errors. We decided to go further about these 138 rows. All of them have value 0 in all the cuisine types (CUI\_). Further forward, they could be dropped.

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. Most of the users belong to a younger age group.

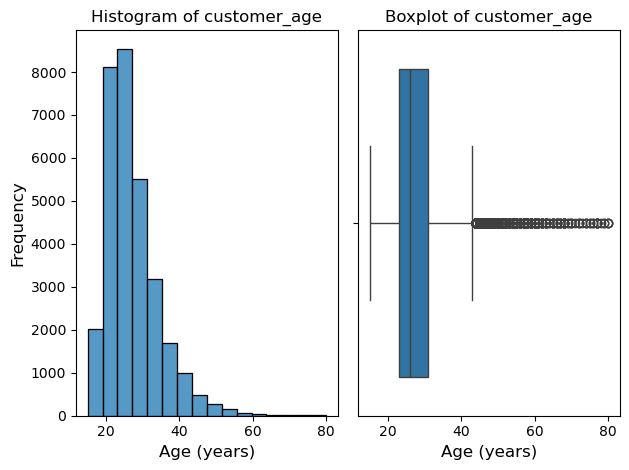


Figure 1

For the **customer\_region** we wanted to see the main target regions. The three main clients are from the regions **8670**, **4660** and **2360**.

Looking now for the **vendor\_count** variable’s boxplot (Figure 2) , we can see two outstanding outliers. We believe that’s not an error but client who just like to diversify when it comes to different vendors.

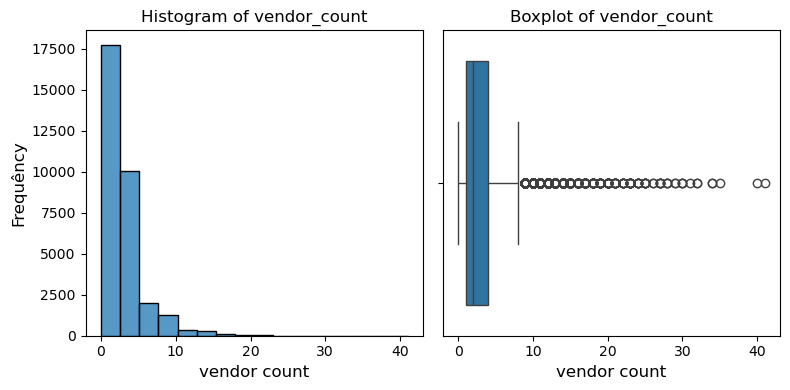


Figure 2

Analyzing the **product\_count** variable, we saw an outstanding outlier (Figure 3). This extreme point could negatively impact our analysis and visualization. Our idea is to adjust it to a more reasonable high value.

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

Descrição gerada automaticamente

Figure 3

Now, let’s look at two variables at the same time, **first\_order** and **last\_order**. Looking at both histograms (Figure 4) we could see contrasting shapes between them. While new customer acquisition slowed down over time, existing customers were retained. In the first order histogram, most customers made their first purchase early in the dataset timeline, with fewer joining over time. In the last order histogram, many customers remained active or re-engaged toward the end, suggesting strong retention or successful reactivation efforts.

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

Descrição gerada automaticamente

Figure 4

Looking at the **last\_promo** variable, we interpreted that the rows with value **( - ),** indicates customers who did not use any promotions. Most of our users didn’t use any promotions (52.5%).

At the **payment\_method**, the majority use card as their payment method (63.2%).

Looking at all **CUI\_** variables, the cuisine with the most money spent is Asian being followed by the American and Street food / Snacks.

Next, we analyzed the number of orders placed on each day of the week, the **DOW\_** variables. The data indicates an increase in orders throughout the week. Peaking on Sunday. This could reflect a behavior where people are more likely to shop during the weekend.

As for the **HR\_** variable, we checked that in the hour 0, all rows take the value 0, indicating either that there were no orders placed at midnight or that was an error in collecting data for this variable. Orders peak at hours 17 and 11 are likely aligning with lunch and early dinner times or after work hours. In contrast, order volumes are lowest during early morning when people typically don’t order food.

For the **is\_chain** variable, we can see some outliers represented in Figure 5. These outliers are unlikely to be data errors, probably representing clients who simply decided to make more orders.

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

Descrição gerada automaticamente

Figure 5

# Creation of New Variables

**Customer\_time** is a variable that represents the duration of each customer’s time with the delivery service. We get this variable by doing **last\_order** minus **first\_order**. Because we have missing values in the first\_order variable, customer\_time will have the same number of missing values. The mean is approximately 35 days.

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

Descrição gerada automaticamente

Figure 6

Taking a look at the histogram (Figure 6), we can see that most users have a **customer\_time** of zero, indicating that they only used the application once and never ordered anything again.

**Order\_hour\_spread** quantifies the variability in order counts across different hours of the day for each customer to find a pattern. A low value indicates that orders are consistently placed at certain hours of the day, suggesting a regular purchasing pattern. A high value indicates the opposite. This variable has 0.73% of missing values.

FALAR ALGUMA COISA DOS GRAFICOS??

**Order\_day\_spread** is similar to the order\_hour\_spread but instead of working with hours, it works with the days of the week. It has 0.43% of missing values.

FALAR ALGUMA COISA DOS GRAFICOS??

**Order\_count** represents the total number of orders for each customer.

**Intensity\_of\_activity** quantifies how active a customer is by calculating the average number of orders they placed per day. The variable has 0.33% of missing values.

GRAFICO

**Total\_Spended** represents the total amount of money spent by each customer.

GRAFICO

**Diversity\_cuisine** is a variable that measures how many different types of cuisines a customer has ordered. With that we conclude that most of the users only order from 2 different types of cuisines.

**Customer\_loyalty** tells us how diverse a customer’s ordering behavior is. A low value suggests a preference for some vendors. A high value indicates a willingness to try new vendors. This variable takes values between 0 and 1. The mean is 0.8 so it suggests that most of the users has no preference in some specific vendors when they order.

**Age\_category** categorizes individuals into distinct age groups. We set 4 type of age groups:

* Young [ 15, 20 [
* Young-Adult [ 20, 30 [
* Adult [ 30, 50 [
* Senior [ 50, 100 ]

The main group is **Young-Adult** with 63.1%.

GRAFICO

# Relations between Variables

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]