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](#_Toc181569758)

[2. Data Description 1](#_Toc181569759)

[3. Analysis of Variables 2](#_Toc181569760)

[Creation of New Variables 4](#_Toc181569761)

[Incoherence Checking 5](#_Toc181569762)

[Relations between Variables 5](#_Toc181569763)

[1.1.1. Level 3 title 7](#_Toc181569764)

[1.1.1.1. Level 4 title 7](#_Toc181569765)

[Bibliographical References (Optional, Not included in page limit) 9](#_Toc181569766)

[Appendix A (Optional, Not included in page limit) 10](#_Toc181569767)

[Annexes (Optional, Not included in page limit) 11](#_Toc181569768)

# 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 |

(\*) Originally the metadata implied that the variable **is\_chain** should be boolean. The description of the variable 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. We’ll justify this decision in the Analysis of Variables chapter.

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.

# Analysis of Variables

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 id’s, so we set this variable as the index.

Next, we have the **customer\_age**. In our analysis of the data. As we can see in *Figure 1* by examining the histogram and boxplot, we can see that the age distribution has a positive skew. A significant number of outliers skew toward older age groups, those are not errors but some older users using the application. Most of the users belong to a younger age group.

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

Descrição gerada automaticamente Uma imagem com texto, captura de ecrã, diagrama, Gráfico

Descrição gerada automaticamente

*Figure 1 – Plots customer\_age Figure 2 – Bar plot customer\_region*

For the **customer\_region** we wanted to see the main target regions. The three main clients are from the regions **8670**, **4660** and **2360**. The category ‘ – ‘ means unknown and we should decide if we keep it or change it.

Looking now for the **vendor\_count** and the **product\_count** variables’ histogram *Figure 3* and *Figure 4*, we can see a positive skew revealing a concentration at lower counts. Both variables have their outliers, as seen in unlikely to be data errors, representing clients who made more orders/products. It is remarkable that product\_count variable has an outstanding outlier that could negatively impact our analysis and visualization.

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

Descrição gerada automaticamente Uma imagem com captura de ecrã, texto, Gráfico, file

Descrição gerada automaticamente

*Figure 3 – Plots vendor\_count Figure 4 – Plots product\_count*

Now, let’s look at two variables at the same time, **first\_order** and **last\_order**. Looking at both histograms (*Figure 5*) we could see contrasting shapes between them. 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, captura de ecrã, Gráfico

Descrição gerada automaticamente Uma imagem com texto, captura de ecrã, Gráfico, diagrama

Descrição gerada automaticamente

*Figure 5 – Histograms first/last\_order Figure 6 – Bar plot last\_promo*

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%), followed by Delivery Promotion with 19.7% of the clients.

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.

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

Descrição gerada automaticamente Uma imagem com texto, captura de ecrã, Saturação de cores, Gráfico

Descrição gerada automaticamente

*Figure 6 – Bar plot payment\_method Figure 7 – Bar plot DOW\_*

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 (as we can see in *Figure 8*) 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.

Uma imagem com texto, captura de ecrã, Saturação de cores, file

Descrição gerada automaticamente Uma imagem com captura de ecrã, diagrama, file, Gráfico

Descrição gerada automaticamente

*Figure 8 – Bar plot HR\_ Figure 9 – Plots is\_chain*

For the **is\_chain** variable, it takes values between 0 and 83 (not Boolean) so the data given to us does not correspond to the original description. We decided to check its distribution, seen in *Figure 9* , and we noticed that it’s very similar to the vendor\_count’s distribution. Our conclusion is that the variable itself gives us the number of times the costumer ordered from a chain restaurant. Also, it has the same problems as the vendor\_count.

# Creation of New Variables

In our work, we decided to create new variables to help approach the problem and the relationship between the original variables.

**Customer\_time** is a variable that represents the duration of each customer’s time with the delivery service. We get this variable by subtracting last\_order by first\_order. Because we have missing values in the first\_order variable, customer\_time will have the same number of missing values. Looking at the histogram (Figure 6), we can see that most users have a **customer\_time** of zero, indicating that they only used the application one day.

**Order\_count** represents the total number of orders for each customer. We get it by doing the sum of the values of HR\_0 to HR\_23 or DOW\_0 to DOW\_6.

**Intensity\_of\_activity** quantifies how active a customer is by calculating the average time interval between two orders. This variable is obtained by dividing customer\_time by order\_count. The variable has 0.33% of missing values derived from first\_order.

**Total\_Spended** represents the total amount of money spent by each customer by doing the sum of the values from the CUI\_ columns.

**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, 80 ]

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

# 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. We decided to go further about these 138 rows and all of them have value 0 in all the cuisine types (CUI\_), is\_chain and order\_count variables. Right now, we will interpret them as errors.

The last case is that the last order can’t come before the first order. None of them reject this rule.

# Relations between Variables

Logo

Description automatically generated

Figure 0.1 – 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 0.1 – 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]