Data Mining Project

Master in Data Science and Advanced Analytics

**NOVA Information Management School**

Universidade Nova de Lisboa

ABCDEatsInc.

Data Mining Project

**Group 31**

Student1 Name, 20001111

Student2 Name, 20002222

Student3 Name, 20003333

Student4 Name, 20004444

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

[1. Introduction 1](#_Toc176981912)

[2. Section Heading Level 1 1](#_Toc176981913)

[2.1. Level 2 Title 1](#_Toc176981914)

[2.1.1. Level 3 title 2](#_Toc176981915)

[2.1.1.1. Level 4 title 2](#_Toc176981916)

[3. Another Section Heading 2](#_Toc176981917)

[Bibliographical References (Optional, Not included in page limit) 3](#_Toc176981918)

[Appendix A (Optional, Not included in page limit) 4](#_Toc176981919)

[Annexes (Optional, Not included in page limit) 5](#_Toc176981920)

# Introduction

The purpose of this project is to **utilize all the data** collected from the company ABCDEats.inc and transform it into **meaningful information** to reach **substantiated conclusions**. In the following report you will find an exploratory analysis and visual representations of selective parts of the data, which we believe **show significant patterns of behavior** of the company’s customers. ABCDEats.inc can use this knowledge to potentially cut down operational costs, increase operational efficiency and extract detailed customer characteristics that can help the marketing department in every aspect of it.

In the following report you will find a comprehensive explanation of the data and all its components before and after the preprocessing stage. The descriptive statistics of the dataset’s variables and added features. Finally, an exploratory analysis of the patterns and bivariate relationships in the dataset. Please bear in mind that all the referenced graphs and figures can be found in Annex A.

# Data and variable description

Our data consisted of **31888 observation**s and **56 variables**. Each observation corresponded to a customer of the food delivery service company ABCDE Eats. You can find an overview of all variables in the following table (Table1).

# Data preprocessing

To ensure data quality and reliability, some preprocessing techniques were implemented before any statistical analysis. The variable “Customer\_id” was identified and set as the primary key (index). The primary key indicated **13 duplicate observations**, along with their respective variables, and subsequently, they were removed. With the pre-mentioned alterations, the data set was left with 31875 rows and 56 columns.

**Inconsistencies were found** between the Days of the Week (**DOW**) and Hours of the Day (**HR)** columns, suggesting that **HR was underrepresented.** DOW was utilized to fill in the missing data in the HR variable, more specifically in the “HR0” column.

In addition, the **categorical variable “last\_promo”** also contained **missing values**. It was reasonable to assume that no promotion was used, so the missing values were filled in with “no\_promo”.

Apart from the HR and “last\_promo” variables, **the data set contained missing values** across 6 other variables although in relatively small amounts. Particularly, “customer\_age” with 0.46% of all NaNs in the data set, “customer\_region” with 0.28%, and “first\_order” having  0,07%. It is worth noting that there were no rows entirely filled with missing values. Therefore, to maintain the integrity and accuracy of the data, we decided to keep the missing values for further exploratory analysis. For a detailed description of all missing values, please see the table below.

For the purposes of consistency and avoiding redundancy, we created a function which goes through the above mentioned steps to “clean” the data. A detailed overview of the function can be found in the submitted Python Notebook “Function”.

### Descriptive Statistics

**Numerical Variables**

Descriptive statistics offer a thorough summary of the main features of a dataset. The customers’ ages range from 15 to 80 with an average age of approximately 28 years old. The largest portion of our customers are aged between 23 and 31. The age distribution seems to be positively skewed, with all 1083 outliers in the upper bound. Since these outliers represent 0.03% of the entire age dataset, we believe that strong conclusions cannot be reached about customers over the age of 43.

The mean number of vendors per customer is about 3.1, suggesting that customers typically order from three different vendors, although some highly engaged users order from a wide variety of vendors. On average customers order around 6 products, and a standard deviation of 6.95 indicates significant variability in the number of products ordered by different customers. This is supported with the range in product count being from 0 to 269, suggesting that some users are very active compared to others. On average, customers have ordered from chain restaurants about **2.82 times**, suggesting that while chain restaurants are popular, customers don't exclusively rely on them. For a more detailed overview of the descriptive statistics please see the table below.

**Categorical Variables**

The **customer data is spread across 9 regions**, of which one is compiled of the missing data in that variable (customers with an unknown region). **The most active ones** are region **8670** containing 31% of the orders, followed by **4660** with 30.4% and **2360** with 28.1% (Figure\_).

The data set shows that 52.5% of orders were placed without the use of a promotion code, while 47.5% used one of the types: **Delivery** (19.7%), **Discount** (14.1%), and **Freebie** (13.7%). Since we had no knowledge of when the discounts were used or what they signify, we belived it would be more efficient to **focus solely** on the number of customers that **used or did not use the promotion** (Figures\_,\_).

# added features

To facilitate a more in-depth analysis, we created additional variables which were added as separate columns to the dataset, each focusing on the respective customer:

“**delta\_day\_order**” - Displays the time passed between the first order and the last order; “**tot\_value\_cui**” – Displays the total amount of money spent on the app across all options;

With the use of these two new variables, the following analysis were possible: the average number ofproducts a customer orders per day (“**order\_freq**”); the average monetary value of orders made by a customer per day(“**value\_freq**”); the average monetary value a customer spends per order (“**avg\_order\_value**”).

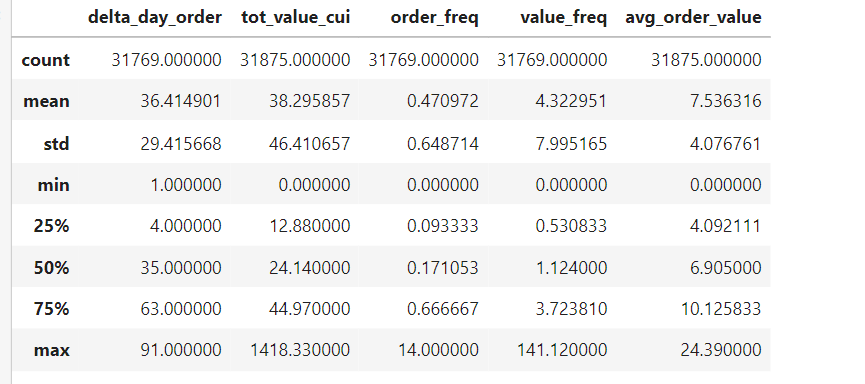
The analysis of these variables gave us a better insight into our customer’s behavior. On average a **customer places their first and last order** over a span of about **36 days,** and with a **high standard deviation** we could see that the company may **have both casual and regular users**.

The **average customer** spends around **38.3 units** of currency overall, however with a **wide range of up to** **1418.33 units**, we can assume that there may be **a group of high-value customers** contributing disproportionately to the company’s order revenue.

Customers make around **0.47 orders per day** on average, indicating they use the service somewhat frequently but not daily. However, there is a small group of customers which place orders more frequently.

On average, customers spend around **4.32 units** per day, however some spend up to **141 units/day**, indicating a very high value contribution from some users. The **high standard deviation** supports the fact that **daily expenditure fluctuates a lot** between customers.

The **average order value** is around **7.54 units**, which is a fair amount, assuming that it reflects typical meal prices. The **modest** **standard deviation** indicates that while most orders fall within a smaller range, there are occasional higher-value orders. For a more detailed overview of the descriptive statistics of the added features please see the table\_ below.



# Exploratory analysis

#### Significant Patterns and Bivariate Relationships

**Focus on Customer orders throughout the week**

As shown on Figure \_ . , as the weekend approaches, customers place a higher number of orders, reaching the peak on Thursdays (15.5%), Fridays (14.9%) and Saturdays(16.1%). On Sundays, the number of orders drops significantly by 21.08%. If we separate the data in weekdays and weekends, the mean order values are the same, which takes into account that weekends only include data of two days(figure\_).

**A graph with blue bars and red line

Description automatically generated**

Assuming that this phenomenon gives us a valuable insight into our cutomer’s behavior, we decided to delve in deeper and relate it to other variables like: “**customer\_region**” and “**customer\_age**”.

**Orders by Region Throughout the Week** -A heatmap of the two variables (Figure\_) shows that almost all regions place a higher number of orders towards the end of the week, particularly on Thursday, Friday, and Saturday, with a drop during the week that begins on Sunday, precisely like the overall distribution previously seen. Only **regions 8550 and 8370 are exceptions**. Region **8550 has very few data points**, so we cannot hold any strong assumptions. On the other hand, the **number of orders** in region **8370** **is dispersed throughout all days** of the week with peaks reached on Monday, Friday, and Saturday. From the data given we can conclude that **75% of the regions present follow the same pattern**. Their **most active days** of the week are **Thursday, Friday and Saturday**.

**Orders by Customer Age Throughout the Week** – The heatmap of these two variables (Figure\_) shows that the order placement across days of the week changes with the customer's age. Customers between the ages of 16 and 40 tend to focus their orders between Thursday and Saturday, with a particular emphasis on Saturdays. Starting from the age of 40, customers also begin to focus their orders on Sunday and Monday. The exception to this pattern appears to be customers aged 15, who also show a strong tendency to place orders on Sunday and Monday. It is also important to point out, that due to the outliers

**Focus on Customer Orders Throughout the Day**

Customers tend to order more in the morning and in the evening, with the most popular time being 11am am and 5pm. Besides that, there is a significant drop in the number of orders placed at dawn and late in the night. In addition, at lunchtime, there is a relatively low decrease in the number of orders placed. For a visual representation please see Figure\_. To be even more precise we created a line chart graphing the normalized cumulative orders over the course of a day (Figure\_). The general trend is that sales start at a low level at the beginning of the day (12am), with a modest increase in the first few hours, and from a certain point (10am to 11am), the growth rate accelerates significantly until 7pm. After that, the growth slows down and eventually stabilizes around 10pm. This could be useful for optimizing operating hours or promotional campaigns during peak activity periods.

**Orders By Region Throughout the Week –** When we look at the orders per region (Figure\_) we can see that not all regions follow the same pattern. In regions 2630, 2490, and 2440, we see that the distribution follows the general trend previously presented, with the majority of orders placed around 11 a.m. and 5p.m. Regions 4660 and 4140 also align with the general distribution but show a drop in order volume during the late morning (around 11 a.m.).Regions 8370 and 8670 exhibit a similar pattern, with a focus on orders placed during the dawn, peaking at 3 a.m., and again in the morning, with a peak at 10 a.m.

**Orders By Customer Age Throughout the Week** – By the heatmap that we created (Figure\_) we can see that the age of the customers does not affect the time they place their orders. The age groups have relatively similar distribution, which follows the general pattern previously mentioned. Only after the age of 42, when the outliers come in, do the peak order hours start to vary by 1 or 2 hours. After the age of 65, we can observe that customers place more orders during dawn. However, since the data is limited and the sample size for this age group is small, we cannot draw definitive conclusions.

**Focus on Customer Orders of Particular Cuisines**

By comparing the distribution of the maximum order per cuisine and the total value spent per type of cuisine the **Asian** **stands out** in both metrics, followed by the American and Street Food cuisine. All three show high spending per order and high overall demand. On the other hand cuisines like Chinese and Japanese only have a high maximum order, suggesting that there are individuals (outliers) that tend to spend much more than the rest in that type of cuisine. (Figure\_, Figure\_)

By creating a box plot for each cuisine (Figure\_), we could notice that there was an extreme outlier in every type. Note here that we needed to disregard the zero values, to be able to make the plots readable. By looking at all the outliers in each region, per type of cuisine, it was apparent that every extreme outlier was a part of either region 4660, region 8670 or region 2360 (Figure\_).

**Regional Spending per Type of Cuisine** – Almost each region tends to favor a specific type of cuisine. The two most popular ones are Asian and Other, which are the first choice for 3 regions separately. On the other hand, some regions exhibit a broader range of high values across multiple cuisines, suggesting a more diverse preference to types of cuisines.

# Appendix A