Data Mining Project

Master in Data Science and Advanced Analytics

**NOVA Information Management School**

Universidade Nova de Lisboa

ABCDEatsInc.

Data Mining Project

**Group 31**

Klimentina Gilevska, 20240747

André Alexandre Loureiro Lourenço, 20240743

Matilde de Sousa Loreto Street Domingues, 20240523

Leonardo Di Caterina, 20240485

Fall/Spring Semester 2024-2025

Table of Contents

[1. Introduction 1](#_Toc181305379)

[2. Data and variable description 1](#_Toc181305380)

[3. Initial Data exploration 1](#_Toc181305381)

[4. Data preprocessing 2](#_Toc181305382)

[4.1.1. Descriptive Statistics 2](#_Toc181305383)

[5. Added Features 3](#_Toc181305384)

[6. Exploratory analysis – Patterns and Relationships 4](#_Toc181305385)

[6.1.1. Focus on Customer orders throughout the week 4](#_Toc181305386)

[6.1.2. Focus on Customer Orders Throughout the Day 4](#_Toc181305387)

[6.1.3. Focus on Customer Orders of Particular Cuisines 5](#_Toc181305388)

[Appendix 6](#_Toc181305389)

# 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 Table 2.1 below.

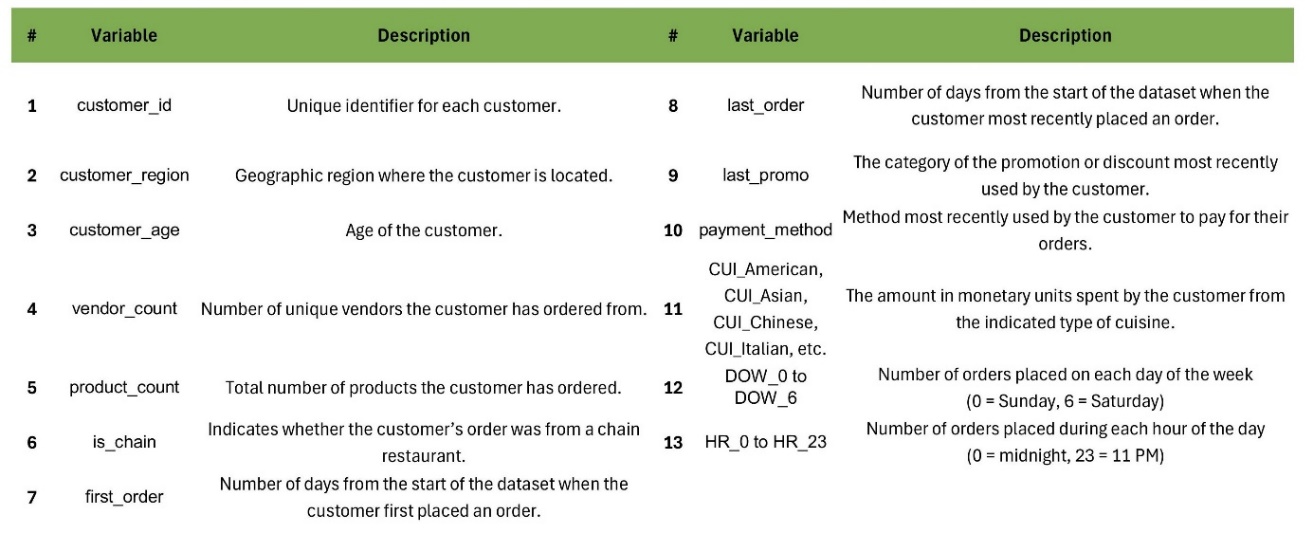


Table 2.1 – Dataset Variable Description

# Initial Data exploration

The given dataset contains customer IDs with its respective variables. There are categorical and numerical variables which the payment method, promotions used, weekly and daily order routine of **31888 customers over** a 90-day period. A monetary value of 122068.44 was spent across 139263 orders choosing between 15 types of cuisines provided by at least 41 vendors, some of which are chain restaurants. All customers are spread over 8 different regions (over 3 different cities).

The raw data contains 1998 NaNs spread across the variables: *first\_order (5%), customer\_age (36%), and HR\_0(58%)(Figure 1, Appendix).* However, with a more in-depth analysis, we can notice missing values in *last\_promo* and *customer\_region* reaching a number of 19188 NaNs. In addition, duplicate rows were found in the *customer\_id*. The previous, the NaNs and other inconsistencies will be addressed in the “Data Preprocessing” chapter.

# 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 *HR\_0* 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”***.

Another inconsistency was found in the ***is\_chain*** variable. We have three possible ways of interpreting it, namely the **number of orders**, **number of products** or **number of vendors** from chain restaurants, yet in all cases its value was **bigger** than the total order count, which makes its credibility questionable. We could also treat **is as a boolean** (true for values bigger than zero, otherwise false) **however there are discrepancies** even then. In future work we can assess the importance of this variable.

After analyzing the distributions of the customer regions we noticed an unequal representation, with 3 regions having most of the data, making it hard to draw conclusions for the other regions separately (Figure 2, Appendix). Knowing that the data is divided in three cities we decided to **aggregate the regions** **based on** the first number of the **region code**. To corroborate our assumption, we can see that after our division the **population distribution is balanced in all three cities** (Figure 3, Appendix). In addition, we first looked at the **average age distribution between cities** and saw that it was more consistent than the one **between regions** (Figure 4; Figure 5, Appendix). Lastly, when looking at the total revenue by cuisines, we notices far less sparsity after aggregating the data by cities (Figure 6; Figure 7, Appendix).

### Descriptive Statistics

**Numerical Variables**

The customer age ranges 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 99% of the data under the age of 52 (Figure 8, Appendix). Having this in mind, we cannot draw any solid statistical conclusions for customers above that age.

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. For a more detailed overview of the descriptive statistics please refer to Table 1 in Appendix.

**Categorical Variables**

The **customer data is spread across 8 regions**, of which one is compiled of the missing data in that variable (customers with an unknown region). As previously stated, we divided the regions into City 2 containing 34.2% of the population, City 4 with 33.1% and City 8 with 32.7% (Figure 9, Appendix).

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 if they were used cumulatively, we believed it would be more efficient to **focus solely** on the number of customers that **used or did not use the promotion** (Figure 10, Appendix).

Lastly, by looking at the *payment\_method* variable we suspect that it doesn’t carry much information for our current analysis. It only gives us the last payment used, and it doesn’t exclude other ones. However, we can mention that the use of the Card is significantly prominent (20153), followed by the Digital (6098) and lastly Cash (5624) method (Figure 11, Appendix).

# Added Features

To facilitate a more in-depth analysis, we created additional variables which were added as separate columns to the dataset, focusing on the habits of each customer: **1.*customer\_city*** *–* Groups the *customer\_region* by the first number of the region identification code (to resolve the problem of disproportionate representation of customer regions); **2.*used\_promo*** – A Boolean feature (“True” if the customer has used the promotion, otherwise it is “False”); **3.*order\_count*** *–* Total amount of orders each customer made (sum of DOW columns); **5.*delta\_day\_order*** - Displays the time passed between the first order and the last order, hence the permanence on the app oof each user (difference between the *last\_order* and *first\_order);* ***6.tot\_value\_cui*** – Displays the total amount of money spent on the app across all options by each user (sum of all CUI columns);

With the use of these new variables, the following analysis were possible: **4.*avg\_product\_by\_order*** –average number of products in an order (*product\_count* divided by order count); **7.*order\_freq , 8. value\_freq, 9.product\_freq –*** *give us an insight on the frequency of these features ((order\_cunt,tot\_value\_cui,product\_count)/delta\_day\_order);****10.avg\_order\_value***, **11.avg\_product\_value** - the average monetary value a customer spends per order or per product (*tot\_value\_cui/(**order\_count, product\_count))*

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 Table 2 in Appendix.

# Exploratory analysis – Patterns and Relationships

### Focus on Customer orders throughout the week

As shown in Figure 12 in Appendix, 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 almost the same and that metric takes into account that weekends only include data of two days (Figure 13, Appendix).

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 City Throughout the Week** –The general trend is that in every city the peak orders are in the weekends, including Friday. More specifically in City 2 and City 8 have the highest amount of orders on Saturday, and curiously City 4 on Thursday (Figure 14, Appendix).

**Orders by Customer Age Throughout the Week** – The heatmap of these two variables (Figure 15, Appendix) shows that there is a general trend across all ages with orders places on Thursday and Saturday. Also, we see a progression of orders on Sunday as the customer age increases. As mentioned before, for customers above the age of 52 the data is scattered, and we cannot draw any conclusions.

### 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 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 16 in Appendix. To be even more precise we created a line chart graphing the normalized cumulative orders over the course of a day (Figure 13, Appendix). 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 City Throughout the Day -** A distribution of orders throughout the day (Figure 17, Appendix) shows a similar trend between *City 2* and *City 4*, with a one hour shift, having peaks around noon and in the late afternoon. A specular trend is shown by *City 8* with a higher trend in the very early morning and right before noon.

**Orders By Customer Age Throughout the Day** – By the heatmap that we created (Figure 18, Appendix) 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. The highest number of orders are placed between 10am and 12pm and 3pm and 7pm. After the age of 50, the data becomes scattered and 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 19; Figure 20, Appendix)

**Regional Spending per Type of Cuisine** – Each City shows a different trends, both in the type of cuisine and the distribution between them. Namely, City 8 the dominant cuisine is Asian, in City 4 we have Italian and American, and City 4 doesn’t tend to favor any group (*CUI\_Others – is assumed to be a group of unmentioned different cuisines*) (Figure 21 – Radars, Appendix) .

# Conclusion

Our analysis gave us a valuable insight into ABCDEats’s customers and their behavior. We can conclude that customer engagement rises further along the week, reaching its peaks on Thursdays, Fridays, and Saturday, while orders significantly decline on Sundays. This trend is consistent across different cities and age groups, suggesting that promotional efforts could be effectively timed to align with these peak periods. In addition, the general pattern of peak order frequencies is in the morning and evening, with specific times such as 11 AM and 5 PM being particularly popular. Using this information, the company can increase operational efficiency and marketing campaign effectiveness.

Lastly, we wanted to state the importance of understanding regional preferences for the types of cuisines. Each city reveals favoritism towards a specific cuisine, with Asian having the highest orders placed out of all cities. This information can be potentially used to create tailored promotions for each specific City.

# Appendix

A graph with text on it

Description automatically generated

Figure 1 – Description of Missing Values in Raw Data

A pie chart with numbers and a number in the middle

Description automatically generated

Figure 2 – Distribution of Population by Regions Aggregated in Cities by Color

A pie chart with numbers and a few different colored circles

Description automatically generated with medium confidence

Figure 3 – Distribution of Orders Across Cities

A green rectangular object with red lines

Description automatically generated with medium confidence

Figure 4 - Distribution of Customer Age per Region

A green rectangular object with white text

Description automatically generated

Figure 5 - Distribution of Customer Age per City

A screenshot of a graph

Description automatically generated

Figure 6 – Orders by Region and Type of Cuisine

A screenshot of a graph

Description automatically generated

Figure 7 - Orders by City and Type of Cuisine

A group of green and red graphs

Description automatically generated

Figure 8 – Distribution of Age Compared to Features

A table with numbers and a number of objects

Description automatically generated

Table 1 - Description of Numerical Variables

A graph of a bar

Description automatically generated with medium confidence

Figure 9 – Population Count by City

A pie chart with different colored circles

Description automatically generated

Figure 10 – Promotion Distribution

A graph with green rectangular bars

Description automatically generated with medium confidence

Figure 11 – Payment Method Distribution

A screenshot of a computer screen

Description automatically generated

Table 2 – Descriptive Statistics of New Features

# A graph of green rectangular bars Description automatically generated with medium confidence

Figure 12 – Orders by Days of the Week

A graph with lines and a green line

Description automatically generated

Table 13 - Normalized Cumulative Orders Throughout Hours of the Day

A green squares with black text

Description automatically generated

Figure 14 – Orders by City and Day of the Week

A green and white chart

Description automatically generated

Figure 15 – Orders by Age and Day of the Week

# A graph of green rectangular bars Description automatically generated with medium confidence

Figure 16 – Orders by Days of the Week

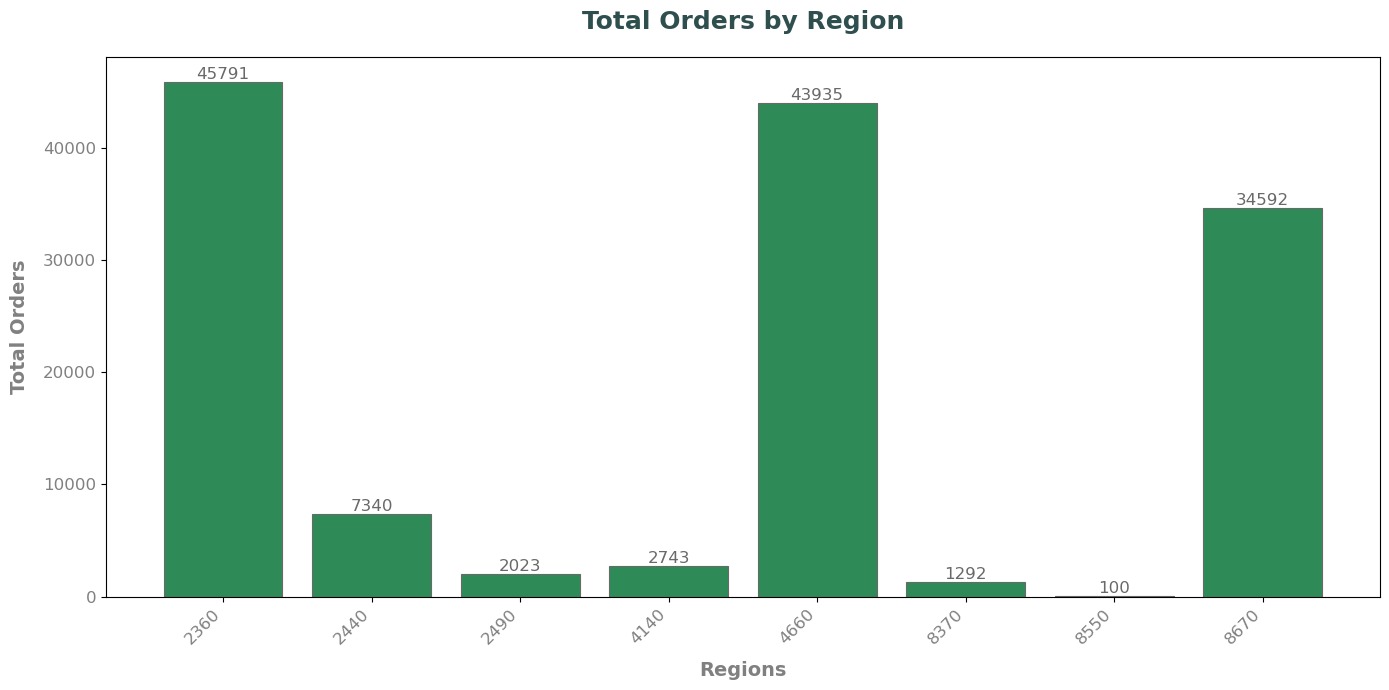


Table 3 – Number of Orders Per Region

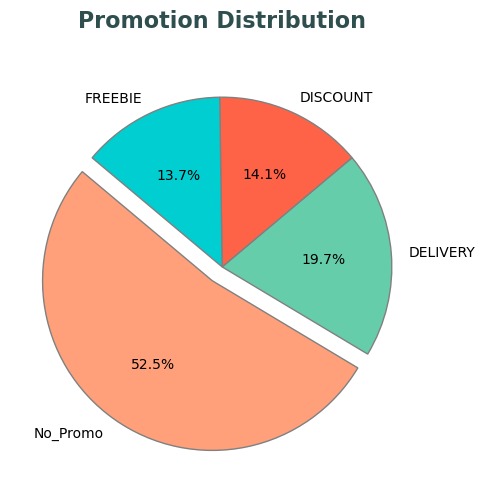


Figure 1 – Promotion Distribution

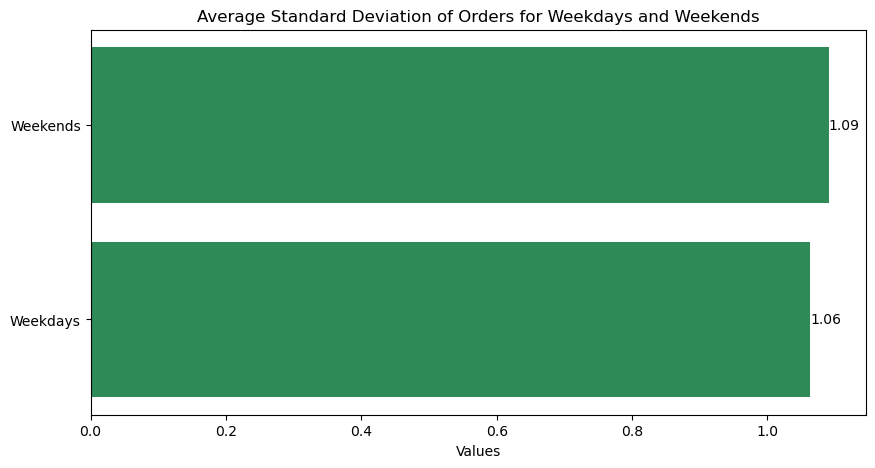


Table 4 – Average Standard Deviation of Orders per Weekdays and Weekends

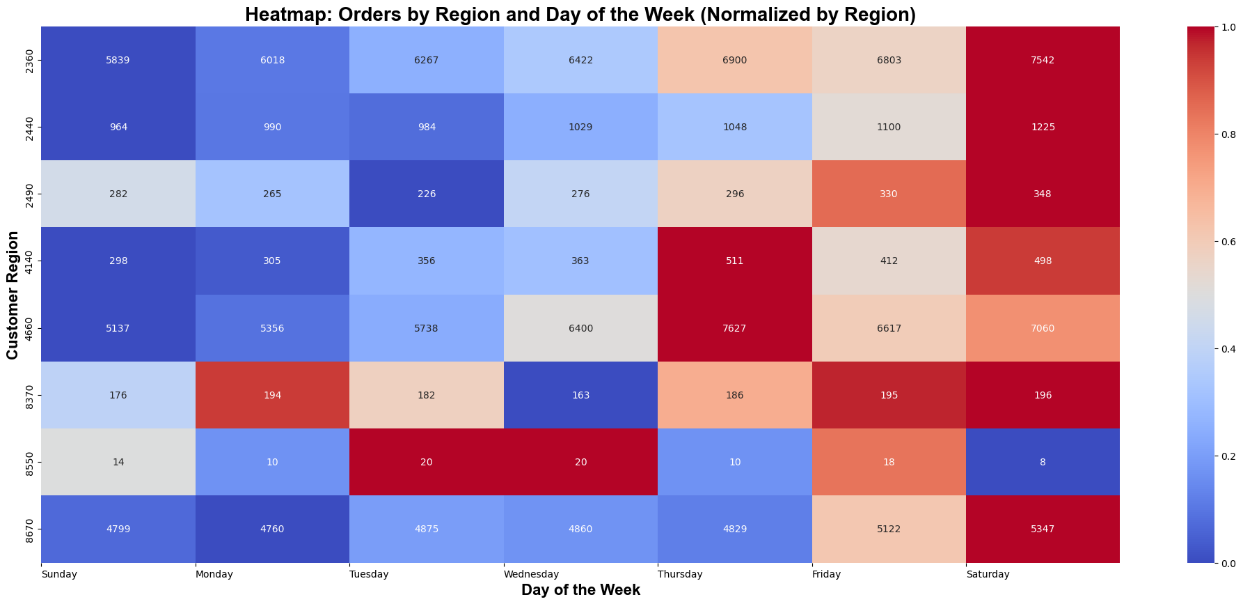


Figure 2 - Orders by Region Throughout the Week

Figure 3 - Orders by Customer Age Throughout the Week

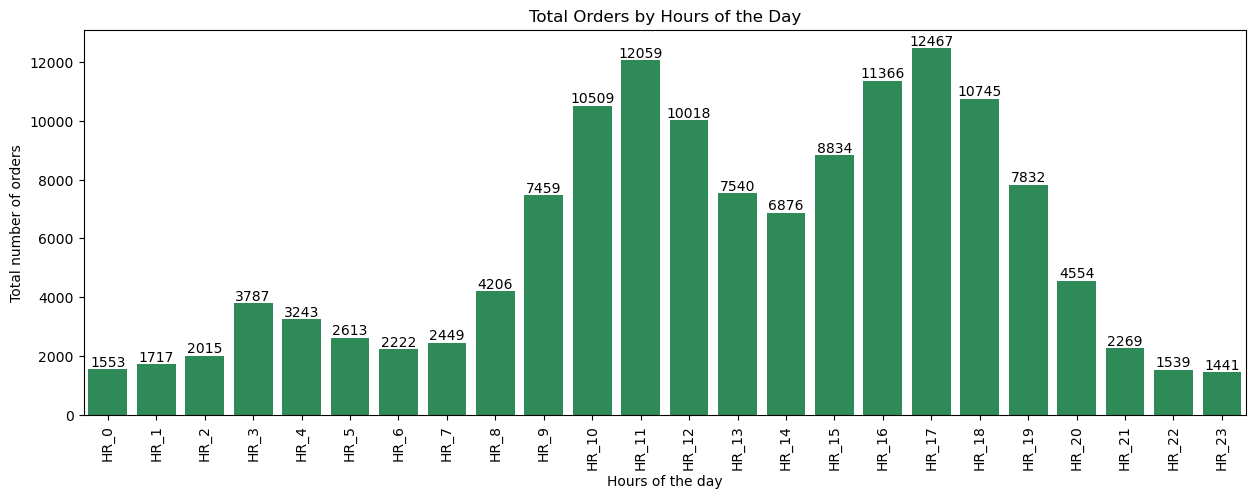


Table 17 – Orders Throughout Hours of the Day

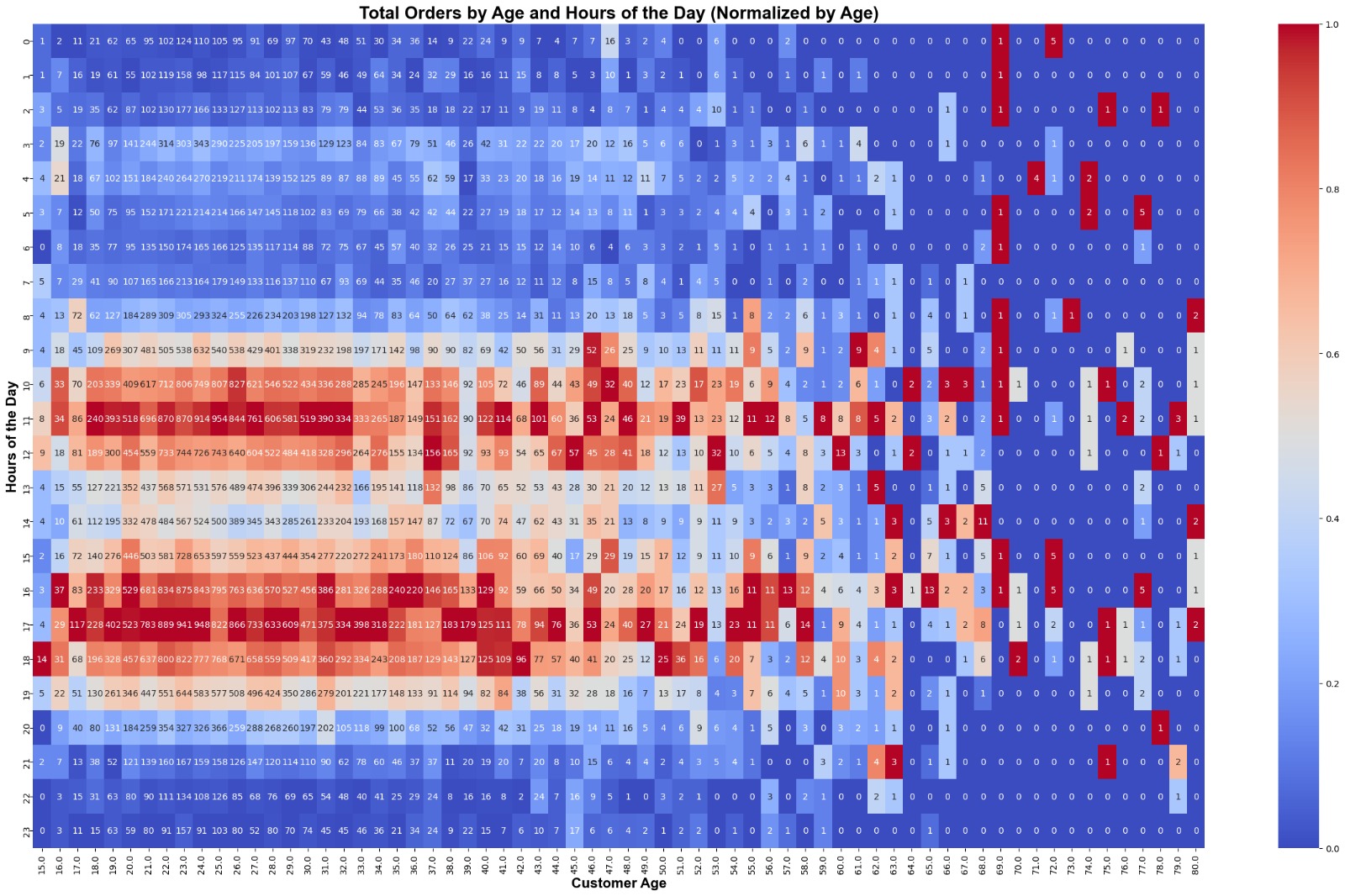


Figure 18 - Orders By Customer Age Throughout the Hours of the Day

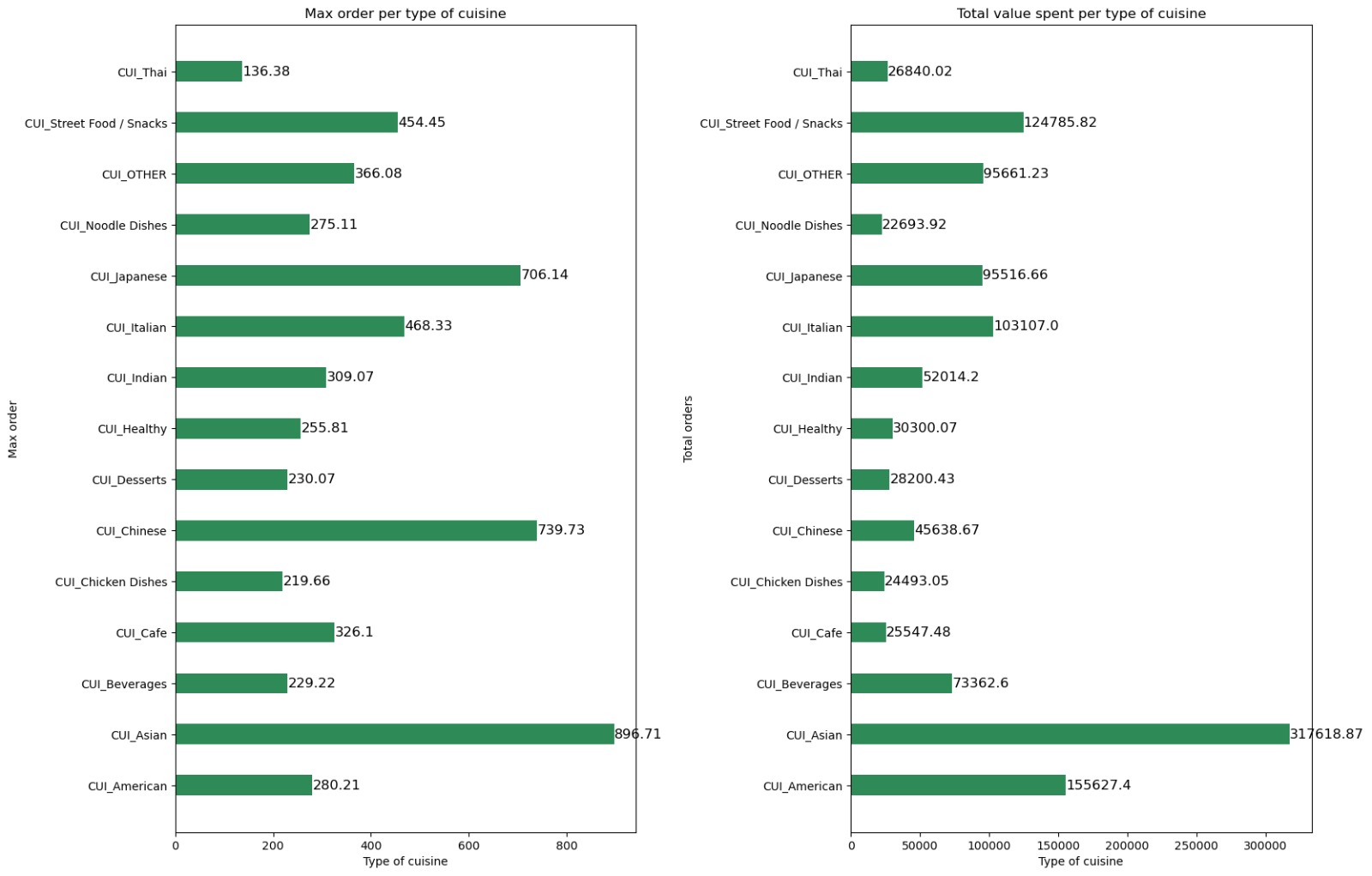


Figure 19 – Highest Order Value per Type of Cuisine Table 20 – Total Value Spent per Type of Cuisine

A graph with a point on the center

Description automatically generated with medium confidence

Figure 21 – Spectrum of Distribution of Orders per Cuisine