DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc. (1st Part)

Group 37

André Silvestre, 20240502

Filipa Pereira, 20240509

Umeima Mahomed, 20240543

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION & OBJECTIVE

Consumers are increasingly selective about the businesses they support, making it essential for companies to understand customer preferences to adapt strategies and drive growth. For ABCDEats Inc., a fictional food delivery service, the goal is to develop a data-driven strategy for various customer segments.

The first part of the project focuses on an in-depth exploratory data analysis (EDA) of customer data collected over 3 months across 3 cities. This includes summarizing statistics, identifying trends and patterns, anomalies and exploring relationships between features. The report will also focus on creating new features to enhance analysis and use visualizations to communicate findings effectively.

2. EXPLORATORY DATA ANALYSIS

To maximize the potential of the provided dataset and achieve the proposed objectives, we chose to apply the *CRISP-DM* methodology, which is defined and summarized in **Annex A**. [1][2] Based on this approach, the first part of the project focuses on the first two phases of this methodology—*Business Understanding* and *Data Understanding*—and begins the third phase, *Data Preparation*, with the creation of new variables. These three phases will be discussed throughout the following subsections, where we perform initial data analysis, individual and multivariate analysis.

2.1. Basic Data Exploration

The dataset comprises 31,888 rows (each representing a unique customer) and 56 columns. A preliminary check for missing values reveals that three variables contain null values – *customer_age*, *first_order* and *HR_0* – though none exceed 4% of the dataset.

In addition, we also checked missing data by rows and observed that most rows (94%) have no missing values, and at most, only 2 variables are missing (0.09% of the rows), indicating a minimal level of incompleteness across individual entries. Based on this, no rows were removed due to missing values.

Overall, the distributions of our variables are right-skewed (except for *last_order*) and display a leptokurtic pattern (**Figure A2**). In the **Figure A1**, we observe a notable relationship between each pair of the variables: *vendor_count, product_count*, and *is_chain*. In the **Figure A3**, while visualization is challenging, we can see that from 1 AM to 6 AM, there is a higher deviation of orders in regions 8670, 8370, and 8550 compared to other areas, regardless of the last payment method used. This is likely due to a more nightly lifestyle or work patterns.

2.2. Consistency Check

Several consistency checks were performed to ensure data integrity and accuracy. We started by checking and removing 13 duplicate rows. We also examined the data types of each column to verify their appropriateness. Overall, the variables are all in the correct data types, but we will examine them in more detail during the individual analysis of each one. Additionally, we checked for any hidden missing values by examining the unique values. It was found that the value '-' appeared in two columns: *customer_region* and *last_promo*.

Further consistency checks included verifying that there were no cases where *first_order* was later than *last_order*, ensuring a consistent order timeline. In addition, we found 18 instances (0.06%) where *vendor_count* exceeded *product_count* and 75 cases (0.24%) where *is_chain* exceeded *product_count*. These discrepancies are likely due to data entry errors, however further investigation may be required to determine the cause.

2.3. In-depth Exploration

This section will explore each variable to find more relevant patterns, giving an outline of possible explanations for them.

2.3.1. Order Count

During the dataset analysis, it became evident that a variable was required to store the number of orders placed by each customer. The order count could be calculated by summing the columns DOW_0 to DOW_6 or HR_0 to HR_23. However, since the columns referring to the hours had missing values, it was deemed more appropriate to use the days of the week.

In total, 139,263 orders were placed, averaging 4.4 orders per customer over three months, with a maximum of 94 by a single customer. The median of 3 suggests the mean is slightly inflated by a few high values, and it also may suggest lack of usage of this platform by the customers, given the timespan. **Figure A4** shows a long, heavy right tail, while **Table A1** confirms this pattern, with a skewness of 4.5 and kurtosis of 37.01, indicating a strongly right-skewed, leptokurtic distribution.

Notably, 138 registered customers had no orders placed. It became evident that this was an anomaly, as these customers had other columns that indicated activity. This flagged a data inconsistency, and the rows in question were removed. In parallel, the 18 cases where *order_count* exceeded *product_count* matched those where *vendor_count* also exceeded *product_count*, and we found no instances of *order_count* being less than either *vendor_count* or *is_chain*.

2.3.2. Customer Region

Given that there are 3 cities, we decided to assume that postal codes within a city or metro area tend to be numerically close to each other, so we group them as follows: **Region 2**: 2360, 2440, 2490; **Region 4**: 4140, 4660 and **Region 8:** 8370, 8550, 8670 (**Figure A5**). This approach, informed by **Table A3**, further supports the division into three regions under the assumption of a well-distributed and balanced consumer base. About 1.4% of the rows have '-' instead of a code which for now we will treat as "Unknown" but in the 2nd part we intend to use imputation techniques to fill them in.

Furthermore, in **Figure A6** and **Table A4**, it is notable that *Region 2* has the highest *order_count* values, followed by *Region 4* and *8*, where the latter has a significant discrepancy since the highest value is 52.

2.3.3. Customer Age

The age distribution (**Fig. A7**) of the customer base is predominantly young, as expected [3], with a median age of 26, which is notably below the global median [4]. The distribution shows a slight right skew, meaning that while the core user base is younger, there is a smaller portion of older users as well. The kurtosis of 4.09 suggests a concentration of ages around the median with some outliers,

indicating a focus on young adults but with some appeal to older customers too. This younger demographic may be drawn to the service due to higher digital literacy and busy lifestyles that reduce time for cooking. The minimum age recorded is 15, which aligns with the policies of some food delivery apps, such as Uber Eats [5] that allow users under 18 to place orders. (**Table A1**)

Ages were then categorized into equal-width ranges – 15-28, 29-41, 42-54, 55-67, 68-80 – for clearer demographic analysis. However, despite these groupings, age did not significantly differentiate user behaviours within the dataset, suggesting other factors may better explain customer trends. [6] (**Table A5**)

2.3.4. Vendor & Product Count

On average, customers order 5.7 products per transaction from a little over three unique vendors. Considering the averages and up to the 75th percentiles, we can infer that most consumers do not use the app frequently, just like we saw in the analysis of the order count variable. Lastly, the maximum counts of 41 vendors and 269 products, along with their respective kurtosis values, suggest the presence of outliers who engage extensively with the platform. (**Table A6** and **Table A7**)

2.3.5. Is Chain

This variable doesn't align with its description, we believe it represents the number of orders placed at chain restaurants. It's better to adjust the metadata rather than convert this variable to binary, as binary would only show if at least one chain restaurant order was placed (1) or none (0).

Despite most consumers placing relatively few orders, the maximum count of 83 suggests that probably some users significantly favoured chain restaurants, which motivated us to calculate the percentage of chain orders in each customer's order count. We discovered that 41.4% of customers ordered between 90-100% of their orders from chain restaurants, indicating a strong preference. It is important to note that cases where customers made only one order, and it was from a chain restaurant, contribute significantly to the 100% figure (**Table A8**). For 19.04% of customers, where *is_chain* exceeds *vendor_count*, it implies they repeat orders from at least one chain restaurant. Likewise, for 9.45%, where the difference between *order_count* and *is_chain* exceeds *vendor_count*, they repeat orders from at least one non-chain restaurant.

2.3.6. First & Last Order

A significant portion of consumers placed their first order within the first month and the same happened with the last order within the final month, showing a symmetrical pattern in the distribution, with a left-tail for first orders and a right-tail for last orders (**Figure A8** and **Figure A9**). To better understand these variables, we also created *days_between_orders* and *days_between_orders_per_order* (**Figure A10**), where the first measures the time between the first and last order and the second uses this last variable and divides it by the *order_count.* In this way, the frequency of orders can be calculated.

While looking at **Table A1** we can conclude that on average, customers placed orders every 8 to 9 days and at least 22,2% (**Figure A10**) of consumers placed only one order, as indicated by having the same first and last order date. When observing the mean value for the *days_between_orders_per_order*, we can notice that it is relatively low given the dataset's characteristics, likely due to the high number of

customers who placed only 1 order. It also indicates to ABCDEats that there is room for improvement in their marketing campaigns, customer attraction, retention and a need for better communication with consumers (feedback).

2.3.7. Last Promo & Payment Method

In last promo we considered that the presence of *'-'* means *'NO PROMO'*, thus, around 53% of the customers didn't use a promotion on their last order. The most popular promotion was free delivery (20%), followed by discounts and freebies, each accounting for about 14%. (**Figure A11**) Furthermore, the preferred payment method for the last order was by card, holding around 63% of customers (**Table A9**). The **Figure A11** and **Figure A12** (where all last promo methods were grouped a binary variable: 'Promo' or 'No Promo') suggest that customers with lower order counts used a promotion on their last order. This may indicate that these customers placed an order only because they received a promotion, which could help explain the company's challenges with retaining certain customers.

2.3.8. CUI

There are 15 unique cuisine categories. In every cuisine category, some consumers did not make any purchases within the three-month observation period. In fact, for each type of cuisine other than *CUI_Asian* and *CUI_American*, at least 75% of consumers in our database did not spend on that type of food (**Table A1** and **Figure A13**). To aid the analysis of this variable we also created *CUI_Total_Amount_Spent, CUI_Most_Spent_Cuisine* and *CUI_Avg_Amount_Spent* per customer, where we found out that the most spent cuisines were the Asian and American (**Table A10**), perhaps indicating a preference or higher prices.

To explore if customers showed high variety-seeking behaviour we created the metric *CUI_Total_Food_Types*. Overall, we observed at **Table A11** and **Figure A13** low percentages, indicating a less food-enthusiastic customer base. It's important to note that the limited number of orders may constrain the variety in food choices. Supporting this, a query of *CUI_Total_Food_Types* revealed that 14.74% of customers who ordered more than once chose only one type of food, and among those with three or more orders, 6.5% still ordered just one type. This trend suggests a preference for consistency in food choices among part of the customer base. This conclusion is further enriched with **Figure A14**.

While looking at **Figure A16**, we can notice a similar pattern across all cuisines, indicating that the type of cuisine does not influence the last promotion that the customer had.

2.3.9. DOW - Days

These variables have similar distributions. For each day of the week, at least 50% of the dataset did not make any orders. On average, spending is highest on Saturdays, while Sundays see the lowest average spending, although the difference is insignificant. (**Table A12**)

2.3.10. HR - Hours

Customers, generally, spend the most on the application at 5 PM, followed closely by 11 AM, while other hours see significantly lower order volumes. This is likely due to post-work convenience and late-night cravings. Targeted promotions and operational adjustments during these peak times could boost

engagement and efficiency. Interestingly, there is at least one customer who made 52 orders at 8 AM over these three months. (**Table A1** and **Figure A17**)

2.4. Relationships between Features

In this phase, we calculate the correlation between features to understand underlying patterns and dependencies.

2.4.1. Numerical VS Numerical

To calculate the correlations of these features, we used the Pearson's Correlation, where we can highlight the correlation of *product_count* and *vendor_count* (0.83), *vendor_count* and *is_chain* (0.76) and *product_count* and *is_chain* (0.83). This goes in accordance with what was stated before, since the *vendor_count* and *is_chain* refer to the food establishments. In the case of *product_count* and *vendor_count*, this can be due to people buying one food item/menu from one establishment, and the more products bought, the more vendors chosen. The same goes for *product_count* and *is_chain*. (**Figure A18**)

2.4.2. Categorical VS Categorical

In this section, we have 3 eligible features: *customer_region*, *last_promo* and *payment_method*, where none of them had a strong correlation (using the Cramer's V correlation test). (**Figure A18**)

2.4.3. Numerical VS Categorical

In this last section, it was calculated, using Eta Squared, the correlation between numerical and categorical features, having reached very low correlation values. (**Figure A18**)

3. FINAL CONSIDERATIONS & NEXT STEPS

In this phase, we've gathered several key insights, primarily recognizing that ABCDEats has a low variety-seeking customer base with modest order volume over a three-month period. Additionally, as anticipated with technology adoption, the customer demographic skews younger.[1]

Next, we'll further analyse this dataset by imputing missing values in fields like *customer_age*, *first_order*, and *HR_0*, where, in this last case, we can use the days of the week to impute deterministically these missing values. We will also consider PCA to reduce dimensionality and create new features to refine clustering accuracy in features such as *CUI*, *DOW* and *HR*.

Then we will use clustering methods like K-means, DBSCAN, or hierarchical clustering to group customers by behaviour, allowing us to identify unique customer segments and support targeted marketing. Finally, we'll evaluate the performance of these clustering models to ensure robust results.

[1] Throughout this project, we utilized AI tools (e.g. *ChatGPT* and *Github Copilot*) to assist in code development and to help summarize repetitive conclusions derived from the data analysis. However, all AI-generated information was carefully reviewed and validated by the team to ensure accuracy and relevance.

BIBLIOGRAPHICAL REFERENCES

[1] Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly.

[2] Waqar, M. (2023, October 26). *Unlocking CRISP-DM: Your Path to Data Science Success*. Medium. https://medium.com/@mwaqarbatalvi/mastering-the-crisp-dm-framework-your-path-to-successful-data-science-projects-56f15d6f4c54

[3] Start.io. (2021, October 14). *Start.io | Food Delivery App Users: What Mobile Data Tells Us*. Start.io - a Mobile Marketing and Audience Platform. https://www.start.io/blog/food-delivery-app-users-what-mobile-data-tells-us-about-foodies-and-how-to-catch-your-audiences-attention/

[4] Ritchie, H., & Roser, M. (2019, September 20). *Age Structure*. Our World in Data. https://ourworldindata.org/age-structure

[5] Uber. (2024). *Keep Safe*. Uber. https://www.uber.com/us/en/safety/uber-community-guidelines/keep-safe/

[6] Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.

APPENDIX A

Table A1 – Descriptive statistics for the numerical features.

Variables	count	mean	std	min	25%	50%	75%	max	skew	kurtosis
customer_age	31010	27,51	7,16	15	23	26	31	80	1,56	4,09
vendor_count	31737	3,12	2,77	1	1	2	4	41	3,03	15,66
product_count	31737	5,69	6,96	0	2	3	7	269	5,71	92,91
is_chain	31737	2,83	3,98	0	1	2	3	83	4,98	48,33
first_order	31631	28,4	24,07	0	7	22	44,5	90	0,76	-0,49
last_order	31737	63,76	23,18	0	49	70	83	90	-0,94	-0,12
CUI_American	31737	4,9	11,67	0	0	0	5,71	280,21	5,43	54,57
CUI_Asian	31737	10	23,6	0	0	0	11,86	896,71	6,64	108,37
CUI_Beverages	31737	2,31	8,49	0	0	0	0	229,22	8,13	107,50
CUI_Cafe	31737	0,8	6,44	0	0	0	0	326,1	19,10	571,57
CUI_Chicken Dishes	31737	0,77	3,67	0	0	0	0	219,66	15,81	575,59
CUI_Chinese	31737	1,44	8,21	0	0	0	0	739,73	34,73	2477,72
CUI_Desserts	31737	0,89	5,27	0	0	0	0	230,07	14,32	380,72
CUI_Healthy	31737	0,95	5,84	0	0	0	0	255,81	14,85	377,88
CUI_Indian	31737	1,64	7,46	0	0	0	0	309,07	10,82	218,73
CUI_Italian	31737	3,25	11,27	0	0	0	0	468,33	9,51	191,25
CUI_Japanese	31737	3,01	10,2	0	0	0	0	706,14	15,64	768,33
CUI_Noodle Dishes	31737	0,72	4,55	0	0	0	0	275,11	18,54	685,05
CUI_OTHER	31737	3,01	9,79	0	0	0	0	366,08	8,68	148,36
CUI_Street Food / Snacks	31737	3,93	15,58	0	0	0	0	454,45	7,77	100,16

CUI_Thai	31737	0,85	4,44	0	0	0	0	136,38	10,77	183,05
DOW_0	31737	0,56	1,02	0	0	0	1	16	3,32	18,69
DOW_1	31737	0,57	1,05	0	0	0	1	17	3,63	24,29
DOW_2	31737	0,59	1,05	0	0	0	1	15	3,29	19,36
DOW_3	31737	0,62	1,07	0	0	0	1	17	3,25	19,21
DOW_4	31737	0,68	1,09	0	0	0	1	16	3,03	16,68
DOW_5	31737	0,66	1,07	0	0	0	1	20	3,17	19,52
DOW_6	31737	0,71	1,17	0	0	0	1	20	3,22	18,66
HR_0	30573	0	0	0	0	0	0	0	0,00	0,00
HR_1	31737	0,05	0,32	0	0	0	0	14	10,93	228,25
HR_2	31737	0,06	0,35	0	0	0	0	12	10,84	206,49
HR_3	31737	0,12	0,5	0	0	0	0	11	6,96	75,09
HR_4	31737	0,1	0,44	0	0	0	0	14	8,00	119,16
HR_5	31737	0,08	0,36	0	0	0	0	7	6,15	53,19
HR_6	31737	0,07	0,33	0	0	0	0	8	6,94	73,52
HR_7	31737	0,08	0,38	0	0	0	0	15	11,61	282,55
HR_8	31737	0,13	0,64	0	0	0	0	52	27,68	1706,13
HR_9	31737	0,23	0,73	0	0	0	0	23	6,77	92,52
HR_10	31737	0,33	0,89	0	0	0	0	25	6,48	84,26
HR_11	31737	0,38	0,96	0	0	0	0	36	6,92	115,68
HR_12	31737	0,32	0,84	0	0	0	0	26	6,25	80,87
HR_13	31737	0,24	0,64	0	0	0	0	14	4,73	41,07
HR_14	31737	0,22	0,6	0	0	0	0	13	4,82	41,85
HR_15	31737	0,28	0,74	0	0	0	0	23	5,71	74,07
HR_16	31737	0,36	0,88	0	0	0	0	22	4,61	39,06

HR_17	31737	0,39	0,95	0	0	0	0	20	4,93	44,94
HR_18	31737	0,34	0,9	0	0	0	0	24	5,77	64,06
HR_19	31737	0,25	0,8	0	0	0	0	35	8,70	186,34
HR_20	31737	0,14	0,59	0	0	0	0	36	13,94	550,28
HR_21	31737	0,07	0,35	0	0	0	0	11	8,52	129,23
HR_22	31737	0,05	0,3	0	0	0	0	15	13,11	365,01
HR_23	31737	0,05	0,28	0	0	0	0	13	12,21	301,37
order_count	31737	4,39	5,09	1	2	3	5	94	4,54	37,01
days_between_orders	31631	35,57	29,39	0	4	34	62	90	0,20	-1,37
days_between_orders_per_order	31631	8,51	7,89	0	1,5	7,18	12,75	44,5	1,12	1,34
last_promo_bin	31737	0,47	0,5	0	0	0	1	1	0,10	-1,99
CUI_Total_Amount_Spent	31737	38,46	46,44	0,37	13,02	24,2	45,18	1418,33	4,81	53,83
CUI_Avg_Amount_Spent	31737	10,31	7,86	0,37	5,07	8,22	12,92	104,32	2,44	10,98

Table A2 – Descriptive statistics for the categorical features.

Variables	count	unique	top	freq
customer_id	31737	31737	1b8f824d5e	1
last_promo	31737	4	NO PROMO	16693
payment_method	31737	3	CARD	20099
customer_region_buckets	31737	4	2	10757
CUI_Most_Spent_Cuisine	31737	15	Asian	7021

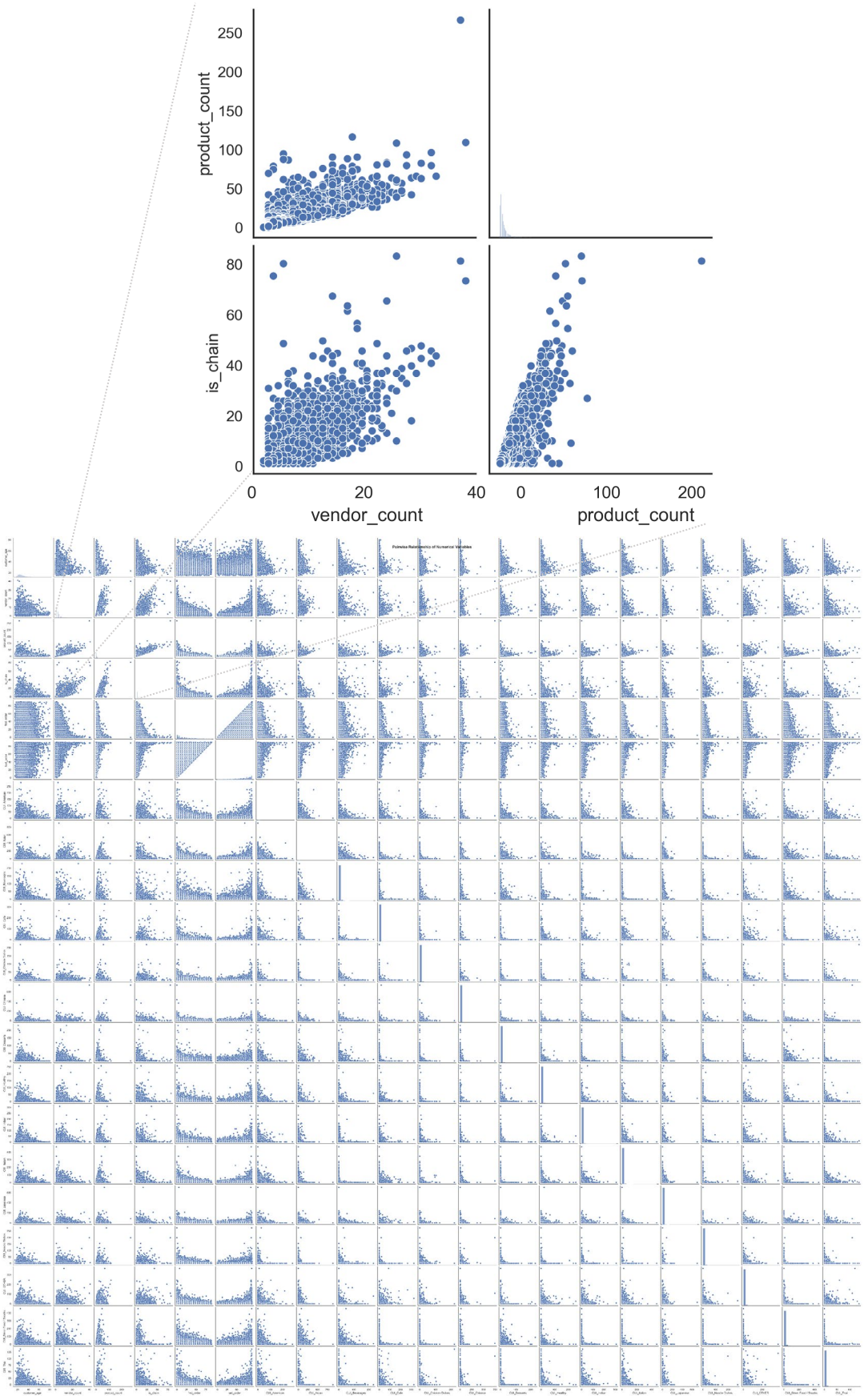

Figure A1 – Pairwise relationship of numerical variables.

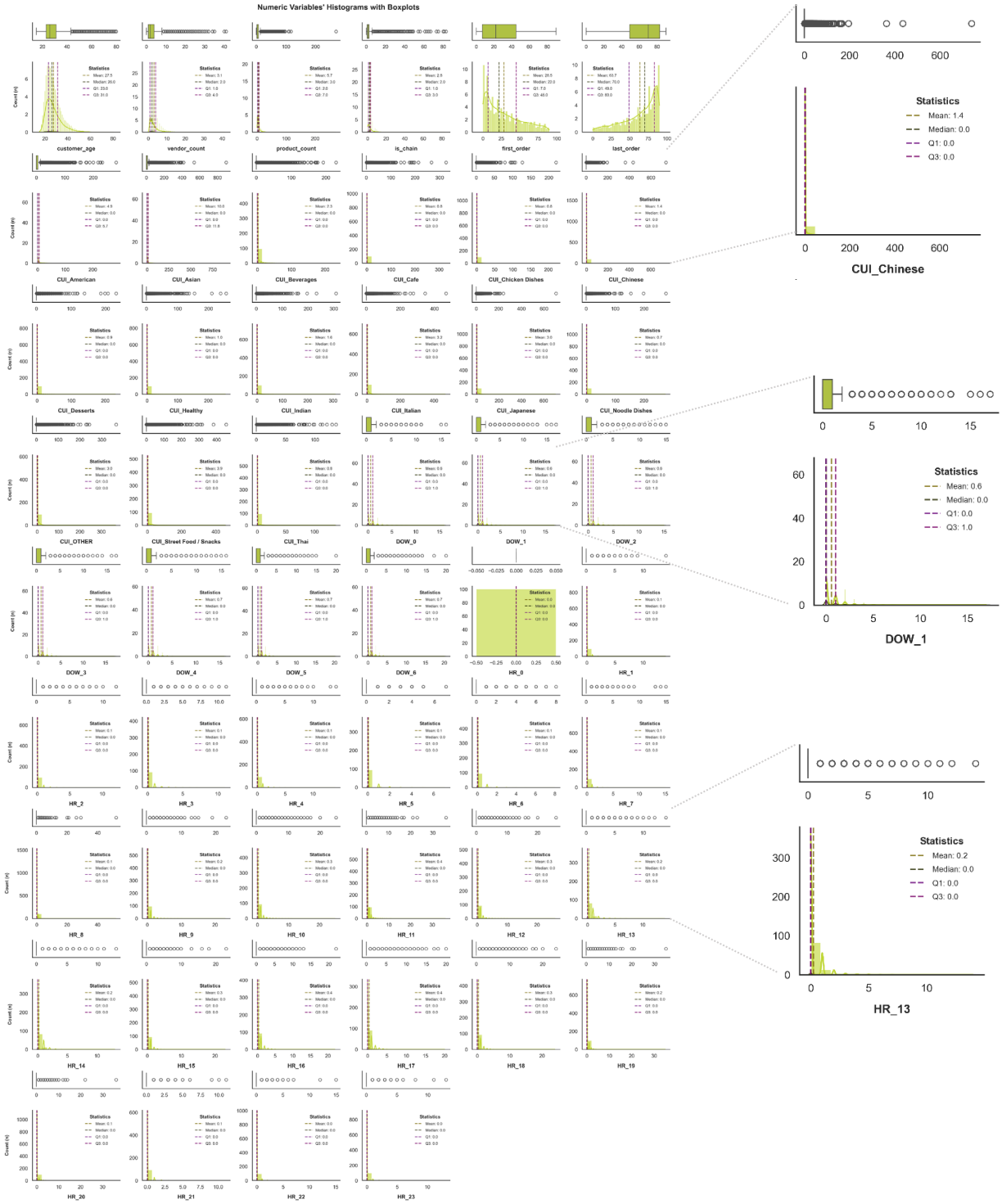

Figure A2 – Numeric Variables' Histograms, KDEs and Boxplots.

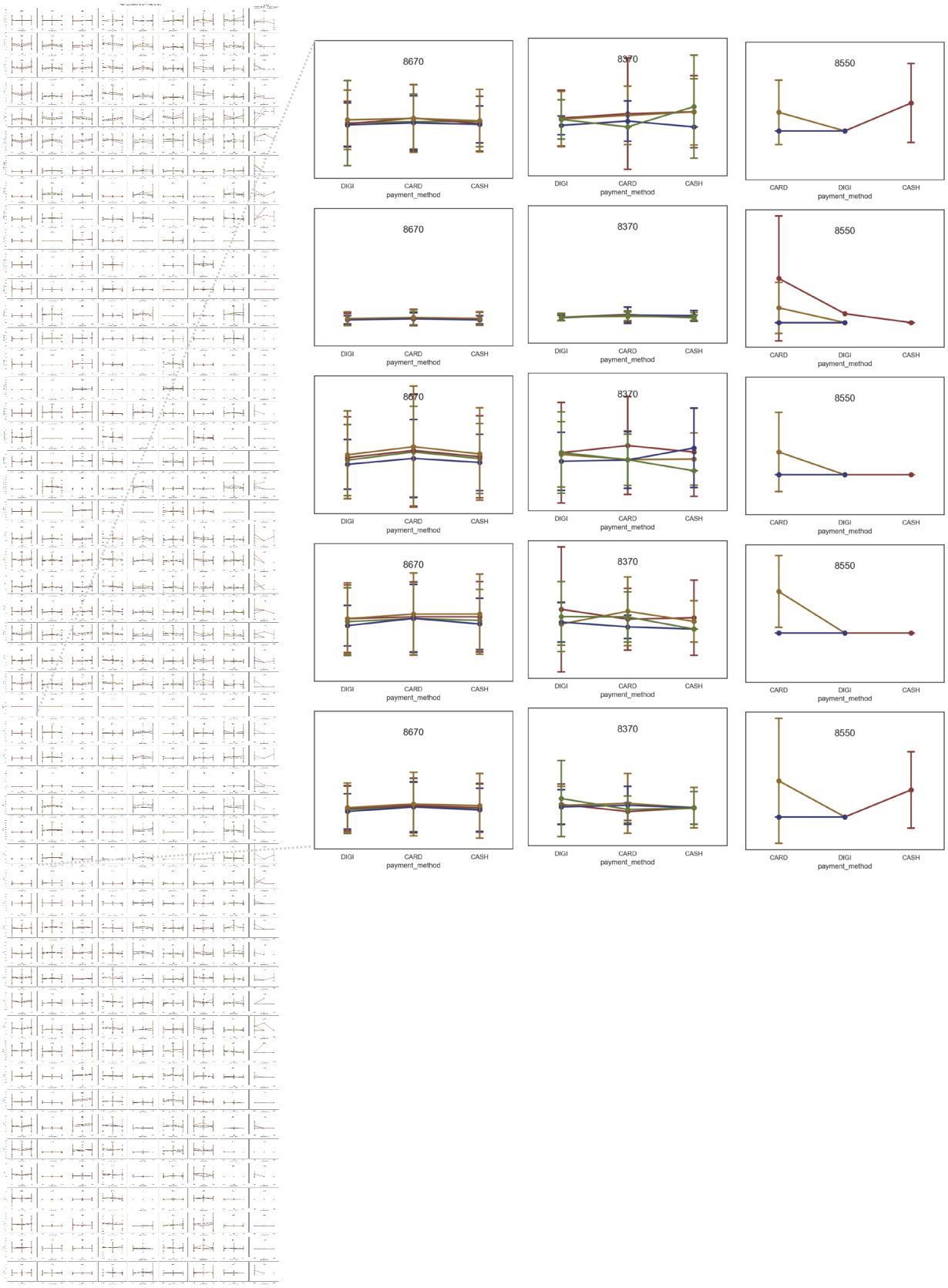

Figure A3 – Three-way ANOVA for each metric variable by last_promo, customer_region and payment_method.

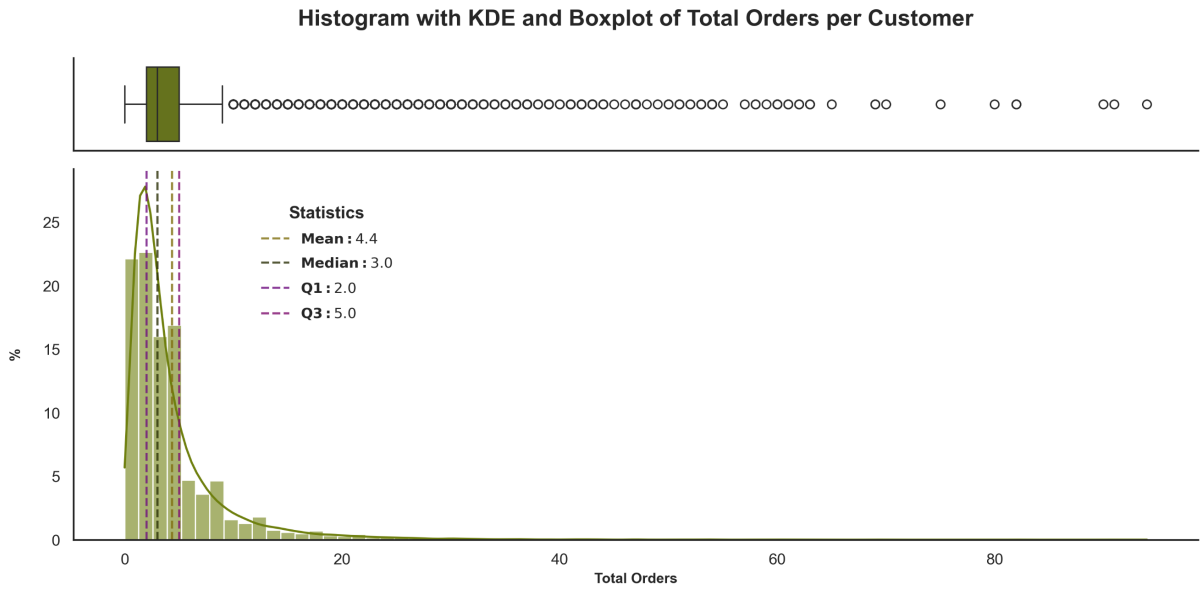

Figure A4 – Histogram, KDE and Boxplot of Total Orders per Customer.

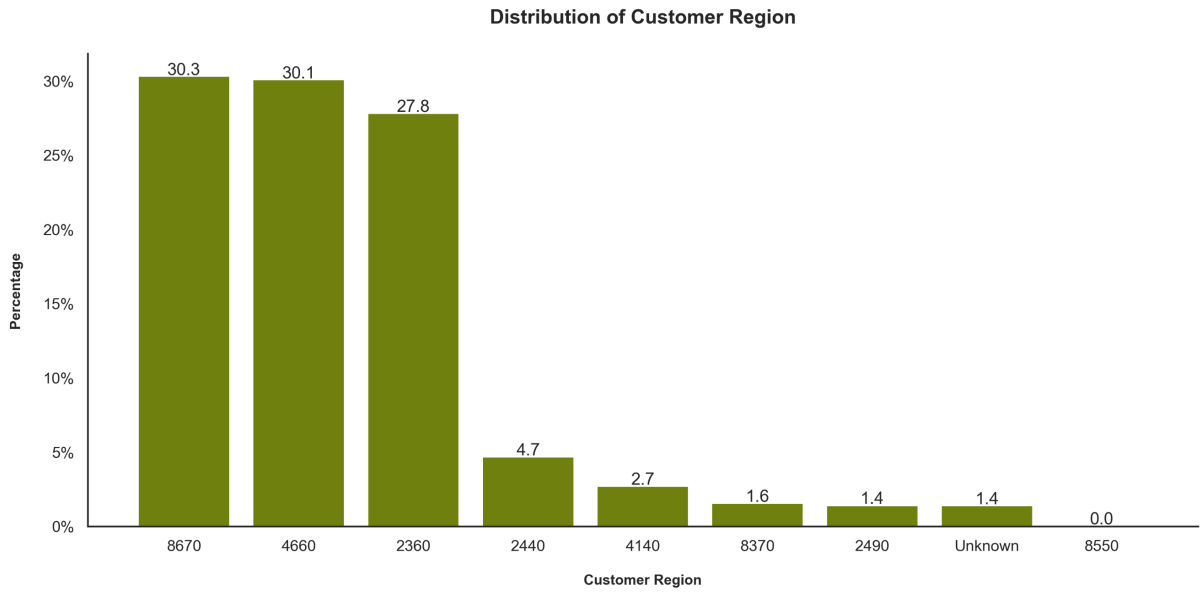

Figure A5 – Bar chart of the Customer Region.

Table A3 – Absolute and Relative Frequency Table of *Customer_Region* and *Customer_Region_Buckets*.

	Customer Region	n		%	
2	**2360**	8829	10757	27,82	33,89
	2440	1483		4,67	
	2490	445		1,4	
4	**4140**	857	10407	2,7	32,79
	4660	9550		30,09	
8	**8370**	495	10131	1,56	31,92
	8550	13		0,04	
	8670	9623		30,32	
U	**Unknown**	442		1,39	

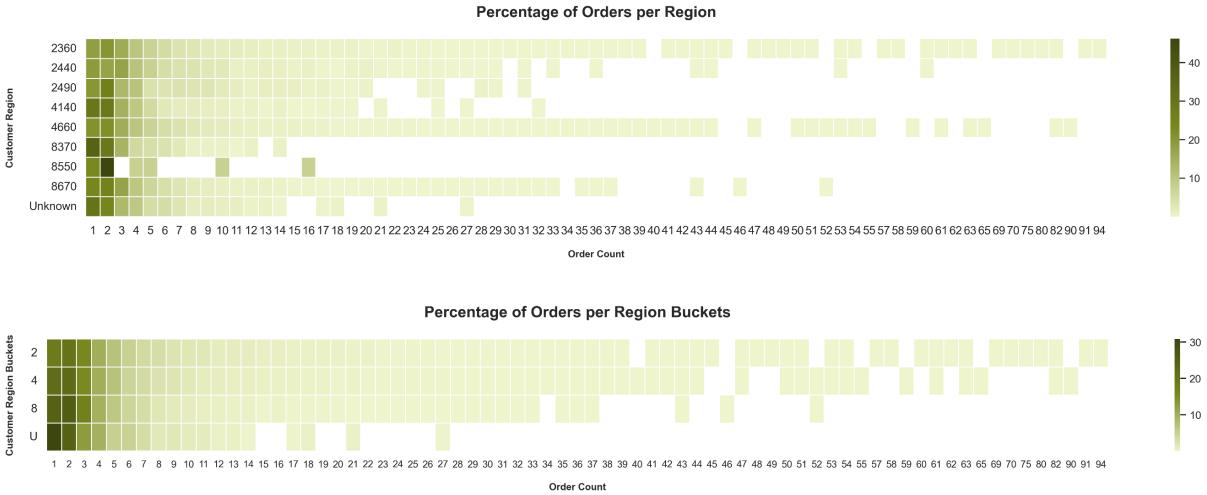

Figure A6 – Heatmap with percentage of Orders per Region and Region Buckets.

Table A4 – Percentage of Order Counts across Customer Region Buckets.

	1	2	3	4	5	6	7	8	9	+10	Total
2	18,7	20,3	15,8	10,2	7,3	5,2	3,8	3,2	2,3	13,1	100
4	21,9	22,4	15,3	10,1	7,2	4,6	3,6	2,6	2,2	10,2	100
8	24,5	25,6	17,3	9,7	6,4	4,3	3,3	2,0	1,5	5,3	100
U	30,8	24,4	12,7	9,7	5,2	5,0	3,4	1,6	1,8	5,4	100

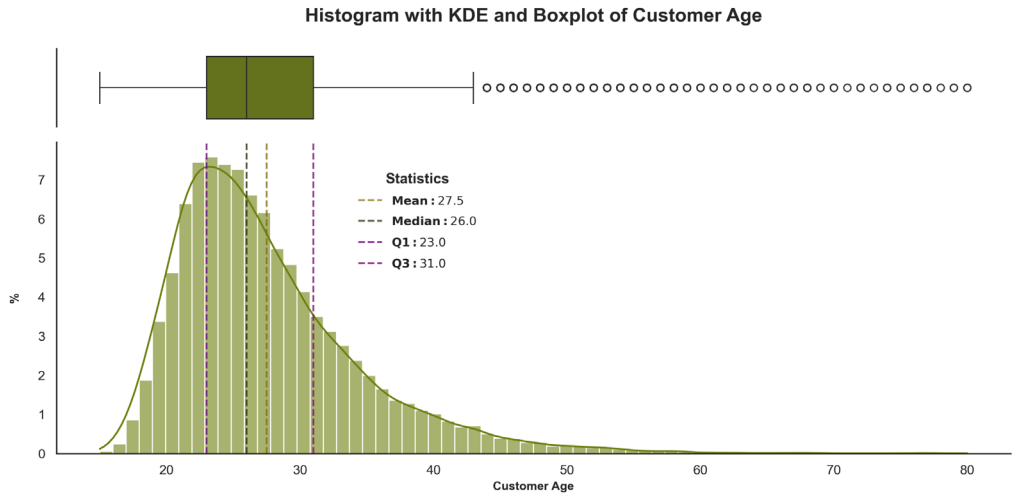

Figure A7 – Histogram, KDE and Boxplot of Customer Age.

Table A5 – Absolute and Relative Frequency Table of *Customer_Age_Group.*

Customer Age Group	n	%
15-28	20202	63,65
29-41	9300	29,3
42-54	1322	4,17
55-67	149	0,47
68-80	37	0,12

Table A6 – Absolute and Relative Frequency Table of *Vendor_Count.*

Vendor Count	n	%	% cumulative
1	9059	28,54	28,54
2	8547	26,93	55,47
3	5173	16,30	71,77
4	2984	9,40	81,18
5	1875	5,91	87,08
6	1166	3,67	90,76
7	816	2,57	93,33
8	594	1,87	95,20
9	396	1,25	96,45
+10	1127	3,55	100,00

Table A7 – Absolute and Relative Frequency Table of *Product_Count.*

Product Count	n	%	% cumulative
0	18	0,06	0,06
1	4690	14,78	14,83
2	6282	19,79	34,63
3	4970	15,66	50,29
4	3404	10,73	61,01
5	2382	7,51	68,52
6	1852	5,84	74,35
7	1359	4,28	78,64
8	1086	3,42	82,06
9	923	2,91	84,97
+10	4771	15,03	100

Table A8 – Absolute and Relative Frequency Table of the column *is_chain* in relation to the *order_count.*

Percentage Group	%
[0-10]%	19,20
]10-20]%	2,03
]20-30]%	2,68
]30-40]%	5,72
]40-50]%	11,19
]50-60]%	2,66
]60-70]%	6,03
]70-80]%	5,69
]80-90]%	3,44
]90-100]%	41,36

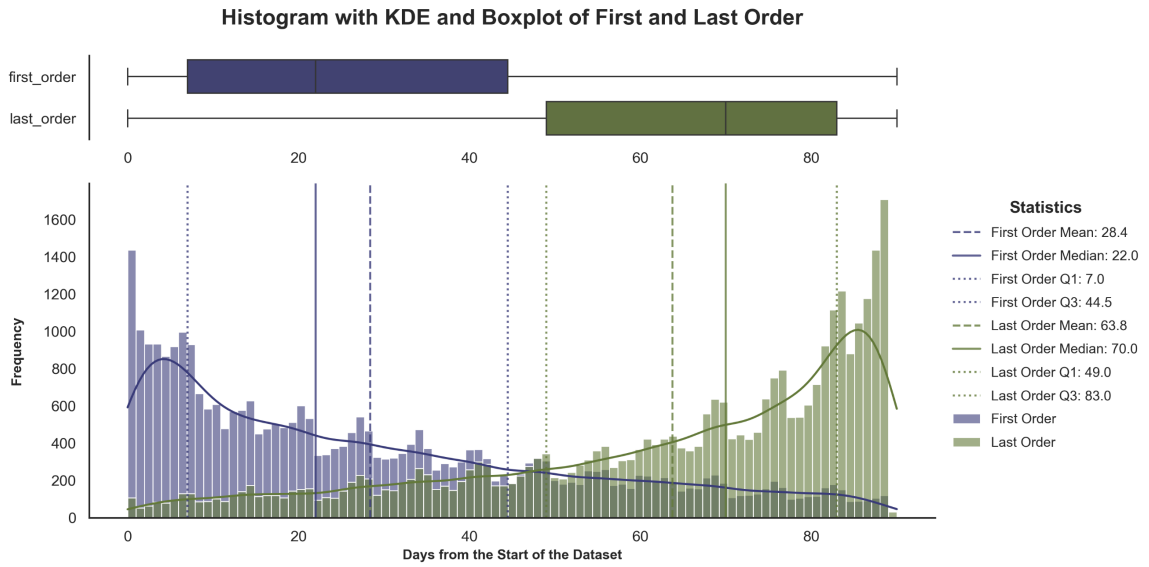

Figure A8 – Histogram with KDE and Boxplot of First and Last Order.

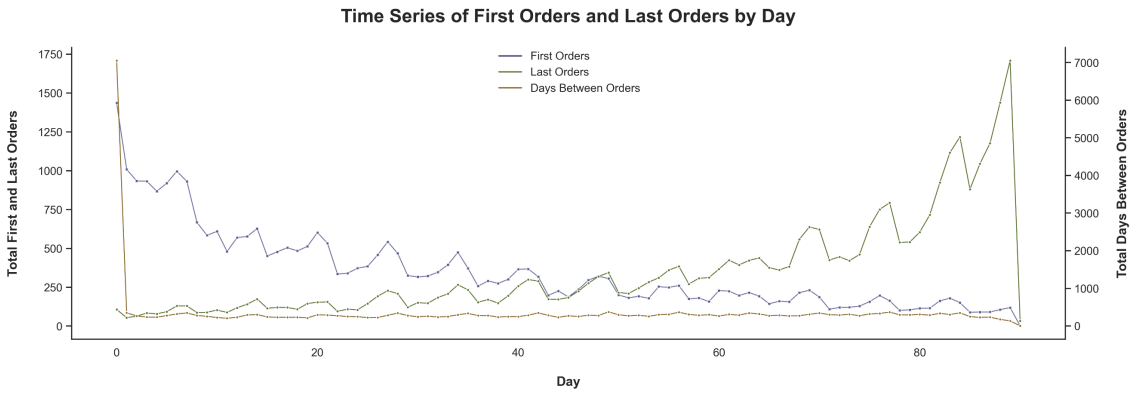

Figure A9 – Time Series of First Orders and Last Orders by Day.

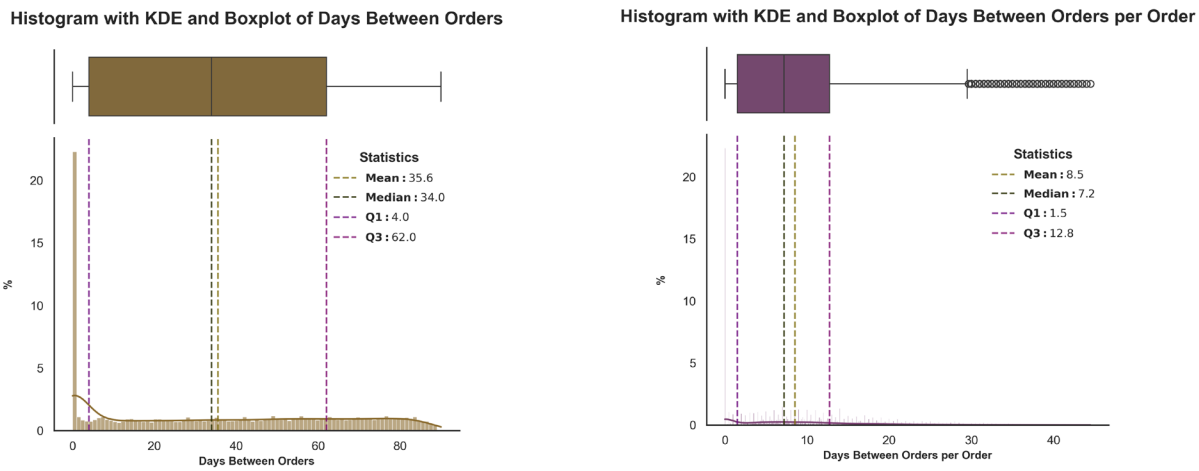

Figure A10 – Histograms with KDE and Boxplots of Days Between Orders and Days Between Orders per Order.

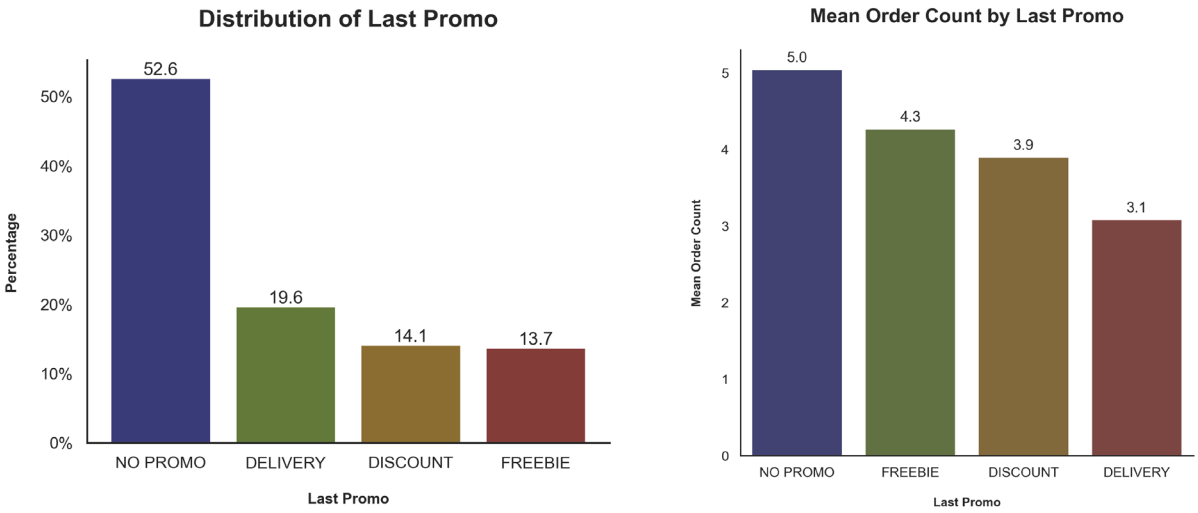

Figure A11 – Bar chart of the Customer Region Percentage and Mean Order Count.

Table A9 – Absolute and Relative Frequency Table of Payment Method.

Payment Method	n	%
CARD	20099	63,33
DIGI	6060	19,09
CASH	5578	17,58

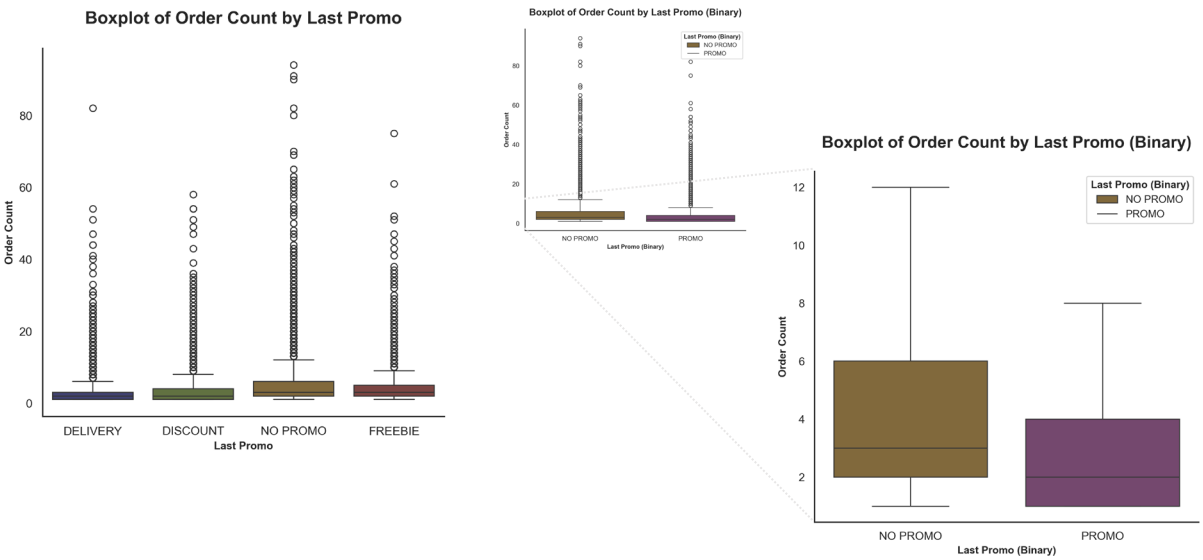

Figure A12 – Box Plot of Order Count by Last Promo and Last Promo Binary (with and without outliers).

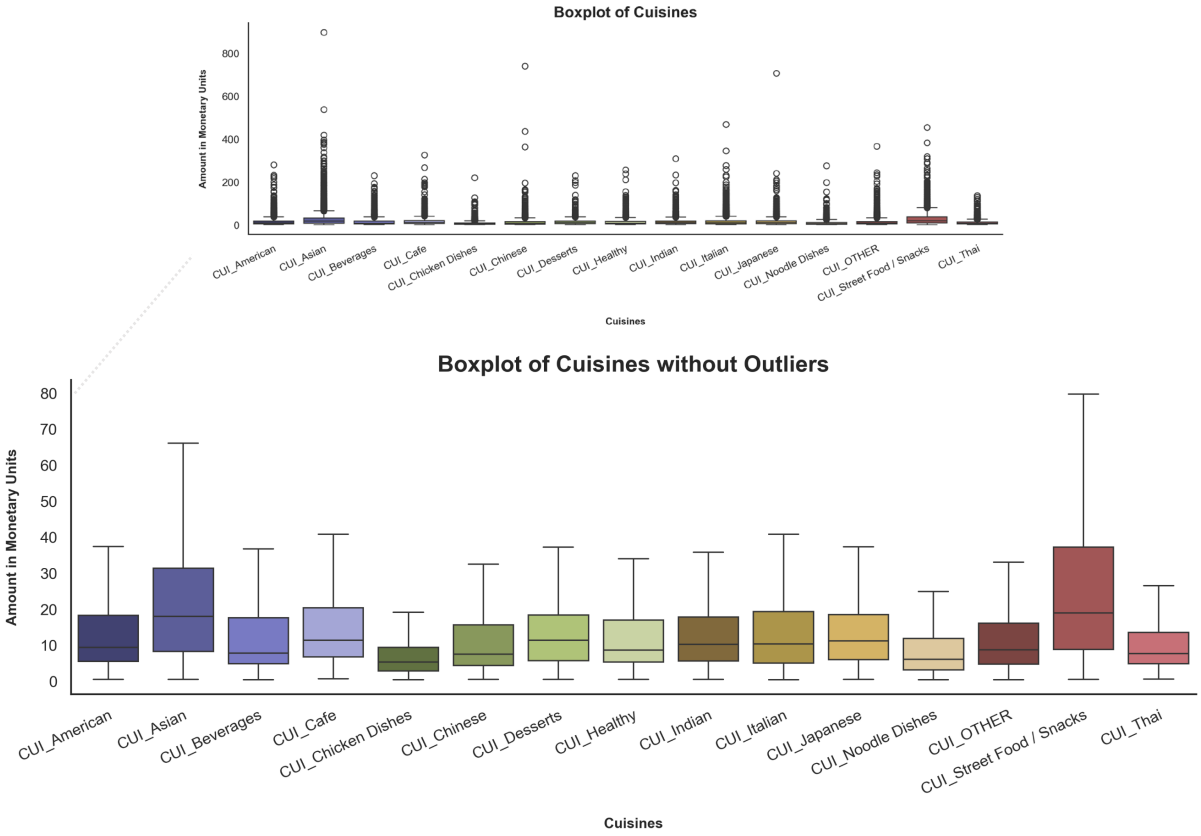

Figure A13 – Boxplot of Cuisines with and without outliers.

Table A10 – Absolute and Relative Frequency Table of Most Spent Cuisine.

Cuisine	n	%
Asian	7021	22,12
American	4170	13,14
OTHER	3134	9,87
Italian	3059	9,64
Japanese	2384	7,51
Street Food / Snacks	2289	7,21
Beverages	2058	6,48
Indian	1689	5,32
Chinese	1330	4,19
Chicken Dishes	1132	3,57
Noodle Dishes	886	2,79
Thai	754	2,38
Healthy	723	2,28
Desserts	621	1,96
Cafe	487	1,53

Table A11 – Absolute and Relative Frequency Table of Total Number of Different Types of Food.

Number of Different Types of Food	n	%	% Cumulative
1	11587	36,51	36,51
2	9804	30,89	67,4
3	5177	16,31	83,71
4	2564	8,08	91,79
5	1266	3,99	95,78
6	677	2,13	97,91
7	359	1,13	99,04
8	177	0,56	99,6
9	68	0,21	99,81
10	34	0,11	99,92
11	20	0,06	99,98
12	3	0,01	99,99
13	1	0	99,99

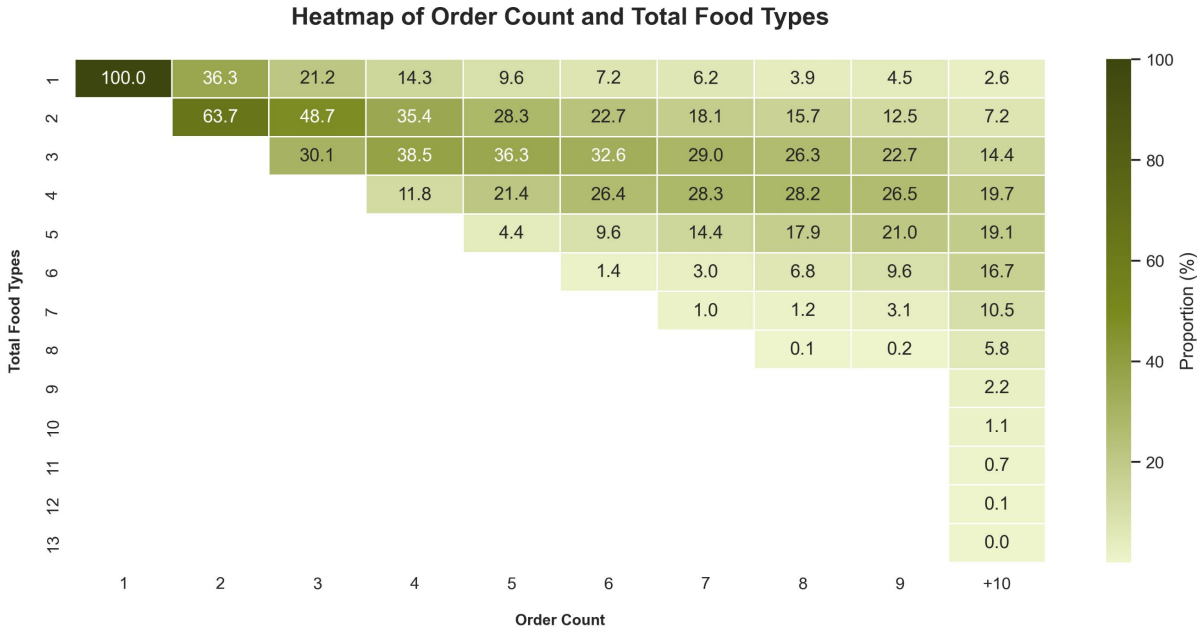

Figure A14 – Heatmap of Order Count and Total Food Types.

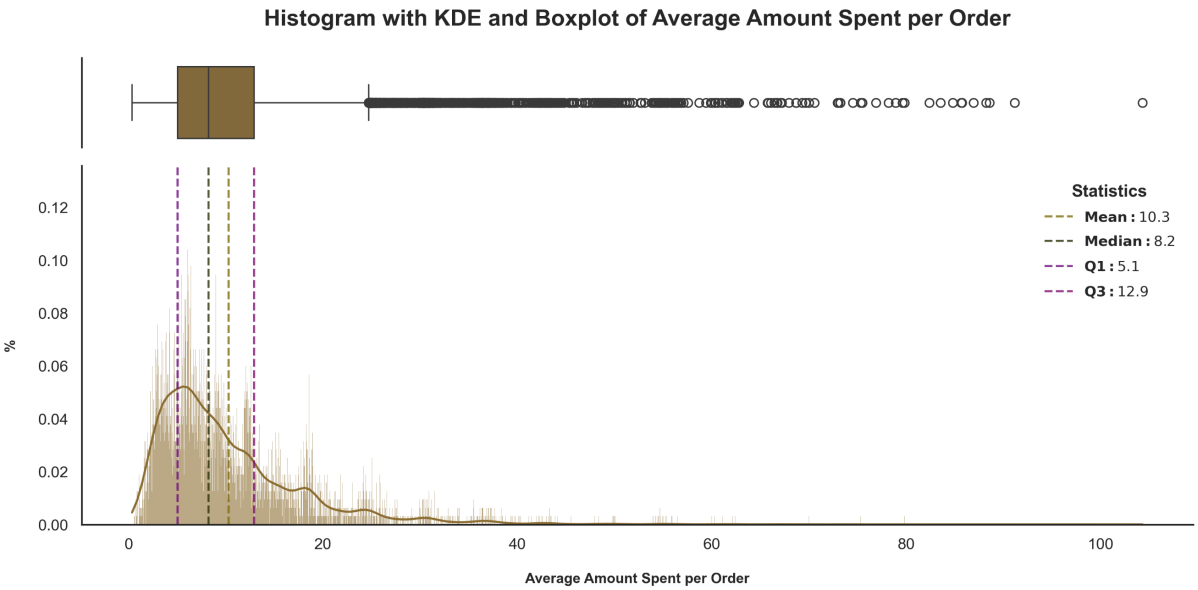

Figure A15 – Histogram with KDE and Boxplot of Average Amount Spent per Order.

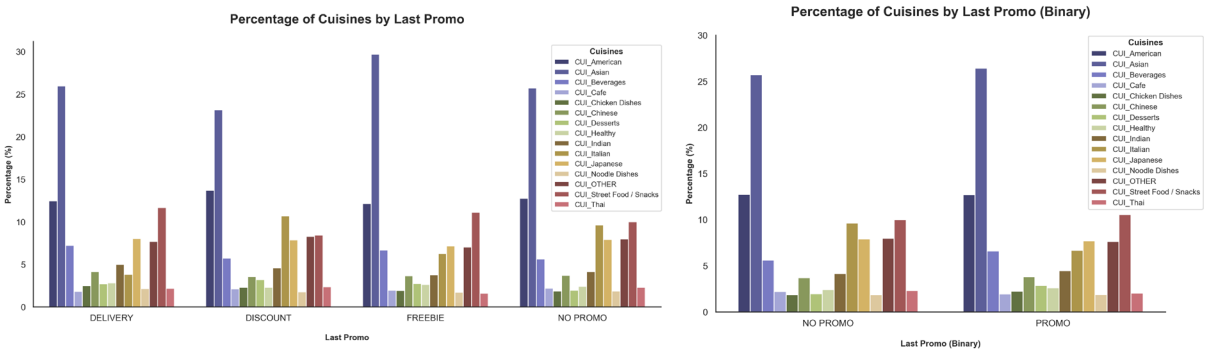

Figure A16 – Bar chart with percentage of Cuisines by Last Promo and Last Promo Binary.

Table A12 – Absolute and Relative Frequency Table of Orders by Day of the Week.

Day	Orders	%
Sunday	17720	12,72
Monday	18091	12,99
Tuesday	18836	13,53
Wednesday	19743	14,18
Thursday	21607	15,52
Friday	20813	14,95
Saturday	22453	16,12

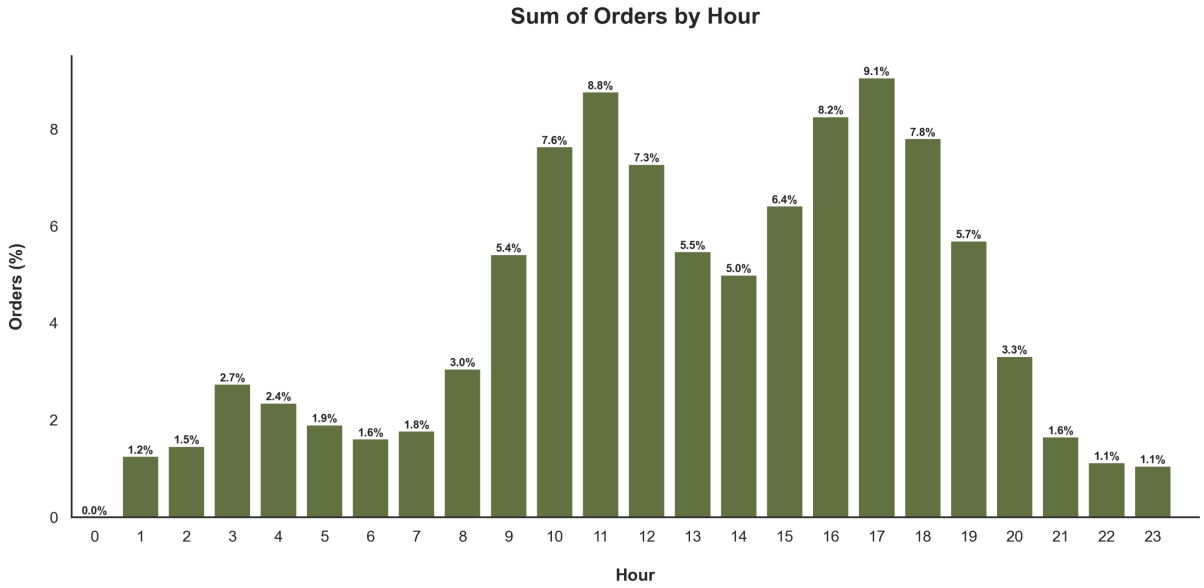

Figure A17 – Bar chart with percentage of Orders by Hour.

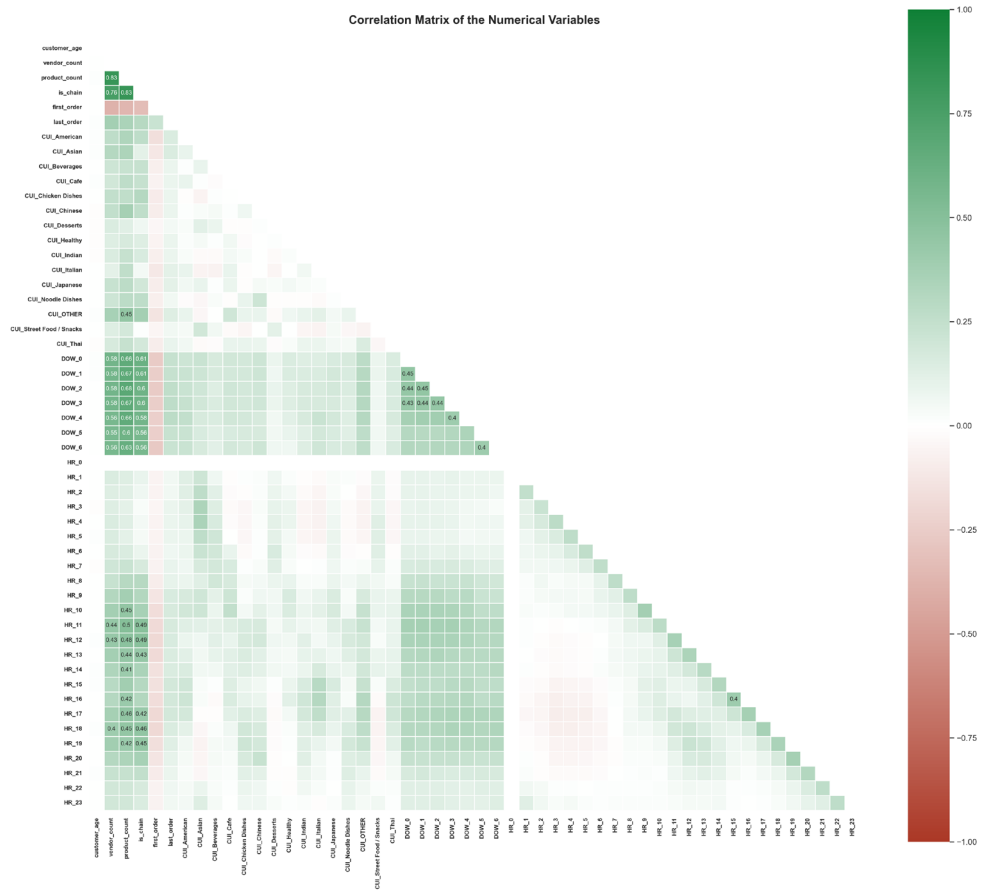

Figure A18 – Pearson Correlation Matrix of the Numerical Variables.

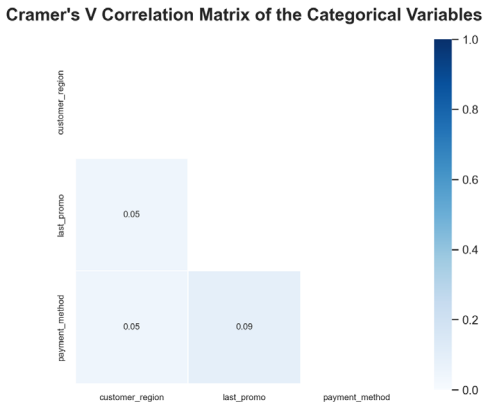

Figure A19 – Cramer's V Correlation Matrix of the Categorical Variables.

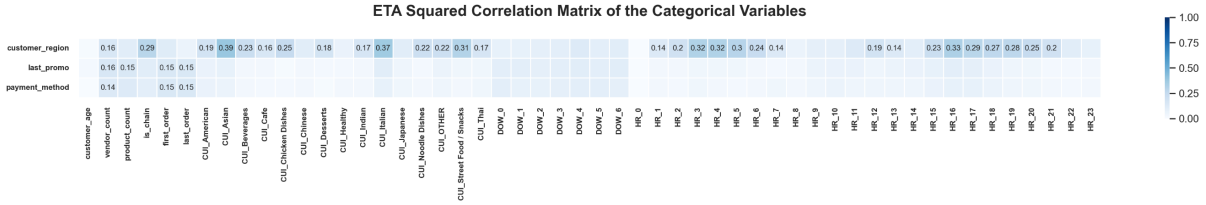

Figure A20 – ETA Squared Correlation Matrix of the Categorical Variables VS Numerical Variables.

ANNEX A

To address the defined objectives, the **CRISP-DM** methodology was employed. This methodology is a dynamic and fluid iterative approach (**Figure A1**) and relies on the exploration and refinement of its different stages to achieve better results, thereby supporting more effective and meaningful decision-making. By definition, this methodology comprises six phases: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, and *Deployment*, with various tasks within each phase that will be explored in this project. [1]

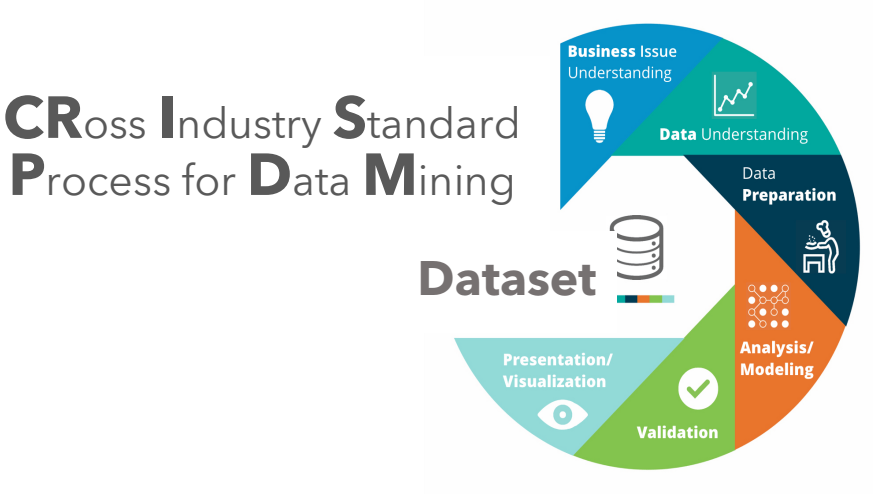

Figure A1 – CRISP-DM Methodology Cycle.
[Adapted from [2]]