Data Mining Project

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

ABCDEats Inc.
Exploratory Data Analysis (EDA)

Group 05

Daniel Caridade, 20211588

Gonçalo Teles, 20211684

Gonçalo Peres, 20211625

João Venichand, 20211644

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. Introduction

In this Data Mining project, our team was selected to act as consultants for ABCDEats Inc. This company works as a food delivery service partner with several restaurants. Our main goal within the scope of this project is to develop a customer segmentation that integrates different perspectives, providing a deep understanding for a better marketing strategy. This report presents an exploratory data analysis to ABCDEats Inc. customer's data.

2. Missing Values

Analysing and handling missing data is essential in any Data Mining initiative, due to the fact that most clustering algorithms, such as hierarchical clustering and K-Means clustering, rely on computing distances to function. Hierarchical clustering groups observations based on proximity [1], while K-Means clustering aggregates observations around a central point or centroid [2], this way when an observation has missing values in any of its features, calculating distances becomes impossible.

In the dataset provided by ABCDEats Inc., only 3 out of 56 features contain missing data: *customer_age* has 727 missing values, *first_order* has 106 missing values, and *HR_0* has 1165 missing values. It is not particularly surprising to see missing data in *customer_age*, as this feature contains sensitive customer information. Overall, 1968 customers have missing values in at least one of these features.

To understand the nature of these missing values, we analysed whether they are Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR) [3]. Based on the patterns observed, there is minimal overlap in missing values across these features, suggesting they are likely **missing completely at random**. Specifically, no customer has missing values in all three features, only 2 customers have missing values in both customer_age and first_order, while 27 customers have missing values in both customer_age and HR_0, and just 1 customer has missing data in both first_order and HR_0.

3. Descriptive Statistics

Computing descriptive statistics is one of the most important sections to accomplish our goal, by enabling us to uncover trends, spot anomalies, and identify potential issues with the data. The numeric features descriptive statistics revealed that the minimum value for the feature *customer_age* is 15 years old, indicating there are customers that are minors. This is important because based on GDPR (General Data Protection Regulation) data belonging to people younger than 16 years old cannot be treated without parental consent. For this reason, those clients will be removed from the analysis. Another relevant statistic is that the average age of our customers is twenty-eighth years (28) old, telling us that the company serves mainly young adults. Regarding consumer behaviour, on average each customer orders almost six items per purchase and the cuisine type with the highest average customer spending is Asian cuisine, being also the one where a customer made their highest spending (896 monetary units). Additionally, Noodle dishes are on average the type of cuisine where the customers spend less money. The days of the week when the highest amount of orders was placed were Friday and Saturday, indicating that the company have higher affluence in the weekends

and should consider allocating more workers in these days. Moreover, Saturday is the day when customers place more orders and The peak order time was 8 a.m., with 52 orders, while no orders were recorded at midnight, likely due to missing data. The top three order hours, following 8 a.m., were 5 p.m., 11 a.m., and 4 p.m., highlighting a surge in orders around lunchtime and mid-afternoon.

Regarding categorical features, we noticed that ABCDEats Inc. serves only serve customers that are from 9 regions, and the region which has the highest customer base is region 8670 (representing **30.61%** of our customers). At the date of the last purchase there are four different types of promotions used being the most common one "-" which our team understood as customers not utilizing any promotion, this behaviour represents **52.52%** of the customers. Finally, there were 3 different payment methods, being credit card the most common with **63.22%** of customers using it as their preferred method for payment.

4. Incoherences

The next step in our Exploratory Data Analysis (EDA) was to identify any inconsistencies, or violations of data assumptions, as these can introduce bias in analysing ABCDEats Inc. customer segments. While few inconsistencies were found, those present included the presence of 13 duplicate entries that were promptly removed from the dataset, and 138 customers with no transaction record, suggesting that they never interacted with the company, as these people did not order anything or spend any money they will be excluded in Data-Preprocessing along with all other incoherences found to provide proper insights as this set might skew the results and mislead marketing campaigns.

Additionally, we identified 442 customers who belong to an unknown region, indicated by the value of "-" in feature *customer_region*, which clearly is not a valid name for a region. This ambiguity creates analysis problems when based on a regional point of view and calls for further probing or cleansing of data to classify these customers further. One additional insight that we drew during this exploration of the data was that from our, 31888 customers, 1436 of them were customers that only made one purchase on the first day that the provided dataset was created, meaning those are very likely lost customers.

5. Univariate Feature Exploration

5.1. Numeric Features

Through careful examination of our numeric features using histograms, boxplots and general statistics, we were able to draw key insights about the distributions and patterns in our data. The feature *customer_age*, for instance, resembles an almost normal distribution despite its slight rightward skew (Appendix 1). However, after applying the Anderson-Darling test, which is a modification of the Kolmogorov-Smirnov test that tests if a sample comes from a population with a specific distribution [4], we concluded that this feature wasn't normally distributed for all usual significance levels. We repeated the test without outliers and reached the same conclusion, meaning our customers tend to be older than the 27.5 year average, which is to be expected as people above 27.5 years are more likely to have the economical means to order food more frequently.

Addressing other features, all of them have a right skewed distribution except *last_order* (Appendix 2)*,* meaning that people continue ordering more than once because if they didn't the distribution would be more similar to the right skewed distribution of *first_order* feature.

The final insights we obtained from the univariate exploration of numeric features regards outliers. All features present outliers except for *first_order*, *last_order* and *HR_0,* this last one only possesses either missing values or the value 0. This large presence of outliers is a bit worrisome since a lot of scientific papers have already proven that outliers negatively affect the quality of created clusters, that is the primary goal of this project [5].

5.2. Categorical Features

For the univariate analysis of categorical features, our team created bar plots to analyse the percentage distribution of each feature. Notably, the feature *is_chain* shows that **71%** of the client base falls into the values 0–3, indicating a large concentration of customers within these groups. Regarding promotions, the most often used one was "Delivery", used by **19.7%** of the consumers. The "Discount" promotion came in second with **14.1%** of customers using it, while "Freebie" was the least popular with **13.7%**. This pattern shows that consumers clearly favour certain deals, especially when it comes to delivery services, which may enhance customer satisfaction.
When it came to payment options, a significant proportion of clients, **63.2%** paid their bill using their credit card (Appendix 3). This overwhelming preference for card payments emphasizes the importance of optimizing the credit card payment process to enhance customer satisfaction.

According to the region's customer belong, **88.3%** of our clientele is concentrated in just three of the eight regions, which shows a large market concentration in just **37.5%** of the accessible regions (Appendix 4). In particular, regions 8670, which included **30.6%** of customers, 4660 which included **30%**, and 2360 that included **27.7%**, had the largest customer representation. With this focused customer base, ABCDEats Inc. has the chance to improve its marketing efforts in underrepresented areas in order to increase its customer base and boost market penetration overall.

6. Additional Feature Exploration

Regarding the analysis of additional features, our team created 3 features: *order_lifetime* which represents the number of days between the customer's first and most recent orders, calculated as the difference between the values in *last_order* and *first_order*, *D_orders* and *H_Orders* which indicate the total number of orders placed on each day of the week and during each hour of the day respectively. Regarding feature *order_lifetime,* which was the main focus of our investigation, we discovered that, 7187 clients have only ever placed one order with the business. Additionally, examining the distribution of this feature, we observed that it is less skewed than the *first_order* feature, and presents the same right-tail tendency. Interestingly, the distribution of *order_lifetime* takes on a more uniform shape when customers who made only one purchase were removed, with a notable reduction only seen for values longer than **80 days** (Appendix 5).

Variables *D_orders* and *H_Orders* revealed that the values in these two features are not equal, even though they should be, due to missing values in the *HR_0* feature, underscoring the significance of addressing data completeness for accurate analyses. Both *D_orders* and *H_orders* retained the

expected right-skewness present in the individual features that comprise them, while their distributions showed more outliers than the individual features, suggesting multivariate outliers, since this increase in outliers was the result of the inclusion of multiple features in a single one.

7. Multivariate Analysis

7.1. Pairwise relationship of Numerical features

When exploring our data, it is crucial to explore features not just individually, but also paired with others to understand the relationships they possess. Firstly, we analysed the pairwise relationship of numerical features. Through a careful visualisation of spearman correlation values and scatterplots between our numerical features, we were able to get more insights about the way our features are connected to each other.

Our original numerical features did not present very high correlation between each other (bigger than 0.8) aside from *vendor_count* and *product_count,* which makes sense since the more vendors we order from, the more products we are expected to order. Bringing the features created in the previous section (*D_orders*, *H_orders* and *order_lifetime*) to the analysis, we observe that those have high correlations with themselves and other features (Appendix 6). Curiously, *D_orders* and *H_orders* don't have a perfect correlation, their correlation is of (0.99), this is due to the missing values in HR_0. This correlation is to be expected since they both represent the amount of orders done by a customer. Both *D_orders* and *H_orders* are highly positively correlated with *product_count*, *vendor_count* and *order_lifetime,* which means the more the number of orders we make, the more products we order, the more vendors we order from and the longer we have been a customer. This last correlation shows that when people order from ABCDEats Inc., they tend to continue ordering, which is a great sign. If this wasn't the case, the amount of time we've been a client wouldn't be related (positive correlation) to the times we order.

Lastly, our team identified multiple multidimensional outliers in the data. Features *product_count, CUI_Italian*, *CUI_Japanese*, *CUI_NOODLE_Dishes* and *CUI_Asia* have at least one multidimensional outlier with all other features. Additionally, there are multidimensional outliers between *CUI_Healthy* and *CUI_Chiken_Dishes* (Appendix 7). The presence of multidimensional outliers means that either there is erroneous data in our dataset that was wrongly recorded since a client shouldn't be able to have certain values for two or more features, or there is a client that is that unusual (Bill Gates effect) [6].

7.2. Comparing categorical features

Regarding the comparisons between categorical features, the analysed stacked columns charts only derive conclusions for relationships with the *is_chain* feature. Since the metadata provided by ABCDEats Inc is a bit ambiguous about this feature, with the description of this feature resemble a binary, but after analysing it, it proved to not be like it, presenting several numbers that seem to act more as codes than numbers with any arithmetic connection between them. For this reason, our team is considering *is_chain* as being a categorical feature. These codes would represent the last chain the customer ordered from (if the customer never ordered from any chain, the value of this

feature would be 0). The evidence for this assumption will be presented in the relationships this feature has with the other categorical features.

When *is_chain* is analysed vs *last_promo,* we can see that there are certain values for the feature *is_chain* where promotions are never used, are always Delivery or always Freebie. There are some other values for *is_chain* (e.g. 25, 28) where the promotion is never Freebie, in some others is Freebie or no discount (e.g. 36, 40), and in other values is Delivery or no discount (e.g. 48). This seems to indicate a connection between the features where for certain values of *is_chain* some promotions are or aren't available, showing that the numbers in *is_chain* act more as categories (codes) than numbers.

This is exacerbated when exploring the relationship between *is_chain* and *payment_method* (Appendix 8), where we can see that there is also a connection between the number in *is_chain* and the possible payment methods.

Lastly, if we take the assumption that the *is_chain* numbers beside 0 represent the last chain ordered from, it would make sense that when compared to the *customer_region* feature, there are certain customer regions that only appear for certain values of *is_chain*. This would show that some chain restaurants aren't available in some areas, therefore can't be ordered from those regions.

7.3. Comparing categorical features vs continuous (or discrete) features

There are very few conclusions our team was able to derive from comparing categorical with numerical features. The insights we found were that the first customers of the company and the more recent orders tend to be pay by credit card and do not use any promotion in their most recent purchase in the company.

8. References

[1] Nielsen, F. (2016). Hierarchical Clustering. In: Introduction to HPC with MPI for Data Science. Undergraduate Topics in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-319-21903-5_8

[2] Unsupervised K-Means Clustering Algorithm | IEEE Journals & Magazine | IEEE Xplore. (n.d.). Retrieved 28 October 2024, from https://ieeexplore.ieee.org/abstract/document/9072123

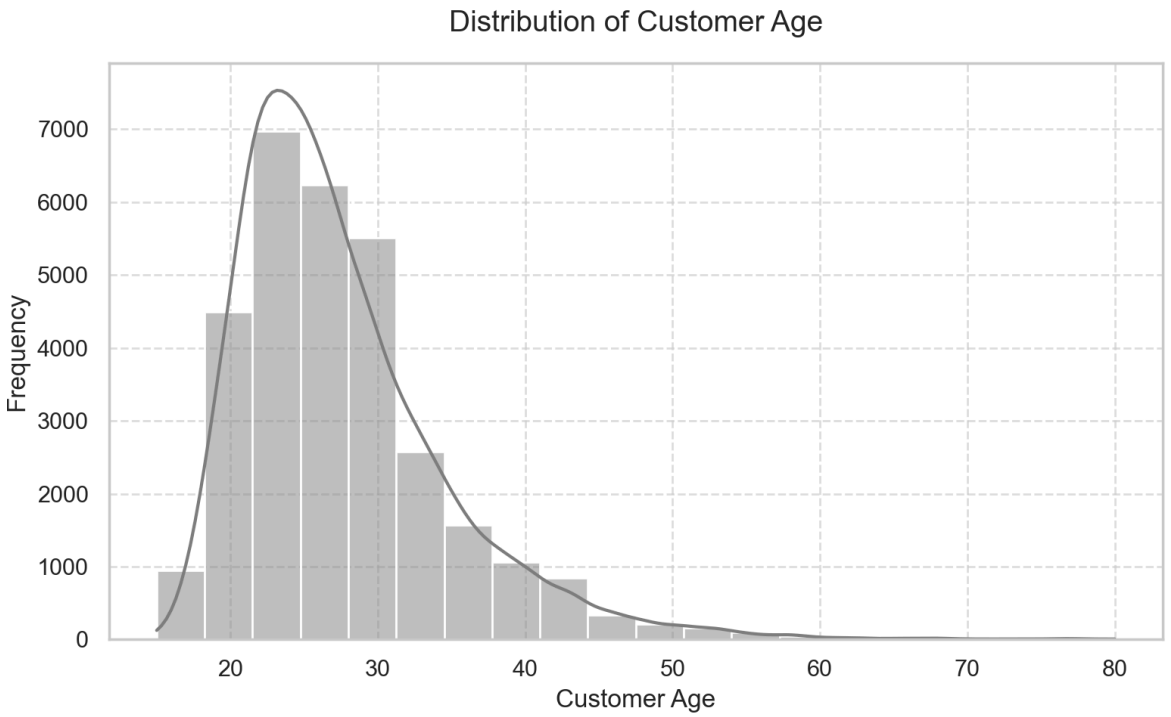
[3] Bhandari, P. (2021, December 8). Missing Data | Types, Explanation, & Imputation. Scribbr. https://www.scribbr.com/statistics/missing-data/

[4] *1.3.5.14. Anderson-Darling Test*. (n.d.). Retrieved 28 October 2024, from https://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm

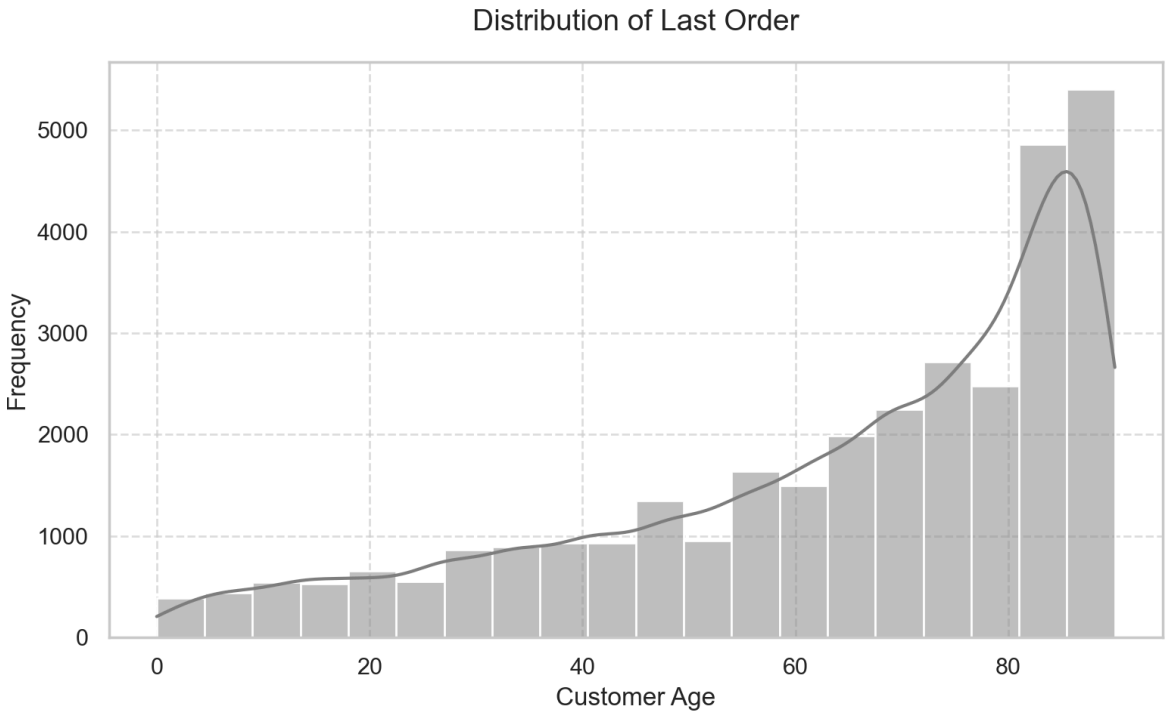
[5] Nowak-Brzezińska, A., & Gaibei, I. (2022). How the Outliers Influence the Quality of Clustering? *Entropy*, *24*(7), 917. https://doi.org/10.3390/e24070917

[6] The "Bill Gates Effect" in Google Analytics and how to get rid of it with Power BI. (2016, March 4). https://community.fabric.microsoft.com/t5/Community-Blog/The-Bill-Gates-Effect-in-Google-Analytics-and-how-to-get-rid-of/ba-p/21769

APPENDIX

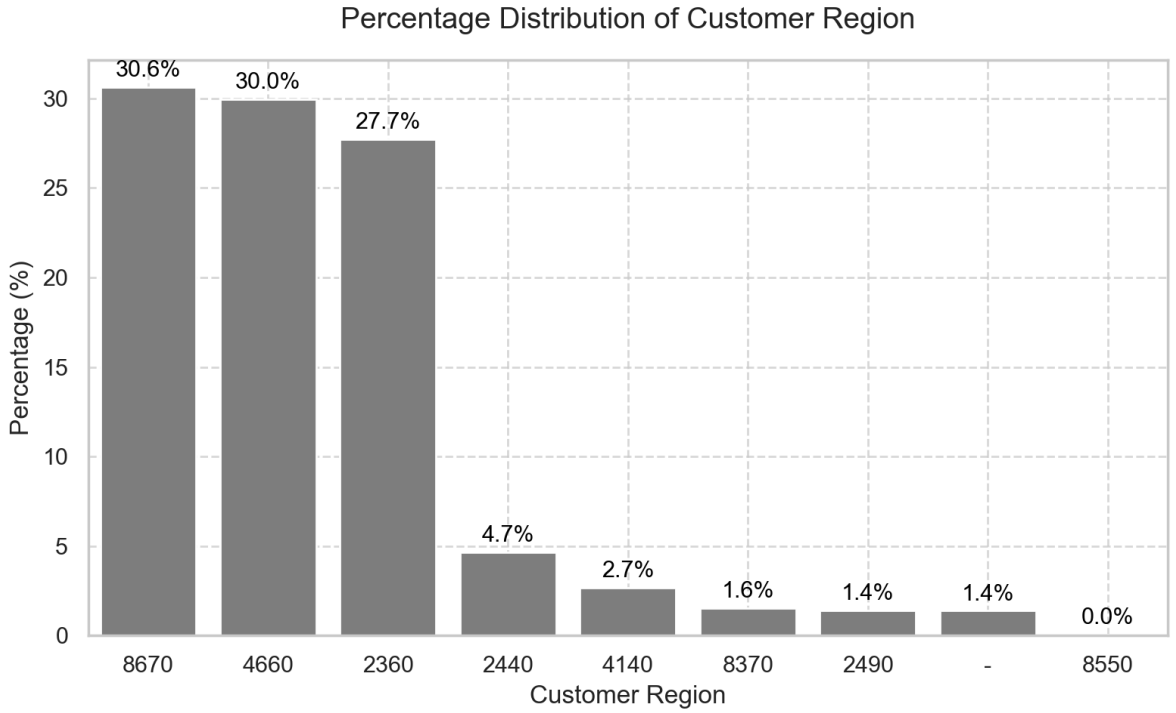


Appendix 1: Histogram of feature *customer_age*

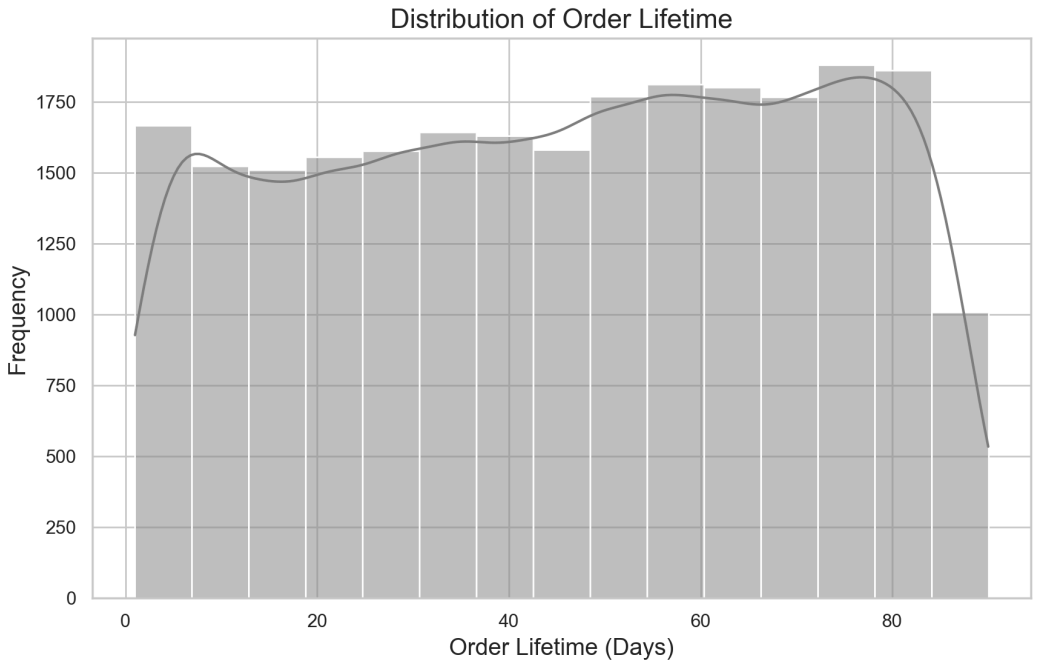


Appendix 2: Histogram of feature last_order

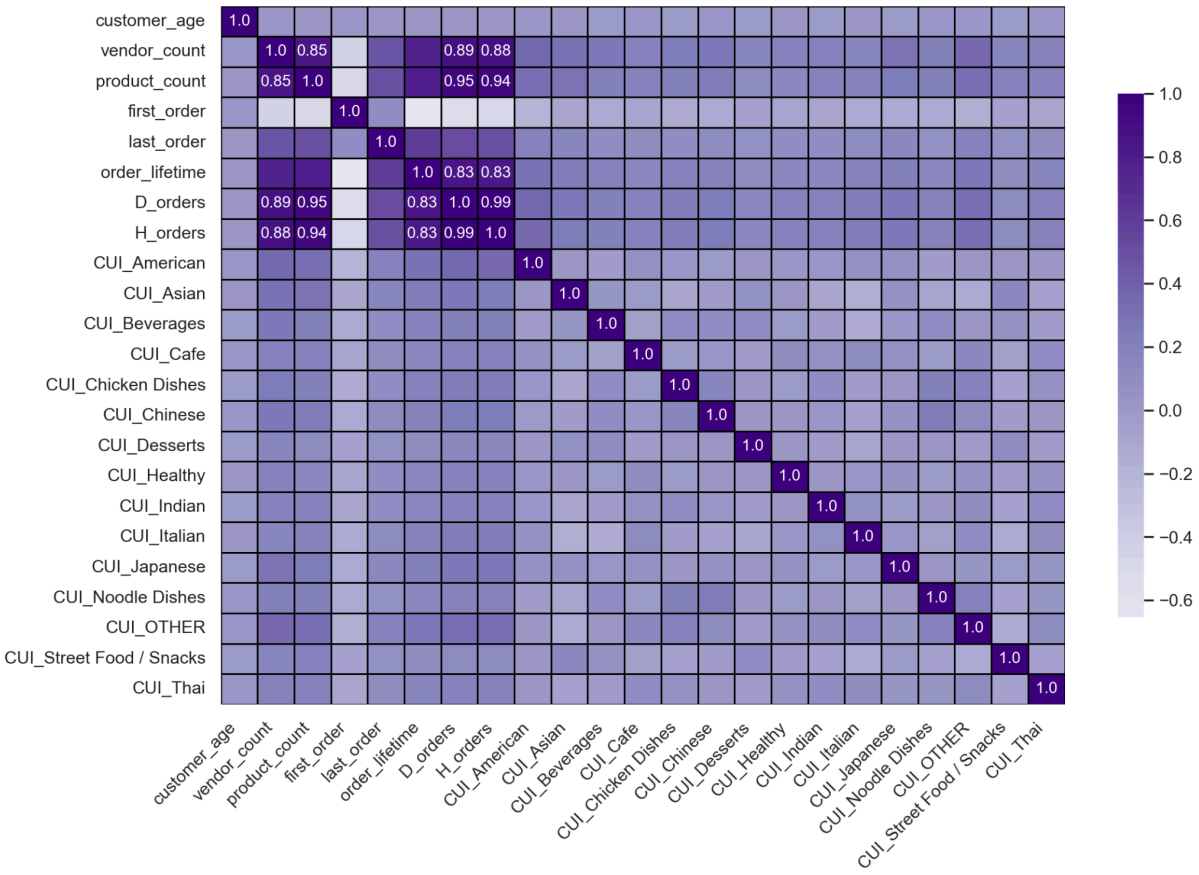
Appendix 3: Bar plot of feature *last_promo*



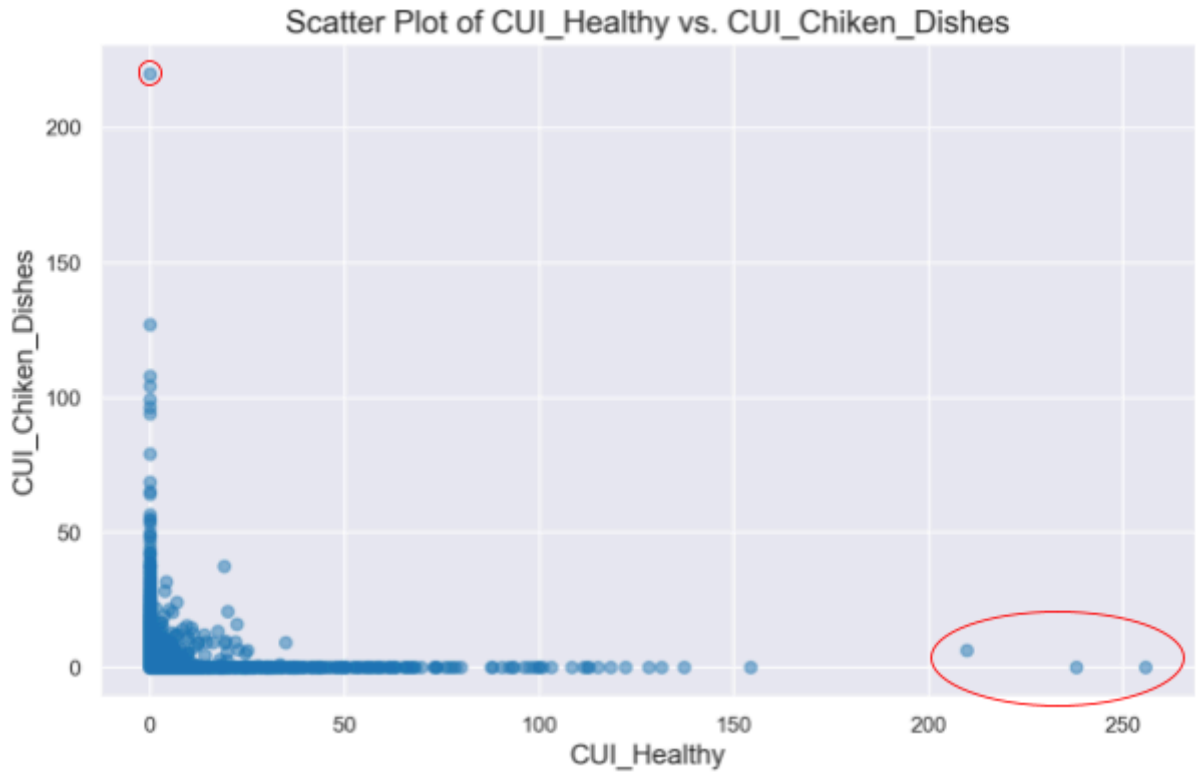
Appendix 4: Bar plot of feature *customer_region*



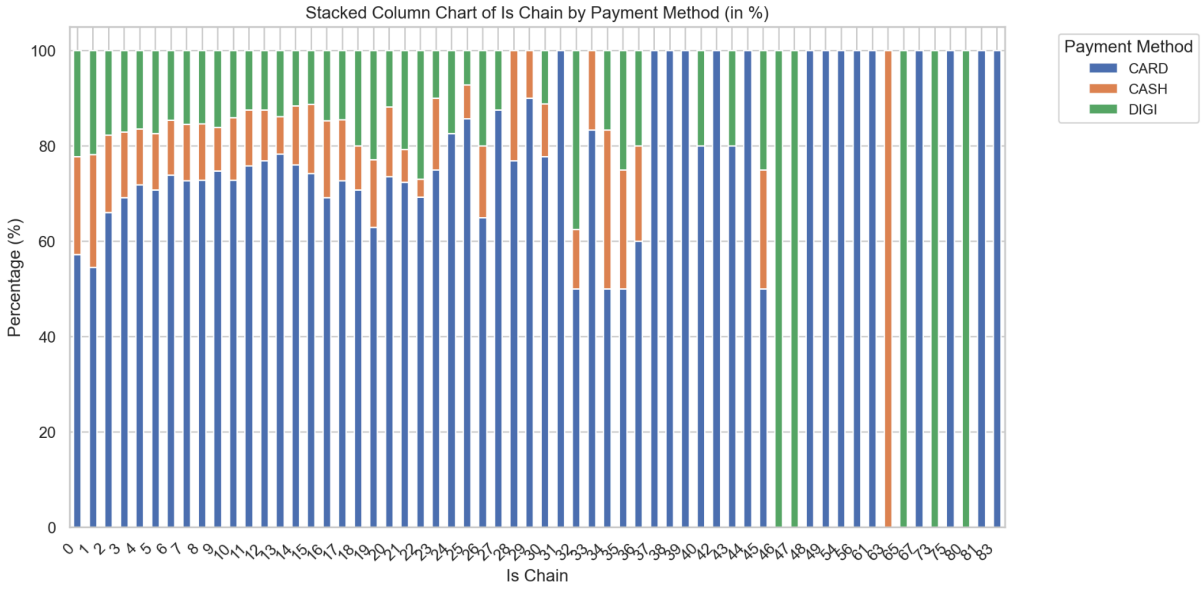
Appendix 5: Histogram of feature *order_lifetime* for customers with multiple purchases



Appendix 6: Correlation matrix of numeric features and the additional features



Appendix 7: Multivariate outliers between CUI_Chiken Dishes and CUI_Healthy



Appendix 8: Payment Method Preferences by *is_chain* value