In general, the analysis is thorough and the ideas are quite interesting.

However, be careful about maintaining an academic tone in your writing.

Taking as an example Page 16:

"Note below how the boxplots of our variables are in absolute disarray"

This is not suitable for academic writing for several reasons:

1. Subjective language: The term "absolute disarray" is subjective and based on personal interpretation. Academic writing should be objective, relying on clear, precise descriptions of the data.

2. Vague terminology: "Disarray" is a vague and informal word that doesn't specifically describe what is happening with the boxplots. Use specific language to ensure clarity and precision.

3. Imprecise tone: Maintain a formal and neutral tone. Words like "disarray" can sound overly informal and unprofessional in this context.

A more appropriate way to phrase this would be: "Note below how the boxplots of our variables exhibit significant variation," which is clear, objective, and precise.

T

Analytics

chool

Exploratory Data Analysis

Group 69

Paulo Martins, r2015469

Beatriz Fonseca, r20201599

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. INTRODUCTION

This exploration seeks to provide the reader with a deep understanding of the main dynamics at play in the sales data for ABCDEats, for the given three-month period.

To present our findings, the data in question underwent preprocessing, including but not limited to, missing value imputation and data inconsistency correction, which is described in detail in Appendices A. The same appendix shows how we engineered variables for the purpose of enabling exploration such as aggregating modal features derived from the available customer information. While the log transformations therein enabled flagging potential outliers, and careful inspection allowed us to deduce a whole host of potential non-customers, in the shape of one time buyers.

Following data treatment, our analysis focused on quantifying the general trends and patterns in the data making use of basic aggregations, histograms, scatterplots and correlations, this analysis can be consulted in Appendices II where it is explained thoroughly. Lastly, we combined the information we gathered in the previous stages to create the visualizations that allowed us to capture the essential information regarding the data.

The resulting visualizations, which will be the focus of the discussion below, follow the key dimensions in sequence Revenue, Geography, Time, and Cuisine. They can be consulted in the Annexes.

2. REVENUE

2.1. Measuring inequality across revenue percentiles with Lorenz Curve

As can be seen in the Annex Figure 24, the Lorenz Curve that measures the inequality between the top and bottom spenders is evidence that a relatively small proportion of customers is responsible for the vast majority of our revenue, reporting an extremely low GINI coefficient. When then analysed over time, as we did in Annex Figure 25, accrued customer average daily revenue rapidly increased up a peak of 1600 at day 49, proceeding to free fall and ending period close to zero.

Should add a reference

Then as we plotted the same graph for orders and products, see Annex Figures 26 and 27, were greeted with the essentially same distribution shape, and deduce therefore, that revenue is being driven, not by the businesses ability to capture an increasing amount customer value, as that would have led to a different shape for these curves, *ceteris paribus*, but by an increase in customer base. This might have been of consequence were it not for the fact that we then deduce that the customer base is essentially shrinking.

Reference how this is calculated

2.2. Realized Customer Lifetime value vis. Expected Customer Lifetime value

The assumption that the above could be due to seasonal sales was promptly dismissed, grounded on the evidence provided by realized lifetime value (RLTV), as of day of first order for the period, is was acquired. Explicitly, customers acquired within the first ten disproportional share of the total value generated through operatio

The idea that the above patterns could be explained by seasonal sales trends was quickly ruled out, as shown in Annex Figure 28. Instead, the data indicates that a customer's realized lifetime value (RLTV), measured from the day of their first order, depends on when they were acquired. Specifically, customers acquired within the first ten days of the period contribute a disproportionately large share of the total value generated.

To know, with the calculation of RLTV, net of Expected customer lifetime value (CLTV), as seen in Annex Figure 29. To measure how customer value, from customers registered on distinct days is being captured relative to the mean realization. Unfortunately, this quantity decays very fast after the first few days of the period, crossing zero and becoming negative, as we approach the end of the dataset.

Edit for clarity

To further dissect this, we ~~purged the dataset of all~~ removed customers that failed to place more than one order - considering them detractors ~~—~~ we also, netted the set of regular customers, from those that we have previously identified as outliers in some sense, and then calculated the same quantity as before for the resulting regular customer dataset. For these customers the result was a negative lifetime value relative to the mean expected lifetime value, indicating finally that the average customer for the period grossly underperformed the expectations. See Annex Figure 30 for further details.

We also filtered out customers previously identified as outliers and calculated the same metric for the remaining regular customers.

When finally plotting only for one-time customers we finally see the amount of customer value that is being left on the table and conclude it as truly large – see Annex Figure 31.

This initial analysis sets the tone for the rest of our report, as it underpins the segmentation effort, with the need to find our sources of revenue and focus our efforts on them. No doubt the correct segmentation and targeting of customers is at fault, and to that end clustering can help immensely. Our discussion will now shift towards understanding the different dynamics at play between customers and service.

3. GEOGRAPHY

Beginning with demography, each city is well represented in the data~~,~~ full stop total sales vary quite substantially by city - see Annex Figure 32 - as city '8' was responsible for slightly over 45% of sales, '4' just 34%, and '2' accounting only for 20%. Regions within each city do not all contribute equally, with regions '2400', '4140', '8370' '8550' accounting together for less than 5% of sales.

As per Annex Figure 33 we find that customers in regions can be more or less well separated into two distinct populations by the average of the log of average amount per product of the average customer in the dataset~~,~~ full stop this informs us that essentially there are two groups of customers with respect to spending power that can be present in any given region - exceptions are 4140 and 8370, ~~were~~ where (?) all customers belonged to the above average group. As common sense would have it, geography is intrinsically linked with cuisine preferences. As such, we tackle information in the next entry.

4. TIME

Focusing our analysis on the time dimensions. Starting with day of Week, Annex Figure 34 shows us that customers placed more orders on Thursdays, Fridays and Saturdays, and less orders on Sundays, Mondays, differentiating start of week customers from end of week customers. Annex Figure 35 that plots the histogram of orders placed at different hours shows demonstrates that customers tend to make their purchases around “11:00”, ”17:00” with some opting to buy at “3:00” in the morning, in a clear trimodal pattern.

4.1. Value counts of day/hour pairs

Our analysis then shifted to quantifying the day hour pairs that showed the most significant activity during the period. To this end, we first generated a heatmap, for order counts proportionally distributed across the set of days and hours when a customer made purchases, as seen in Annex Figure 36.

4.2. Relating Cramer's V with Pearson's R

Perhaps 4.2 should be 4.1.1 ?

To measure the significance of these counts we then created a set of Boolean columns measuring the presence of orders for a given day of week, and hour – refer to Annex Figure 37, and calculated the Cramer V to measure association between pairs. These values were found extremely similar to their pairwise Pearson's R counterparts - see Annex Figure 38 - which led us to formulate that ~~"the knowledge that customers tend to place orders at a particular time, is sufficient to conclude that those customers likely placed more orders than customers that placed orders at hours for which not as many customers tend to place", i.e.,~~ customers that order for lunch will likely do so recurringly, whereas those that order at late at night will likely not.

This led us to conclude that knowing customers tend to order at specific times is enough to infer that those customers are likely to place more orders than those ordering at less popular times. For example, customers who order at lunchtime are more likely to order repeatedly, while those ordering late at night are less likely to do so.

This is confusingly phrased; the statement following it is much easier to understand.

4.3. Cramer's V Weighted Value Counts Heatmap

To achieve our final goal, we performed pairwise multiplication of the original value count heatmap, with the Cramer V association values for day and hour – as seen in Annex 39. The resulting matrix of quantities shows those areas for which there exist both, association, and a relevant number of orders, thus highlighting actionable hour pair combinations, on which marketing efforts can be useful.

The resulting matrix highlights areas with both strong associations and a significant number of orders, pinpointing actionable day-hour combinations where marketing efforts can be most effective.

5. CUISINE

~~Digging into the proverbial meat of the report, A~~nnex Figure 40 shows the frequencies of different cuisine orders, grouped and stacked by customer region. At the cuisine level, Asian and American food represent clear favourites on a per order basis, with Other, Japanese, Italian and Beverages, in close contention. We note that region '4140' is the only region where other, Italian and American have any expressio~~n – so its demography could be similar to the Brooklyn, New York,~~ in the same vein, customers in region '8370' ordered 'Asian food', and 'Street Food', ~~for analogous demographic argument as prior.~~

The larger regions are ~~then~~ more homogenous in terms of order frequency. '2360' has orders in all cuisine~~s, more or less proportionally with level~~, '4660' is missing expression in terms of 'Noodle Dishes', 'Desserts', 'Chinese' and 'Chicken Dishes', while '8670' is missing for 'Chicken Dishes', 'Indian', 'Cafe' and 'Noodles'.

shows

Annex Figure 41 ~~then repeats the plot, but for~~ the sum of the total amount spent by cuisine in each regio~~n,~~ here we see that Asian food~~,~~ now dominates the chart and Food and Snacks then closes in on American food.

full stop

~~On a higher level, the serie~~s of Annexes from Figure 42 through to Figure 56, shows three-way ANOVAs for total amount spent by cuisine, with discount type on the x axis, hued with "above or below average chained restaurant consumption", grouped by city. ~~Some findings were interesting, and we highlight a~~

~~few. For instance,~~ American Food customers that ordered from chained restaurants made proportionally more purchases using FREEBIE than those that ordered from non-chained. The same can be said for region 8 with Beverages and DELIVERY, as well as BEST. In the case of Asian food, we see a slight preference for the BEST discount. Noodle Dishes highlights this preference for BEST even more, with DELIVERY the difference between above average chain for this region and discount showing the most diverging behaviour, the rest we plan on using to guide our clustering efforts.

Before introducing a new concept, you should define it or reference its definition.

Given the number of different cuisines, ~~and as customers showed more than one preference in our aggregations – see Annexes Figure 57 –~~ we use the modal flags for customer top cuisines one and two, to find the total amount per cuisine, discriminated by these customer preferences, proceeding to aggregating the amounts - see Annex Figure 58. ~~The findings were thought provoking, as, while~~ many cuisines like Chicken Dishes, Indian and Noodle Dishes~~,~~ showed almost perfect twin bell curves, which implies that the nature of the orders made by customers likely differed only in monetary amount, while the overall basket of products for choice remains mostly constant. On the other hand, several cuisines like American, Asian and Street Food, show a very distinct behaviour, as modal patterns emerge clearly in their histograms, at more or less regular intervals.

Abduction leads us to conclude that this has one of two likely re
the existence of cheaper 'entry' products, which customers th
gravitate towards – *e.g.,* Pho, Chop Suey, Pizza, and for that rea
in the lower amount range; or, alternatively, b. demand for th
items, and the peaks are a function of the relation between pro
price threshold for a product manifesting in multiple copies of t
more than one order, or several.

We propose two possible explanations:

a) the histograms may reflect the presence of cheaper "entry" products, which first-time customers of a cuisine often choose—such as Pho, Chop Suey, or Pizza—resulting in peaks in the lower price range;

b) alternatively, demand for the cuisine could be driven by a few popular "star" items. These peaks might result from a combination of customer preferences and price thresholds, leading to multiple orders of the same product either in one order or across several.

Lastly, we considered the relationship between days due, days as customer, percentage of chained orders and average days to order, grouped by number of cuisines – which yielded the pattern dense scatter plot visible in Annex Figure 59. The first thing that must be stressed is that these scatterplots clearly represent rotations of the same five-dimensional object in that hyperplane. Note how the covariate of days due with average days to order draws a triangle, but days due with days customer, then rotates this triangle to reveal a prism. This happens because, as we had identified previously, our initial customers have shown the highest propensity towards consumption, this is confirmed once more by days customer covariate average days to order, where customers acquired later show a much higher average days to order, curiously following specific slopes.

consider typesetting feature names in italic to make it easier to identify that they are feature names.

With a slightly different flavour the relation between percentage of order made to chain restaurants and number of orders follows a mesmerizing pattern, with values being drawn from what looks like the projection of a pair of focal points, that collide with ”0”, ”1” or a center value close to ”0.6”.

Add a legend in the figure instead

~~Perhaps in a curious display of mathematical beauty, the only reason why this is all interpretable is due to the hue that we have selected, which as stated before represents with darker values the increasing number of cuisines ordered by the customer. Critically,~~ these darker hues are associated with a higher number of orders, a lower average order, and a higher preference for chained restaurants. Given everything that we know about this dataset we are left with the natural conclusion that the number of cuisines, as it represents a more distinct behavioural pattern that is distinct from the other features, might hold on of the keys towards understanding how to drive customer value further.

Based on this analysis we can conclude that the distinct behavioral pattern exposed by the number of cuisines makes it a critical feature to use in order to understand customer value.

6. FINDINGS

Our initial findings support the belief that a very good clustering solution can be achieved, given that there exist a plethora of dynamics at play, ranging from differentiation in propensity towards consumption, regional preferences, existance of well defined patterns of customer behaviours, as well as, a rich distribution of variables that drive demand at different levels and for diferente customer types. This effort seems now absolutely crucial, as a key distinctive quality of the dataset was how less loyal and less propense towards consumption new customers are turning out to be. Perhaps the correct segmentation will unlock a better marketing strategy, and that will be the end goal of our final report.

Our findings suggest that a good clustering solution can be achieved due to the clear differences evident in customers' consumption habits, regional preferences, and behavior patterns.

Based on our analysis, it is evident that newer customers appear to be less loyal and less likely to make purchases; developing an appropriate segmentation is therefore critical to enable us to create an appropriate marketing strategy to address this.

BIBLIOGRAPHICAL REFERENCES

Uber Eats for Business. (n.d.). *10 types of sales promotions businesses should run*. Uber Eats. Retrieved November 4, 2024, from https://merchants.ubereats.com/gb/en/resources/articles/10-types-of-sales-promotions/

Add a section summarizing new features (incl. formula)

APPENDIX A – DATA WRANGLING

Before we start any of our exploratory processes, we know from the metadata that we have a column named **customer_id** which contains unique identifiers for each customer. This feature is in hexadecimal format, so we decide to convert it to an integer feature and plot its distribution only to observe a uniform distribution which tells us the values are indeed either random, or randomly sampled, or both.

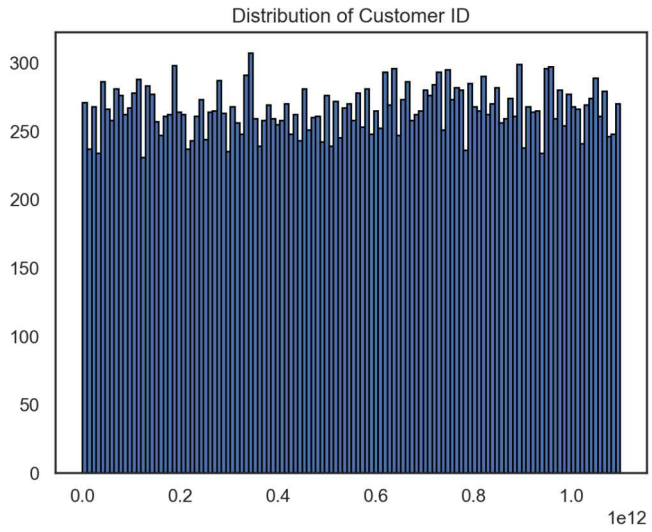


Figure 1 - Distribution of Customer ID

We abstain from performing any further analysis and set it as our data frame index, after removing the duplicate values of this feature.

Starting the process of analysing our data, we quickly realise we do not have a lot of missing values, ~~which is good news to us.~~ Only three of the 56 features had missing values:

- **first_order:** that we filled with the value 0, assuming the data for these customers had been collected before this 3-month period.
- **HR_0:** which we replaced with the difference between the sum of values of all DOW features and the sum of values of the other HR features.
- **customer_age:** we attempted to find a way of filling these missing values, but we were unable to find any evidence of linear association between this feature and all others. We plotted this feature with some others and found that they followed the same distribution as that of **customer_age**, and we performed a statistical test which showed no evidence of linear association with any other features, since most p-values were too large and those that were not, the values for correlation are insignificant.

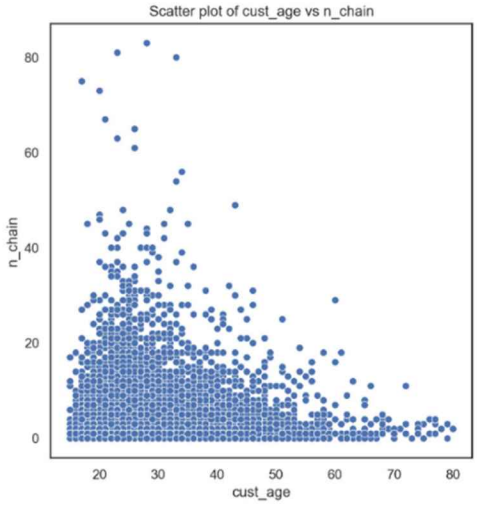


Figure 2 - Scatter Plot of Customer Age and Chain Restaurants

Looking at the unique values of each feature, we found some undescriptive values: '-' in **customer_region** and **last_promo**, which could mean "unknown region" and "NO PROMO", respectively. Furthermore, the feature **is_chain** was described in the metadata as a binary feature, but it has 60 different values – so, we assume the provided metadata is wrong and that this feature holds information about the number of orders made in chain restaurants.

Regarding the consistency of the data, we observed no negative values in any feature, but we still need to assess the consistency regarding duplicate values and relational impossibilities. As for the duplicate values, after we set **customer_id** as the index the dataset gets some duplicate values which we remove; as for relational impossibilities, there were some rows that did not have any values different from 0 for the features representing the amount spent (**CUI** features) and orders made (**DOW** and **HR** features) and so we decided to exclude these values from the subsequent analysis as well.

Moving on to the weird values, we performed some statistics as an attempt to fix them:

- **customer_region:** we fill the values of '-' with 8670 and join region '2440' with '2490', as both cases fail to demonstrate statistical evidence of belonging to different populations; we create a variable city, with value corresponding to the first digit of each **customer_region**; and we give no focus to region '8550' in further exploratory data analysis as it only has 13 observations and is very difficult to group with another due to lacking a statistically representation of the population.
- **last_promo:** according to the conducted t-tests, we have evidence to reject the hypothesis that it belongs to the distribution of any other type of promotions; however, if these values were to mean that no promotion had ever been used by a customer, it would be very difficult to fathom this would be the case for sixteen thousand customers, near half of our data. As an attempt to find a more meaningful value to this category, and because we feel setting this value as 'OTHER' would be unfitting due to the large amount of observations under this category, we decide to call it 'BEST'.

Why?

What makes "BEST" a better label than "OTHER"?

We conclude this part of the analysis by setting the correct datatypes for our features.

APPENDIX B – INITIAL EXPLORATION

In this section we start by defining new features that will help us with the analysis and visualization of our data and follow to analyse some aggregating statistics.

Aggregations

Metric features

The company has a very young customer base with a mode of 23, median of 26, and mean of approximately 28 years, its distribution is as a result very skewed to the right and somewhat significantly leptokurtic, as can be confirmed by looking at its quantiles, and in fact, 98% of customers are below the age of 47, while the oldest is 80.

Most of this customer base showed itself loyal to a relatively small number of vendors, with half not placing orders from more than two vendors, this behaviour somewhat extends itself to the products purchased from said vendors, with median value of products bought at 3. In this case, however, we see that the mean is higher than the median, but the 95th percentile does not go beyond 18, indicating that we have extreme outliers, which the skewness and kurtosis appear to confirm.

Customers placed on average 4 orders during the quarter, but concerningly, half of the customer base, made only 3, amounting to one order per month. This distribution is also extremely right skewed, and leptokurtic, meaning that a very small number of customers are responsible for a large proportion of orders.

If we consider **n_chain** to be the count of purchases made in chain restaurants i.e. a fraction of **n_order**, we see that most customers make ~~relatively~~ about two thirds of their purchases from chained restaurants; this is more or less confirme~~d,~~ by the explicit calculation of **per_chain_order**, which measures the fraction of orders placed by customers, ~~that purport~~ to chain restaurants.

Total amount spent shows the business depends on high spenders, with its mean (38.43) being much higher than the median value, and extremely high values at the higher percentiles, along with monstruous variance and kurtosis (note the max amount of 1418.33). Of all the aggregate amounts, the average per product is the most well behaved with the maximum value being roughly three times the median. Curiously, the average amount per vendor differs from that of average amount per order, indicating that there is some relation between higher spending consumers and specific vendors.

Moving on to **first_order**, it shows that 50 percent of the customer base placed its first order during the first three weeks, while the next 25 percent made orders in the three weeks proceeding. Thus, we conclude the remaining 25 percent of the current customer base was acquired in the last month and a half of operations (approximately 6 weeks), constituting a dramatic slowdown. This can be cause for concern if we look at the information about the previous propensities for small numbers of customers to place large orders, as this makes the company hostage to a small number of cash cows~~, to which it is then forced to make concessions, in exchange for loyalty – i.e. in a traditional PESTEL analysis sense, we can say that in such a scenario the company risks having its costumers gain leverage over the business, and reducing overall profit margins.~~

blue means add a full stop here

Green means good!

red means remove or rewrite

orange means rephrase

Looking at **last_order**, we can assert that 75% of customers made their last purchase within the last 40 days, with 50% in the last 20; this is good news, as at the very least, it shows that the decreasing trend in customer acquisition is not accompanied by an increasing one in customers making their last purchases.

To better understand this relation between first and last order, we inspect average days to order, which measures, on average, how many days passed between each of the customers' order; we see that the median and mean are more or less in agreement, at 7 days, but there is great dispersion around this behaviour with high variance.

Finally, ~~and oddly curious, is~~ the modality of **days_cust** at 0 which is in total disagreeance with the balance we made note of between first and last order, as under normal circumstances we would expect that if those two quantities are in the balance, **days_cust** ought to follow a uniform distribution, with mean roughly at the day 45; the fact that it doesn't might imply that a significant portion of customers made one time purchase~~s, for specific reasons.~~

Day Features

Aggregating over days of week, highlights a few key points: 1. that there is a clear trend towards orders being placed on the days leading up to the weekend, 2. due to an increase in variance, we can also deduce that not all weekends are the same.

Hour Features

Similarly aggregating over the hours, highlights a predictable concentration around lunch and dinner, with a gentle trough at the mid-afternoon mark. And a rather conspicuous point in the morning.

Non-Metric Features

Unfortunately, we don't gain much insight from these features.

Histograms

Metric Features

~~Our initial suspicions about the skewness of the data, are now in full display, as~~ many of our features show clear right tail, sometimes with very sparse values. More interestingly we note that:

- **per_chain_order** shows a sort of self-similar behaviour centered around approximately 0.5.

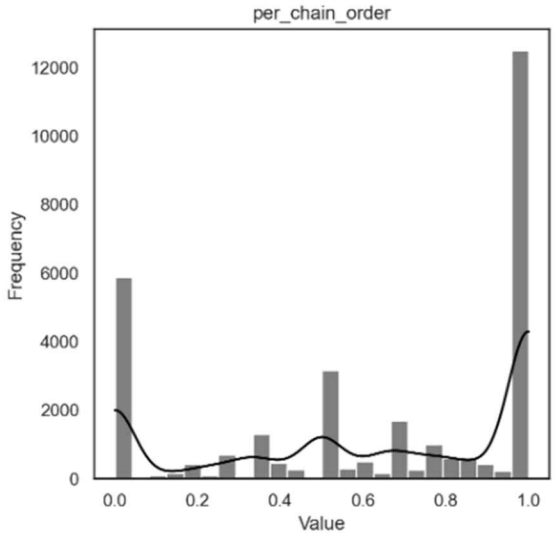


Figure 3 - Histogram of Percentage of Orders from Chain Restaurants

- Almost one third of customers placed a single order, because they were customers only for one day, which is visible in days_cust, and avg_days_to_order.

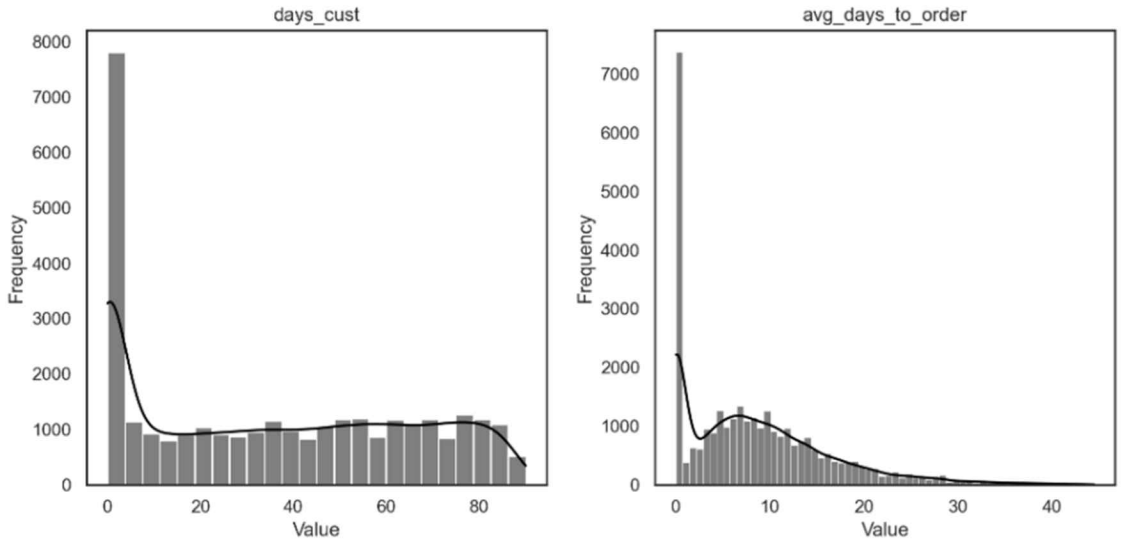


Figure 4 - Histograms of Days as Customer (left) and Average Days to Order (right)

- The variables about average order and product show very consistent spikes, in such a way that it leads us to believe that these might not purport to the same overall populations. But to begin speculating, a population of customers that chooses products and vendors based on very well-defined prices, infers either a subgroup with a very high sensitivity to product/price mix or fraud.

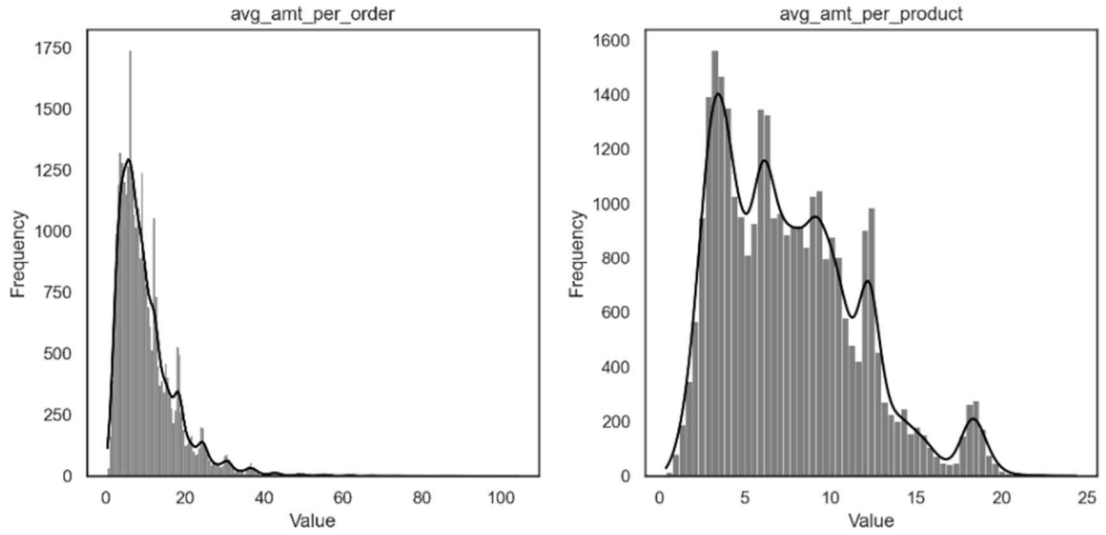


Figure 5 - Histograms of Average Amount per Order (left) and per Product (right)

When we correct for one-time purchases, we see that certain distributions like avg_amt_per order and per product become more locally well behaved i.e. smoother. This leans into the idea that these customers are taking advantage of pricing, or product when they make their first purchase through the service, i.e. their need for the service might be driven by perception of advantage. We will test this later by checking price sensitivity, by comparing these customers with promotions.

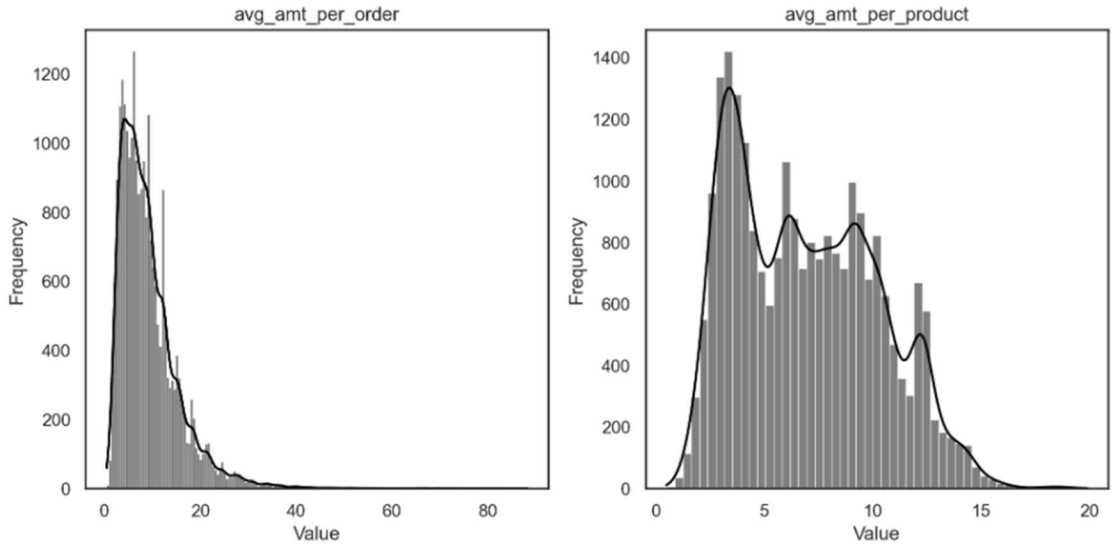


Figure 6 - Histograms of Average Amount per Order (left) and per Product (right) (Excluding One-time Customers)

Day Features

Once we describe the data in terms of at least one day, and for customers that are not just one-time, we see that customers that make more orders over the period are slightly associated with higher values of days of the week.

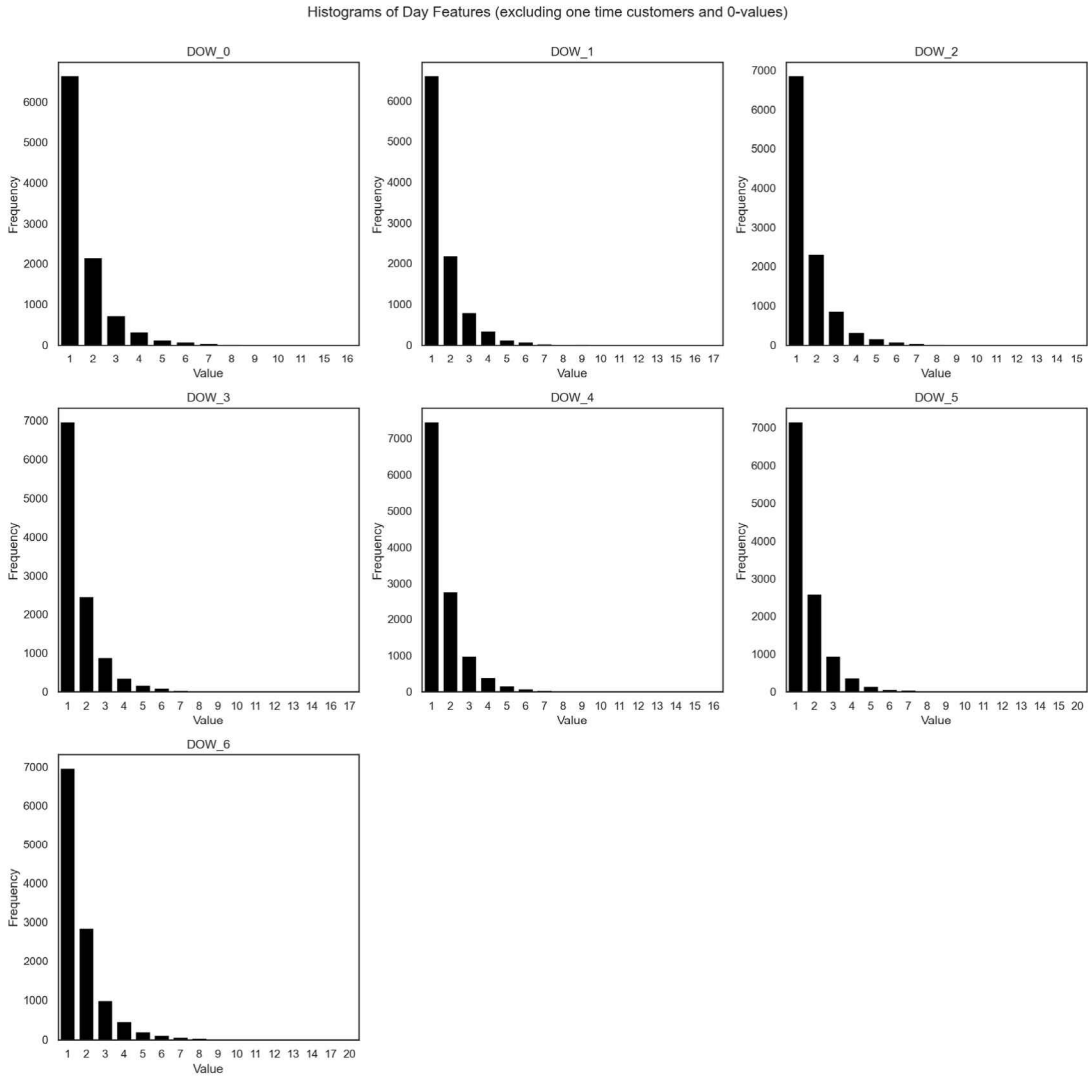


Figure 7 - Histograms of Day Features (Excluding One-time Customers and 0-values)

When we aggregate the values we see the impact of one-time consumers and that most customers that are not one-time customers have made orders in between 2 and 3 different days of the week.

Figure 8 - Count of Customers for Days with Orders (Excluding One-time Customers)

Hour Features

Most people did not make a purchase at most hours, however, the fact that the values ranging from 1 to 3 are more highly populated for certain hours, does imply again that there is a clear preference for orders to line up with meal hours. We need to however take into consideration that food needs to be prepared and delivered, and that customers might account for this when they place an order, thus the order placement in our records likely reflects this perceived lag; as orders begin as early as 10, which on its own could be understood as breakfast, but if we consider that an order process initiated at 10:45 and finalized and placed at 10:55 - which would fall onto the 10H bracket - that then takes 45 minutes to reach the customers door. This means that the customer is having lunch between 11:40 and 12:00, which is a more habitual, if albeit slightly early hour for lunch.

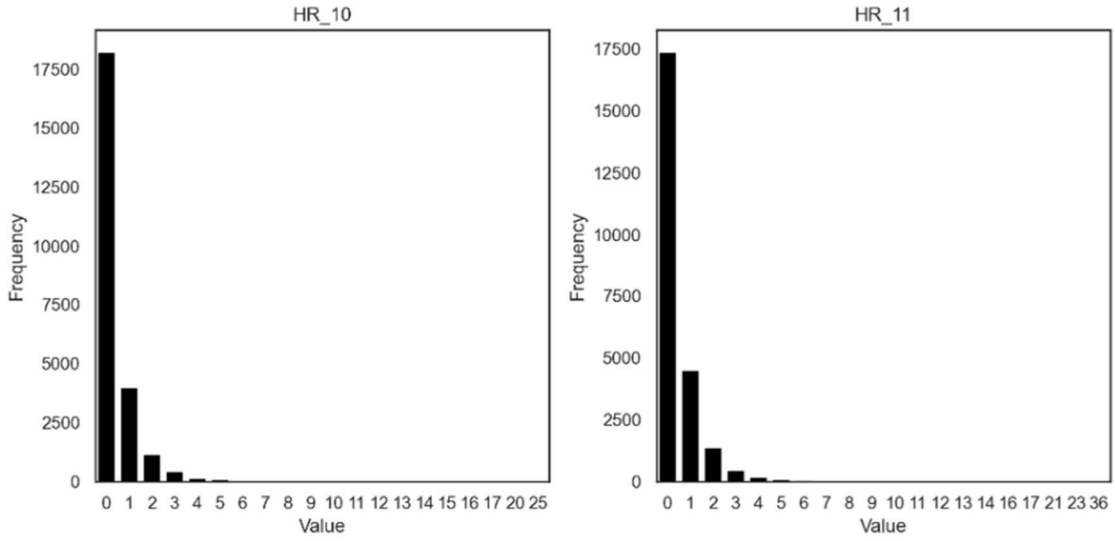


Figure 9 - Distribution of Orders for Hours 10 and 11

Most customers that have made more than one order at a particular time have done so at meal hours. This implies that our more regular customers are to be found in the subset of those that observe habit, and plan ahead for their mealtime. Curiously, HR_5 is the one that is associated with a smaller relative frequency gap between one and two orders, which implies that knowing that a customer placed an order at 5 in the morning, gives us greater confidence that they will have done so, more than once. This makes intuitive sense, if we account for a. party goers, b. night shift workers.

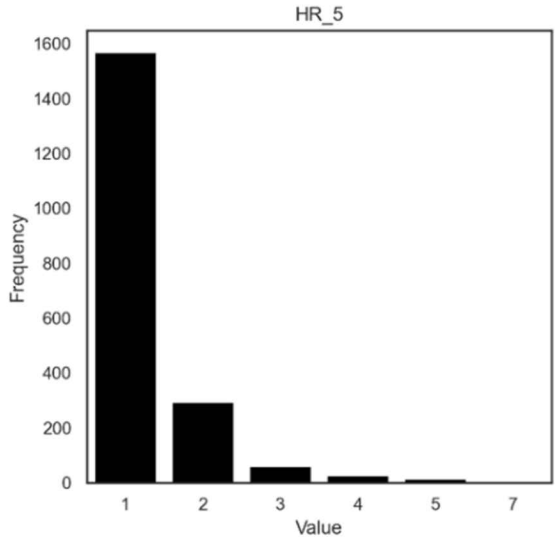


Figure 10 - Distribution of Orders for Hour 5 (Excluding One-time Customers)

When we net customers that made only one purchase, the distribution of this va
more apparent. Customers tend to concentrate the orders that they make around
Ironically, this further reinforces the idea that we need to consider purging our r
values, at least for the purpose of understanding the average customer, if not all

When we exclude customers who made only one purchase, the distribution of this variable becomes much clearer. Customers tend to concentrate their orders around 2 to 4 specific hours. Interestingly, this further supports the idea that we should consider removing these values from our main dataset, at least when analyzing the average customer, if not entirely.

Based on our analysis we

~~Strictly speaking, with the evidence thus collected we have evidence to~~ believe these values likely represent either **a.** people that wanted to try out the service; **b.** people that were trying to take advantage of a one-time deal, on installation of the service, on a product, etc., but that otherwise do not wish to continue using the service; **c.** fraudsters creating multiple accounts for the purposes of **b**. Of course, there is the risk that we are removing customers that legitimately belong to the customer base, but without any further way of filtering both situations, we feel it makes sense to put these aside.

Moreover, there are considerations regarding if these values are even worth considering as part of our clustering, as to be fair, it is trivial to build them as a group and just append them to our clusters, and in fact, we might just find that without them our algorithms that depend on distances might have an easier time with other groups.

This raises the question of whether these customers should be excluded in our clustering, as doing so could improve the results of distance-based algorithms.

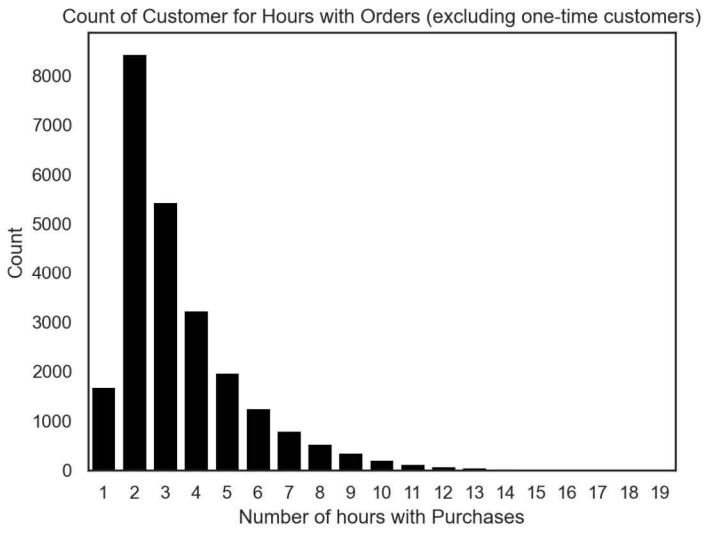


Figure 11 - Count of Customers for Hours with Orders (Excluding One-time Customers)

Cuisine Features

Removing the 0-values improves interpretability, but it is still a difficult endeavour. We c
patterns particular amounts spent, which might be evidence of
purchased, and then the total amount spent in each kitchen
multiplication operation. So*, in limine* if this logical abduction ma
that the histograms for which the distributions show what appear to be several mo
purports to customers strongly preferring a particular set of products within that spe
cuisine. This then further implies that customers that opt for this type of cuisine have a p
this product at all levels of total aggregate spending.

Good insights, but explanation could be simplified.

Removing the 0-values makes the data easier to understand, but it's still challenging. We see that certain spending patterns might show customers buying a few products, with the total amount spent just reflecting this. If this is correct, the histograms with multiple peaks likely mean that customers strongly prefer specific products within that cuisine. This suggests that people who choose this cuisine tend to favor these products, no matter how much they spend.

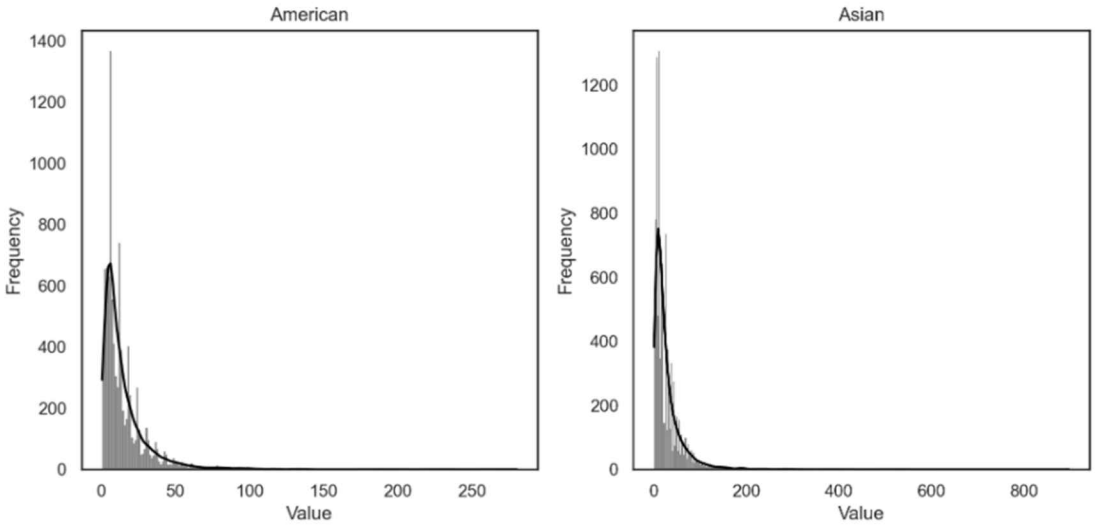


Figure 12 - Histograms of Total Amount for American (left) and Asian (right) Cuisines

Non-Metric Features

Pay methods by one-time customers were highly irregular. It is also true that for these customers a much greater emphasis on promotions was present, and since their last promotion is their only promotion, we can be certain that this was the promotion used for the purchase. Lastly, regions 2400, 4140 and 8370 see a smaller number of one-time customers.

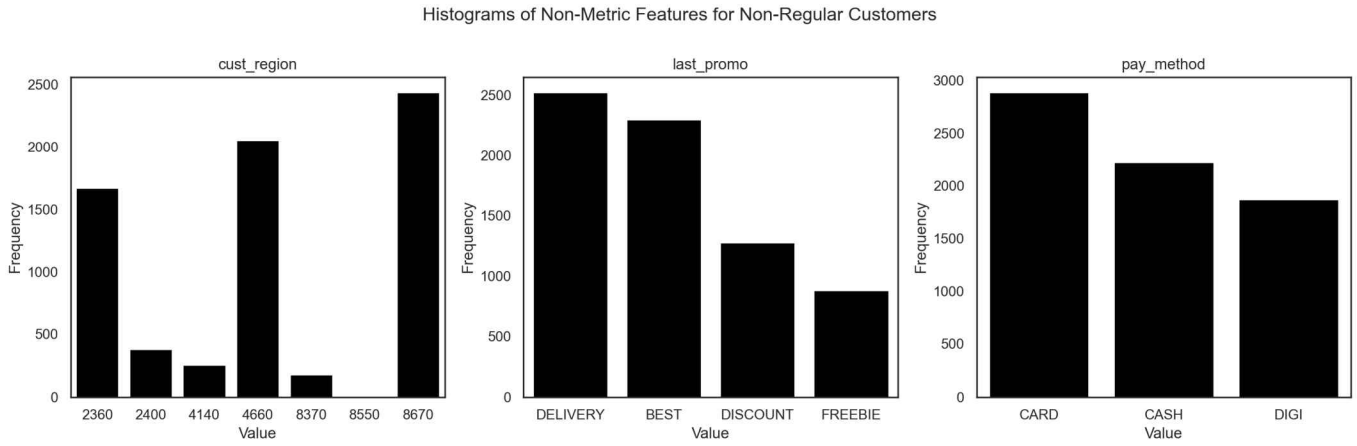


Figure 13 - Histograms of Non-Metric Features for Non-Regular Customers

Flagging Outliers

~~Note below how the boxplots of our variables are in absolute disarray.~~

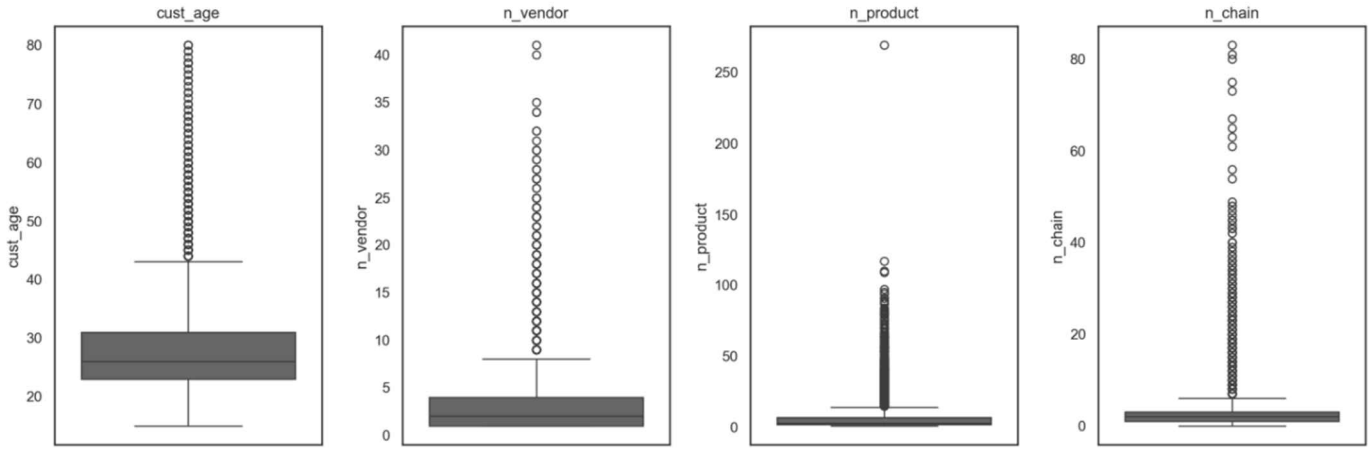


Figure 14 - Boxplots of Customer Age, Number of Distinct Vendors, Number of Distinct Products and Number of Orders in Chain Restaurants (in order)

To this end, it is not so much that we wish to eliminate our outliers, as this is not the time for that, but certainly create flags based on certain types out liers. The log transformation offers a very robust way to find such intervals.

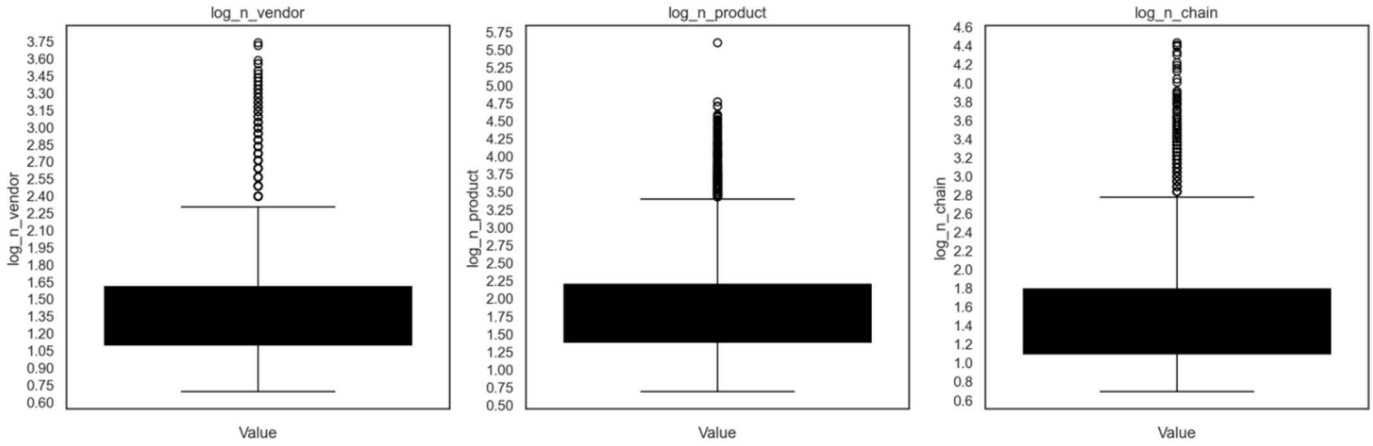


Figure 15 - Boxplots of the Log of Number of Distinct Vendors, Number of Distinct Products and Number of Orders in Chain Restaurants (in order)

Creating Slices

Since we are not really removing outliers, we can be more creative in our approach, we will attribute flags based one the following above IQR behaviours as follows:

- foodie - **n_vendor, n_product, n_order**
  - "Experiment with many vendors, order many products, and place many orders"
- glutunous - **avg_per_order, total_amt, n_chain**
  - "PLace large orders, spend a lot of money, mostly in chained restaurants"
- loyal - **avg_per_vendor**, **CUI**
  - "Spend a lot on each vendor, and spend a lot in a type of cuisine"

Investigating Proportion of Outliers by Customer Age

As we have not yet found a feature that was correlated with **cust_age**, we decided to check the distribution of outliers by customer age. We verify that the outliers follow an empiric distribution that is very similar to the one of customer age, meaning that the likelihood of being a customer is not a function of the age.

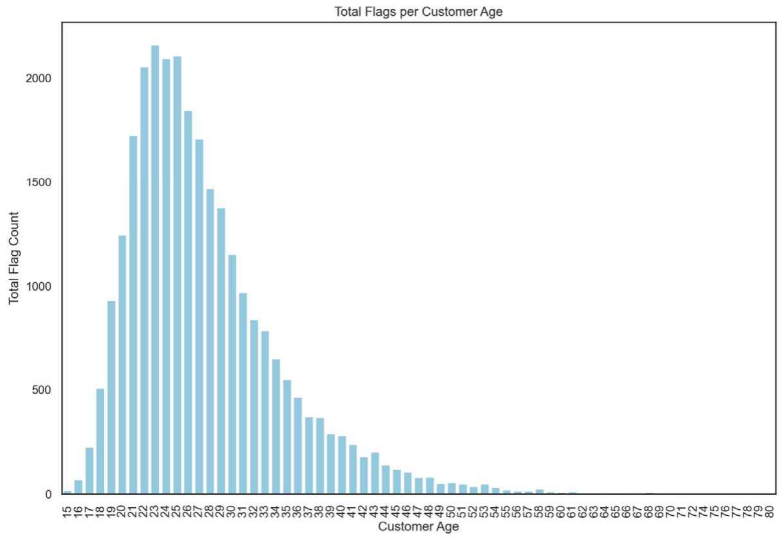


Figure 16 - Count Plot of Customer Flags per Customer Age

Looking for Potential Customer Discriminating Variable Parameters

Later on, during the project we are going to look at more refined strategies to discretize our data. In the meantime, we tested different values for variables that we believed would end up being interesting. In particular, we found that customer region and city are very good discriminants of the general propensity of customers to make purchases, carrying over through various variables. ~~Below, the prettiest~~.

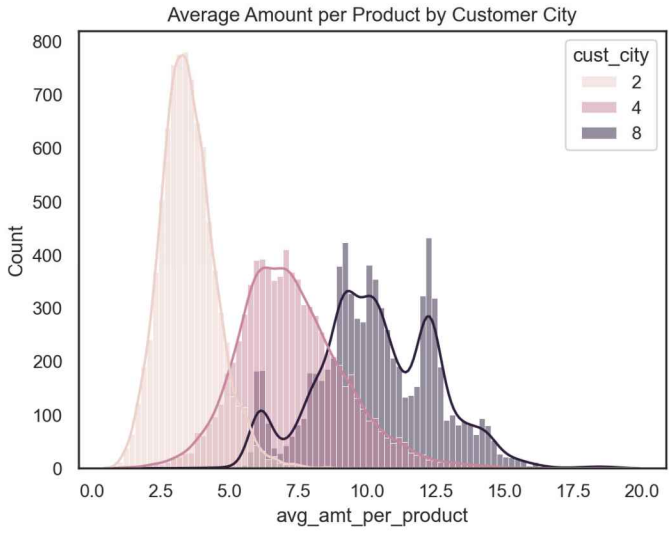


Figure 17 - Histogram of Average Amount per Product by Customer City

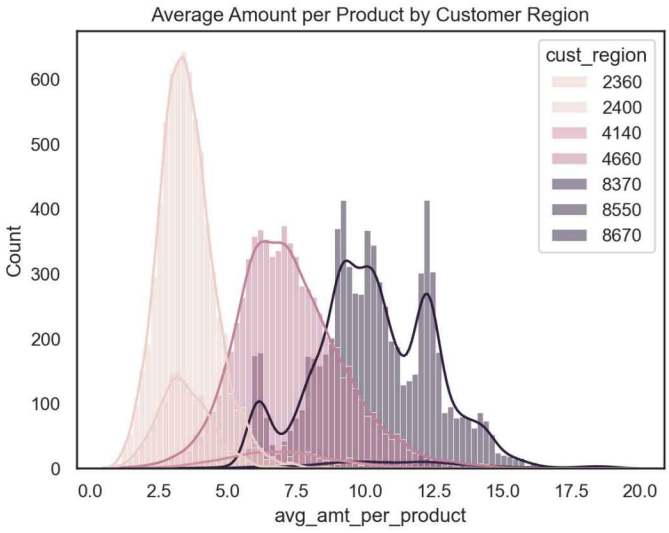


Figure 18 - Histogram of Average Amount per Product by Customer Region

APPENDIX C – CORRELATION MATRICES

Metric, Day, Hour and Cuisine Features

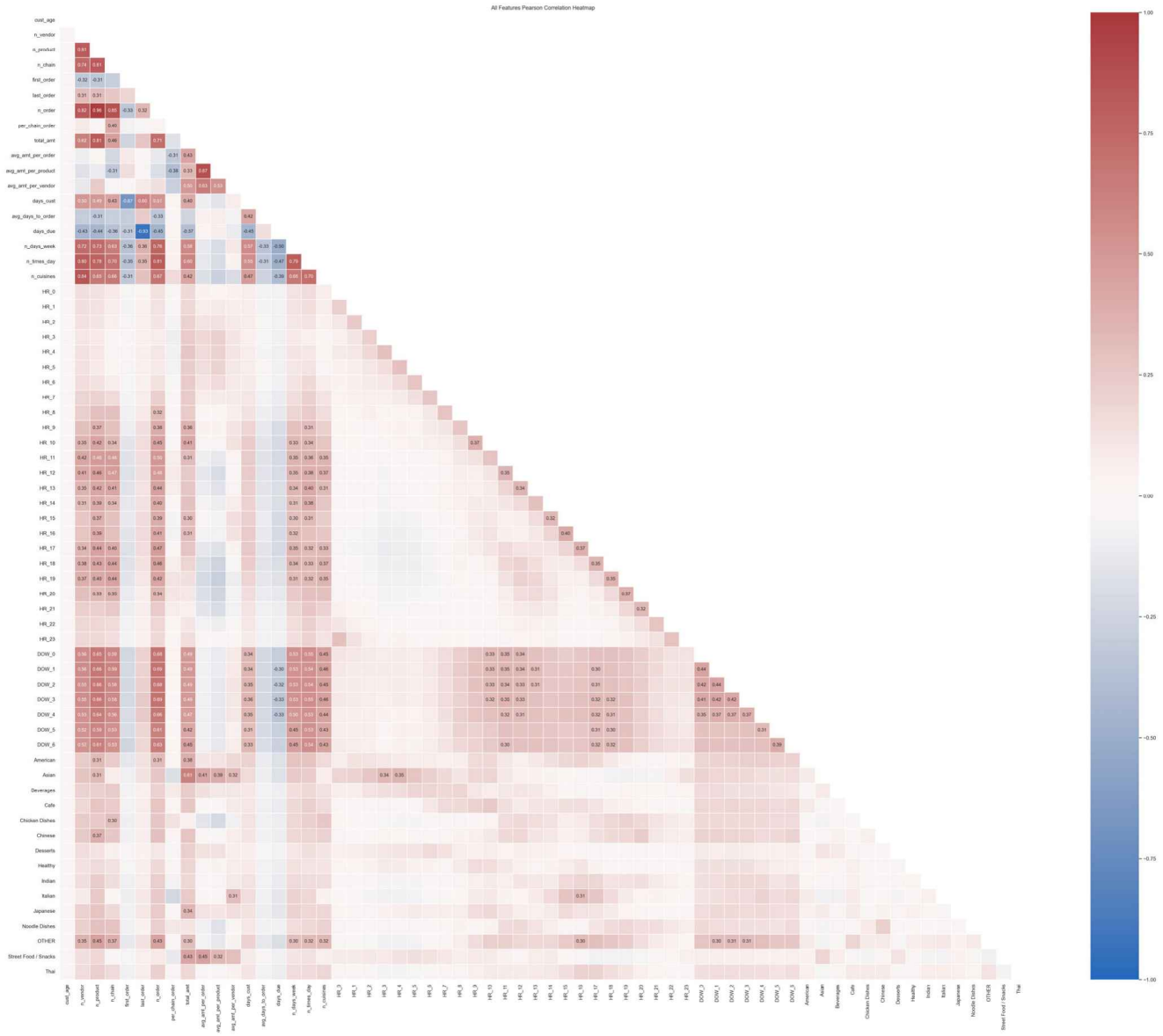


Figure 19 - Correlation Heatmap for All Features (Excluding Non-Metric)

The number of vendors, products and orders made in chained restaurants seem to be fairly correlated among themselves, which could mean that customers that order more products, order them from more distinct vendors and often from chained restaurants.

The total amount spent and the number of orders directly correlate with the number of products, as to be expected since the more products are bought, the more money is spent, and more likely it is for the products to have been bought on different occasions.

First order is inversely correlated with the number of vendors, products, purchases in chained restaurants and orders and the total amount spent - since it is a first order it makes sense that these values would be at their lowest.

The longer a person has been a customer, the more orders tend to be placed and the bigger the number of vendors and products. However, the average days between each order is also bigger, which could mean that the customers are not buying as regularly, i.e. this is a strong indicator that propensity towards consumption tends to deaccelerate with time.

Regarding the hours, we can see two order spikes around lunch and dinner time, with the lunchtime spike starting rather early which can indicate the orders of our workers that want to make sure their lunch is delivered in time for their lunch break. In addition, the afternoon hours seem to show autocorrelation between lagged hours, this means that if we know that someone usually buys at a certain time in the afternoon, it is likely they have made purchases at the previous hour - this likelihood increases with the number of orders.

Regarding the weekdays, it seems the number of orders has a lower correlation with Friday and Saturday and that the days as customers is also less correlated with these two weekdays - this could mean our customer base tends to order more on the weekdays.

Regarding the cuisines, the Asian cuisine shows high correlation with the total amount spent by a customer, implying that higher spenders tend to order Chinese food; the Chicken dishes seem to be ordered a lot from chain restaurants; the OTHER cuisine gets a lot of orders with a bigger number of products and from chain restaurants and is mostly consumed during the afternoon.

High spenders mostly spend their money on Asian, Street Food/Snacks, American and Japanese dishes, while the overall product volume is greatest in OTHER, Chinese, American and Asian. The cuisine on which customers tend to spend more money per order is on Street Food/Snacks. OTHER is highly correlated with 4 days of the week - Monday to Thursday. To sum up, when looking at the cuisines and time, 4 o'clock in the afternoon seems to be a great time to eat some pasta.

Regarding our customer base (i.e., people that have ordered more than once), we conclude we have found our 9 to 5 workers, as the days due inversely correlates with the working days of the week - this means that for those customers for which this correlation holds, we can say that they tend to order more often and below the average days to order threshold for which we would expect the average customer to place a new order.

Non-Metric Features

We performed the chi-square statistic for these features and found out that all of them are associated. details?

APPENDIX D – MULTIVARIATE ANALYSIS

Three-Way ANOVA

We tried to understand the customer spending habits according to their city, payment method and last promotion. We notice that city 2 tend to buy inexpensive products, while city 8 tends to buy more

expensive products. When it comes to the payment methods and promotions, it does not appear to be any difference across the cities or themselves.

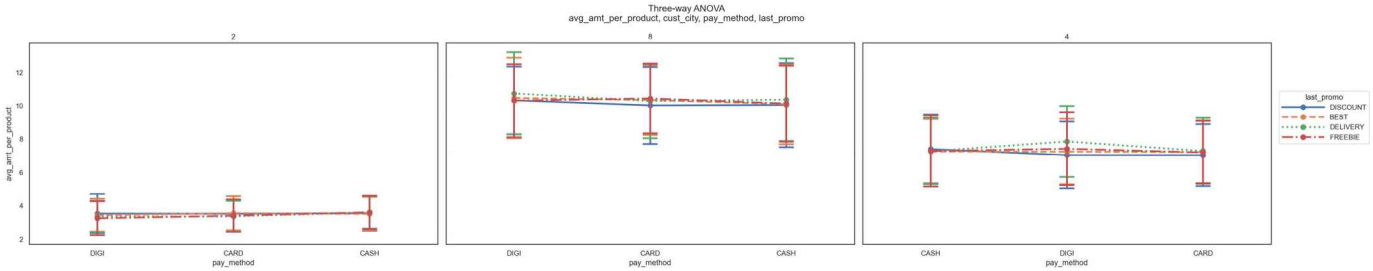


Figure 20 - Three-way ANOVA of Average Amount per Product, Customer City, Payment Method and Last Promotion

We also plotted the three-way ANOVAs for each of the cuisines for each customer city with last promotion as the x-axis value and the line colours represent if more than 50% of the orders were made in chain restaurants or not. These plots can be found in the Annexes Figures 42 to 56.

Pair Plots

We plotted the pair plots for all features with themselves but found no meaningful information worth mentioning.

Vendor Count and Customer Age

We verify that across the ages between 15 and 55, the average vendor count remains fairly stable nearing 4 vendors. However, from that age onwards, this value follows a very erratic pattern.

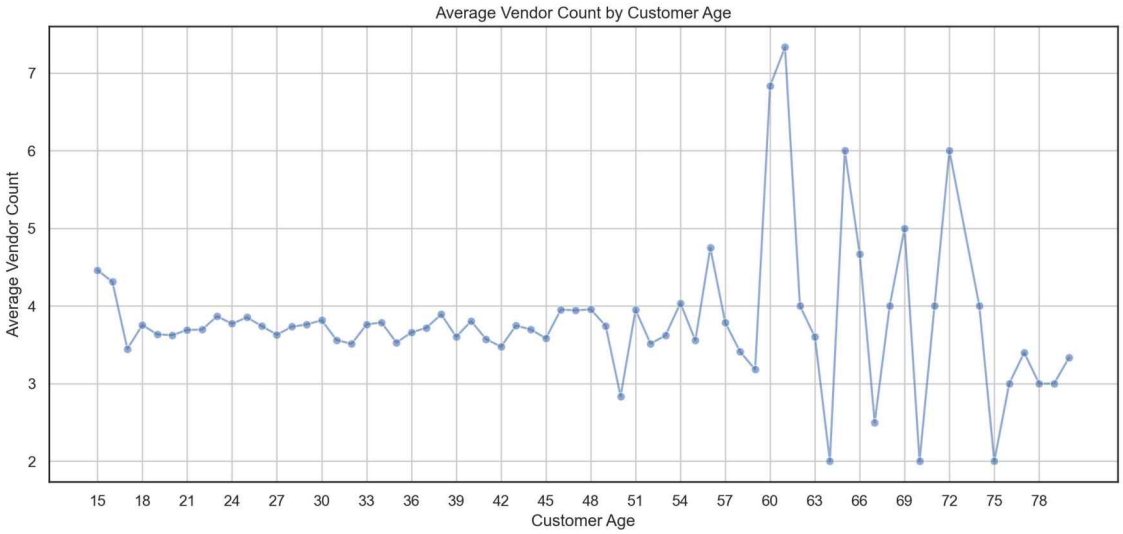


Figure 21 - Line Plot of Average Vendor Count by Customer Age

Over Time Analysis

The distribution of products sold and number of orders over time follows a similar distribution to that of the revenue mentioned in the main part of the report, which tells us the restaurants did not change their product offer and that the revenue is only increasing because we sold more products over more orders.

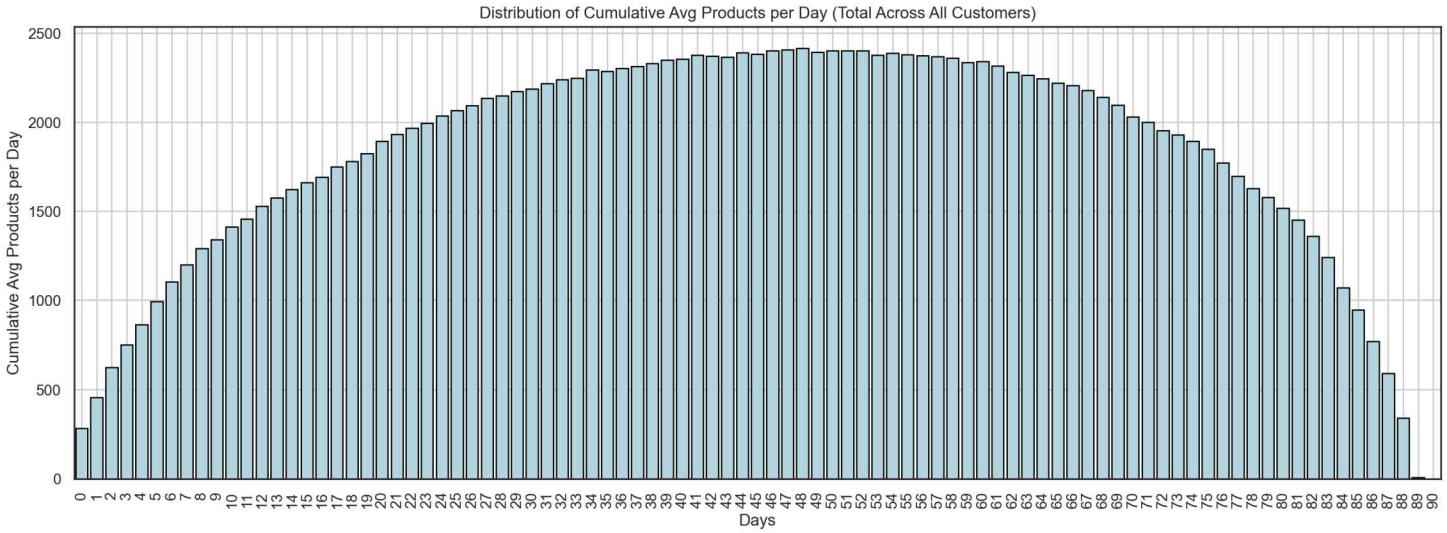


Figure 22 - Distribution of Cumulative Average Products per Day (for All Customers)

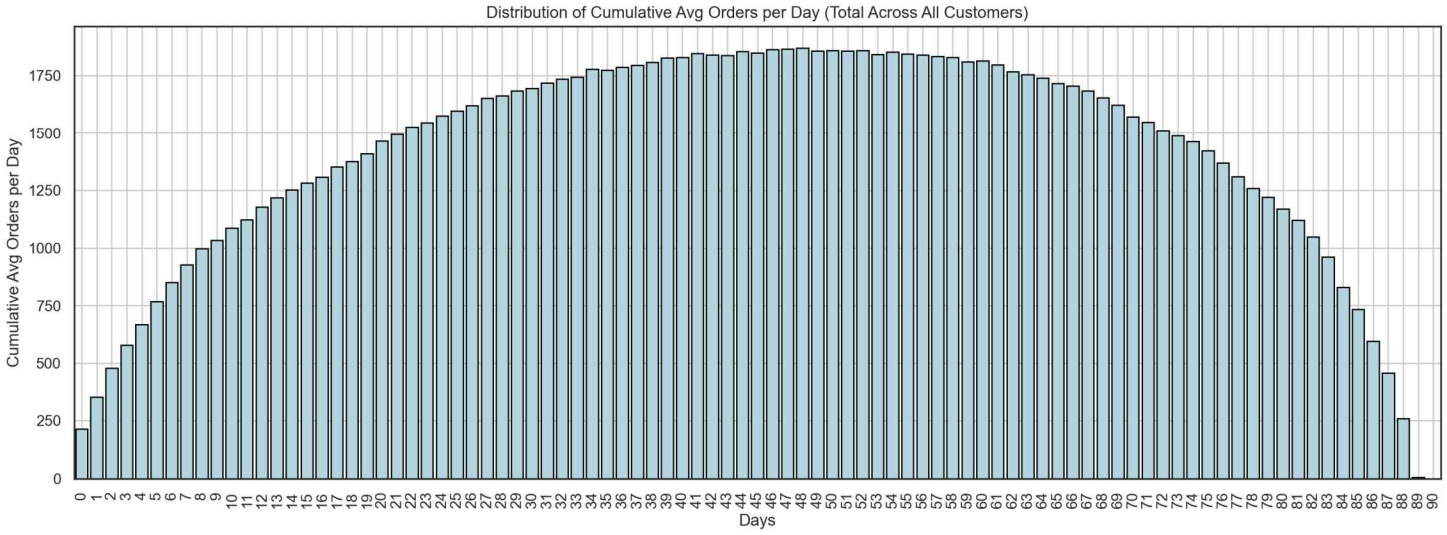


Figure 23 - Distribution of Cumulative Average Orders per Day (for All Customers)

ANNEXES

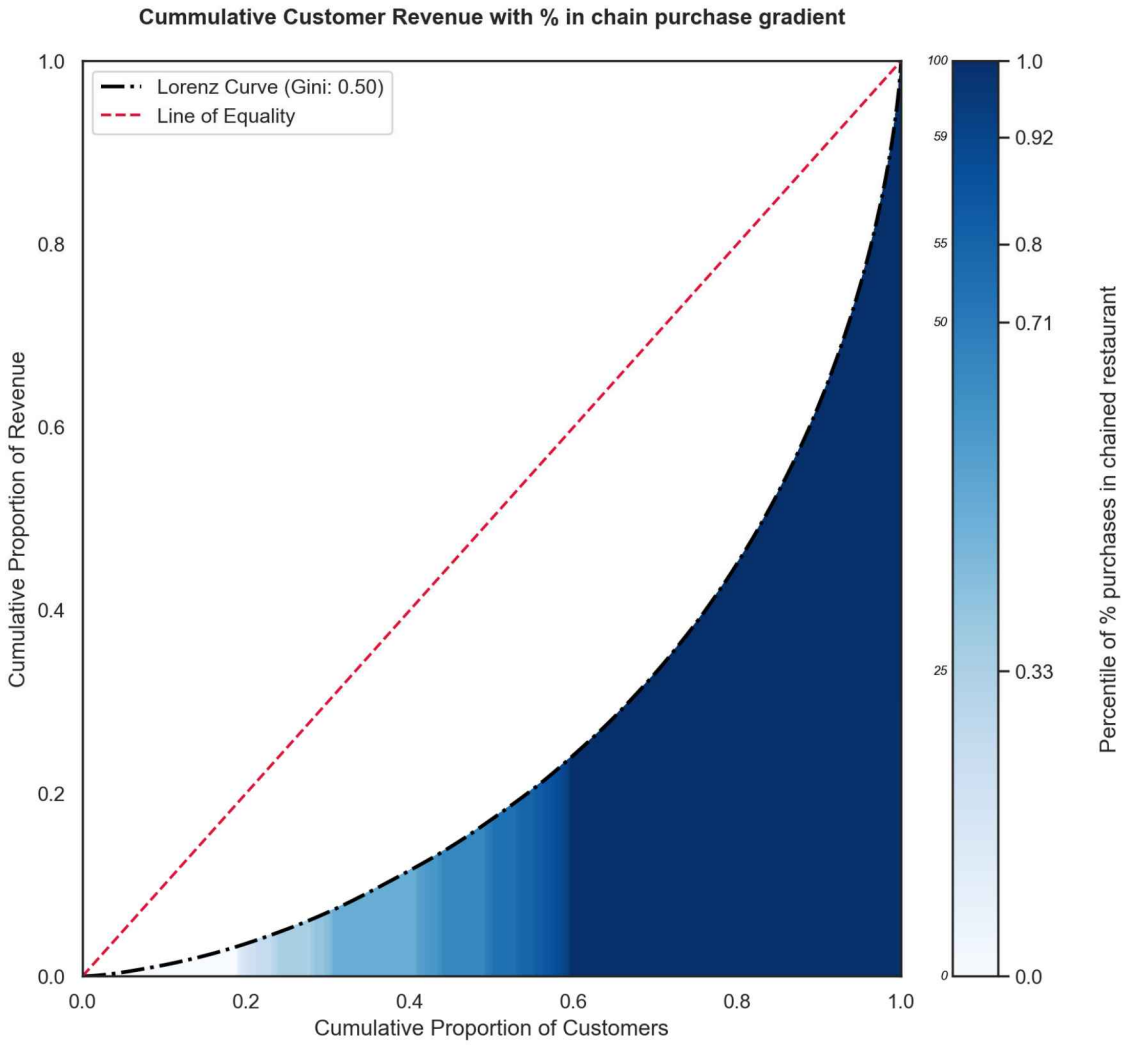


Figure 24 - Lorenz Curve of Cumulative Customer Revenue with Percentage of Orders in Chain Restaurants Purchase Gradient

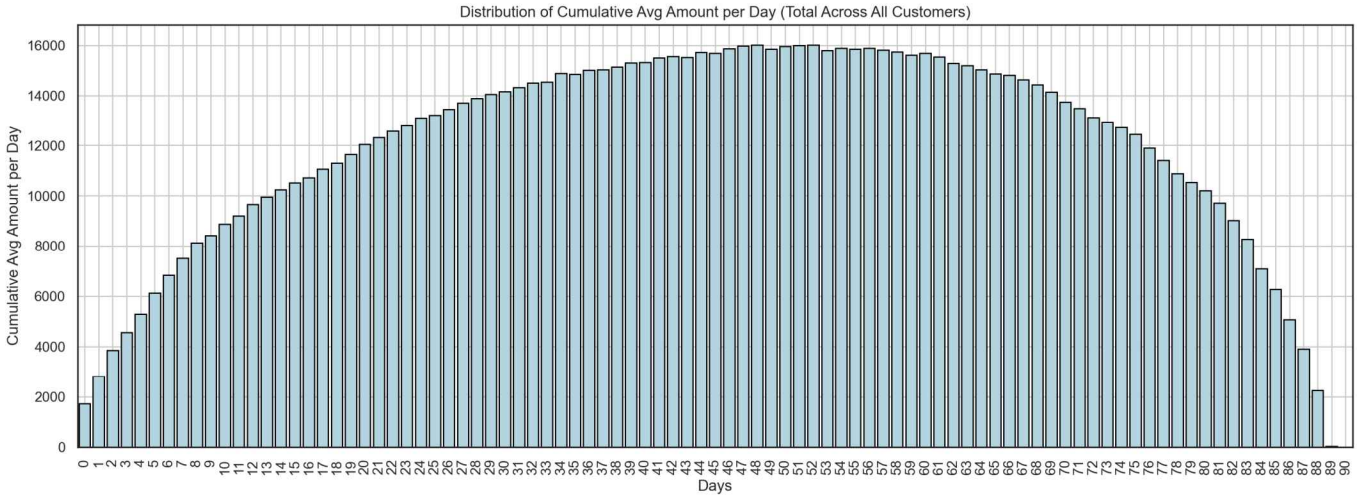


Figure 25 - Distribution of Cumulative Average Amount per Day (for All Customers)

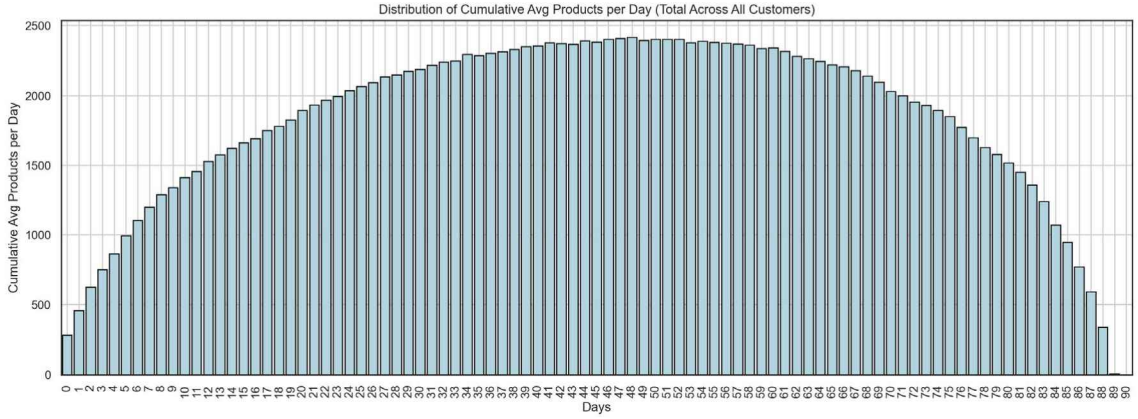


Figure 26 - Distribution of Cumulative Average Products per Day (for All Customers)

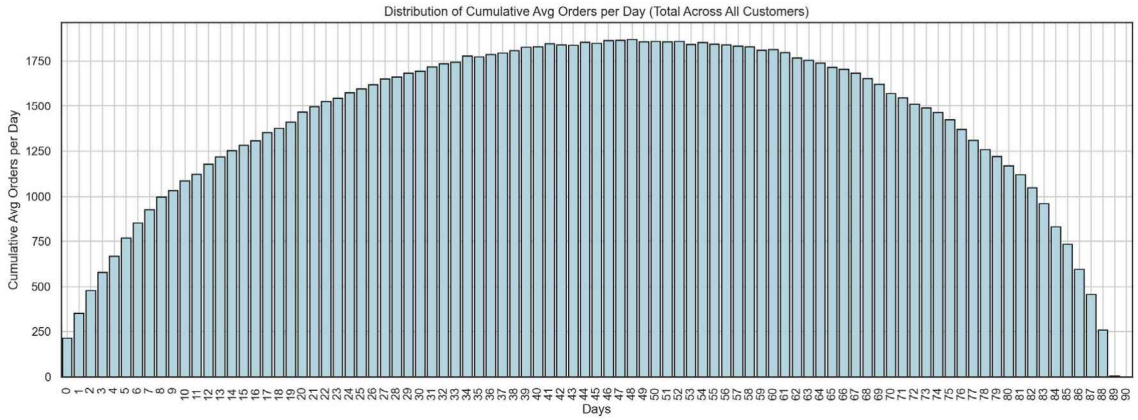


Figure 27 - Distribution of Cumulative Average Orders per Day (for All Customers)

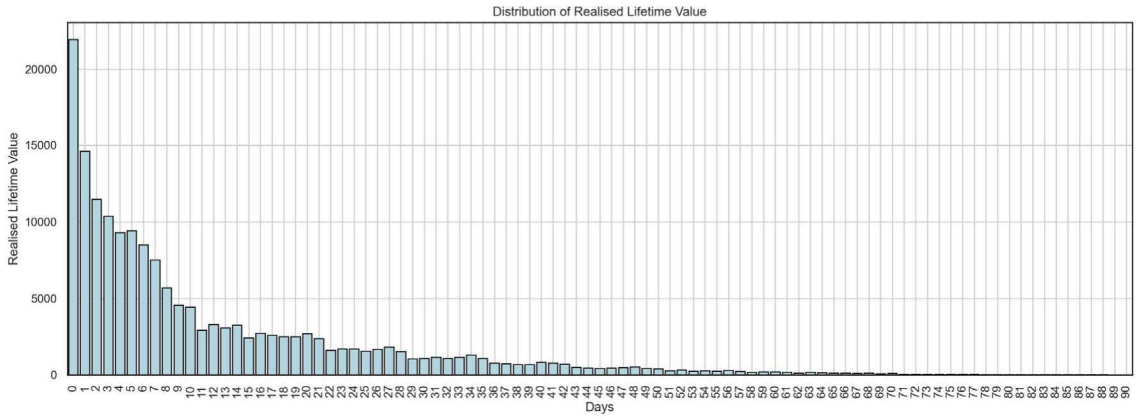


Figure 28 - Distribution of Realized Customer Lifetime Value

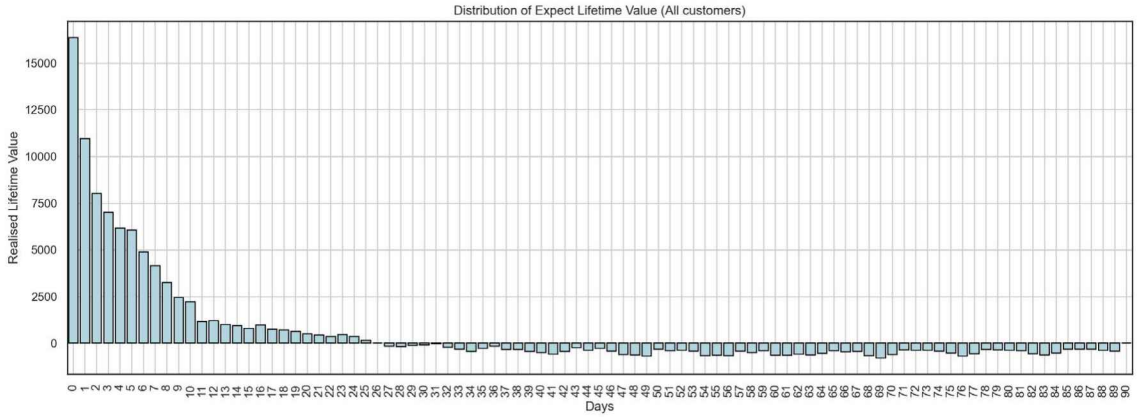


Figure 29 - Distribution of Expected Customer Lifetime Value

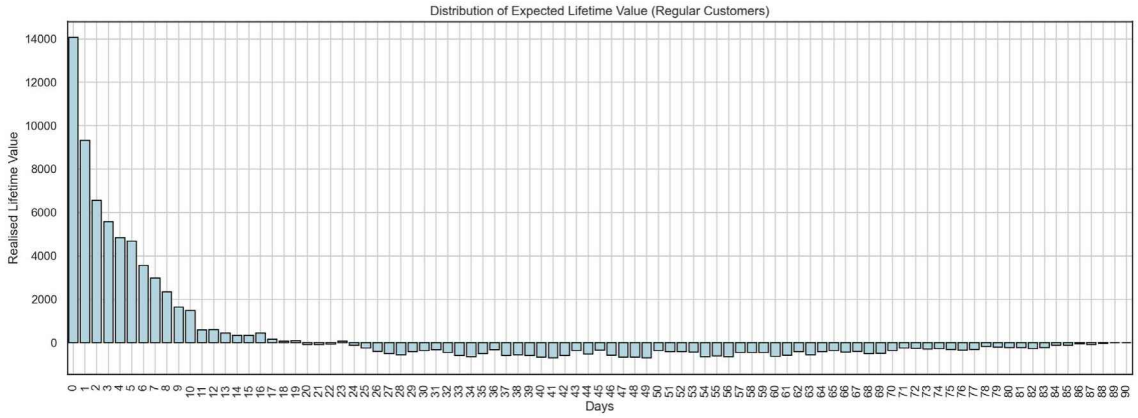


Figure 30 - Distribution of Expected Customer Lifetime Value (for Regular Customers)

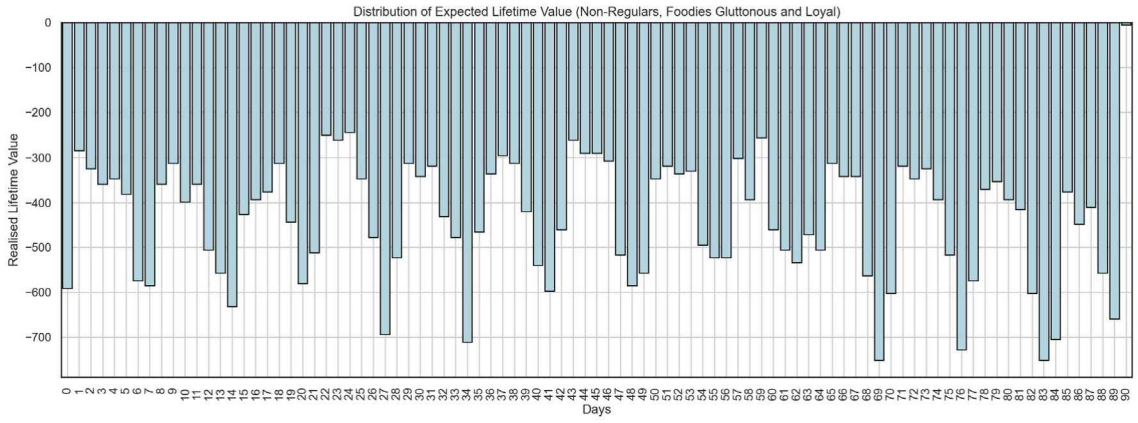
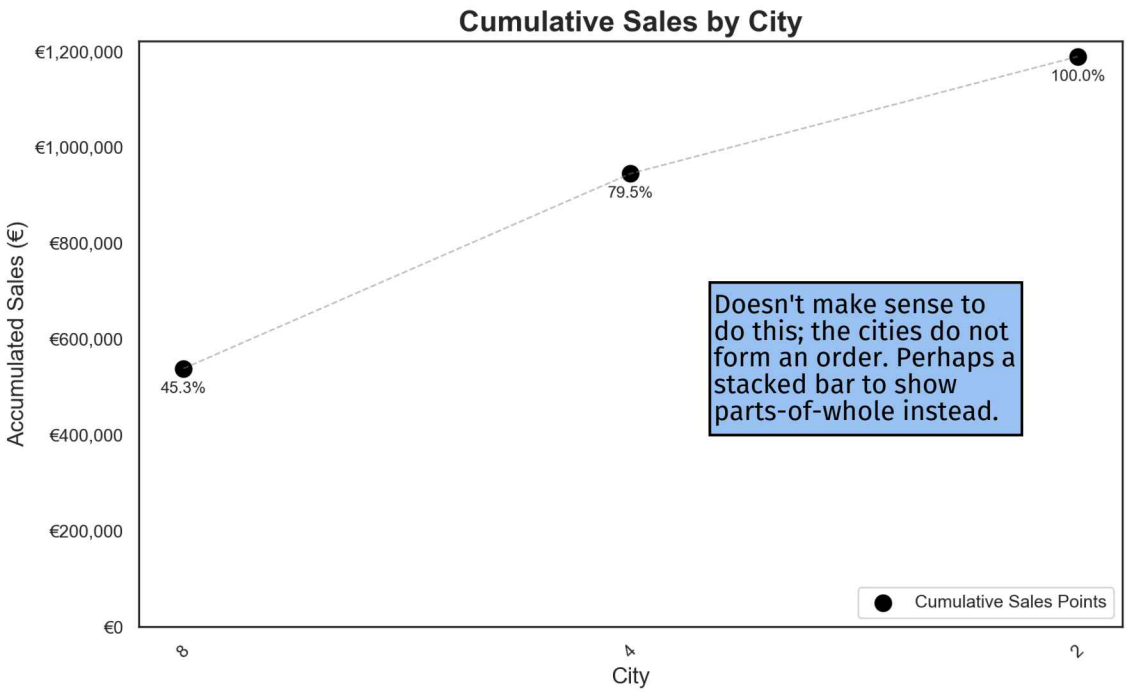


Figure 31 - Distribution of Expected Customer Lifetime Value (for One-time Customers)



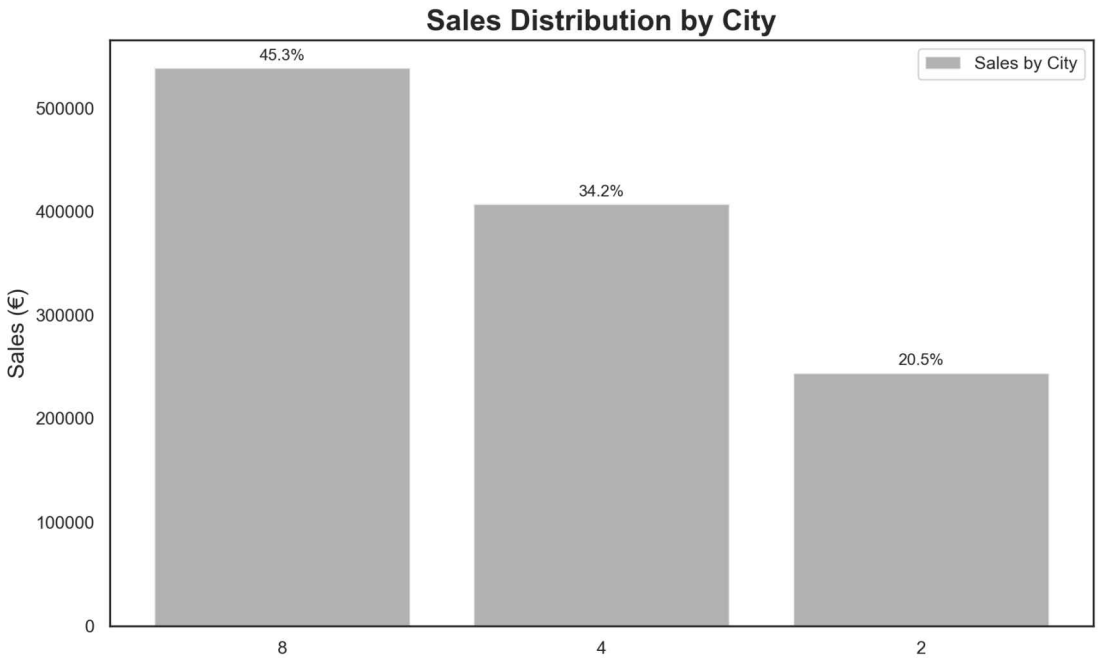
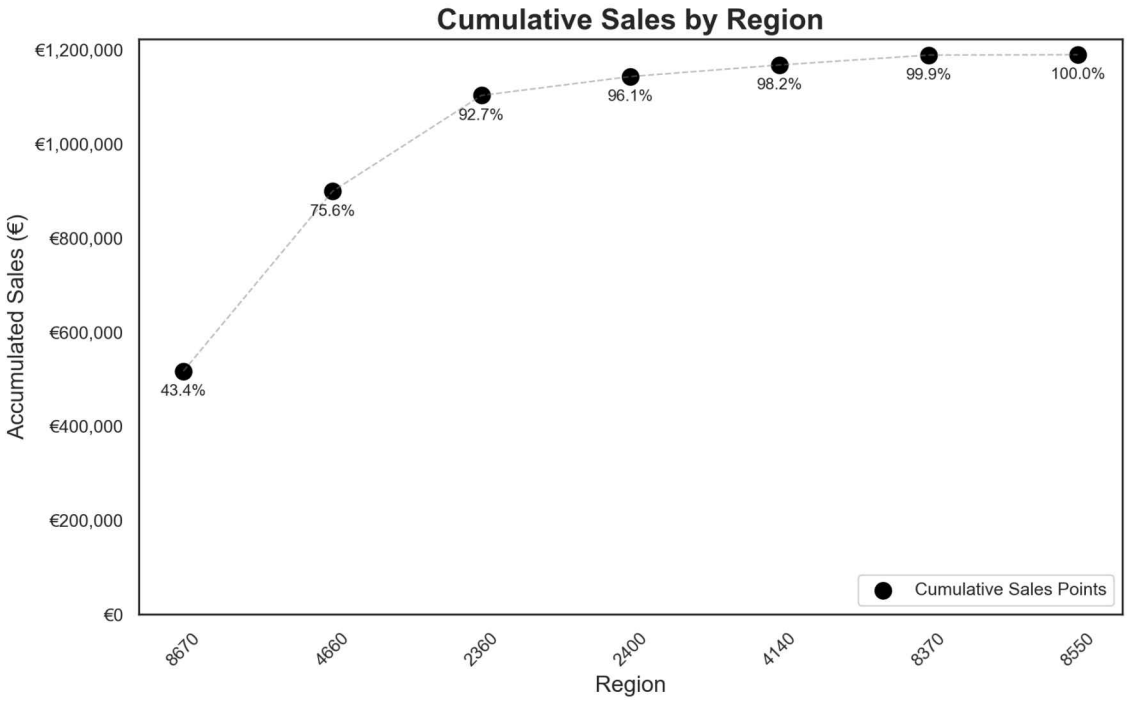


Figure 32 - Cumulative Sales and Sales Distribution by Customer City



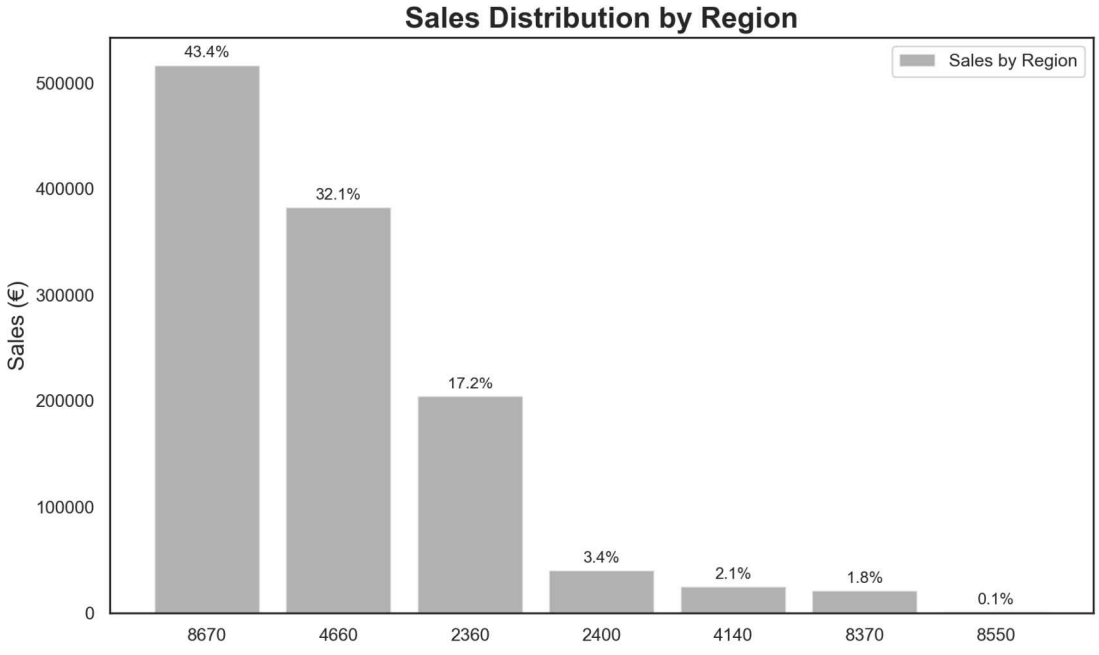


Figure 33 - Cumulative Sales and Sales Distribution by Customer Region

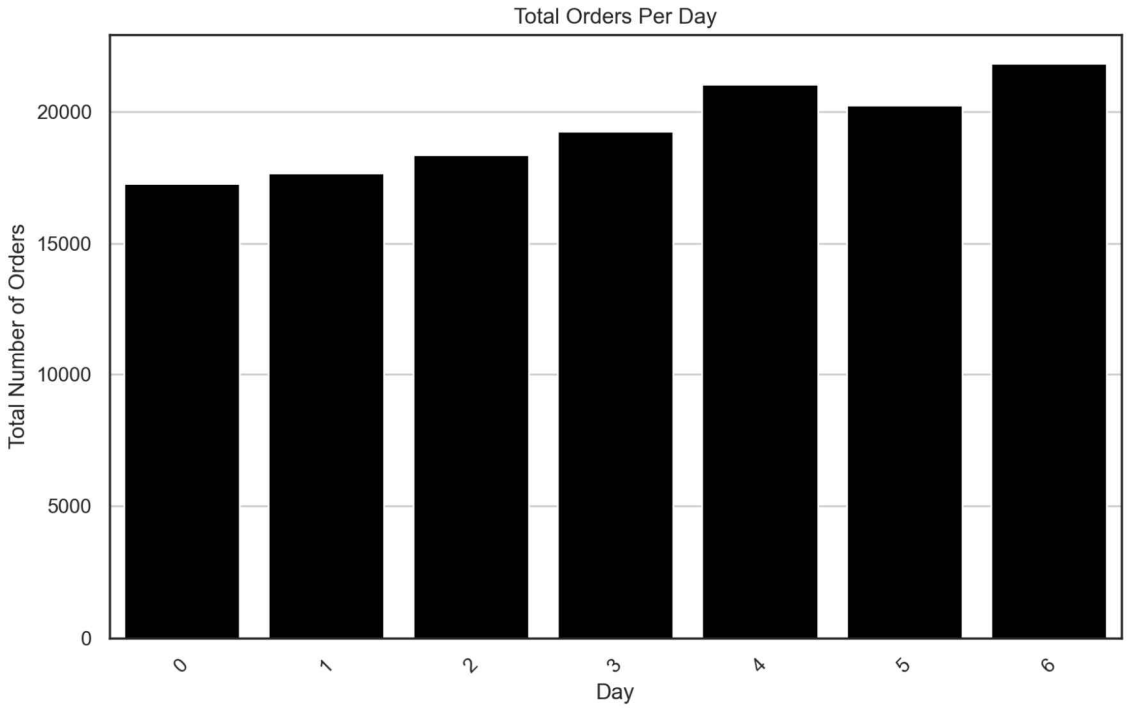


Figure 34 - Histogram of Total Orders per Day

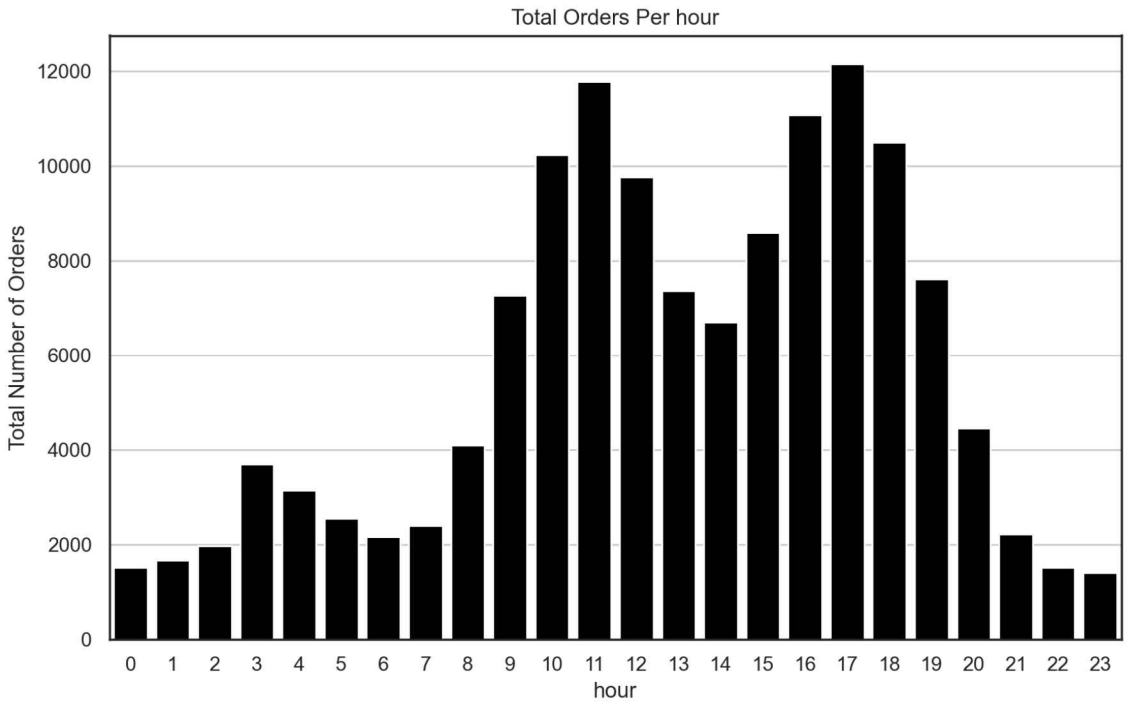


Figure 35 - Histogram of Total Orders per Hour

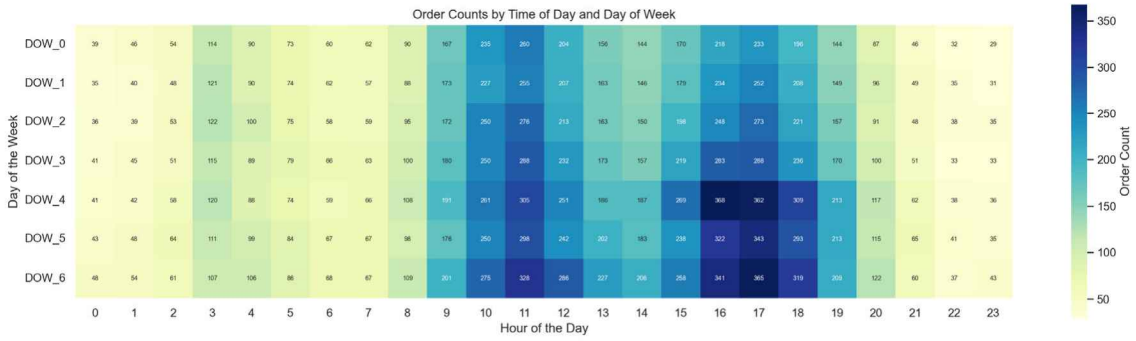


Figure 36 - Order Counts by Time of Day and Day of Week

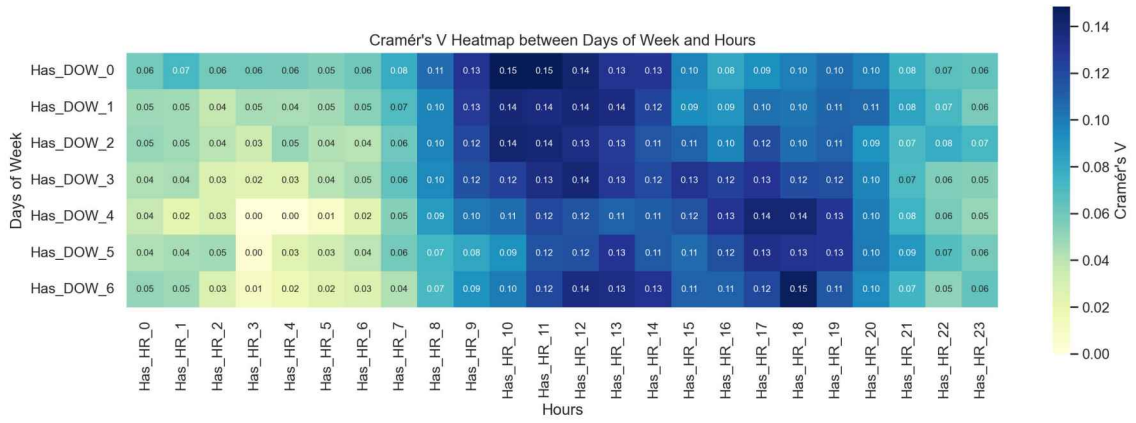


Figure 37 - Cramér's V Heatmap between Days of Week and Hours

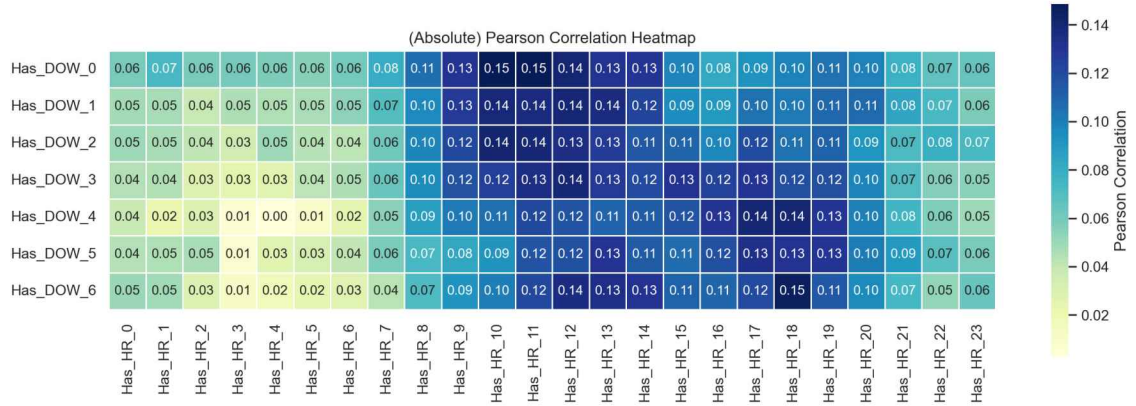


Figure 38 - (Absolute) Pearson Correlation Heatmap

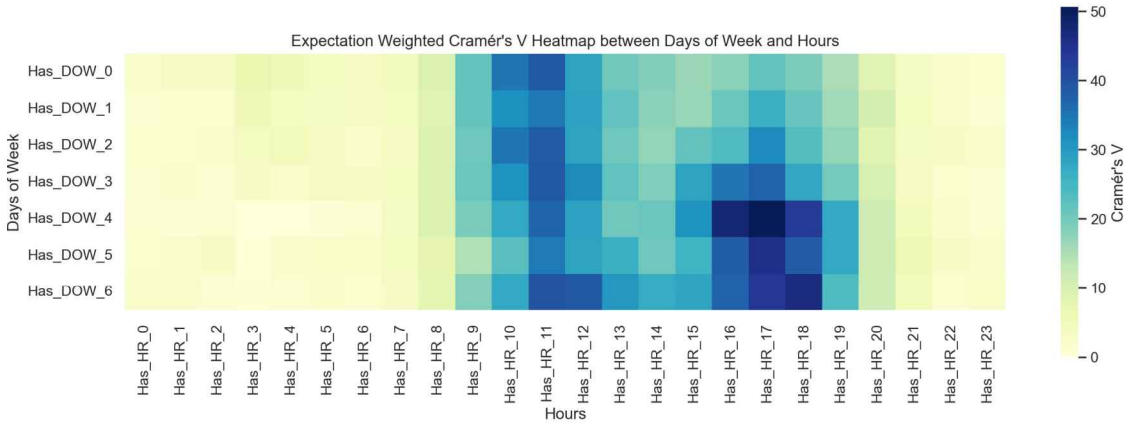


Figure 39 - Expectation Weighted Cramér's V Heatmap between Days of Week and Hours

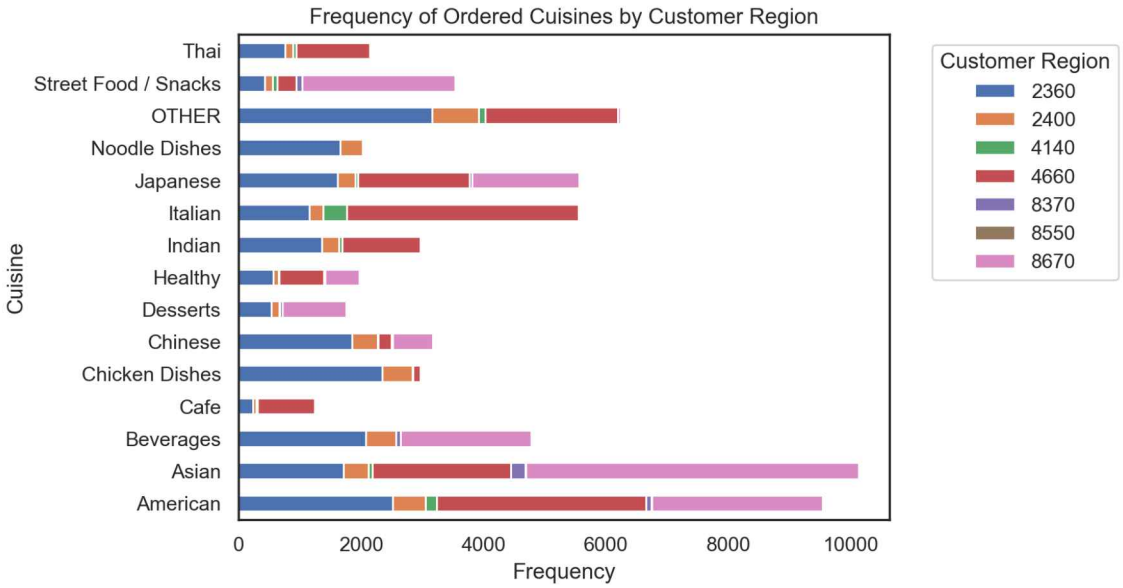


Figure 40 - Frequency of Ordered Cuisines by Customer Region

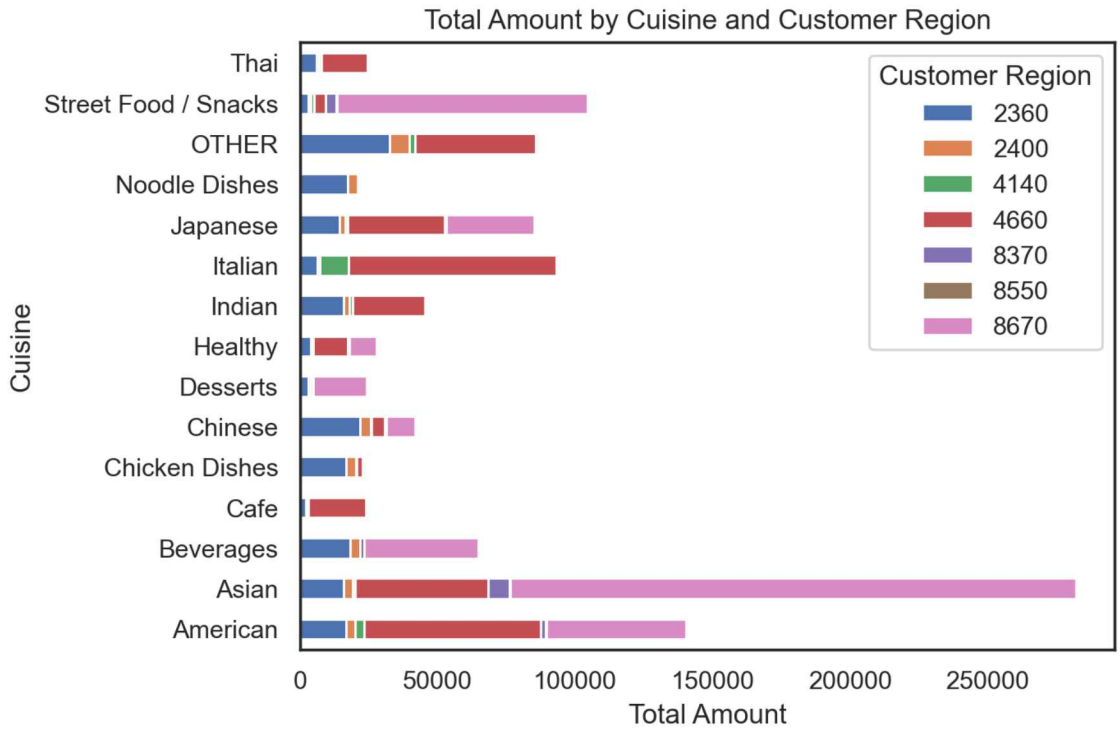


Figure 41 - Total Amount by Cuisine and Customer Region

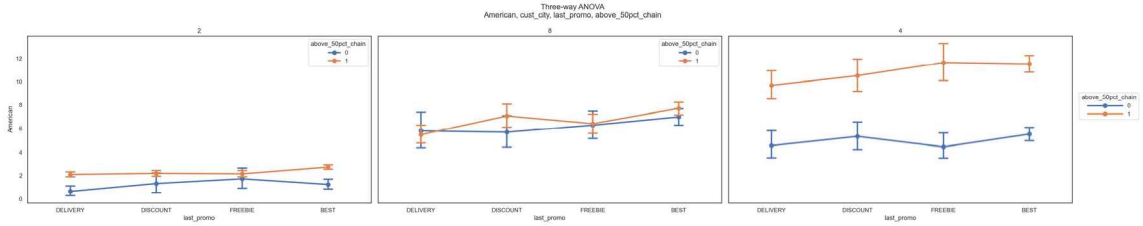


Figure 42 - Three-way ANOVA American Cuisine

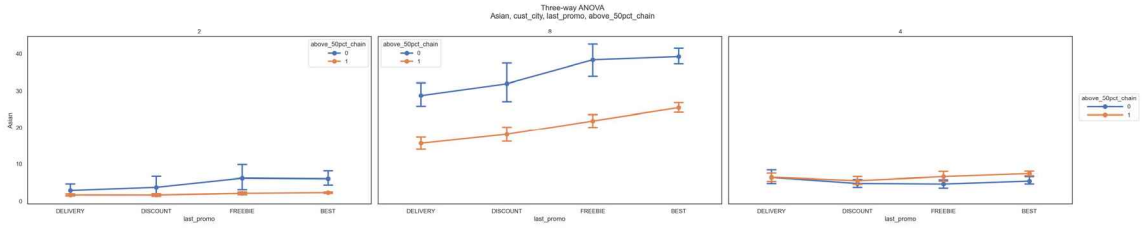


Figure 43 - Three-way ANOVA Asian Cuisine

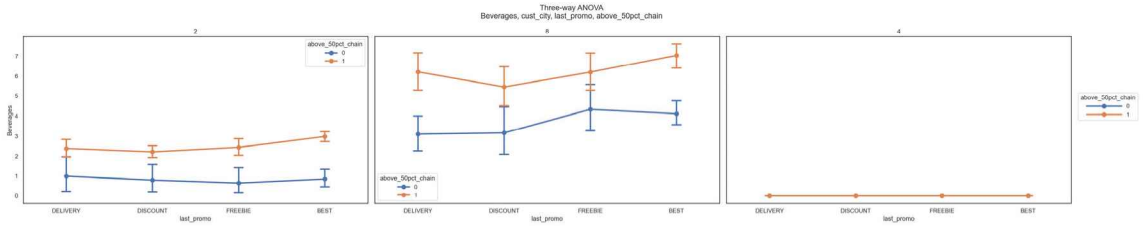


Figure 44 - Three-way ANOVA Beverages

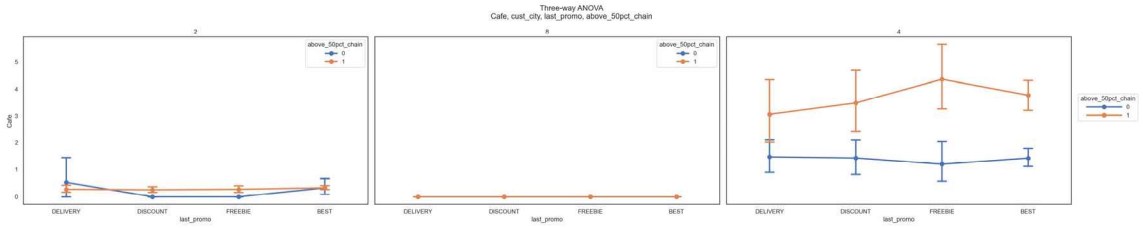


Figure 45 - Three-way ANOVA Cafe

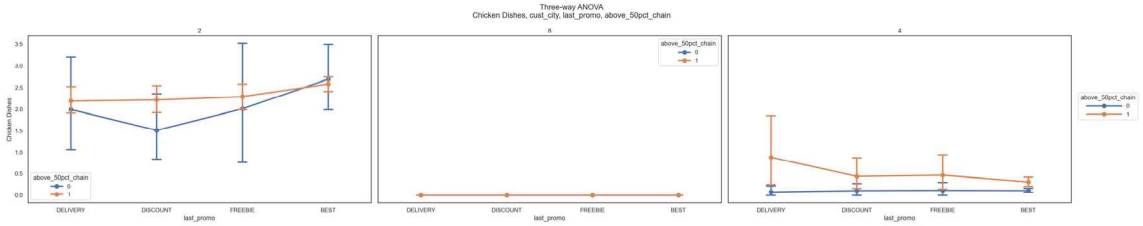


Figure 46 - Three-way ANOVA Chicken Dishes

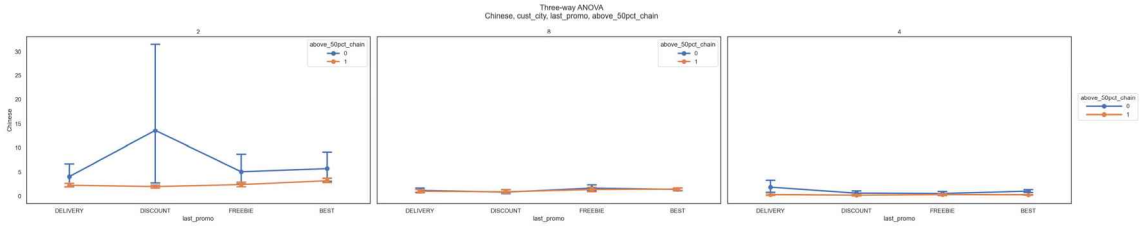


Figure 47 - Three-way ANOVA Chinese Cuisine

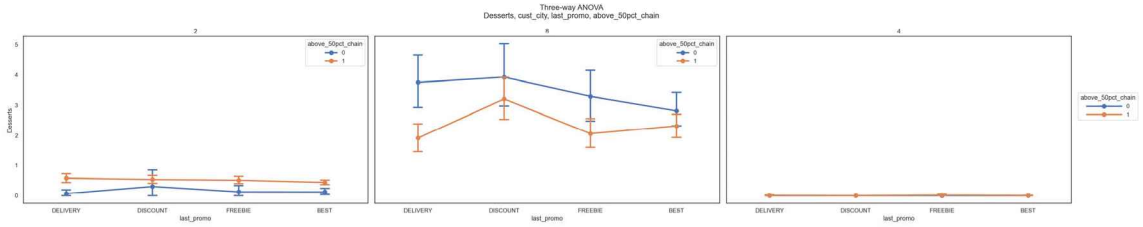


Figure 48 - Three-way ANOVA Desserts

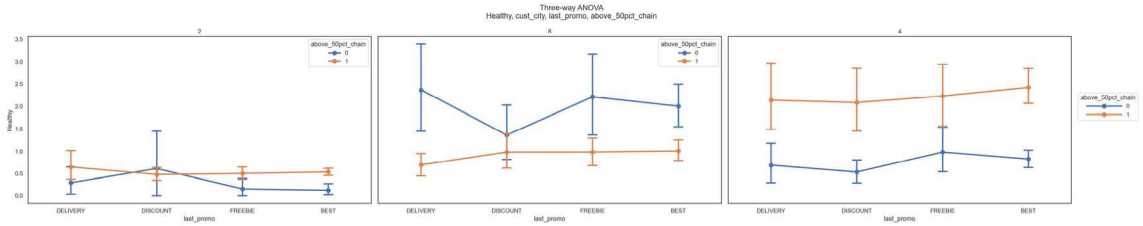


Figure 49 - Three-way ANOVA Healthy

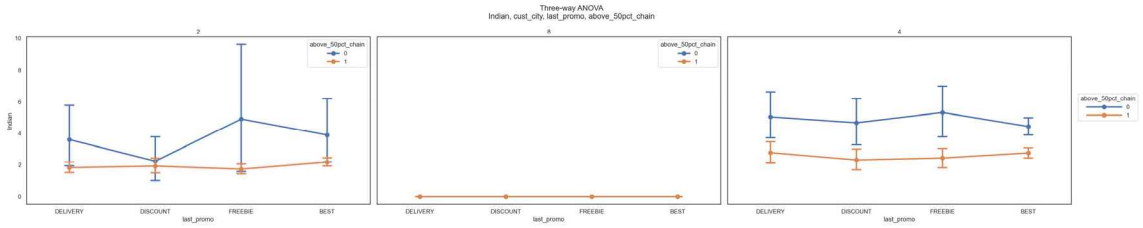


Figure 50 - Three-way ANOVA Indian Cuisine

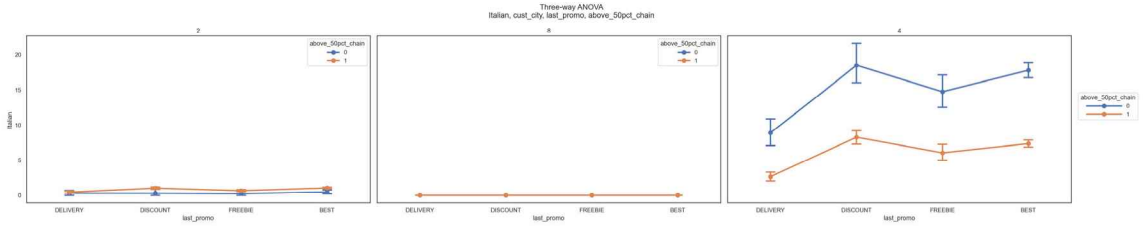


Figure 51 - Three-way ANOVA Italian Cuisine

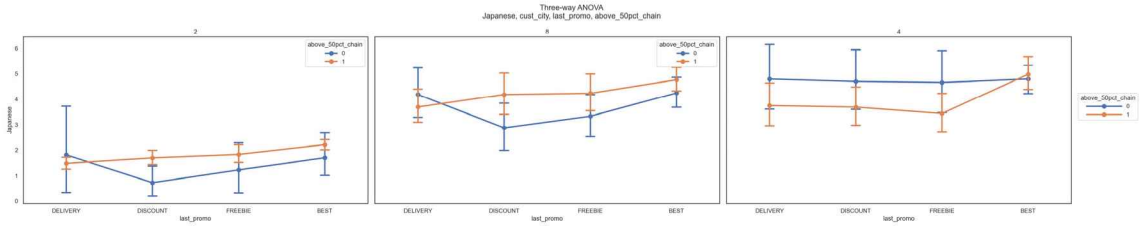


Figure 52 - Three-way ANOVA Japanese Cuisine

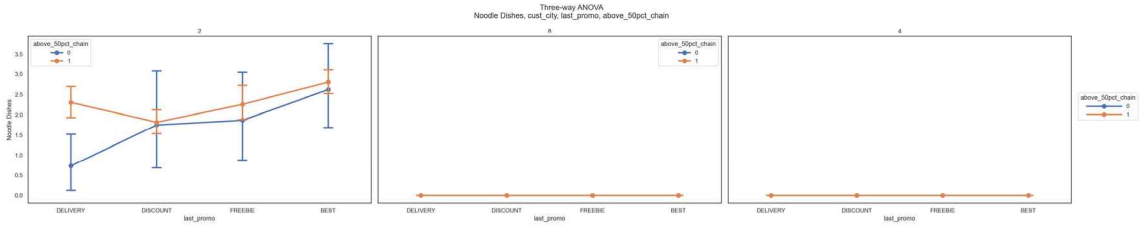


Figure 53 - Three-way ANOVA Noodle Dishes

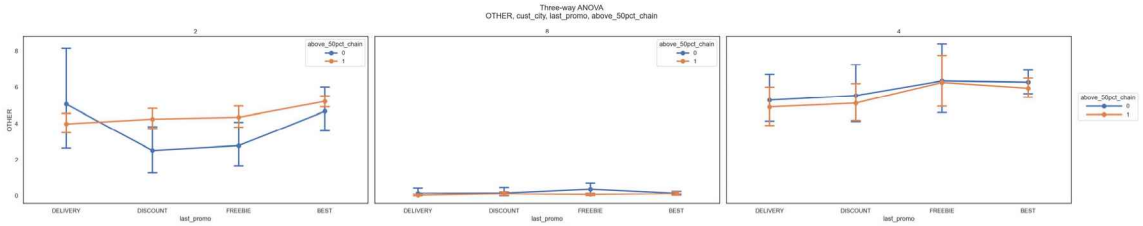


Figure 54 - Three-way ANOVA OTHER Cuisine

Figure 55 - Three-way ANOVA Street Food/Snacks

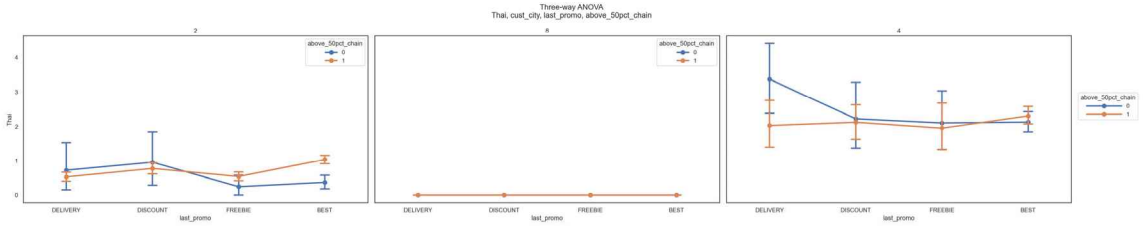


Figure 56 - Three-way ANOVA Thai Cuisine

	count	mean	std	min	25%	50%	75%	max
American	30945.0	4.903148	11.670353	0.0	0.0	0.0	5.72	280.21
Asian	30945.0	9.987160	23.587214	0.0	0.0	0.0	11.87	896.71
Beverages	30945.0	2.307313	8.501437	0.0	0.0	0.0	0.00	229.22
Cafe	30945.0	0.796570	6.420199	0.0	0.0	0.0	0.00	326.10
Chicken Dishes	30945.0	0.770766	3.662677	0.0	0.0	0.0	0.00	219.66
Chinese	30945.0	1.438397	8.250136	0.0	0.0	0.0	0.00	739.73
Desserts	30945.0	0.881988	5.266091	0.0	0.0	0.0	0.00	230.07
Healthy	30945.0	0.956433	5.849520	0.0	0.0	0.0	0.00	255.81
Indian	30945.0	1.639088	7.484511	0.0	0.0	0.0	0.00	309.07
Italian	30945.0	3.255055	11.330662	0.0	0.0	0.0	0.00	468.33
Japanese	30945.0	3.008330	10.218632	0.0	0.0	0.0	0.00	706.14
Noodle Dishes	30945.0	0.717123	4.540086	0.0	0.0	0.0	0.00	275.11
OTHER	30945.0	3.012379	9.621473	0.0	0.0	0.0	0.00	243.18
Street Food / Snacks	30945.0	3.924271	15.526880	0.0	0.0	0.0	0.00	454.45
Thai	30945.0	0.836217	4.381337	0.0	0.0	0.0	0.00	136.38

I know you can do better, I expect a proper table here instead of a screenshot :)

Figure 57 - Cuisine Features Statistics

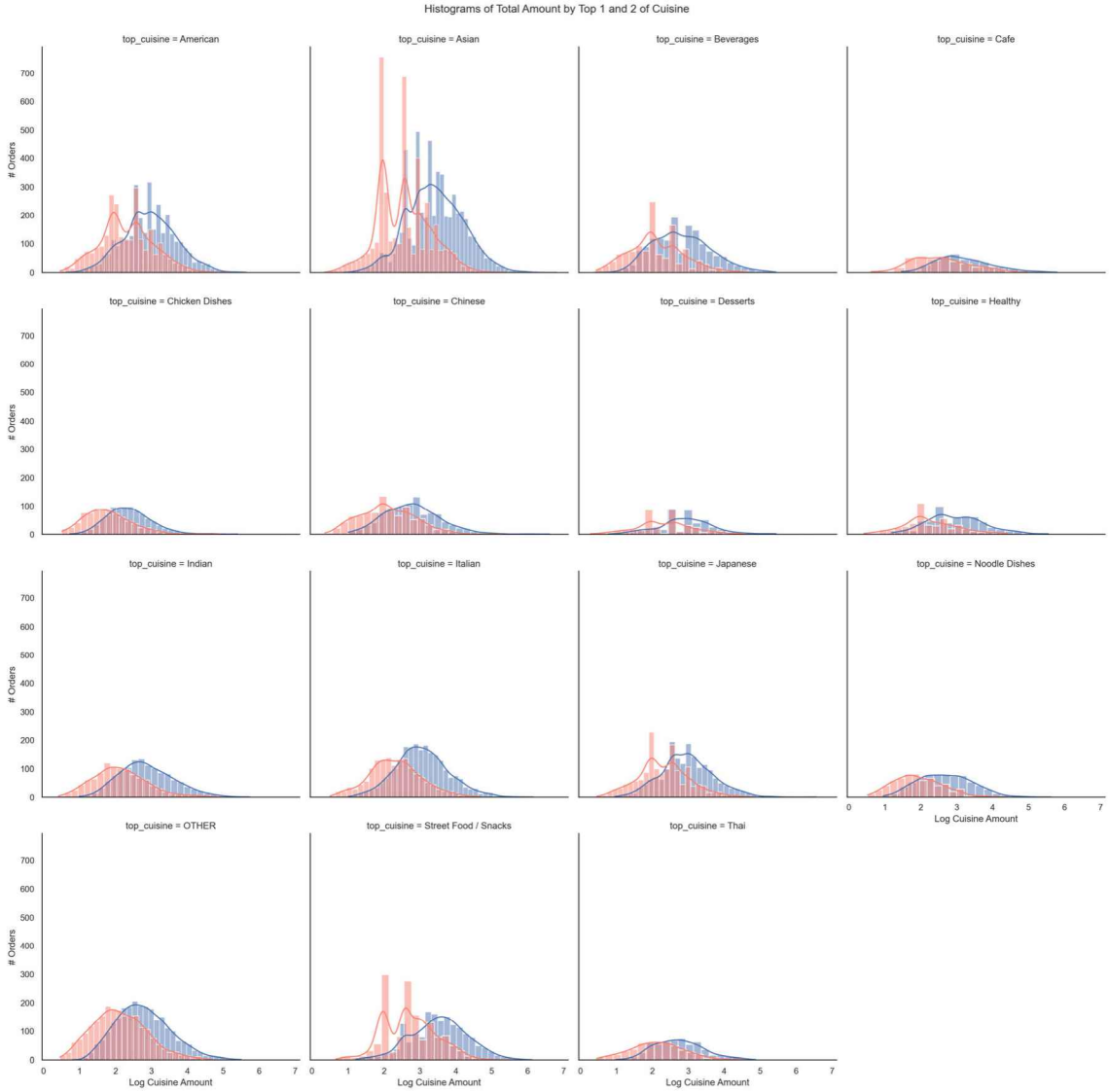


Figure 58 - Histograms of Total Amount by Top 1 and 2 Cuisines

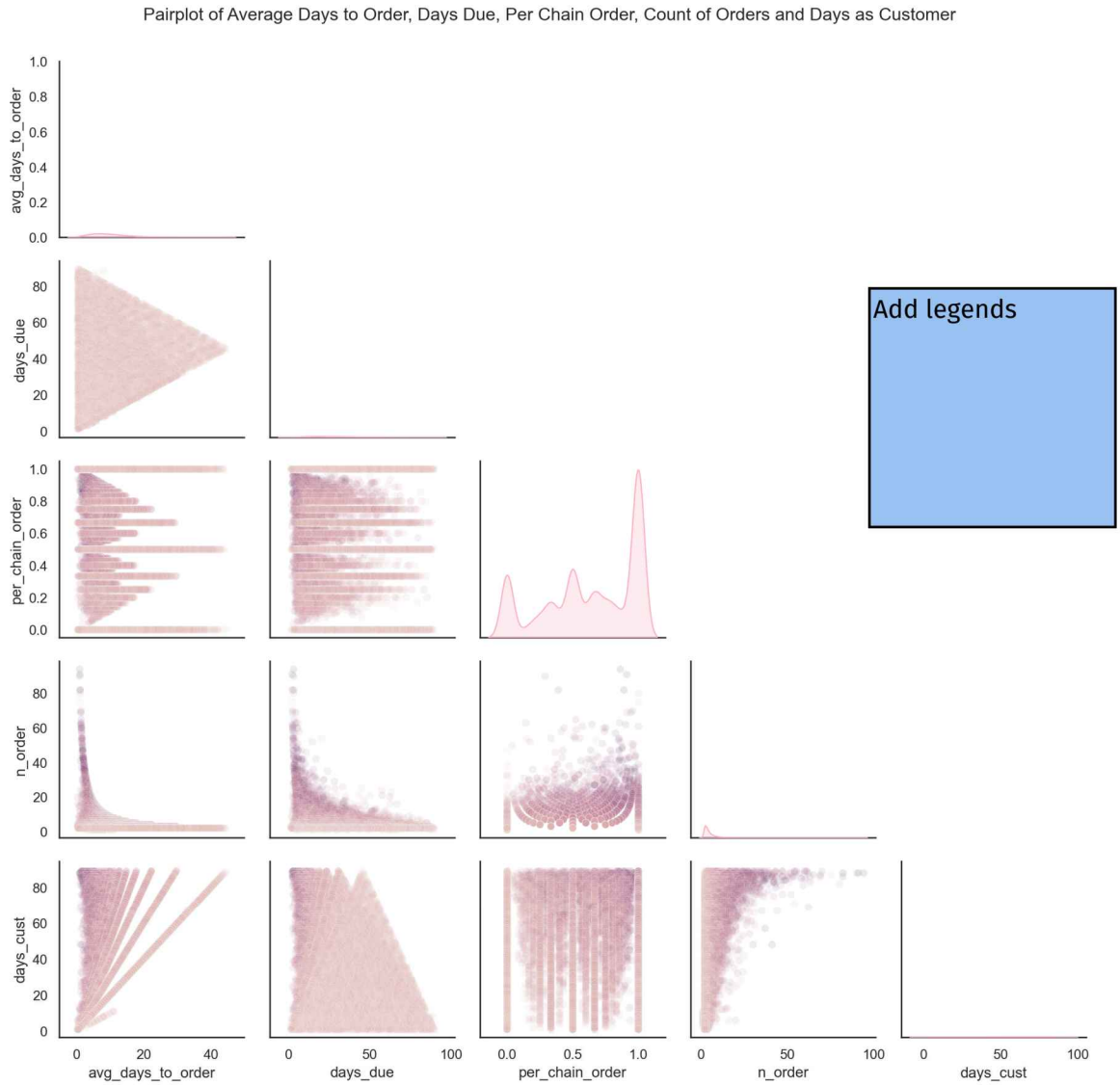


Figure 59 - Pair Plot of Average Days to Order, Days Due, Per Chain Order, Count of Orders and Days as Customer