Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: MARKET BASKET ANALYSIS

Ana Carolina Ottavi, number: 20220541

Carolina Bezerra, number: 20220392

Duarte Girão, number: 20220670

João Pólvora, number: 20221037

Luca Loureiro, number: 20221750

Group Q: OptimaDataConsulting

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2023

INDEX

1. EXECUTIVE SUMMARY

C's Asian Food Brand in Cyprus has been facing challenges due to the increase in competition in the restaurant business. Therefore, we prepared this report, based on the CRISP-DM methodology[1], where we performed several technical analyses on the sales data provided by the client, addressing the key business questions and providing our insights to maintain the client's profit margin and continuous growth due to increasing competition and customers' changing habits. Therefore, an exhaustive exploratory data analysis and data visualization were performed to understand and demonstrate customer patterns and preferences based on patterns and trends in customer behavior, differences between dine-inn and delivery customers, product offerings, and consumption tendencies, in order to be able to create strategies such as the creation of set menus and the introduction of new products and finally revert the actual situation C has been facing. We created a set of recommendations and solutions based on our results that are in line with C's goals. In order to better fulfill the demands of our customers, these insights and suggestions will help guide decision-making and optimize product offerings.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

Due to the intense competition in the restaurant industry, Cyprus' C's Asian food brand is having difficulty sustaining both growth and a profit margin. The client therefore provided their transactions data to better understand customer trends and preferences in order to address these difficulties.

2.1. BACKGROUND

The restaurant business has become increasingly competitive, therefore, companies like C's Asian Food must adapt to changing market dynamics as the adage "if I build it, they will come" no longer holds true. When developing plans, C needs to consider a number of factors, including how dine-in and delivery clients differ, how to evaluate the available products, and how to spot trends in consumer behavior. Through this project, a set of menus will be developed, new items will be introduced, alternative products will be understood, and cross-selling and up-selling will be encouraged.

2.2. BUSINESS OBJECTIVES

C's main business objectives are to maintain profitability, guarantee continuous growth, and improve the competitive position of its Asian food brand in Cyprus by taking advantage of its sales data and customer insights in order to make decisions to optimize its products, strategies, and customer patterns of consumption and preferences.

2.3. BUSINESS SUCCESS CRITERIA

Three significant factors have been identified in order to assess the success of C's Asian food brand. These criteria include enhanced product offers and targeted marketing campaigns to maximize profit

[1] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0, Step-by-step data mining guide. SPSS Inc.

margins, consistent sales, and customer base expansion in order to keep businesses relevant and achieve high levels of customer satisfaction.

2.4. SITUATION ASSESSMENT

By March 29th, we must come up with a company solution leveraging information from client transactions that satisfies a set of goals. To find patterns in customer purchasing behavior that can be used to enhance menu selection, pricing tactics, and stock control, a market basket analysis is the suggested approach. However, there are significant hazards involved in performing a market basket study, including small or non-representative sample sizes and a lack of statistical skill among restaurant managers or owners.

2.5. DETERMINE DATA MINING GOALS

The main data mining goals of our project are: creation of menu sets, the introduction of new products, the understanding of substitute products, the recommendation of cross-selling and other possible results depending on the findings.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

The section's main goal was to examine the dataset for C's Asian food brand, which included 84,109 rows of sales transaction data and 12 columns. The dataset contains information on the products sold, quantities, prices, and clients. In order to comprehend the variables and their unique characteristics, the summary table was studied, giving preliminary insights for each of the features.

Features	Takeaways
DocNumber	This column has 11,147 unique values out of 84109 total rows, which suggests that there are some duplicates, which is possible since each DocNumber represents one transaction in the invoice. The most common document number is "TK0110053522018" and it appears 46 times. There are also some missing values (represented by NaN and NaT) in this column. There might be a possibility that a DocNumber that begins with “TK” is a code used for dine-inns and the code “TKD” is a code used for deliveries.
ProductDesignation	It was possible to disclose that there are 255 different possibilities for Product Designations, which suggests that there are many different products being sold. The most common product is "MINERAL WATER 1.5LT" and it appears 7,061 times. Most customers ordered one or more of the following products: Mineral Water 1.5lt, Egg fried rice, Spring Roll, Delivery Charge And Sweet Sour Chicken. There are no missing values in this column. There is a product designated “Delivery Charge”, which belongs to the family “EXTRAS” and its Total Amount ranges from 0 to 1.7..
ProductFamily	This column has 27 unique values out of 84,109 total rows, which suggests that there are relatively few product families being sold. There are 27 product families, with Starters

	being the most frequent (14,148 times), followed by Drinks, Rice, Meat and Extras. There are no missing values in this column. Both “WITH” and “HOLDS” categories don’t have a price associated.
Qty	There are 24 unique values for this feature. Its mean and quartiles have close values, which can suggest a normal distribution. Most people order 1 unit of each product, which indicates most customers purchase a small number of items per transaction. The mean quantity per transaction is 1.26 and the standard deviation is 0.92, which suggests that most transactions involve selling one unit of a product, however, there is still some variability in the quantity column. The minimum quantity is 1 and the maximum quantity is 53. Additionally, we can tell that 25%, 50%, and 75% of the data for the feature fall into this category, indicating that the distribution is heavily right skewed.
TotalAmount	The average total amount per transaction is 9.83, with a standard deviation of 20.66, which suggests that there is a wide range of transaction values, with some transactions being much larger than others. The maximum transaction amount was 3,000 and the minimum was 0. Since the range between the first and third quartiles is small, the data is tightly clustered around the median, indicating a low variance. The 25%, 50%, and 75% of the data fall between 3 and 12.6, indicating that the distribution is heavily skewed to the right. However, there seems to be a wrong punctuation for separating the whole number from the fractional part. There are 374 unique values registered for this feature in the dataset provided. Typically, the restaurant earns lower revenues from food deliveries than they do from in-person dining. Moreover, based on the “TotalAmount”, it’s possible to gather the cities where the customers came from that brought more revenue, which are Egkomi, Strobolos, Lakatameia and Leykosia.
InvoiceDateHour	This column has 11,146 unique values out of 84,109 total rows, which suggests that there are many different transaction dates and times. From our summary statistics table, we can verify that our dataset, in terms of this column, extends from the 1st of January of 2018 at 19:12:12 to the 31st of December of 2018 at 22:45:17. The busiest day in terms of hours is the 24th of December of 2018, which is Christmas Eve. It makes sense that more people come during the holidays. There are no missing values in this column.
EmployeeID	We identified 7 different employees that were working at the restaurant. Employee 2 was the one who took the most orders because they had the largest proportion of transactions. This was then confirmed when it was discovered that this individual was also connected to more sales overall.
IsDelivery	This column contains a binary variable indicating whether the purchase was a delivery (1) or not (0). This column has a mean value of 0.37 and a standard deviation of 0.48, which suggests that most transactions are not deliveries, but some are. 25%, 50%, and 75% of the data are zero, indicating that the vast majority of transactions are not deliveries. Most transactions take place in the restaurant, and a smaller portion of transactions are deliveries. We can conclude that in deliveries, there's a tendency for clients to personalize their requests with additional products. Products such as drinks aren't usually requested in deliveries. Starters are also more requested when dining physically in the restaurant rather than through deliveries.
Pax	This column has 45 unique values out of 84,109 total rows, suggesting that this restaurant can foster different types of meal settings. The mean value for this feature is

	3.30 and a standard deviation of 3.69, which suggests that most transactions serve a relatively small number of people, but there is some variability. The minimum value is 0, and the maximum is 200. The 25%, 50%, and 75% of the data fall between 1 and 4, indicating that the distribution is right skewed. Some records have Pax equal to 0, indicating an employee did not record the number of customers for the order. Customers tend to visit restaurants in pairs or in groups of four, which may suggest that these are the most common group sizes for in-person dining.
CustomerID	The dataset includes transactions from customers with IDs ranging from 0 to 69,101. Customer ID 0 may represent customers who choose not to create a client account with the restaurant. There are 84,109 total observations and 2,316 unique values, and there are no missing values. The mean for this feature is 12,633.06, with a standard deviation of 21,952.07. The minimum CustomerID is 0, indicating some transactions were not associated with a CustomerID, and the maximum is 69,101. There are transactions where 'CustomerID' has no value, which means that it is a Dine-inn situation, not a delivery. The data suggests that the most customers (41%) visit the restaurant chain between 5-10 times.
CustomerCity	This column holds 31,248 values, with 17 unique values represented. The city "EGKOMI" appears the most with a count of 9,423. There is missing data indicated by NaN. All duplicates encountered (52,861 occurrences) are related to Dine-inns.
CustomerSince	This column holds 29,142 values, with 2,109 unique dates represented. The earliest date recorded is from 2005-12-06 at 15:00:00 and it appears 1,164 times, indicating long-term customers of the restaurant. The latest date is from 2019-09-262019-09-26, 20:18:09.877, but it only appears once, and the hour format seems to be incorrect. The column contains missing data, indicated by NaN. Most of the duplicates in this column (52,861 occurrences) are related to Dine-inns. A filter was applied to exclude CustomerSince values outside of the range of the dataset, which accounted for 793 rows.

3.2. DATA PREPARATION

The goal of this step was to choose the necessary data to move on to the modeling stage. Following earlier insights acquired in the Data Understanding section, we began by removing records that were not necessary. We excluded impossible values from the dataset, notably instances where clients were registered after the date of their existence or when invoices were created before a client's registration date, in order to assure data authenticity and accuracy. As a consequence, after this cleaning procedure, 0.9891 percent of the data was retained.

The "EmployeeID column has been dropped from the dataframe. This choice was made because of the feature's high cardinality and lack of pertinent information. The amended dataframe is now better suited for further analysis and modeling. Then, we defined the following features to proceed to the modeling phase: "Qty", "TotalAmount", "IsDelivery", "Pax", and "DocNumber", as well as "ProductDesignation", "ProductFamily", "CustomerID" and "CustomerCity".

During the data processing phase, we found that the "CustomerCity" column in our DataFrame contains incorrect city names. We made an effort to combine these misspelled names with their proper counterparts to tackle this issue. In order to do this, we conducted an online research on the official

websites of the in question. Using the official Lakatameia website, we changed "Lakatame" and "Lakstameia" to the accurate spelling of "Lakatameia." In a similar manner, we changed "Strobolo" to "Strovolos". Also, we observed that the same city, "Egkomi," had two distinct lines, each with a different sort of data. We proceeded to repair the "Egkomi" spelling error and delete all the spaces from the "CustomerCity" column. These modifications improved the "CustomerCity" column's precision and consistency in our DataFrame, allowing us to move on with more trustworthy data analysis and modeling. The "TotalAmount" column was changed from an object data type to a float data type, and the dots in place of the commas were used for decimal units. Moreover, the fields "InvoiceDateHour" and "CustomerSince" were changed to datetime data types. Throughout the following stages of the data analysis process, this conversion enables improved data manipulation and analysis. We also displayed the distribution ***(Figure 1)*** of customers by city and a map ***(Figure 2)*** to visualize the restaurant locations***.***

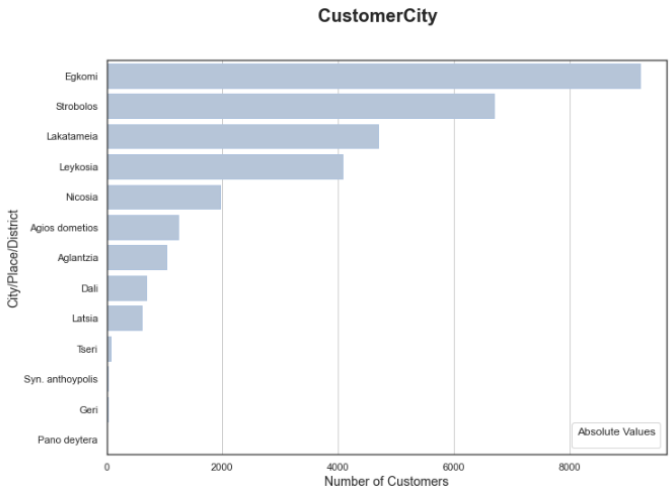


Figure 1 **-** " Bar plot to show the number of customers in each city "

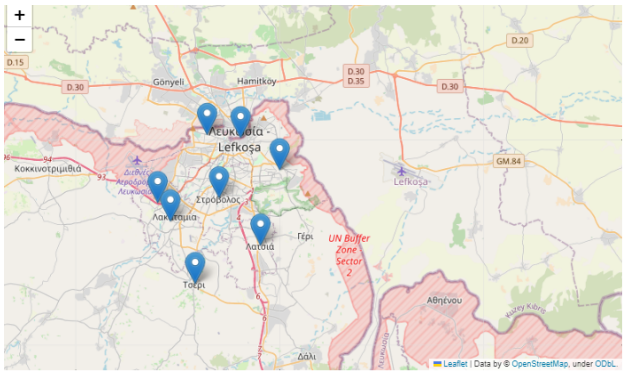
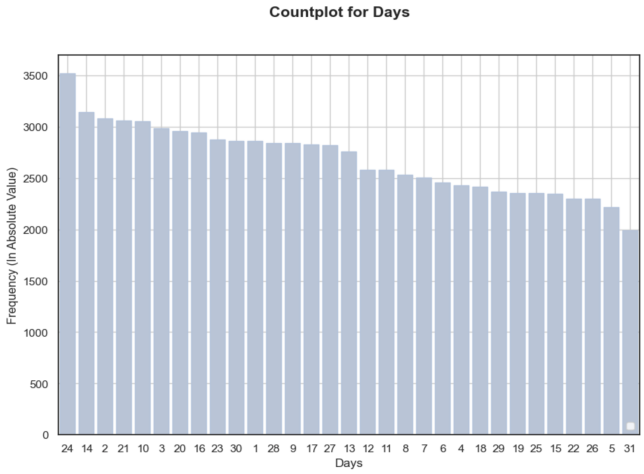


Figure 2 **-** "Map to visualize the restaurant locations"

The Boxplots (***Figures 3*** in the appendix) were created to provide a more in-depth visual comprehension of the metric features and an even better knowledge of outliers. The boxplot for "Qty" has a short box and long whiskers, which may imply that the data is skewed. The outliers indicate there are some transactions whose quantity is above the central tendency. The right-skewed nature of "TotalAmount" and "Pax" is the same. There are several transactions whose amount is higher, and "Pax" per transaction is higher than the central tendency. Also, we performed a manual outlier removal

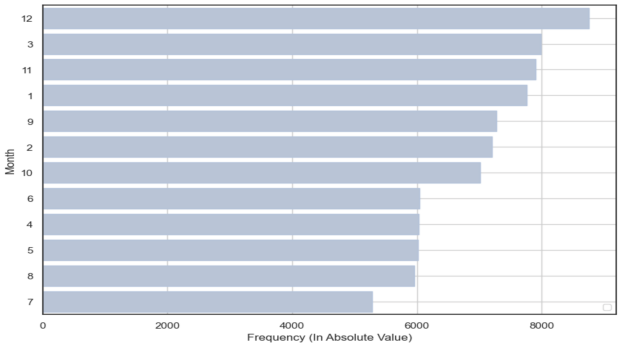
as a result of our visual analysis. Based on the criteria, 99.96% of the data was retained: "Pax = 70," "Qty = 20," and "TotalAmount = 1000." We also compared the outcomes with the Interquartile Range (IQR) removal approach to try a different strategy. The percentage of data left after the IQR method was 85.195%. We chose the manual procedure since it offered more precise data removal. In order to identify redundancies between variables, we created a heatmap based on the Spearman correlation (***Figure 4*** appendix). There is a significant negative correlation of -0.9 between "Pax" and "IsDelivery", so we could have dropped one of them, but we consider both variables relevant for the analysis, so we are going to keep both.

The decision was to introduce four new variables in order to understand the business from a different perspective and to examine client buying habits based on particular days of the week and hours of the day. These variables consist of: weekday Invoice, weekend Invoice (a binary variable) and the hour of the day when the invoice was issued is represented by the value hour of day.



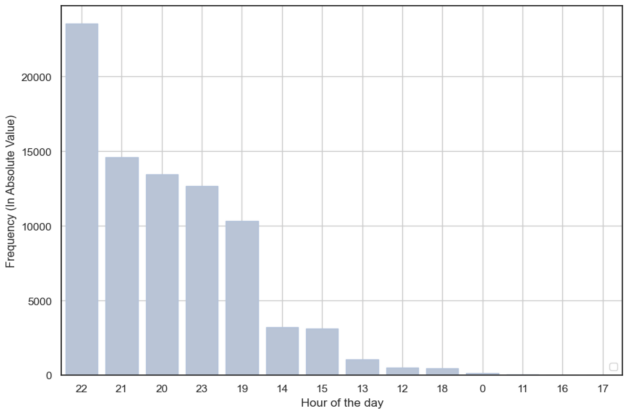
The distribution frequency of the invoices over the course of the months is depicted (***Figure 5)***. With more than 3500 cumulative invoices, the 24th day is by far our best day. More so, the restaurant had more than 3000 cumulative invoices per day on the 14th, 2nd, and 21st. The 26th, 5th, and 31st had the fewest total bills. The fact that the 31st day doesn't fall on every month of the year explains why it's by far the worst day.

Figure 5 **-** " Invoices frequency per day of the month"



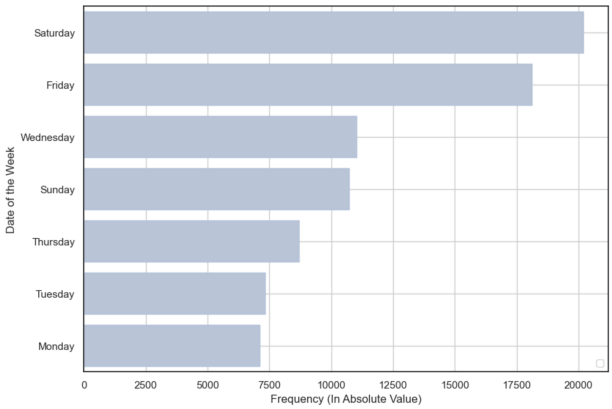
The graph shows us the invoices' distribution frequency throughout the months of the year (***Figure* 6)**. The December month is the best by far, with more than 8000 cumulative invoices. March, November, and January are the following best months, where the restaurant almost achieved 3000 cumulative invoices per month. July and August have the fewest invoices.

Figure 6 **-** " Invoices frequency per month"



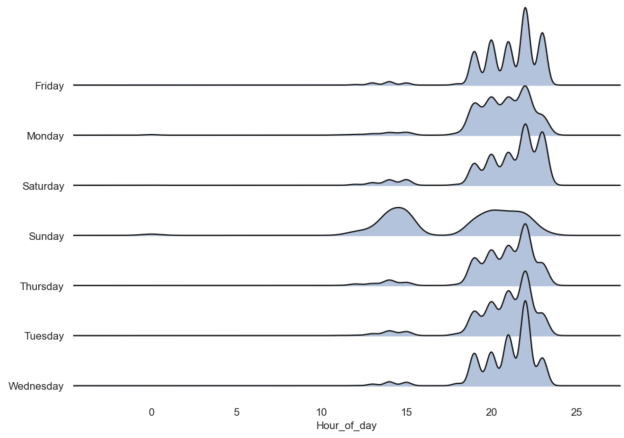
The distribution frequency of the invoices (***Figure 7)*** during the course of the day is depicted in the graph below. The majority of the bills were generated after 19:00, which indicates that our business has had more activity during those hours. Between 12 and 15 o'clock around lunchtime, there is a negligible difference in the number of invoices sent out compared to those sent out during dinnertime.

Figure 7 - " Invoices frequency per hour of the day"



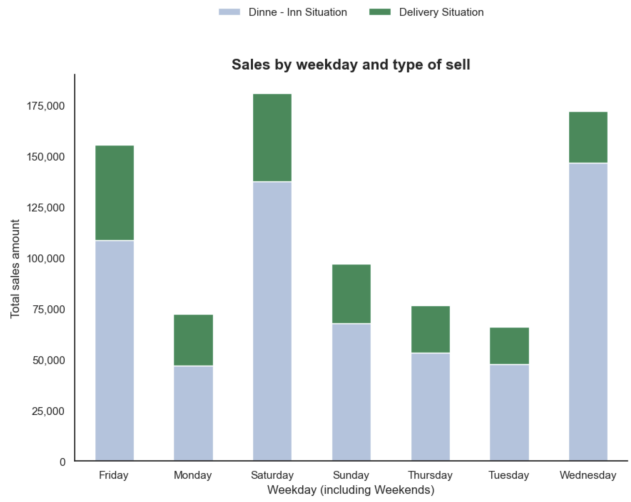
The graph shows us the invoices' distribution frequency throughout the days of the week (***Figure* 8)**. Saturdays and Fridays are the days where the most invoices are emitted, comprising almost 40.000 invoices. Monday is the worst day, with less than 7,500 invoices throughout the year.

Figure 8 - " Invoices frequency per week day"



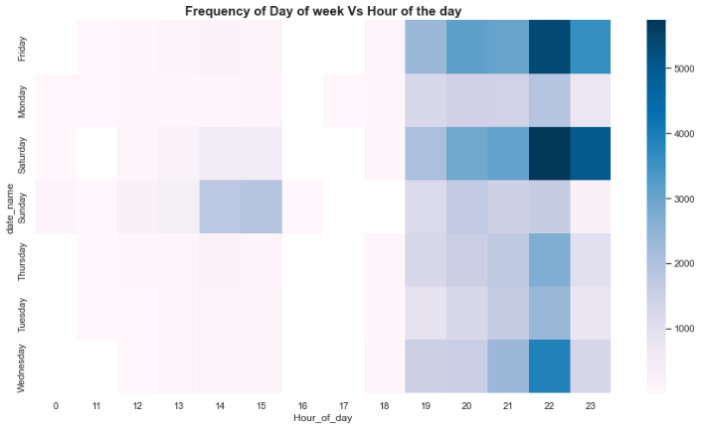
The graph ***(Figure 9)*** shows us the invoices' distribution frequency throughout the day for each weekday. The pile of invoices is pretty similar between the six days of the week, with the exception of Sunday. The majority of the invoices occur during the dinner period.

Figure 9 - " Invoices frequency per weekday and hour of the day."



(Figure 10) shows the total sales amount per weekday, revealing a high discrepancy. The best days in terms of sales are Saturday and Wednesday, where the total amount of sales is above 170000 euros each day. The lowest are Monday and Tuesday where the total amount is not more than 85.000 euros each day. Additionally, the graph provides the distribution of the total amount per day between Dine-inn and Deliveries. Deliveries represent a niche activity for the business.

Figure 10 **-** "Sales by weekday and type of sell"



Also, we used a heatmap, ***Figure 11,*** to understand the patterns of customer visits to the restaurants, and it revealed that the busiest days of the week are Fridays and Saturdays, with Sundays experiencing busy lunch hours.

Figure 11 **-** " Heatmap to understand at which time of the week and at which time of the day, people come to the restaurant"

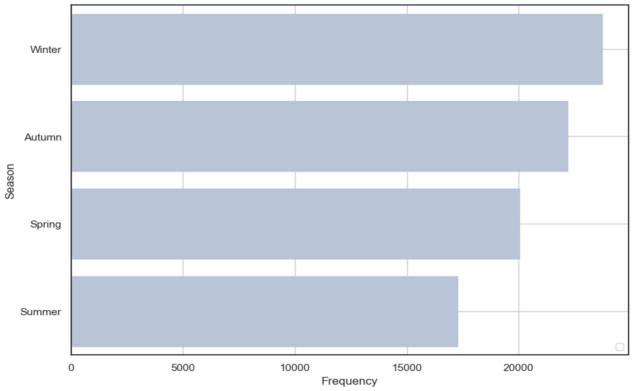


Figure 12 shows the total amount of sales per season. The amount of sales is higher in winter and autumn than in spring and summer.

Figure 12 **-** " Total amount of sales per Season"

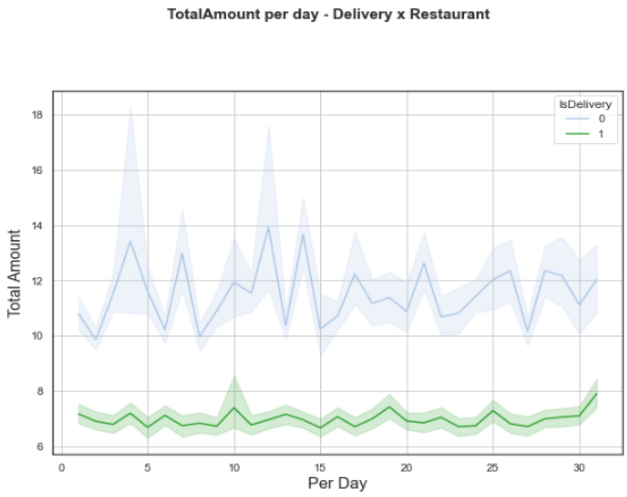
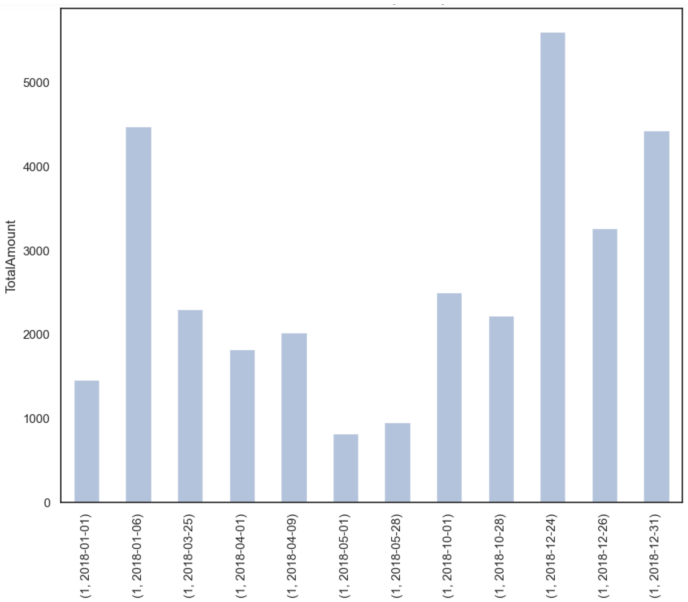


Figure 13 reveals a plot line graph visualizing the total amount of sales per day, enabling us to compare deliveries and dine-inns purchases.

Figure 13 - "TotalAmount per day - Delivery x Restaurant"



The plot (***Figure 14***) shows us that the three biggest holidays in terms of revenue are Christmas Eve, New Year's Eve and Epiphany. The three holidays occur in a calendar window of 13 days.

Figure 14 - " Total revenue on Holidays"

4. MODELING

4.1. ASSOCIATION RULES

To analyze the data, the team used three methods. First, without using any filters, they looked at the key findings by product and by product family. Due to the different behavior patterns of dine-inns and deliveries, they also built two distinct datasets for each. Two products were left out of the deliveries dataset so that we could concentrate on the meal-related items, which were our main concern. Finally, the team saw that Mineral water 1.5lt was a recurring item in the association rules and made the

decision to generate a fresh dataset drawn from dine-inns, but without this product, to better understand the patterns of consumer behavior.

To prepare the data for the apriori algorithm, the code cell uses the pandas "pd.pivot table" function to generate pivot tables. Five pivot tables have been built, each with a different set of goods that are included or excluded, as well as groups of products or product families. The following pivot tables were created to analyze the association rules between products:

First, by product with no restriction; second, by product in the delivery situation, with the exclusion of the "Tsanta" and "Delivery charge" products; third, by product in the dine-inn situation; then, by product in the dine-inn situation, excluding the "Mineral water 1.5lt" product due to high imbalance; and finally, by product family with no restriction.

4.2. MODELING: OVERALL DATASET

The process of mining association rules entails identifying important connections between dataset's variables. The list of items in both the antecedent and consequent are referred to in this manner as a "itemset." The "support" statistic is employed to ascertain how frequently the itemset appears in all transactions. The antecedent and consequent supports each give a separate indication of how frequently they occur. Although "lift" gauges how strongly the two are related, the "confidence" metric estimates the possibility of the consequent happening given the antecedent.

The frequent itemsets are produced by the Apriori algorithm with a minimum support of 5%, and they are then submitted to the association rules function of the mlxtend.frequent patterns module. The outcomes are arranged in descending order according to the selected metric, which is typically support, confidence, or lift.

It is usual to set support, confidence, and lift levels in order to detect robust and statistically significant correlations between variables. We set a lift threshold of 1.5 to include rules where the ratio of observed joint probabilities to their expected joint probabilities under independence is at least 1.5, a support threshold of 10% to include rules where the items are present in at least 10% of transactions, a confidence threshold of 50% to include rules where the probability of the consequent given the antecedent is at least 50%, and a lift threshold of 50% to include rules where the probability of the consequent given the antecedent is at least 50%. These thresholds assist in getting rid of weaker, chance-based restrictions and put more of an emphasis on rules that are more important and could have real-world applications.

To illustrate the link between the support, confidence, and lift metrics in the rules discovered using association rule mining, we built a scatter plot, ***Figure 15***. Each rule is represented as a point in the plot's two-dimensional space, where the x-axis stands for support and the y-axis for confidence. Each point's size varies in direct proportion to the lift metric. The graphic demonstrates that, although the correlation is not particularly strong, there is often a positive correlation between support and confidence, and that rules with higher support tend to have higher confidence.

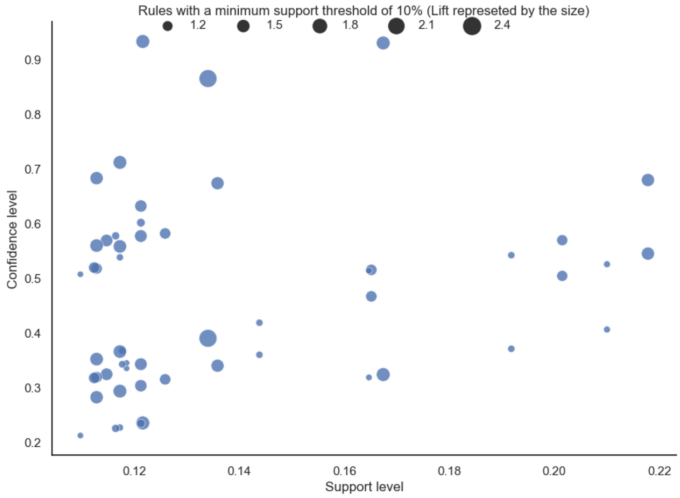


Figure 15 *- "Rules with a minimum support threshold of 10%"*

The relationship between the support, confidence, and lift metrics in the rules discovered using association rule mining was depicted using a bubble plot. The antecedent and consequent are shown on the y- and x-axes, respectively, and the size of the bubbles denotes the lift. The size of the bubbles in the plot's matrix of the rules indicates how strongly the antecedent and consequent are associated.

After utilizing association rule mining to create frequent itemsets from transaction data, we then calculated the number of products per set. The itemsets with exactly two products and a support of at least 0.17, or at least 17% of all transactions, are then filtered. The generated itemsets and their corresponding support values are: (Mineral water 1.5lt, Egg fried rice), (Egg fried rice, Spring roll), (Sweet sour chicken, Egg fried rice), (Mineral water 1.5lt, Spring roll). We also selected rules that have a high lift (greater than or equal to 3) and a high confidence (greater than or equal to 0.7). Finally, a network graph was created to visualize the top 20 rules with high confidence (***Figure 16***).

The most common items in the dataset, according to the association rule mining analysis, are Egg Fried Rice, 1.5LT Mineral Water, Delivery Fee, and Meat Noodles. Each of these frequent itemsets has its most frequent antecedents close to its center; and closer the point, the greater the degree of confidence. These findings support the strong relationship between the Delivery Charge and the Tsanta product. Thus, it is advised against taking either of these factors into account while making business decisions.

4.3. MODELING: DINE-INN

The thresholds for support, confidence, and lift were defined as 10%, 50%, and 0, respectively. Once again, the top 20 confidence rules for the "Dine-inn" are displayed in the form of a network graph. Based on our analysis, Mineral Water 1.5LT and Noodles with Meat have been shown to occur frequently. We displayed the network graph (***Figure 17***) and found that the antecedents with the highest confidence levels were those that were closest to the center of each item, helping us identify the antecedents that were most frequently connected with these items. The closest antecedents to the center of each item are therefore those that are most frequent and likely to occur.

4.4. MODELING: DELIVERIES

In this section, we used the Deliveries Dataset (df_pt_delivery) to explore some metrics as support, confidence and lift. We started by setting the thresholds for support at 10%, confidence at 50% and lift at 0 in order to analyze possible substitutes products. From the generated rules by the Apriori Algorithm, we could check that the most frequent consequents are Sweet Sour Chicken and Egg Fried Rice, due to the fact that these components appear to be the most frequent pairings with a minimum support of 0.05.

In order to see the top 15 confidence rules visually, we drew the network graph (***Figure 18***). Three smaller groupings were visible that were not connected to the larger ones, indicating that these products are not often organized with the other clusters of products.

4.5. MODELING: DINE-INN EXCLUDING WATER

Another analysis we performed in order to explore metrics such as support, confidence, and lift was with the dataset for dine-inn orders excluding water (df_pt_dinne_in_no_water). In this case, we kept setting the same thresholds for support at 10%, confidence at 50%, and lift at 0 to analyze possible substitute products. The most frequent consequents itemsets generated by the Market Basket Algorithm were Egg Fried Rice, Spring Roll, and No Meat. These items appear as the most common combinations in the dataset.

We also plotted (***Figure 19***) a network graph to visualize the top 15 confidence rules. Considering that the closer a node is to the center, the higher its confidence level, Toffee Banana Complementary and Beef BBS were found to be antecedents to both Egg Fried Rice and Spring Roll, highlighting their importance in the dataset.

4.6. MODELING: BY FAMILY

The dataset grouped by family (df_pt_family) we set the thresholds for support at 20% (due to higher granularity, support is also higher, so we increased the support threshold), confidence at 50%, and lift at 0 to analyze possible substitute products and from the generated rules, we had rice as most frequent consequent, with a minimum support of 0.05. A network graph was plotted to visualize the top 15 confidence rules so it was possible to verify that the two closest antecedents are {Holds, Sizzling, and Drinks} and {White Wine, Sizzling, Meat, and Starters}.

4.7. MENU SUGGESTIONS

4.7.1. Menu for Dine-inns

In order to create menu suggestions, we analyzed the frequent product pairs in the dataset. We also investigated the association rules to identify items with high support, confidence, and lift, indicating a strong relationship between the items. In this sense, we measured the quantity of products per set, and filtered the frequent products for those that have two products and a support of at least 0.17, which represents 17% of the total transactions, to identify popular product combinations that frequently occur together and have a high likelihood of being purchased simultaneously.

Next, we analyzed products with high confidence and high support, because that is a sign that many customers purchase together and are highly likely to purchase one item when they purchase another, in order to create new menu suggestions. The analysis did not reveal any substitute products, suggesting that there is no need for product deletions from the menu, as there were no negative lift values. However, we did identify some opportunities for creating new products and menu suggestions:

- **Creation of a new product**: Noodles without meat. There is 100% certainty that customers who purchase "No Meat" also order "Noodles with meat." Furthermore, a lift of roughly 5.4 indicates a significant correlation between these items.
- **Creation of new menu 1:** Spring roll with Egg fried rice (support > 0.2 and confidence > 0.6) and Mineral water 1.5lt (support > 0.33 and confidence > 0.94 with each of the previously referred products).
- **Creation of new menu 2**: Sweet sour chicken with Egg fried rice (support > 0.2 and confidence > 0.6) and Mineral water 1.5lt (support > 0.28 and confidence > 0.93 with each of the previously referred products).

We also have the following recommendations for C's Restaurants:

- **Saturday Lunch Campaign**: We have noticed above that Sunday lunch, represents an outlier both in terms of orders volume but also in proportional to the dinner sales volume. However, we do not verify the same on Saturday. In this sense, we propose a possible campaign: if you order on Saturday Lunch before 4pm, we offer you a discount on your next dinner visit.
- **Best Holidays Promotion**: The three most profitable holidays are separated by 13 days: 24.12 (Christmas Eve), 31.12 (New Year's Eve) and 06.01 (Epiphany). We suggest the adoption of aggressive marketing strategies for those specific days: first, the business hours can be slightly extended since we know that there must be high demand; second, we can create special products (for examples, a discount on the most expensive starters and desserts, since the in a special occasion, the customers may be willing to spend more); and third, think about some partnership nearby the restaurant: festivals, cinema and massage center, for example.
- **Worst Holidays Promotion**: For the worst holidays, we suggest a campaign to attract customers. For example: if you come on one of these holidays, you receive a 50% discount on the next visit in May or June. The holidays are the following: 1.05 (Labor Day/May Day) and 28.05 (Orthodox Pentecost Monday).

4.7.2. Menus for Deliveries

By analyzing the relationships between various products in the dataset, we focused this study on menu recommendations for delivery. We calculated the number of products per set, then filtered for itemsets containing two products and a support of at least 0.1 (i.e., itemsets that represent at least 10% of the total transactions). Next, we filtered for rules with high support (≥0.2) and high confidence (≥0.5), which are considered strong relationships between products. Then we checked the rules with high confidence (≥0.5) and high lift (≥1.5).

Considering that Lift measures how much more likely the consequent is to be bought when the antecedent is bought compared to when the antecedent is not bought, high lift values indicate strong associations between products, therefore, we checked for products that customers tend not to buy

together (substitute products) in deliveries, and we did not find rules with a lift value less than or equal to 1, which means that there are no substitute products in the deliveries dataset.

In attention to our analysis, regarding suggestions for Deliveries, we had the following suggestions to propose:

- **Substitute Products:** there are no substitute products since there is no negative lift, which means that we do not suggest any product deletion from the menu.
- **Creation of new products:** We suggest the creation of a new product of Noodles without meat. If the confidence is 100 percent, it indicates that customers who purchase "No Meat" also order "Noodles with meat," which is logical. The likelihood of ordering noodles with meat is also over eight times higher than it would be if they were independent, according to a lift of nearly eight.
- **Creation of new menu 1:** Sweet sour chicken with Egg fried rice (support>0.2 and confidence>0.6).
- **Increase Prices 1:** On the items Jira pulao and Naan, there is a lift above 4.5 and a confidence above 0.54. Jira pulao is an Indian side dish, whereas naan is an Indian appetizer. Hence, unlike what we proposed in the dine-inn restaurant, there isn't a main dish here so that we may connect the items. Anyhow, it could be possible to advertise a discount on the beginning Naan while raising the cost of the side dish.
- **Increase Prices 2:** There is lift above 8.2 and a confidence above 0.70 on the items extra pancakes - extra sauce, naturally. Here, we could explore a promotion on the price of additional pancakes, along with an increase in the extra source price, for example.

Following, a few additional recommendations are also considered:

- Similarly as we suggested previously for the “Dine-Inn” segment, we also suggest the adoption of the same strategies: “*Saturday Lunch Campaign*”, the “*promotion for the three best holidays*” and the “*promotion for the two worst holidays*”.

5. DEPLOYMENT AND MAINTENANCE PLANS

After modeling and producing the market basket analysis, the next step consists of delivering the results and suggestions through a quarterly report to the main decision-makers in the marketing and financial areas to promote promotions and discounts. The report must explain the model in an appropriate business communication so that the other areas can analyze it and have a good understanding of the process. In addition, the results also can be accessed by Google Sheets immediately. As a suggestion for medium-term implementation, the C informatic system must be updated in order to propose up-selling and cross-selling suggestions to the waiters based on their previous behaviors. After the waiters introduce the initial request, the waiters can come back to the clients and provide additional suggestions to their meal, stimulating upselling and, at some specific moments, such as dessert, encouraging cross-selling.

The model should be executed annually to assess the changes; this way, the model will be recurrently updated year by year and will take into account the seasons, allowing the business to adjust their menus, introduce new products, and promote cross selling through adjusting variables such as prices.

6. CONCLUSIONS

After conducting a comprehensive analysis of the dataset from Asian Food by C, located in Cyprus, it was possible to extract a significant amount of valuable information. Market basket analysis was used to optimize operations and drive profitable improvements for the business. Gaining an understanding of customer behavior was pivotal to developing more effective marketing strategies. This, in turn, led to business proposals based on patterns and ultimately resulted in increased revenue based on the results that are in line with C's goals.

7. REFERENCES

OpenAI. (2021). GPT-3: Language Models are Few-Shot Learners. https://arxiv.org/abs/2005.14165

Li, S. (2017, September 27). A gentle introduction on Market Basket Analysis - Association rules. Medium. Retrieved March 20, 2023, from https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce

Tutorial — NetworkX 2.5 documentation. (n.d.). Networkx.org. https://networkx.org/documentation/stable/tutorial.html

Sharma, R. (2022, July 19). Market-Basket-Analysis. GitHub. https://github.com/sharmaroshan/Market-Basket-Analysis

punits152. (2023). notebook_projects/Market_basket_analysis.ipynb at main · punits152/notebook_projects. GitHub. https://github.com/punits152/notebook_projects/blob/main/Market%20Basket%20Analysis/Market_basket_analysis.ipynb

geopy. (n.d.). PyPI. Retrieved March 29, 2023, from https://pypi.org/project/geopy/

Python Word Clouds Tutorial: How to Create a Word Cloud. (n.d.). Www.datacamp.com. https://www.datacamp.com/tutorial/wordcloud-python

8. APPENDIX

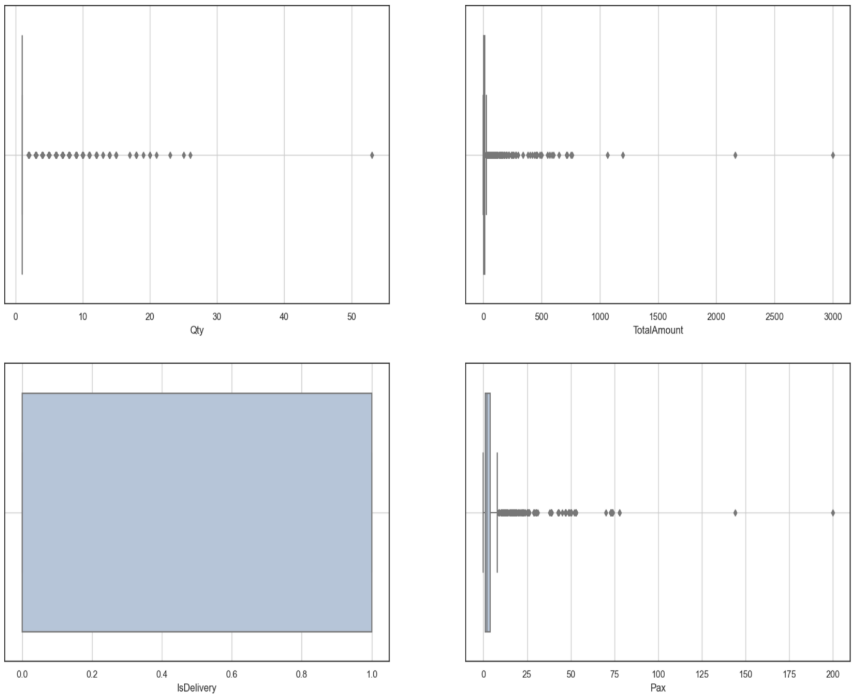


Figure 4 *- "Boxplots for the metric features"*

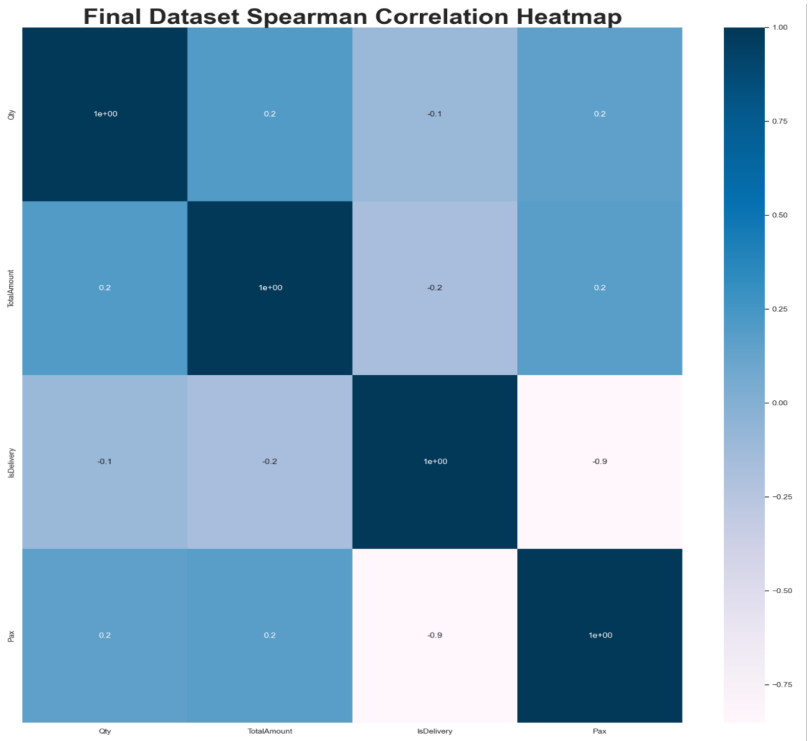


Figure 5 **-** "Dataset correlation Spearman Heatmap"

Figure 16 **-** *"ProductDesignation network graph of the top 20 confidence rules"*

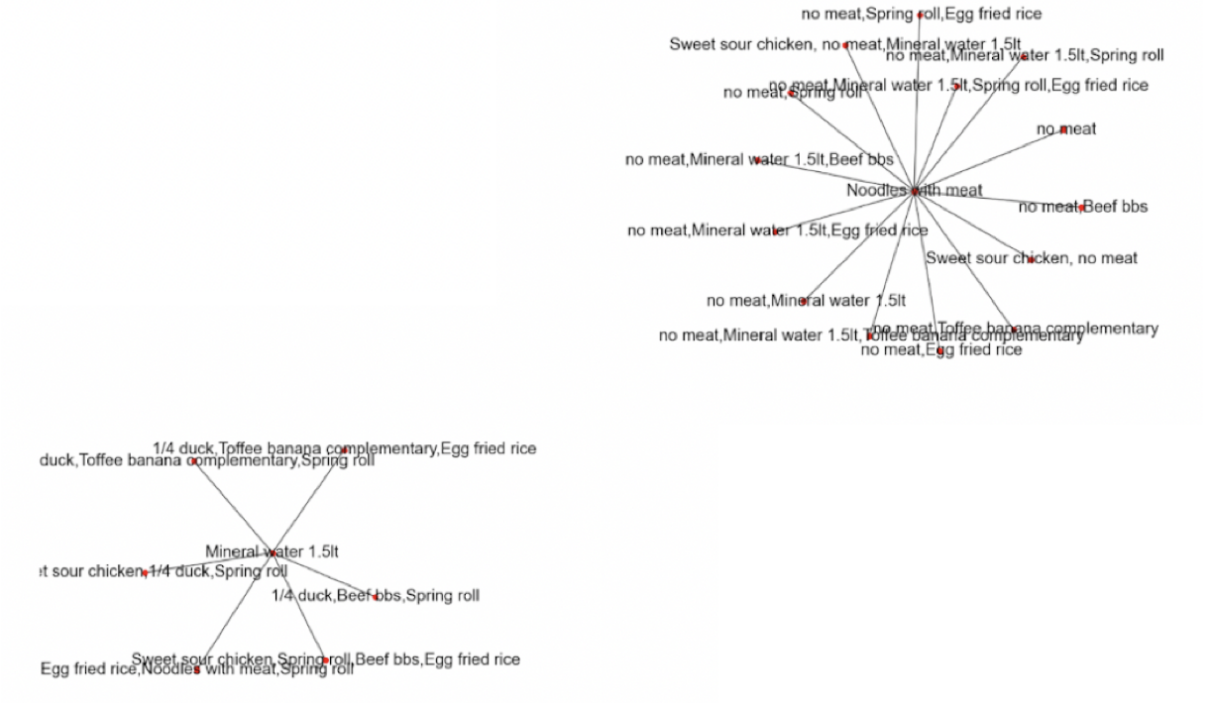


Figure 17 **-** *"Dine-inn network graph graph of the top 20 confidence rules"*

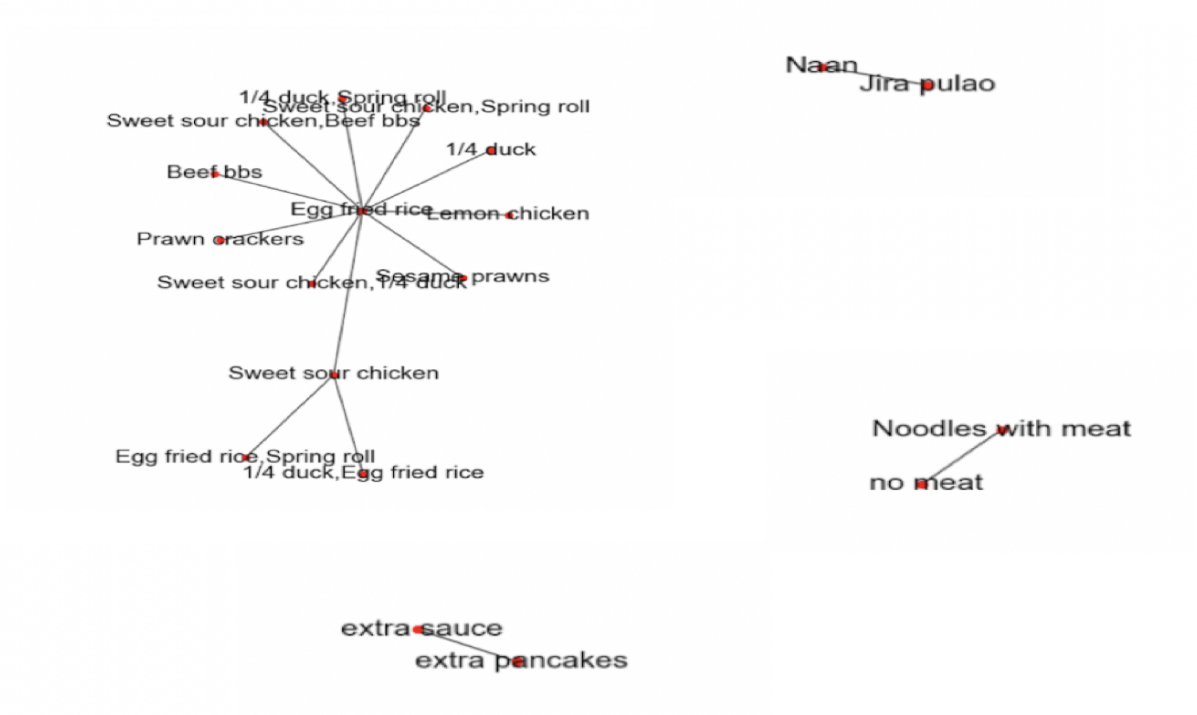


Figure 18 **-** *"Deliveries network graph of the top 15 confidence rules"*

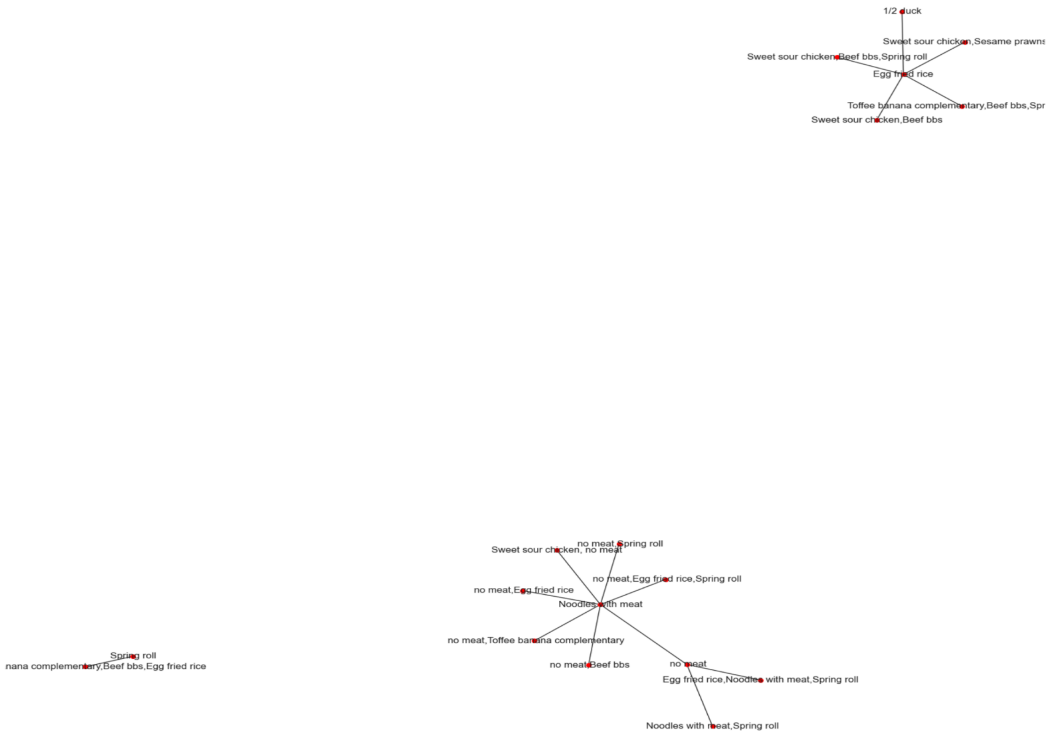


Figure 19 **-** *" Dine-Inns Excluding Water Network graph of the top 15 confidence rules"*