NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: Market Basket Analysis

Ana Miguel Monteiro, number: 20221645

Ana Rita Viseu, number: 20220703

Miguel Cruz, number: 20221391

Rodrigo Brigham, number: 20221607

Sara Galguinho, number: 20220682

Group G

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2023

Index

1. EXECUTIVE SUMMARY

This project aims to develop a Market Basket Analysis tailored for company C, an Asian restaurant in Nicosia, specialising in Chinese gastronomy. Our objective is to uncover meaningful patters and trends in customer consumption behaviour across different situations, gaining valuable insights into their preferences and habits. By doing so, we can suggest new menu sets, introduce new products, and identify substitute items. Additionally, we can recommend cross-selling strategies, segment customers, and offer other targeted business applications and marketing based on our findings.

The restaurant industry has become increasingly ferocious in recent years. With the rise of food delivery services, online reviews and social media platforms, restaurants have experienced intensifying competition and strong changes in customers' habits. Those that can meet these challenges have the opportunity to thrive in today's market, others face tough problems that, in the worst-case scenario, might lead to closure. Company C's restaurant has felt the impact of those adversities, having had obstacles in sustaining its profit margin and growth. It is essential for the restaurant to find innovative solutions to those challenges and remain competitive in the market.

During this project, we followed the process model CRISP-DM, starting by gaining a broader understanding of the business. To do this, we explored Company C's restaurant background, clarified business objectives and defined success criteria for the problem, which will allow us to measure the performance of our findings and suggestions. We also assessed the kind of resources that were available, as well as risks and costs/benefits that this project entailed. As a final step, we determined what were our data mining goals - objectives in technical terms - for this project. Afterwards, we proceeded with data preparation and exploratory analysis. Finally, by applying the Apriori algorithm for identifying frequent item sets in different situations, we developed a Market Basket Analysis that effectively identified meaningful patterns and correlations among customers' purchasing behaviors.

Using the valuable insights obtained through our analysis, we were able to answer the key questions of company C, as well as recommend the creation of two new products (water and noodles), two menus (general and indian) and find distinct combinations and relationships of substitution between products. Important patterns within the customer base were also uncovered. With the two new products, we estimate a potential increase in total revenue per year of around 1%.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BACKGROUND

Company C, created more than two decades ago, owns several restaurants in Cyprus, which differ by concept, location and type of cuisine. One of their oldest brands, specialized in Asian food, particularly Chinese food, is facing difficulties in sustaining its profit margin and growth because of intensifying competition and shifting consumer habits.

By capitalizing on its sales data, the company hopes to understand the main differences between dine-in and delivery customers, determine if their product offering is appropriate and identify important tendencies that may arise from the patterns in consumption. These insights could prove useful in various areas such as creating set menus, introducing new products, understanding substitute products, recommending or promoting cross-selling, customer segmentation and other possible outcomes, depending on the findings.

Thus, by working closely with company C's key stakeholders, our goal is to conduct a thorough data exploratory analysis to find relevant patterns and perform a market basket analysis that takes into account various factors, such as customer ordering patterns, weather conditions, seasons, important events and holidays. This analysis and the resulting insights will be particularly beneficial to the marketing team (if there is one) and to the managers/executives of the restaurant and the chain it belongs to. Additionally, the restaurant will gain a better understanding of its customers, which will ultimately benefit the entire company.

2.2. BUSINESS OBJECTIVES

Our primary goal for this project is to gain a comprehensive understanding of customer ordering patterns and preferences. Through analysing their usual requests, spending habits, and routines, we aim to assist the restaurant and seek to address the following key business questions:

- Identify the main differences between dine-in and delivery customers;
- Determine if the product offering is appropriate;
- Discover interesting patterns that may indicate tendencies within the client base;
- Understand how to improve the profit margin and growth of the restaurant and present recommendations aligned with company C’s objectives.

Some of the expected benefits of this project include:

- Achieving a higher level of customer satisfaction, as measured by surveys or online reviews;
- Obtaining a better understanding of customers' consumption patterns and preferences, leading to more targeted and effective product placement;
- Successfully segmenting customers based on their preferences and behavior, enabling more personalized services and improved customer engagement;
- Developing a product offering that resonates with customers and drives sales growth;
- Introducing new products that meet customers' needs and preferences, resulting in increased sales and customer loyalty;
- Increasing the frequency of cross-selling opportunities, resulting in higher sales revenue.

2.3. BUSINESS SUCCESS CRITERIA

In order to evaluate the success of the project, it is necessary to define concrete criteria that enable us to evaluate its progress and outcomes. Therefore, we established the following business success criteria to be assessed by the relevant stakeholders:

- Increase the profit margin by one percentual point within a year;
- Determine specific distinctions between dine-in and delivery customers;
- Analyze if the product offering is adequate and suggest at least:
    a) 1 new set menu
    b) 1 new product
    c) Two important combinations of products
    d) Two substitute products;
- Find the main patterns and tendencies in the customer base.

2.4. SITUATION ASSESSMENT

To develop this project, we used the file "Case2_AsianRestaurant_Cyprus_2018.txt", a dataset with 12 features/columns and 84,109 transactions/rows provided by one of the most popular restaurants of the chain in Nicosia (includes only the transactions made in 2018). In the dataset, there are 11,147 documents, 255 different product designations and a maximum of 46 rows per document (meaning the maximum number of products registered for the same sale was 46).

To prepare the data for analysis, we resorted to software tools such as Python and some of its libraries: *Pandas*, *NumPy*, *ydata_profiling*, *datetime*, *math*, *mlxtend*. For data visualization, which is essential to communicate insights and present results from our exploratory analysis, we also used libraries like *Seaborn, Matplotlib* and *NetworkX*.

Besides the given dataset, we added new sources of information as we believed it would be an important resource to get a better understanding of the data and to achieve greater results. Firstly, we added information regarding the holidays and the seasons by creating two new features: *IsHoliday* and *Season*. Secondly, we imported two new datasets:

- The file "Nicosia 2018-01-01 to 2018-12-31.csv" is a dataset related to the atmospheric conditions in Nicosia, including daily information regarding the year of 2018. It consists of 6 features: *datetime*, *temp*, *precip*, *windgust*, *windspeed* and *cloudcover*. We removed the column *windgust* because it contained 86.58% of missing values and then added the remaining features to the original dataset through the column *datetime*, leaving us with 4 additional features.
- The file "TOURISM_MONTHLY_ARRIVALS-2018-EN-170119.xls" is a dataset related to the number of tourist arrivals per month (columns) and per country (rows). We decided to keep two rows: the number of arrivals from all countries and the number of arrivals only from China. However, the present information did not have sufficient differences between months (for China) and it is very difficult to understand which factors would be actually caused or impacted by the arrival of tourists. Another problem is that this data corresponds to Cyprus and not only Nicosia, for which we could not find specific touristic data. Therefore, this information was not used in the project.

Finally, we pointed out other relevant dates that we believe could be useful in our analysis. Overall, the original dataset ended up with 6 additional columns/features. The information about all the features is present in Table 1.

Besides resource assessment, it's important to mention the multiple assumptions that we made about the provided dataset:

- It is normal to have rows with the same values for every variable, because instead of the employees registering two of the same item in a transaction, they can register it separately;
- A *CustomerID* of "0" was used when no customer record was assigned to the sale because there are many records with this value, these customers do not have a city associated, and they all correspond to transactions for a dine-in (as customers tend to have a profile associated when asking for a delivery);
- The transactions where the *ProductDesignation* is "FOOD" or "DRINK" are the case of special menus or groups where each person pays a previously agreed price and there is no record of which products were ordered;
- Products with a value of "0" in *TotalAmount* are offers from the restaurant, removal of certain ingredients in a dish, or extra ingredients that have no additional charge.

During the realization of the project, some challenges may arise due to limitations in resources, such as computing power and time. It is also important to consider potential risks, as well as take measures to minimize their impact:

- Data quality issues: The data might contain missing values, incorrect values or inconsistencies that can affect our analysis. To minimize this problem, we resorted to data cleaning techniques, ensuring that the final data was clean, complete, and accurate.
- Unrepresentative samples: The dataset provided might not be representative of the entire population of the restaurant's customers, since it only represents one year of business, which can limit the generalizability of the results. To help this problem, it is good to make sure the data is diverse and representative of the customer base.
- Problems with the implementation of some algorithm: In order to mitigate this risk, we made sure our data was carefully prepared for the algorithm to work without problems.
- Lack of domain knowledge: Without a good understanding of the restaurant industry and the specific context of this business, we could miss important patterns and make incorrect assumptions. To solve this issue, we conducted a thorough additional research to build some domain knowledge.
- The insights and recommendations generated might not be accepted or adopted by the restaurant: To ensure this does not happen, our findings should be communicated with clear and detailed explanations in a way that is easily understandable and actionable for the restaurant's stakeholders.

Finally, this project may involve several direct and indirect costs, such as acquiring and processing the data, computing resources, hiring or training the restaurant staff, and redefining and implementing strategies. On the other hand, there are also potential benefits that the restaurant can expect, like increasing its revenue and improving its operational efficiency, such as inventory and staff management. Additionally, the insights gained can help the restaurant to better understand its

customers and tailor its offerings to their preferences, which can help to increase customer loyalty and retention.

2.5. DETERMINE DATA MINING GOALS

With the aim of evaluating our market basket analysis, it is important to determine the data mining results we are expecting, meaning what are the technical objectives of this project. Considering the already mentioned business objectives, our general goal, and the available information provided, we have defined the following data mining goals:

- Minimum support of 10%: The support metric is defined for itemsets and is the minimum of two other metrics — antecedent support and consequent support. The antecedent support is the proportion of the transactions that contain the antecedent and the consequent support is the support for the itemset of the consequent (Raschka, 2022). Generally, support defines the frequency of an itemset in a database, so having a high support helps identify frequent itemsets.
- Minimum Confidence of 60%: The confidence of a rule is the probability of observing the consequent in a transaction given that the antecedent is also present (Raschka, 2022). Having a high confidence implies a strong association between the sets of items in a rule. This would help the restaurant understand which products are frequently ordered together and assist in menu design.
- Minimum Lift of 1.5: The lift metric measures how much more often the antecedent and consequent of a rule occur together than it would be expected if they were statistically independent (Raschka, 2022). Therefore, having a high lift helps identify itemsets that are more likely to occur together than alone.
- High Leverage: The leverage computes the difference between the observed frequency of the antecedent and the consequent occurring together and the frequency that would be expected if they were independent (Raschka, 2022). Although we consider the first three metrics above to be more relevant, it is also a goal to have a high leverage.
- High Conviction: A high conviction means the consequent is highly dependent on the antecedent (Raschka, 2022). Again, although we consider the first three metrics above to be more relevant, it is also a goal to have a high conviction.

3. METHODOLOGY

To present a solution for the proposed problem, we followed the CRISP-DM methodology, consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The first phase (developed previously in section 2) involves defining the problem and project objectives. In the second phase, data is collected and analysed to gain a better understanding of the available information. The third phase involves cleaning, transforming, and integrating the data to prepare it for modeling. In the modeling phase, various techniques can be used to build and test models. The fifth phase is evaluation, which involves assessing the quality and effectiveness of the respective models. Finally, in the deployment phase, the chosen model is put into action and monitored for performance. It is important to note that the sequence of these phases is not rigid, and we moved back and forth between steps when necessary.

3.1. DATA UNDERSTANDING

As mentioned before, we uploaded new information to get a better understanding of the customers, which culminated in having 6 extra features in our dataset. After these additions, we began to perform an initial exploration of the available information and to try to identify data patterns, structural problems and draw early conclusions. We started by becoming familiar with the customer data and identifying the existing features.

FEATURES	DESCRIPTION
DocNumber	Number of the document. The document number repeats in as many rows as the rows in the original document (invoice)
ProductDesignation	Product designation
ProductFamily	Name of the family of the product. A product can only be member of one only family
Qty	Quantity
TotalAmount	Sale price of the total quantity
InvoiceDateHour	Date and hour when the document was issued
EmployeeID	ID of the employee who issued the document
IsDelivery	Indication if sale was a delivery or a dine-inn (1: delivery, 0: dine-inn)
Pax	Number of persons at the table
CustomerID	ID of the customer (if a customer record was assigned to the sale)
CustomerCity	City of the customer (usually only employed in delivery)
CustomerSince	Date of creation of the customer
IsHoliday	Indication if the date is a holiday (1) or not (0)
Season	Indication of the season of the year
Temp	Mean temperature
Precip	The amount of precipitation that fell or is predicted to fall in the specified time period
WindSpeed	Wind speed (measured 10m above ground in a location with no nearby obstructions)
CloudCover	The amount of sky that is covered by cloud expressed as a percentage

Table 1 – Features/Columns

The dataset used for this analysis contains 84,109 rows and 18 variables. There are 3,923 duplicated lines. Afterwards, we used descriptive statistics and data visualization tools (like pandas profiling) to generate an overall report of the data and discover inconsistencies and potential insights:

- The variable *DocNumber* is a unique identifier for each transaction, and all rows with the same *DocNumber* have the same value for *InvoiceDateHour*, *IsDelivery*, *Pax*, and *CustomerID*. However, not all have the same value for *CustomerCity* and *CustomerSince*, as some of the duplicate rows have missing values; Important patterns within the customer base were also uncovered.
- All rows with the same *ProductDesignation* have the same value for *ProductFamily*, which means there are no products that have been incorrectly placed in two or more *ProductFamily*;
- All rows where there is a value in *CustomerCity* correspond to a value of "1" in *IsDelivery*, meaning only the customers that ordered a delivery have an associated city;
- All our variables had appropriate data types, except for *TotalAmount* that was converted from an object to a float and *CustomerSince*, which was converted from an object to datetime to facilitate it is use later;
- *CustomerSince* and *CustomerCity* are the only variables with missing values;
- The customers whose *CustomerID* is "0" are the same that have missing values in the variable *CustomerCity*. Aditionally, for these customers, we do not have data regarding the time since they have been purchasing our products, as we do not have information regarding the sale.

We can verify that they all correspond to transactions for dine-ins. Customers who request a delivery tend to have a profile;

- For the rows that contain information regarding *CustomerCity,* we can identify "Egkomi" as the most common city in the dataset;
- The average spending per item is 9.83€;
- Winter is the season with most orders placed;
- The restaurant has seven employees, each identified with a unique value for *EmployeeID.* The employee with ID 7 seems to be very recent as he only placed 14 item orders. The employee with ID 2 has been responsible for the most orders having placed 55,586 item orders;
- We have a total of 255 products – from ProductDesignation- and 27 unique values for *ProductFamily*. The most requested *ProductFamily* are the "STARTERS" and the most purchased product is the "MINERAL WATER 1.5LT";
- The 27 categories of *ProductFamily* seem to not be organized in the most effective and accurate way. For instance, we have multiple categories regarding sushi orders ("SUSHI", "NEW SUSHI" and "JAP SUSHI") and wines ("RED WINES", "ROSE WINES" and "WHITE WINE"). We also have some *ProductFamily* categories that correspond to a single *ProductDesignation* like ''SUSHI'' which is always a ''SUSHI BUFFET'' and ''TSANTA'' which corresponds to a bag (which, as expected, is only registered when the order is a delivery). Furthermore, some *ProductFamily* values, like "EXTRAS", have multiple meanings — 1. extra ingredients in a dish; 2. delivery charge; 3. desserts) and some products are categorized inaccurately (for instance, "CARLSBERG 33CL" is in "SPIRITS" instead of "BEERS"). Overall, the *ProductFamily* variable appears to be one of the main sources of potential inconsistencies/incoherences and will have to be reorganized with particular attention in section 3.2.3;
- There are some very high values of *TotalAmount* in the *ProductFamily* ''SOUPS''. In this category, the items with *ProductDesignation* ''FOOD'' and ''DRINKS'' most likely refer to group meals with previously established prices. The maximum value for the number of people sat at a table is 200 and it corresponds to an order where the *TotalAmount* is 3000 and the *ProductDesignation* is "FOOD";
- There are some instances of products with a *TotalAmount* equal to 0. These are most likely offers from the restaurant, removal of certain ingredients in a dish and extra ingredients that have no additional charge;
- As expected, ''DELIVERY CHARGE'' is only placed for customers that request a delivery;
- We have 52 records where the number of people at the table is 0. However, these orders are not deliveries and could be take-away orders where the customer picked up their order in the restaurant instead of having it delivered;
- For orders that are delivered, the *Pax* variable is always defined with the value "1";
- As expected, all transactions are in 2018. However, there are some instances of customers with a *CustomerSince* date value in 2019. This is most likely an error;
- Also as expected, weather related variables like *Temp, Precip, Windspeed* and *CloudCover* are highly correlated with each other and with the variable *Season.*

3.2. DATA PREPARATION

The following actions are done to ensure that the data is as clean and accurate as possible, which is crucial for the subsequent analysis.

3.2.1. Inconsistencies/Incoherences

Firstly, we noticed that the records where the ProductDesignation is “FOOD” or “DRINK” correspond to groups or special menus, and since we do not know which products were ordered, they would not be useful for the association rules. Therefore, we decided to remove them from the dataset. However, we still considered analyzing them separately, as they could offer interesting insights.

Then, we discovered that some customers had transactions in 2018 but were only considered customers in 2019. This issue could potentially cause some problems when analyzing the data, so we decided to address it by turning the values in CustomerSince, where the InvoiceDateHour is lower than CustomerSince, into NaNs. Additionally, we found some customers' date of creation higher than the date of an invoice, so we also turned those into missing values.

Moreover, we noticed that the CustomerID value “0" is used generically, and we were unable to extract any meaningful insights from it. To address this issue, we decided to turn these cells into missing values. Furthermore, we removed offers from the dataset, assuming that these products were not specifically requested but rather given by the restaurant.

We also uncovered that for all IsDelivery equal to “1”, Pax is registered as “1”, but we do not have the information for how many people the food is ordered. Therefore, we decided to transform those values into NaNs. Similarly, having Pax equal to “0” did not provide us with useful information, so we too turned those into missing values for clarity.

Finally, we observed that there were different records corresponding to the same DocNumber (same transaction) and the same ProductDesignation. To address this issue, we decided to merge these records into a single line by doing the sum of Qty and TotalAmount and keeping the other variables constant.

The issues related to the variables ProductDesignation and ProductFamily were not treated in this section. Instead, we will address these potential problems in section 3.2.3. of the report.

As a final check for inconsistencies/incoherences and possible outliers that should be removed, we generated a pandas profile from which we concluded that any “strange” values were normal and made sense given the context of the restaurant.

3.2.2. Data Imputation

We had missing values for the following features: *EmployeeID*, *Pax*, *CustomerID*, *CustomerCity* and *CustomerSince*. We decided to not impute these missing values, considering that these variables will be used solely for visualizations and not for the association rules and that these correspond to a high percentage of the available data.

3.2.3. Feature Creation

To conduct a better market basket analysis, we created new features based on the combination of the original ones. These new variables are possibly more relevant for the problem at hand and allow a more effective analysis.

Because granularity is very important for frequent itemset mining (Clarke, 2023), we created four different variables in order to obtain various and effective levels of granularity:

- *Product* — This variable is a better organized version of *ProductDesignation*. For example, the product "DUCK" was registered as $^1/_2$, $^1/_4$, $^1/_6$ or "whole" duck, and we kept only the value "DUCK".
- *ProductWithQuantity* — This variable combines the product and its respective quantity and it could be useful to obtain insights for quantities in frequent itemsets;
- *ProductCategoryI* — Consists of only six different categories: STARTERS, MAIN, SIDES, DRINKS, DESSERTS and OTHERS;
- *ProductCategoryII* — A less broad version of *ProductCategoryI* with the following categories: STARTERS, STARTERS IND, MAIN, MAIN IND, MAIN VEG IND, SIDES, SIDES IND, ALCOHOLIC DRINKS, NON-ALCOHOLIC DRINKS, DESSERT, DESSERT IND and OTHERS.

Besides, we also created some variables solely for data visualization purposes.

3.2.4. Company's KPIs and Customer/Product Analysis

In order to measure the current overall performance of the restaurant, we calculated the company's standard KPIs, such as the total revenue for the year (805240.79€), total number of customers for which we have a record (2313), the monthly average revenue (67103.4€) and the average yearly revenue per customer (348.14€). We, then, used *Seaborn*, *Matplotlib* and *Plotly* to analyze our customer base and products.

3.2.4.1. Product Analysis

By analyzing Figure 8.1, Figure 8.3 and Figure 8.4, we can identify the main products sold in the restaurant by volume, namely the top four – "SPRING ROLLS", "MINERAL WATER", "EGG FRIED RICE" and the "BUFFET SUSHI". If we consider the different levels of aggregation of our products – *ProductCategoryI* and *ProductCategoryII* - we also visualize that the main dishes and the starters represent the categories with higher values in terms of sales volume (Figure 8.5 and Figure 8.7). It is worth noting that the desserts represent an extremely low volume of sales when compared to all other categories.

When analyzing our products, considering the revenue they generate, we find some different insights (Figure 8.6). As expected, the main dishes represent the primary source of income – namely the "BUFFET SUSHI", "DUCK" and "SWEET-SOUR CHICKEN" – as they sell for a higher price. The "SPRING ROLLS" – a starter - also represent a substantial portion of the revenue as they are the restaurant's most sold product.

If we take into consideration the proportion of customers that acquire the restaurant's products by delivery (39.1%), there are slight differences in our product sales volume and revenue structure. These customers request considerably more ''OTHERS'' category products and rarely request drinks, neither alcoholic nor non-alcoholic (Figure 8.9). It is worth noting that orders delivered generate considerably less revenue than orders that are done in the restaurant (Figure 8.8). However, starters and main dishes continue to represent the higher proportion of generated revenue and sales volume. The delivery charge was only paid 80% of the times a delivery was ordered – which may indicate the existence of no delivery charge promotion.

In total, we have more 200 different products represented in our data. We should consider reducing the number of products, for example, by analysing which sell less than 100 units per year.

3.2.4.2. Customer Analysis

Company C´s restaurant has currently 2,313 registered clients, (all delivery) who have been steadily acquired from 2005 to 2015 and have had a substantial growth from 2016 until now (Figure 8.10). Around 60.9% of the restaurant's orders have been consumed in the restaurant, although the generalization of food delivery apps usage may have had a significant impact in the acquisition of new customers since 2016, which currently represent the remaining 39.1% of orders. Customers spend on average between 6.8€ and 8.2€ per item ordered, with no visible trend associated to the year in which they started coming to the restaurant. It is however worth noting that customers that registered in 2011 are those who spend more on average (Figure 8.11).

In 2018, there seems to exist a seasonal pattern in terms of revenue generated: it is steadily decreasing from January to July – with a slight increase in March – and then growing considerably in the following months until December (Figure 8.13).

If we take into consideration a weekly analysis of the restaurant's activity, we observe that, as expected, the weekends nights and Sunday lunch have a higher flow of activity, as well as Wednesday nights (Figure 8.12). It is also visible that more revenue is generated during the dinner when compared to lunch hours (Figure 8.16).

If we make an analysis on a more micro day-to-day level, we also observe that holidays do not seem to generate more revenue on average when compared to non-holiday days (Figure 8.15), besides the 6th of January and the 26th of December. However, if we take into consideration non-holiday days in which a relevant event is happening (for instance, New Year's Eve, Christmas eve or the Chinese New Year) there seems to be an above average activity (Figure 8.17). If we take into consideration more specific level national/local events like the Cyprus Film Days International Festival or the Cyprus Women's Cup International Football Tournament, there seems to exist no clear pattern as some days are quite above average and others quite below. By taking a look at our top 20 sales days, we observe that only one corresponded to a specific holiday. However, all the rest are very near in time to other events that stimulate the restaurant's activity.

We also analysed how the temperature may have stimulated the restaurant's activity (Figure 8.14). We observed that a temperature of 10ºC-20ºC seems to coincide with the greatest number of transactions and sales. There seems to be no impact of the temperature on the average spending of our customers but this range seems to stimulate the customers' purchasing habits in volume.

Considering only customers that have requested a delivery, we also did an analysis of the cities that order the most and Egkomi and Strabolos are those that request more orders. They also generate the most sales.

3.2.5. Feature Selection

We have decided perform a feature selection, even though this would not affect the modeling, in order to keep the dataset organized. We removed the feature *EmployeeID* since it does not add value to the business objectives that were set in the beginning. We also removed both *ProductFamily* and

ProductDesignation because the categories were not well defined. After, we also removed the variables previously created for data visualization purposes.

3.3. MODELING

To perform a market basket analysis, we used the frequent itemset mining technique, which allows us to discover relationships between different items in a dataset. These relationships are represented in the form of association rules, which are composed of two elements: antecedents — item or group of items that are typically found in the dataset — and consequents — item or group of items that is implicated by the antecedent (Keshari, 2020).

In order to obtain the association rules, we used the Apriori algorithm. It was the first ever algorithm proposed for frequent itemset mining, which was later improved by R. Agarwal and R. Srikant and came to be known as Apriori. This iterative approach consists of two steps (*Apriori Algorithm In Data Mining: Implementation With Examples, 2023*):

1. Join Step — All possible combinations of itemsets are generated by going through each item.
2. Prune Step — The count of each item in the database is scanned. If the candidate item does not meet the minimum support, then it is regarded as infrequent and thus it is removed.

Other algorithms such as FP-Growth and ECLAT tend to be more efficient and faster (Korstanje, 2021) but considering we were able to use Apriori without any problems we did not use a different algorithm.

For the generation of the association rules, we decided on a minimum threshold of 5% for the support. We chose this value based on intuition, since one of the current methods for determining the value of minimum support is based on the intuitiveness of the user (Hikmawati, Maulidevi, Surendro, 2021, p.2). Then, we started by applying the Apriori for all transactions in the dataset (in a general context). Furthermore, in order to look for various tendencies in distinct groups of customers and different conditions, we also applied the algorithm for several cases:

- Indian Dishes — After analysing the algorithm in a general context, we noticed there was a tendency for Indian products to be ordered together, so we decided to generate association rules only for these types of products.
- Dine-in or Delivery — Since one of the key questions company C wants to answer is what differences there are between delivery and dine-in customers, we also generated rules for both types of clients.
- Seasons — As there could be different consumption habits in different seasons, we explored the association rules generated for each season.
- Weekends — There could also be distinct patterns in consumption on weekends, so we applied the Apriori for these specific days.

As explained in section 3.2.3, we created four different variables to use in the association rules to obtain various levels of granularity: *Product*, *ProductWithQuantity*, *CategoryI* and *CategoryII*. We started by using all four variables in the general context. Then, for the Indian dishes, using the two category variables did not make sense, so we only utilized the variables *Product* and *ProductWithQuantity*, in order to obtain insights on the quantities of each item in a frequent itemset. Finally, for each season of the year and for the weekends, we used the variables *Product* and

CategoryII, since the variable *CategoryI* was too broad, and it wasn't justifiable to use the *ProductWithQuantity* in these cases.

For every situation where we applied the Apriori algorithm, the main metrics we used to analyse the sets of rules (besides setting a minimum threshold for the support metric) were the lift and the confidence. We always started by setting a minimum threshold of 70% for the confidence and of 1.5 for the lift, however, upon generating the association rules in each case and for each level of granularity, we sometimes set a lower limit for both, depending on the results.

3.4. EVALUATION

In order to assess the results of the Apriori algorithm, besides using metrics like confidence, lift, leverage, and conviction, we also analyzed the possible insights the generated association rules provided us with. We were able to achieve many of the data mining goals set at the beginning of the project, including association rules with a high confidence (although sometimes lower than 60%), a minimum lift of 1.5 and an overall high leverage and conviction. We did not have a minimum support of 10%, however, as mentioned above. We chose a support of 5%, as we found it a more fitting value for the dataset in question and by testing the insights generated.

For the general case, we gained valuable insights and obtained some relevant association rules, such as:

- {NO MEAT} ⇒ {NOODLES WITH MEAT} with a confidence of 1 (every time no meat was asked for, noodles with meat were also present in the transaction), a support of 0.098 (these items appear together 9.8% of the time), a lift of 6.27 (the probability of observing the items together in a transaction is six times higher than if the they were independent of each other), a leverage of 0.082 (the probability of observing both items in a transaction is 0.082 higher than if they were independent of each other) and a conviction of "inf." (as expected, since we have a confidence of 1);
- {SPRING ROLL, BEEF BBS, EGG FRIED RICE} ⇒ {DUCK} with a confidence of 0.75, a support of 0.058, a lift of 1.75, a leverage of 0.025 and a conviction of 2.3;
- {SPRING ROLL, MINERAL WATER 1.5LT, SWEET SOUR CHICKEN} ⇒ {DUCK} with a confidence of 0.74, a support of 0.072, a lift of 1.73, a leverage of 0.03 and a conviction of 2.21;
- {1.0 SWEET SOUR CHICKEN, 1.0 1/4 DUCK} ⇒ {1.0 EGG FRIED RICE} with a confidence of 0.57, a support of 0.052, a lift of 1.70, a leverage of 0.022 and a conviction of 1.54.

When considering only Indian products, we generated rules like:

- {JIRA PULAO, CHICK KORMA KASHMIRI} ⇒ {NAAN} with a confidence of 0.75, a support of 0.061, a lift of 1.54, a leverage of 0.021 and a conviction of 2.02.
- {NAAN, CHICK KORMA KASHMIRI} ⇒ {JIRA PULAO} with a confidence of 0.71, a support of 0.061, a lift of 1.70, a leverage of 0.025 and a conviction of 2.01.

Considering if the transaction is dine-in or delivery, we derived rules such as:

- {EXTRA SAUCE, EXTRA PANCAKES} ⇒ {DUCK} with a confidence of 0.99, a support of 0.064, a lift of 2.26, a leverage of 0.035 and a conviction of 58.1.

- {EXTRA PANCAKES} ⇒ {EXTRA SAUCE} with a confidence of 0.75, a support of 0.064, a lift of 8.24, a leverage of 0.056 and a conviction of 3.60.

As for the four seasons of the year and the weekends, no major differences or new insights were found when compared to the general case.

4. RESULTS EVALUATION

Our final analysis led us to identify some prominent patterns regarding our customers' preferences, spendings habits and routines, while providing detailed insights and suggestions.

With the respect to our business success criteria, we achieved several of the defined goals: we have determined relevant distinctions between dine-in and delivery customers, we thoroughly analyzed the product offering by suggesting 2 new set menus, 2 new products, various important combinations of products and 2 substitutes products, and we found interesting patterns in the customer base. Concerning our goal of increasing the profit margin by one percentual point within a year, we believe that with the application of our recommendations, regarding the changes in the product offering and the additional suggestions, we can achieve it. As shown below, we expect an increase of 1% in the total revenue per year with the two new products: water and noodles. Having that said, the most noteworthy insights for the three main cases are:

"General Case":

- Top 4 most ordered products are "MINERAL WATER 1.5LT", "DUCK", "EGG FRIED RICE", "SPRING ROLL";
- Interesting relationship between "NO MEAT" and "NOODLES WITH MEAT". We have noticed that all the tables that asked for "NO MEAT" also ordered "NOODLES WITH MEAT".
- The 2 categories of most combined products are "SIDES" and "MAIN";
- A relationship of substitution between "INDIAN" and "NOT INDIAN" dishes.

For this case we have come up with three main suggestions:

- Presenting "WATER" in different sizes, such as 33cl or 50cl. These tend to be more expensive per liter, generating more revenue. If we consider that 30% of the current consumption of water in 1.5L would be substituted for the water in 50cl (corresponds to the sum of orders with 1 or 2 people at the table divided by the number of orders where we do not have missing values) and we would have a profit margin of 40% by substituting three 50 cl bottles for one 1.5L (calculated using data from similar restaurants), ignoring the production costs, we would have an increase of profit of 2780€ (12% increase in the revenue from water);
- We argue that another product should be created: "NOODLES" (plain noodles that are not spicy, since the restaurant already has that, and without meat). If we consider that we would have an increase of 20% in the percentage of noodle related products corresponding to the demand of the new product and its price would be a reduction of 5% of the product "SPICY NOODLES" (using data from similar restaurants), ignoring changes in the demand of the other products, we would have an increase of profit of 6591€ (26% increase in the revenue from noodle related products);

- A new possible menu, named “Special Menu”: 2 “SPRING ROLL” and 1/4 “DUCK” as starters, 1 “BEEF BBS” or 1 “SWEET SOUR CHICKEN” as main, “EGG FRIED RICE” as side and 1 “MINERAL WATER 33/50cl” as a drink. The price of the menu would have to be defined (should be lower than the sum of the prices of the products) and the additional demand coming from the existence of this menu is unclear. Considering that a menu would affect the demand on every product that is part of it, it is difficult to determine a reasonable expectation of the return on the investment with the information we currently possess.

“Indian Dishes”:

For this case we created another possible new menu, called “Indian Menu”: 1 “NAAN” and 1 “JIRA PULAO” as sides, 1 “BUTTER CHICKEN” or 1 “CHICK KORMA KASHMIRI” or 1 “CHICK TIKKA MASALA” as main and 1 “WATER” as drink.

“Dine-in or Delivery”:

- Strong presence of the products “NO MEAT” and “NOODLES WITH MEAT”. Most of the times people asked for “NO MEAT” they were dine-in customers;
- Strong relationship between “SIDE” and “MAIN” in “INDIAN DISHES”;
- People who order “OTHERS”, alcoholic and non-alcoholic drinks also tend to ask for starters and main;
- Most ordered categories are “MAIN” and “STARTERS”;
- Customers who ask for a delivery tend to ask for more “EXTRAS”;
- The combination between “EXTRA PANCAKES” and “EXTRA SAUCE” stands out and is mainly ordered with “DUCK”;
- Customers who ask for delivery usually do not order drinks.

Our suggestions regarding this case are:

- Joining both “EXTRA PANCAKES” and “EXTRA SAUCE” to “DUCK”, possibly as an offer to promote the delivery part of the business. Currently, these two products represent a revenue of 283€ when it comes to delivery customers;
- An online discount could be offered to promote drinks in order to revert the tendency of not ordering them. Thought should be given to which drinks would be under a discount, the percentage of reduction and the time the discount is active. Therefore, with the current information, it is difficult to determine the revenues and costs associated.

5. DEPLOYMENT AND MAINTENANCE PLANS

The deployment and maintenance plan was divided into four different phases:

1. Present the model and its results to the relevant employees (such as executives and tech teams) to ensure we are deriving the expected information and meeting the needs of the users of these insights. If not, go back and rework the model until the needs are met and the insights are actionable.
2. Put the model into production, ensuring compatibility with the current technological infrastructure. The model must also undergo integration and system testing to ensure it is

compatible with the other existing systems within the data ecosystem and the necessary changes are made.

3. Integrate the results within the knowledge management platform so that the relevant employees can access and understand the obtained insights, helping with their decision-making. If this platform does not exist, guarantee that the insights can be easily obtained by the relevant employees following knowledge management best practices.
4. Continuously improve and monitor the performance of the model, considering the changes in the market and customer behavior, and present the insights. This can be done, for instance, every six months:
    a) On a technical aspect, the new transactions that occurred in those six months are added to the dataset and new association rules are generated. Then, we repeat the association rule mining process with the added information.
    b) On a business aspect, an analysis should be done to understand how the restaurant's client base changed. Besides, the suggestions that were implemented should be evaluated to understand their success. For example, we can monitor if the customers are adhering to the two menus, as well as the new products. For the deliveries, we can check if the customers are ordering more drinks than before and if offering extra products to "DUCK" is a successful strategy.

The costs associated to the plan are multiple such as model development and system integration costs. Thus, it is very difficult to determine the precise value without more information.

6. CONCLUSION

In conclusion, we were able to achieve most of our objectives and generate valuable insights to help us understand our customer base. These insights, coupled with the recommendations we made to improve the business, will contribute to its growth and success in the long term.

Moreover, we have also outlined a comprehensive deployment and maintenance plan that aims to monitor the progress of the recommended actions and potentially discover new insights periodically. We are confident that the insights and recommendations, along with the continual development of our model, will help differentiate the hotel from its competitors and increase our profit margin.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

There are several ways to improve our model and the insights we obtained:

- Extracting more features from the customers, such as age or gender, could prove valuable for customer engagement and segmentation;
- Extracting more data about dine-in customers. This could be done by creating an app where they would register and take part in a loyalty program;
- Utilizing data from not only 2018, which would help in generating new patterns and understanding tendencies over time;
- Creating and/or improving automated data pipelines, for instance when it comes to pre-processing, which would help to speed-up and improve the current processes;

- Considering different methods for obtaining the threshold value in support for the Apriori algorithm (Hikmawati, Maulidevi, Surendro, 2021, p.3);
- Testing the Apriori algorithm for the various cases with different support, confidence and lift values;
- Creating new ways of filtering the dataset for the Apriori algorithm, for example by the holidays, hour of the day or product categories;
- Creating more variables to express new levels of granularity, improving the insights in the market basket analysis.

7. REFERENCES

timeanddate (2023). *Holidays and Observances in Cyprus in 2018*. https://www.timeanddate.com/holidays/cyprus/2018

Virtual Crossing Corporation. *Weather Query Builder*. Weather Data Services | Visual Crossing

Republic of Cyprus, Statistical Service. *Arrivals of Tourists by Country of Usual Residence, Annual*. Arrivals of Tourists by Country of Usual Residence, Annual. PxWeb (cystat.gov.cy)

Georgiu, S. (2018). *Nicosia Beer Fest 2018!*. Nicosia Beer Fest 2018! - This September - Be there… (cyprusalive.com)

Cyprus Film Days International Festival (2018). *16TH CYPRUS FILM DAYS INTERNATIONAL FESTIVAL* 2018. 16th Cyprus Film Days International Festival 2018 - Cyprus Film Days

myCyprusTravel (2018). *Cyprus Pride day Parade 2018*. Cyprus Pride day Parade 2018 - My Cyprus Travel | Imagine. Explore. Discover.

Wikipedia (2021). *2018 Cyprus Women's Cup*. 2018 Cyprus Women's Cup - Wikipedia

Raschka, S. (2022). *association_rules: Association rules generation from frequent itemsets*. Association rules - mlxtend (rasbt.github.io)

Keshari, k. (2020). *Apriori Algorithm : Know How to Find Frequent Itemsets*. Apriori Algorithm : Know How to Find Frequent Itemsets | Edureka

Software Testing Help (2023). *Apriori Algorithm In Data Mining: Implementation With Examples*. Apriori Algorithm in Data Mining: Implementation With Examples (softwaretestinghelp.com)

Hikmawati, E., Maulidevi, N.U., Surendro, K. (2021). *Minimum threshold determination method based on dataset characteristics in association rule mining*. Minimum threshold determination method based on dataset characteristics in association rule mining | Journal of Big Data | Full Text (springeropen.com)

Korstanje, J. (2021). *The Eclat Algorithm*. The Eclat Algorithm | Towards Data Science

8. APPENDIX

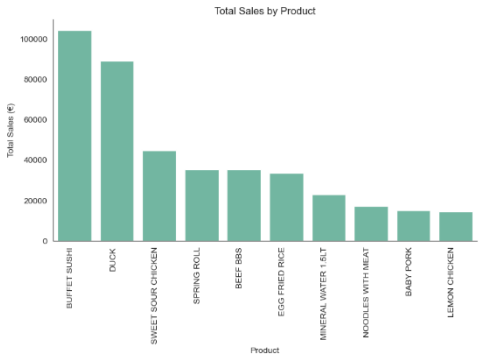

Figure 8.1 – Total Sales Volume by Product2

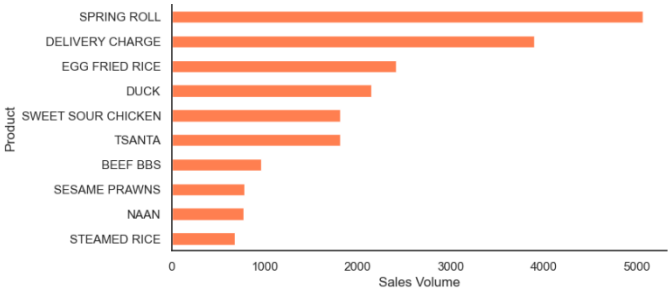

Figure 8.3 – Top 10 Delivery Sales Volume by Product

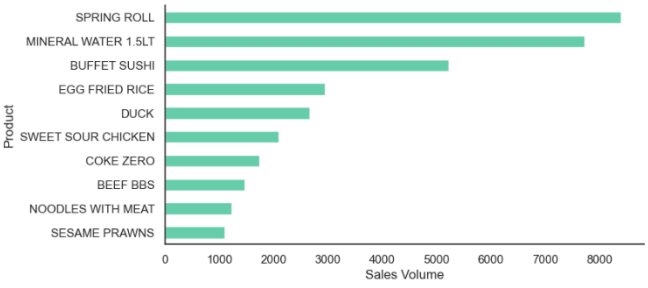

Figure 8.4 – Top 10 Dine-in Sales Volume by Product

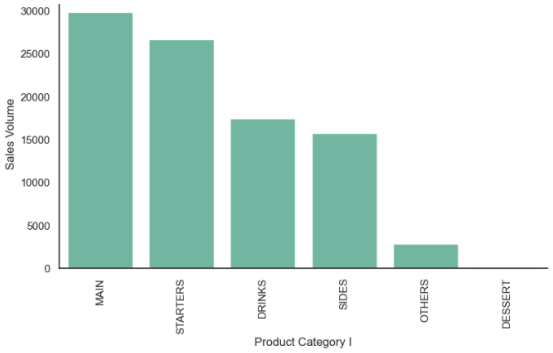

Figure 8.5 – Total Sales Volume by Product Category I

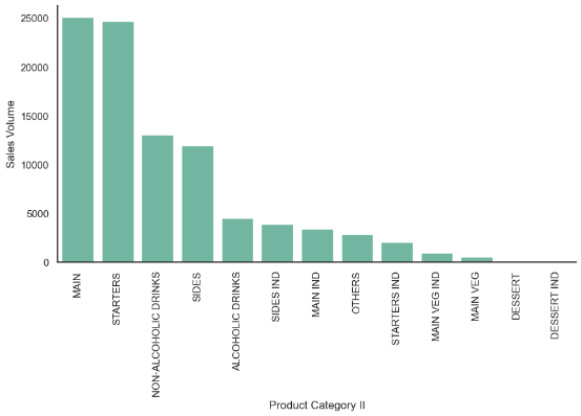

Figure 8.7 – Total Sales Volume by Product Category II

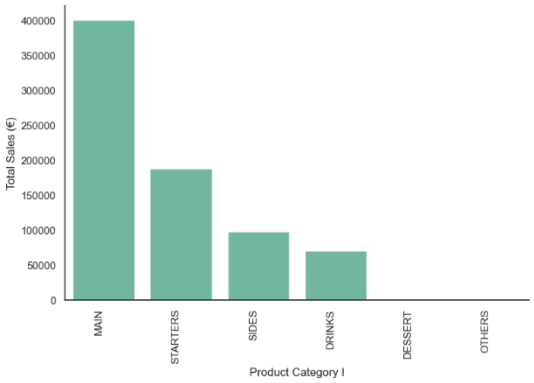

Figure 8.6 – Total Sales by Product Category I

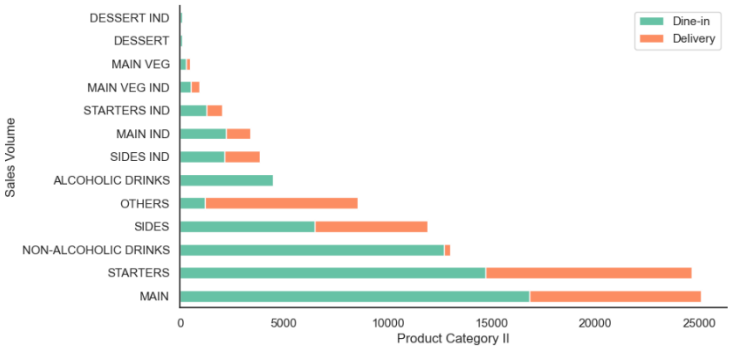

Figure 8.9 – Sales Volume by Product Category II for Dine-in and Delivery

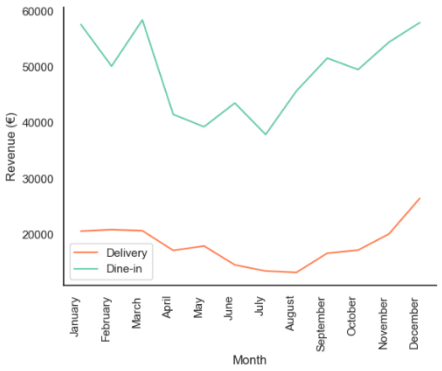

Figure 8.8 – Revenue per Month for Dine-in and Delivery Order

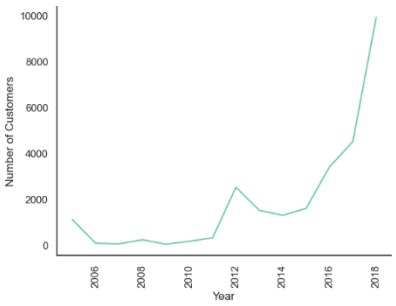

Figure 8.10 – New Customers per Year

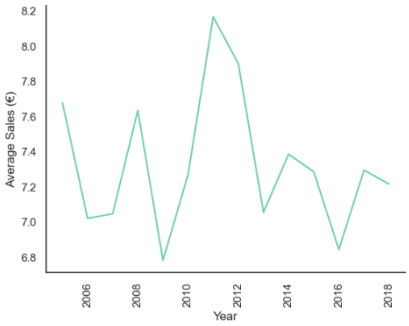

Figure 8.11 – Average Sales by the Year the Customer Was Registered

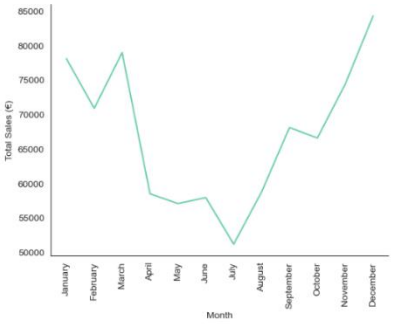

Figure 8.13 – Total Sales per Month

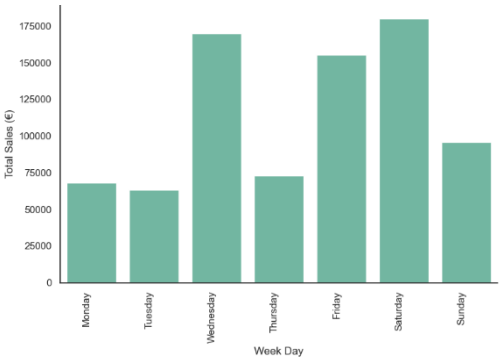

Figure 8.12 – Total Sales per Day of the Week

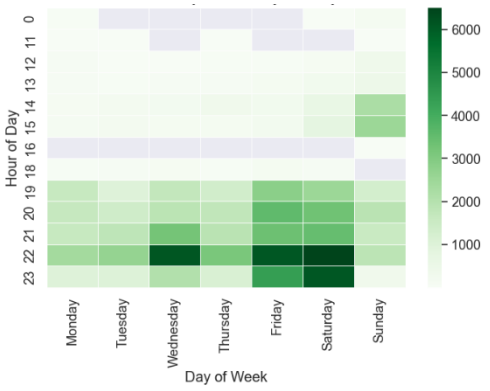

Figure 8.16 – Total Sales Volume by Hour of Day and Day of Week

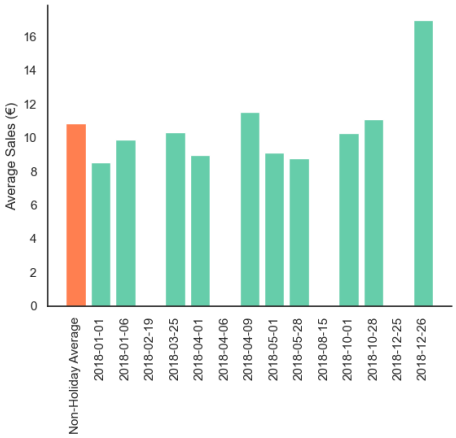

Figure 8.15 – Comparison of Total Sales on Holidays and Non-Holidays

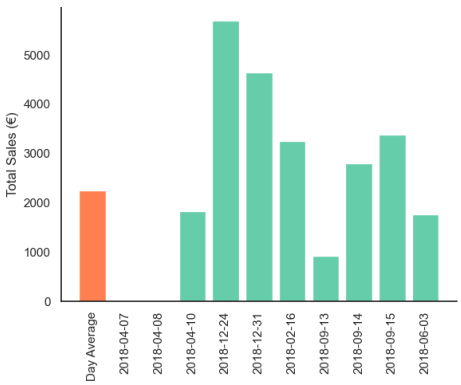

Figure 8.17 – Comparison of Total Sales on Normal and Event Days

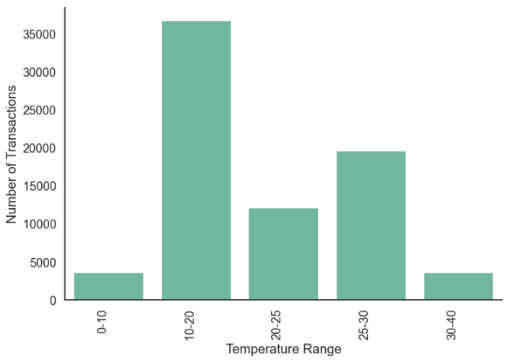

Figure 8.14 – Total Transactions by Temperature Range