BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Instacart Market Basket Analysis

Group F

Lorenzo Pigozzi	m20200745
Nguyen Huy Phuc	m20200566
Ema Mandura	m20200647
Xavier Gonçalves	m20201090

April 19, 2021

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CONTENTS

1. INTRODUCTION 3
2. BUSINESS UNDERSTANDING 3
   2.1. Background 3
   2.2. Business Objectives 3
   2.3. Business Success criteria 3
   2.4. Situation assessment 4
   2.5. Determine Data Mining goals 4
3. DATA MINING PROCESS 4
   3.1. Data understanding 4
   3.2. Data preparation 4
   3.3. Exploratory Analysis 4
   3.4. Market basket analysis with association rules 7
4. RESULTS EVALUATION 8
5. DEPLOYMENT AND MAINTENANCE PLANS 8
6. CONCLUSIONS 9
   6.1. Considerations for model improvement 10
7. REFERENCES 10

1. INTRODUCTION

As a fundamental strategy for any retail company, it is important for consumers' purchasing patterns to be identified to obtain a set of information useful to the sales strategy and product layout, as well as in terms of advertising and marketing. Thus, it is important to identify what are the relationships between the different products, whether they are complementary, substitutes, inferior or superior and of convenience to increase revenues and make logistics more efficient, for example in relation to perishable goods, as well as in the disposition of goods in different areas.

To improve the consumer experience, companies in this field collect data on transactions and customers to know who the customers are who will buy in the future or try the application again and what their consumption patterns are.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Instacart is an American company that provides a grocery delivery and pick-up service via a website or mobile app in the United States and Canada. With a wide range of products, consumers can use this application to better manage their purchases of supermarket goods. A personal shopper is also assigned, and purchases are delivered to customers on the same day they are made on the application or on the website, after being reviewed by employees.

Currently, Instacart uses transactional data to understand which products a user is likely to buy again, try for the first time, or add to their cart next during a session. Although, the lack of analytical capabilities hinders Instacart's ability to take full advantage of this data. Due to the lack of qualified in-house data scientists, the company hired an external consultant to some of the questions they have been struggling to address.

2.2. BUSINESS OBJECTIVES

The main objective of the team of consultants is to deliver a study as complete as possible on the data that is available on Instacart's customers. Several analyzes can be applied to maximize the number of insights and extract valuable information that can be used to improve the consumer experience and increase the company's revenues, facilitating logistics and having a better interaction with customers. In short, to meet what consumers want.

For this, we intend to explore the type of consumer behavior, making segmentation and realizing who they are, what they buy and when they buy. Then, which are the areas that should have a broader range of product offerings, and finally which are the products that can be seen as complementary goods, bought together, and which can be substitutes, rarely bought simultaneously. This analysis will be done using association rules, such as confidence, expected confidence, support and lift.

2.3. BUSINESS SUCCESS CRITERIA

The success of the project can be measured by the variety of correct and accurate analyzes that can be done on the data. For example, at the level of the type of products that tend to be purchased together, or products that are never purchased together because they are eventually substitutes. This study will be done using association rules such as confidence, expected confidence, lift and support. When carrying out this analysis, the expectation is to find some of these relationships between products so that in the future they can even be used for the layout of products in consumer exposure, suggestions and improve the experience, increasing revenues for Instacart and improving logistic operations.

2.4. SITUATION ASSESSMENT

The team has complete data to make an analysis to the consumption patterns of the customers, such as the simplified types of products, departments to which the products belong, day of the week and time of purchase, days since the last purchase, among other data. The data provided was mostly clean, with few insignificant missing values. It is mostly accurate and relevant.

The hotel also provided their full support during the project, as their representative Mrs. Jane was available for questions and assistance during the full length of the project.

2.5. DETERMINE DATA MINING GOALS

In this context, the strategy used in terms of data mining will be a market basket analysis, which will analyze, among other factors, which products are often purchased together, which are substitute or complementary products, as well as a segmentation of customers considering available data. The study's conclusions will later be used to develop the logistics and disposition of goods in the supermarket, its inventory, and to optimize marketing strategies, promotional programs, and recommendation engines.

3. DATA MINING PROCESS

3.1. DATA UNDERSTANDING

The data is composed by four datasets:

- Orders, with 200,000 observations and 6 variables, with information about product and customer ids, as well as order dow, time the purchase was made and days since the last purchase.
- Order_products, with 2019500 observations and 4 features, with information on order and product identification, add to cart and a binary on whether it was reorder or not.
- Products, with 134 generalized product observations, and 3 variables, with information about the name and id of the product and the department to which they belong.
- Departments, with 21 possible departments, including 1 for "missing" and 2 features, containing information about the ID and name of the departments.

The metadata lists the variables and explains their meanings, available here.

3.2. DATA PREPARATION

As part of data preparation, all four datasets were merged into one for the purpose of better data analysis. In the new dataset, each of the 2019501 rows represents a product sold, along with the information about the order it is a part of and the user that bought it. Through dataset exploration, it was found that the only column with missing values was *days_since_prior_order.* With the missing data affecting 6% of the dataset, we still decided to keep the records with missing data as different analysis can still be conducted.

3.3. EXPLORATORY ANALYSIS

Firstly, we explore the data for Instacart products by aggregate the purchasing data. The top 5 products with the most purchases are Fresh fruits (226,039), Fresh vegetables (212,611), Packed vegetable fruits (109,596), Yogurt (90,751) and Packed cheeses (61,502). We can see that the first two products clearly dominated the others. This indicated that the key products of the grocery store are fresh foods

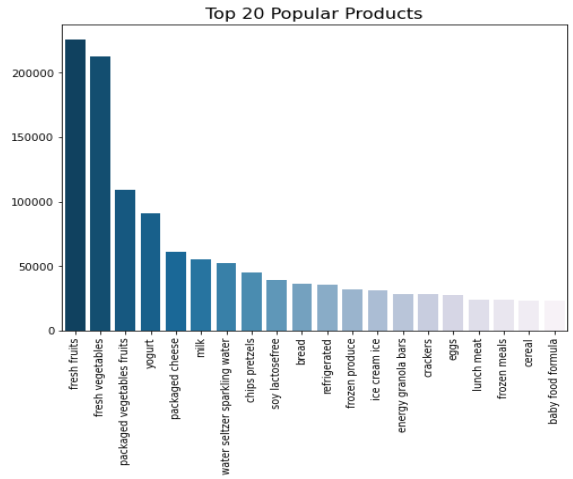


Figure 1. *Amount of purchases by products*

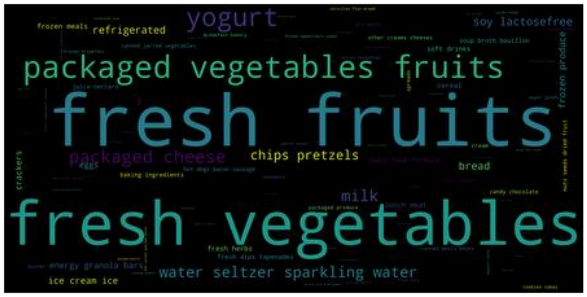


Figure 2. *Word cloud of products' purchases amount*

Secondly, we take a look into customer behavior and orders data such as day of the week when orders occurred (order_dow), hour of the day that orders occurred (order_hour_of_day), number of days since the previous order (days_since_prior_order)

Order amount starts to increase from 8 AM, reaches the peak hours of day at 9 and 15 and then starts to decrease till midnight. During the day, 21% of customers prefer to shop in the morning, 33% mostly shop midday, and 44% like doing their shopping in the evening.

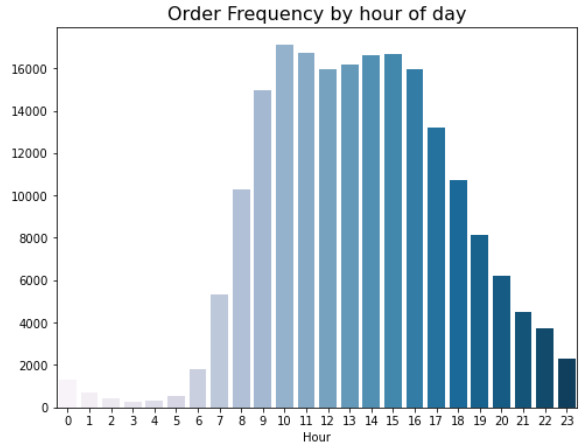


Figure 3. *Distribution of orders by hour of day*

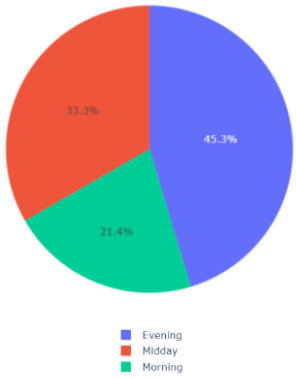


Figure 4. *Percentage by hour*

When it comes to over-all customer behavior, most customers prefer to do their shopping on Sunday, with a significant number of them prefering Monday aswell. However, 45% of customers prefer shopping on weekends.

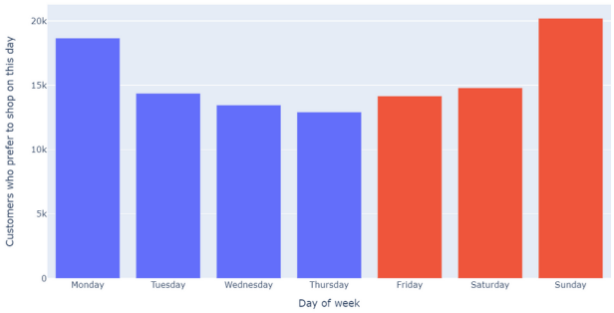


Figure 5. *Orders distribution by day of week*

Frequency-wise, there is a distinct trend of doing grocery shopping weekly and monthly.

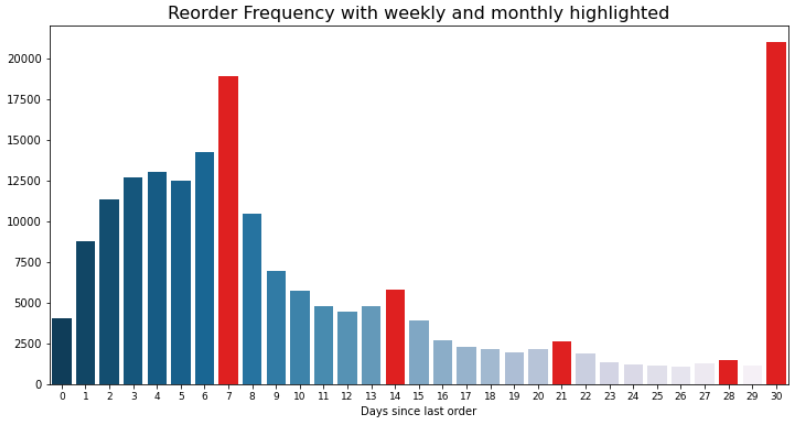


Figure 6. *Order frequency*

The top 5 best selling departments, sorted by number of product purchased, are Produce (588,996), Dairy Eggs (336,915), Snacks (180,692), Beverages (168,126), and Frozen (139,536).

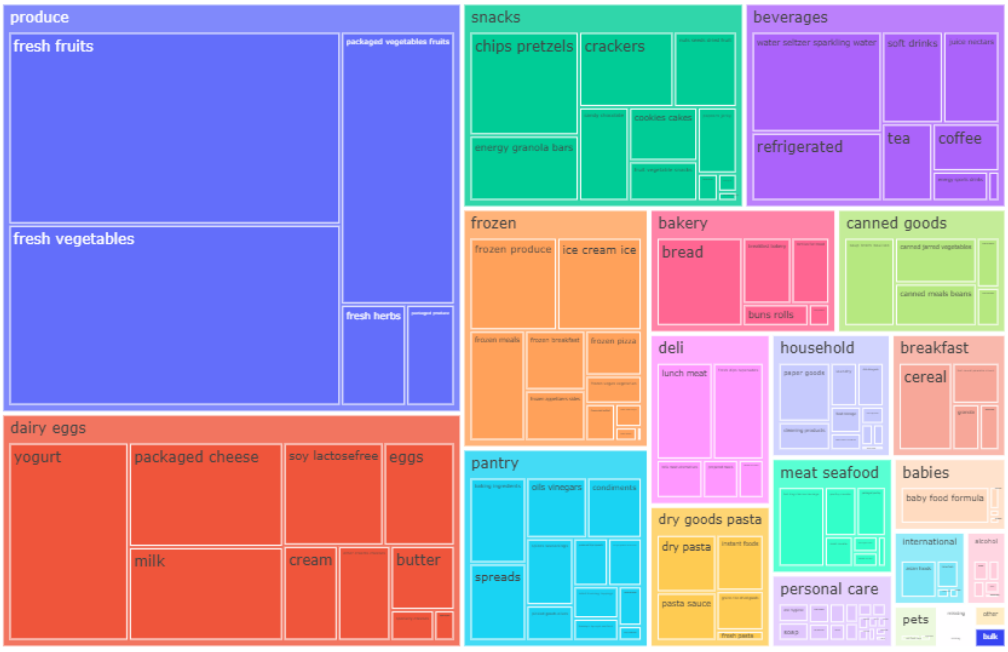


Figure 7. *Top departments*

3.4. MARKET BASKET ANALYSIS WITH ASSOCIATION RULES

In this section, we will use Apriori algorithm to analyze frequent itemset and association rule mining over the transactional data. The Apriori proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those itemset appear sufficiently often in the database, in our analysis, we set 2 thresholds, one of 5% support for complementary products and the other of 1.5% support for substitute products

The first perspective of analysis is the complementary products which are analyzed inter-department. The complementary should be analyzed through all the department as it would be counter-intuitive to analyze restrictedly within each department. The logic is to find the highest lift of top products' pairs in all purchases from all department that have acceptable confidence level of minimum 20%

Antecedents	Consequents	Antecedent support	Consequent support	Support	Confidence	Lift
(fresh herbs)	(fresh vegetables)	0.093005	0.444360	0.078655	0.845707	1.903203
(canned jarred vegetables)	(fresh vegetables)	0.071890	0.444360	0.055055	0.765823	1.723429
(canned meals beans)	(fresh vegetables)	0.069705	0.444360	0.050435	0.723549	1.628295
(bread)	(packaged cheese)	0.163865	0.230995	0.059690	0.364263	1.576931
(packaged cheese)	(bread)	0.230995	0.163865	0.059690	0.258404	1.576931

Table 1. *Association rules of Complementary products*

The table shows the limited top 5 itemsets with highest lift. However, we also provided the dashboard for the business stakeholders to discover more informative rules in their customer's baskets. The interpretation guide of these rules are as follow: If the *[Antecedents]* is part of the basket, also *[Consequents]* will be part of the event *[Confident*100]* % of the times

For e.g.: "If fresh herbs is part of the basket, also fresh vegetables will be part of the event 84 % of the times. "

On the other hand, the substitute products are different from as we find the lowest lift of product pairs, and **within department**. Because we think it would be more effective to analyze the substitute product of the same department as they are likely to be placed next to each other in the store.

Antecedents	Consequents	Support	Confidence	Lift	Department_name
(food storage)	(paper goods)	0.025021	0.211324	0.490873	household
(water seltzer sparkling water)	(refrigerated)	0.070694	0.167094	0.568686	beverages
(fresh vegetables)	(packaged produce)	0.030185	0.050815	0.610449	produce

(nuts seeds dried fruit)	(chips pretzels)	0.045064	0.243690	0.623948	snacks
(soy lactose free)	(milk)	0.061012	0.244840	0.679806	dairy eggs

Table 2. Association rules of substitute products within department

4. RESULTS EVALUATION

The exploratory analysis has carried out some interesting insights about buying patterns and customer behaviors. Such insights can be useful for the store owner to adjust the sales and promotions plan daily and weekly that fit the patterns about day and time. The store owner also can analyze the top sale products to have suitable supply and inventory management.

Regarding the association rules, some of the recommendations have been made and it will be productive to run promotional and marketing campaigns with the help of these rules. The complementary products can be used to not only promote cross-sale opportunities but also further developed into a recommendation system for online customer. Once itemset have been identified as having good 'lift', recommendations can be made to customers in order to increase sales for similar products. On the other hand, substitute products can be analyzed to have action plan when a product is running out of stock or adjust the low sale products.

5. DEPLOYMENT AND MAINTENANCE PLANS

The deployment phase of every company is crucial for the business context. It provides a better idea about the effective usage of the machine learning algorithm created, designing a possible bridge between the business goals and the machine learning goals, applying the algorithm in a real-world context.

In order to improve the accessibility of the results obtained for the company and to maintain them in the future, our team have created a dashboard, in which is possible to analyze the patterns and behaviors of the products purchased.

The dashboard created is composed by three different sections, each one aiming to provide different insights of the study conducted.

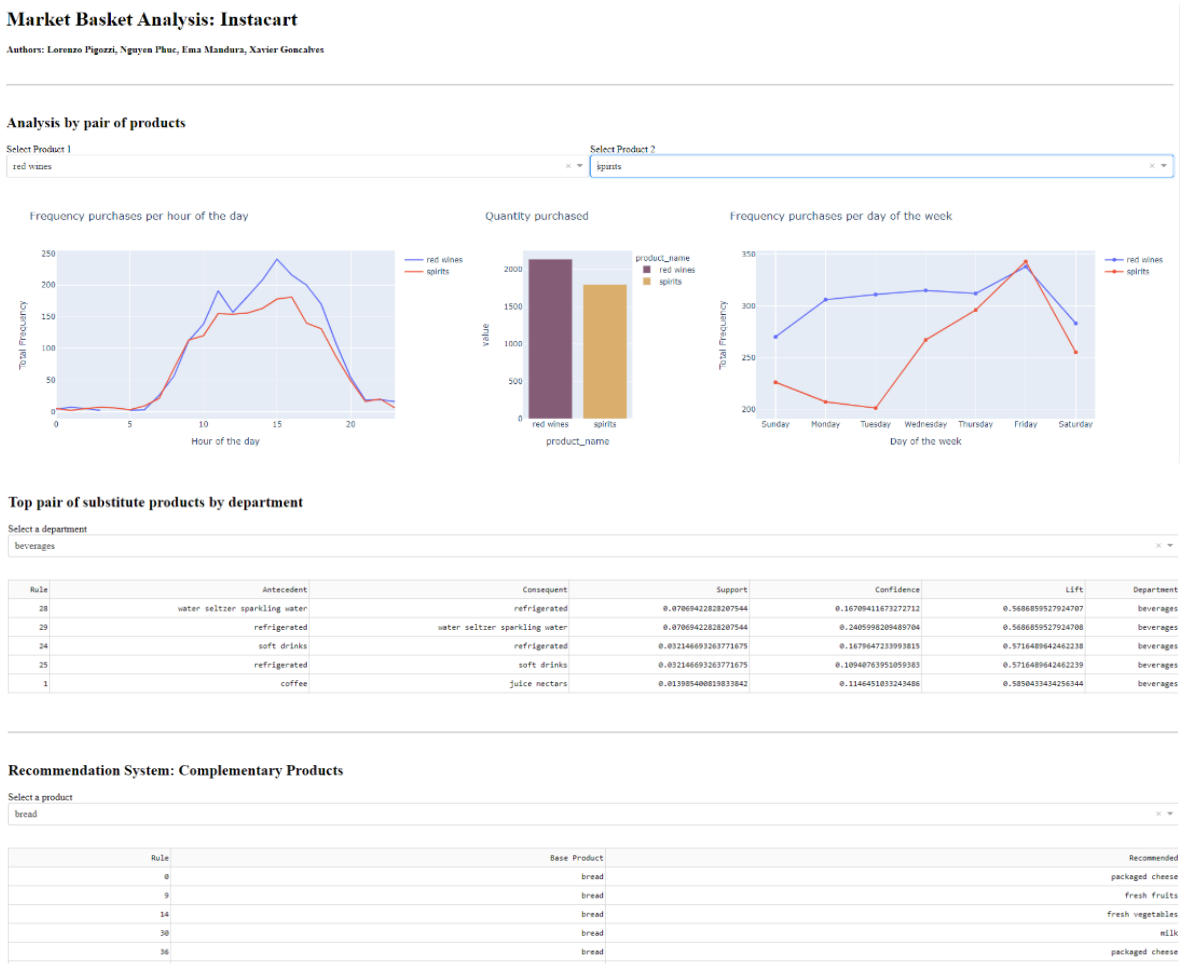
In the first section, it's possible to compare the behaviors of two different products at the same time, through three different plots that are showing respectively the time of the day in which they are purchased, the day of the week and the total amount of purchases of the specific product.

In the second section of the dashboard, it's possible to get insights regarding the substitute products that were identified on the data, based on the techniques applied, in particular the Apriori algorithm. For this insight, we chose to analyze the substitutions based on the departments. Indeed, for the company is important to know which are very similar products that can substitute the others, and that's the reason why we chose the carry out an intra department analysis for the substitutions. Selecting a department through the dropout choice, a table with the best pair of products is displayed, as well as the metrics of the rules, such as Lift, Confidence and Support, and using them the reader will be able to interpret the result and evaluate by himself.

In the last section, we provided a simple recommendation system: a dropout command allows the reader to choose a product and based on that it will be displayed a table with the top complementary products for the selected one. The recommended products that are displayed are selected based on

the strength of the rules defined by the metrics. In this way, it will be possible for the marketing department to suggest some products based on the previous purchases of the customers.

The importance of the deployment plan we're providing is mainly due to the future variability of the results that we have found. Indeed, changing the marketing approach or using some of the rules provided to make recommendation will impact for sure the future patterns and relationships among the products. That's the reason why a tool to maintain the main pipeline of the project is of fundamental importance. In the future the data will be different, as well as some of the rules found, and so the managers will be able to use the dashboard with future data in order to find and analyze the new behaviors of the products.



Rule	Antecedent	Consequent	Support	Confidence	Lift	Department
28	water seltzer sparkling water	refrigerated	0.0706942282807544	0.16709411673727212	0.5686859527924707	beverages
29	refrigerated	water seltzer sparkling water	0.0706942282807544	0.2405998209489704	0.5686859527924708	beverages
24	soft drinks	refrigerated	0.032146693263771675	0.167964723399381 5	0.5716489642462238	beverages
25	refrigerated	soft drinks	0.032146693263771675	0.10940763951059383	0.5716489642462239	beverages
1	coffee	juice nectars	0.013985400819833842	0.1146451033243486	0.5850433434256344	beverages

Rule	Base Product	Recommended
0	bread	packaged cheese
9	bread	fresh fruits
14	bread	fresh vegetables
30	bread	milk
36	bread	packaged cheese

Figure 8. Dashboard

6. CONCLUSIONS

To be able to collaborate with Instacart in the analysis of its information, the study carried out on the existing data should be as deep and objective as possible. The main objectives were that the exploratory analysis was able to define the behaviour of the various consumer segments, the types of products that should have an absence of supply, which are the substitute products, and which are the complementary ones.

Since the dataset is relatively clean and with only one column with missing values, the entire data preparation process has been simplified. After carrying out the exploratory analysis, it was found that

fresh fruits and fresh vegetables were the two most popular products, with a large advantage over the others. The departments that produce consumers' favourite products are also naturally the most sought after, such as "Produce" and "Dairy Eggs". Regarding the time features, there was a great adhesion to Instacart's services during the period between 10 am and 3 pm, and Sunday was the day chosen by most users of the services to make purchases, which seems to be intuitive and not surprising. In addition, it was also relevant to highlight that although there is a range of people who shop regularly during the week, with a few days difference, there is a group of consumers who prefer to shop with a defined periodicity, be it 30 days (purchases monthly) or 7 days (weekly purchases), which are the ones with the highest adherence.

The main conclusions to be drawn using the association rules, after setting the thresholds, are the existence of some combinations of two products that present interesting levels of lift and confidence that will be able to identify potential complementary products. Examples include fresh herbs and fresh vegetables, canned jarred vegetables and fresh vegetables, bread and packed cheese. These relations can help to perceive and establish some suggestions in terms of goods to customers who purchase the previous products, thus increasing the possibility of them coming to purchase the second item.

Relative to substitute products, the analysis was carried out within each department and in terms of product pairs. The products with low lift identified are those that acquire the antecedent, it is more unlikely that the consequent is also acquired, creating a dynamic in which a product is potentially substitute (may even have similar purposes, satisfy the same or similar needs but with some differences). Some more intuitive examples are soy lactose free and milk, nuts seeds dried fruit and chip pretzels or fresh vegetables and packaged produce.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

The data we were provided was of good quality and allowed us to preform valuable analysis. With more time at hands, more insights could be gathered from the data, especially if more records were provided.

7. REFERENCES

[1] Retail Analytics: A Novel and Intuitive way of finding Substitutes and Complements
https://towardsdatascience.com/retail-analytics-a-novel-and-intuitive-way-of-finding-substitutes-and-complements-c99790800b42