Instacart Market Basket Analysis Project

CONTENTS

1. Problem Statement
2. Data used
2.1. Variables used
3. Approach
3.1 General Machine Learning Approach
4. Exploration of Variables
4.1 Initial analysis of data
4.2 Checking the distribution of the variables
5. Determining missing values
5.1 Dealing with missing values
6. Understanding how to work on the data
7. Market Basket Analysis
7.1. Apriori
7.2 Bayesian Updating of a Prior
8. Conclusion

1. Problem Statement

Instacart, a grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them. After selecting products through the Instacart app, personal shoppers review your order and do the in-store shopping and delivery for you.

In this project, we will use the anonymized data on customer orders over time to predict which previously purchased products will be in a user's next order.

2. Data used

The dataset for this problem is a relational set of files describing customers' orders over time. The goal is to predict which products will be in a user's next order. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, there are between 4 and 100 of their orders, with the sequence of products purchased in each order.

Each entity (customer, product, order, aisle, etc.) has an associated unique id. Three files **aisles.csv, departments.csv and products.csv** provide related information.

order_products__*.csv
These files specify which products were purchased in each order. order_products__prior.csv contains previous order contents for all customers. 'reordered' indicates that the customer has a previous order that contains the product. Note that some orders will have no reordered items. We will predict an explicit 'None' value for orders with no reordered items

orders.csv
This file tells us to which set (prior, train, test) an order belongs. We will predict reordered items only for the test set orders.

2.1 Variables used:

- aisle, aisle_id, department, department_id, product_name, product_id.
- order_id,product_id,add_to_cart_order,reordered

3. Approach

3.1 General Machine Learning Approach:

To speak of our model in a nutshell, it is a continuously learned procedure. The shopping patterns of each user entering the system are learned and then used to predict their future purchases.

We can take the history of the prior orders of the users and study the behavior. All the criteria that tend the user to buy or reorder a product are considered.

The machine learning algorithm will now be trained over time with data from the past. Thus, based on the previous trends, it can probabilistically predict the current trend.

We have a huge set of data up to 6 Million observations where each order by a user is an observation. The general idea is to group the orders by the individual users and study the ordering patterns. Some people tend to order ice creams into the evening every alternative day. So, we can say he will order one in the coming Sunday.

A simple illustration is shown below which shows how a basket is interpreted from the user shopping behaviors.

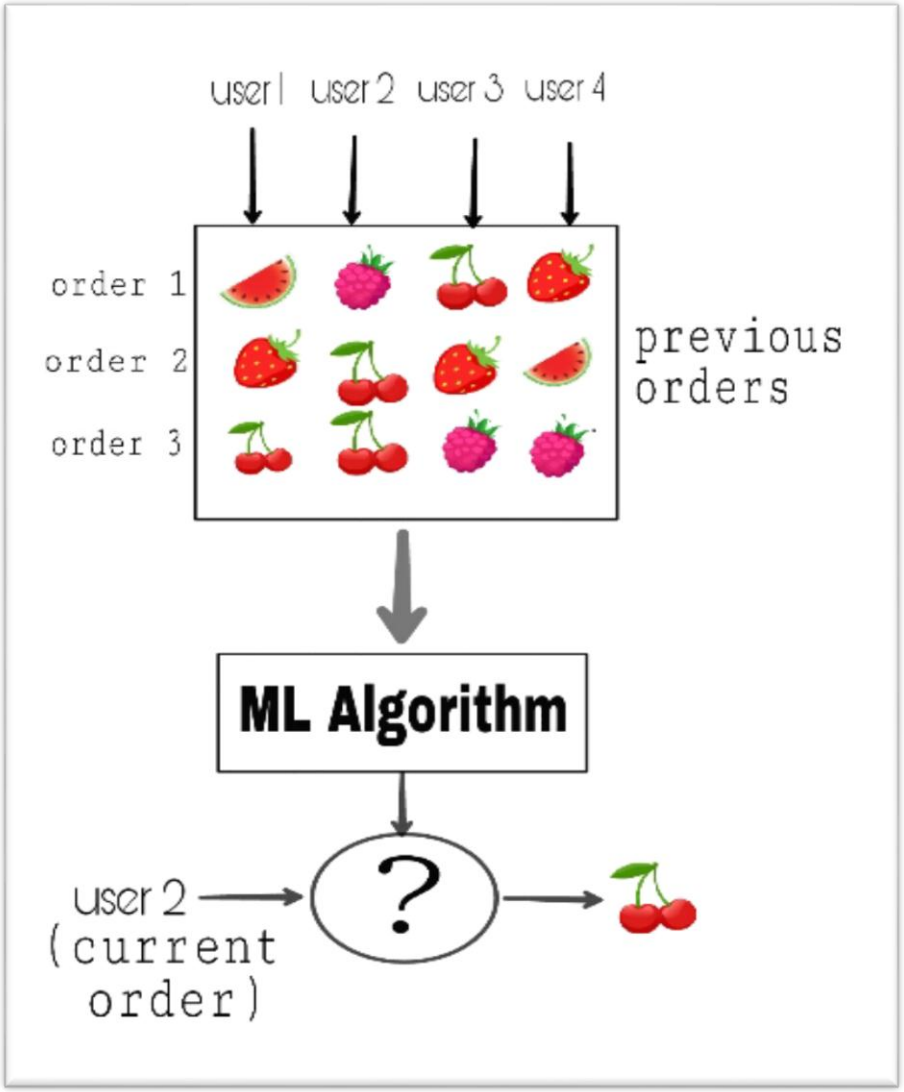

Figure 1 - A General Idea

4. Exploration of Variables:

The "order_product_prior" and "order_product_train" data have 4 variables and 4.9 and 1.3 Million observations respectively. We can extract prior, train and test data from "orders".
The "aisle","department" and "products" give the physical product and the categories it belongs to.

4.1 Initial analysis of data:

Roughly analyzing the data, we observe the distribution of each variable across the range of observations. Below are the distributions of all the observations across around 5 Million data.

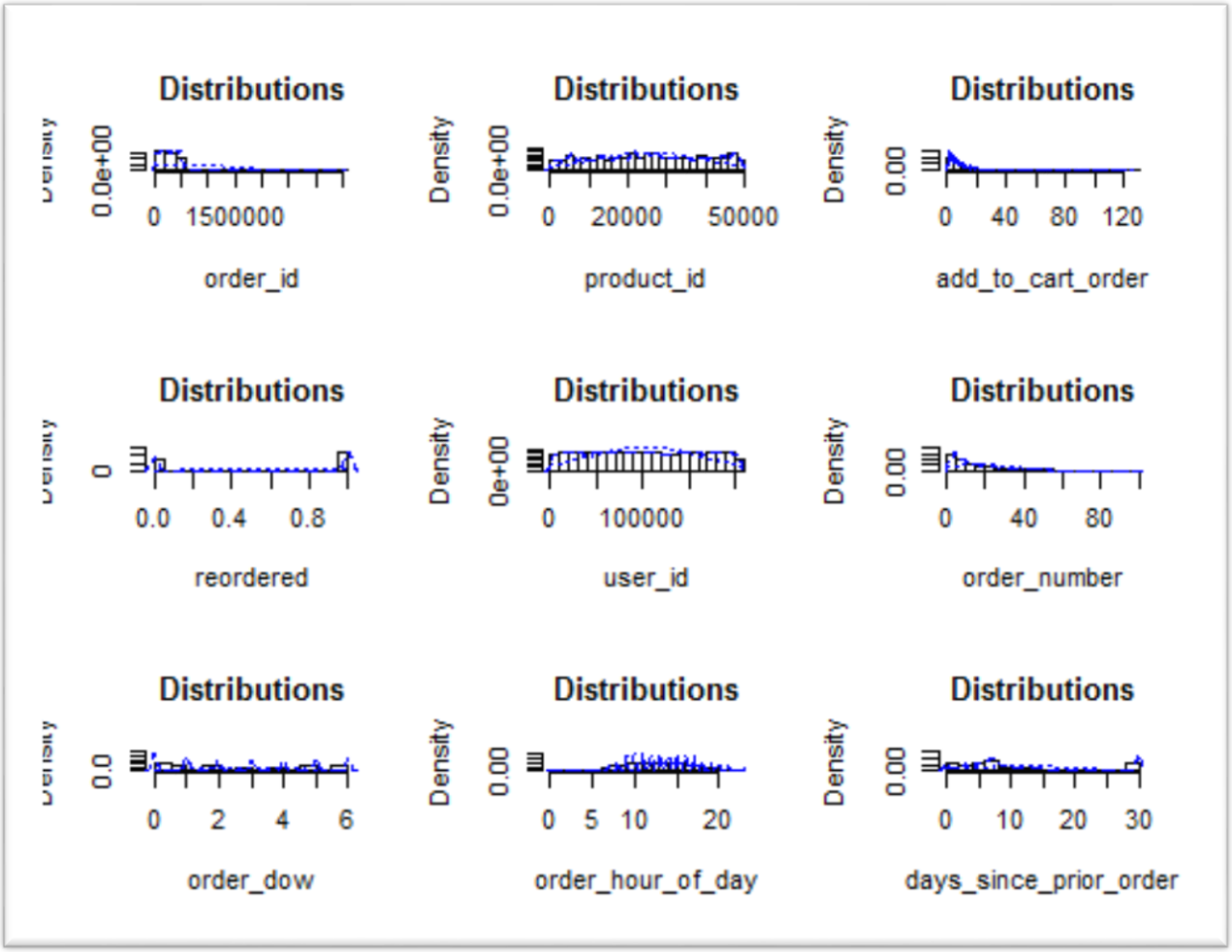

Figure 2 – Distribution of major features

We can already notice patterns like how people order more towards the end of the day and preferably on the weekends(observed on the variable dow – day of the week).

Each feature here contributes towards the problem solution.
For example, nearly 60% of the products are reordered whereas rest are not.

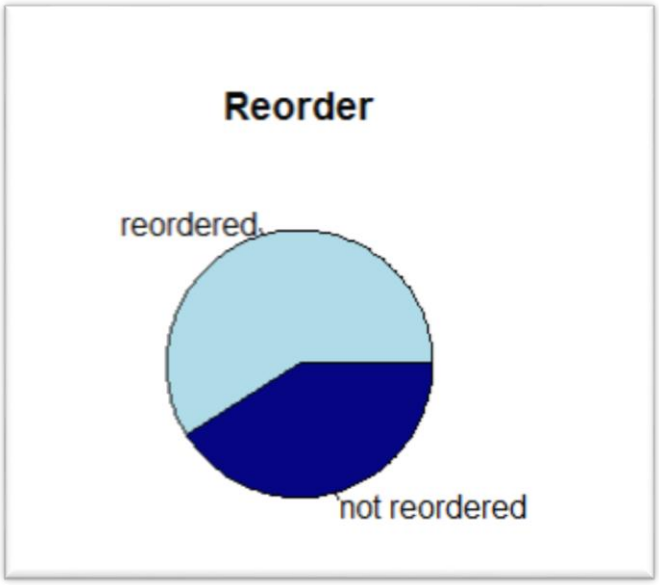

Figure 3 – Reorder rate

4.2 Checking the distribution of the variables:

Histograms were plotted to check the distribution of the variables.

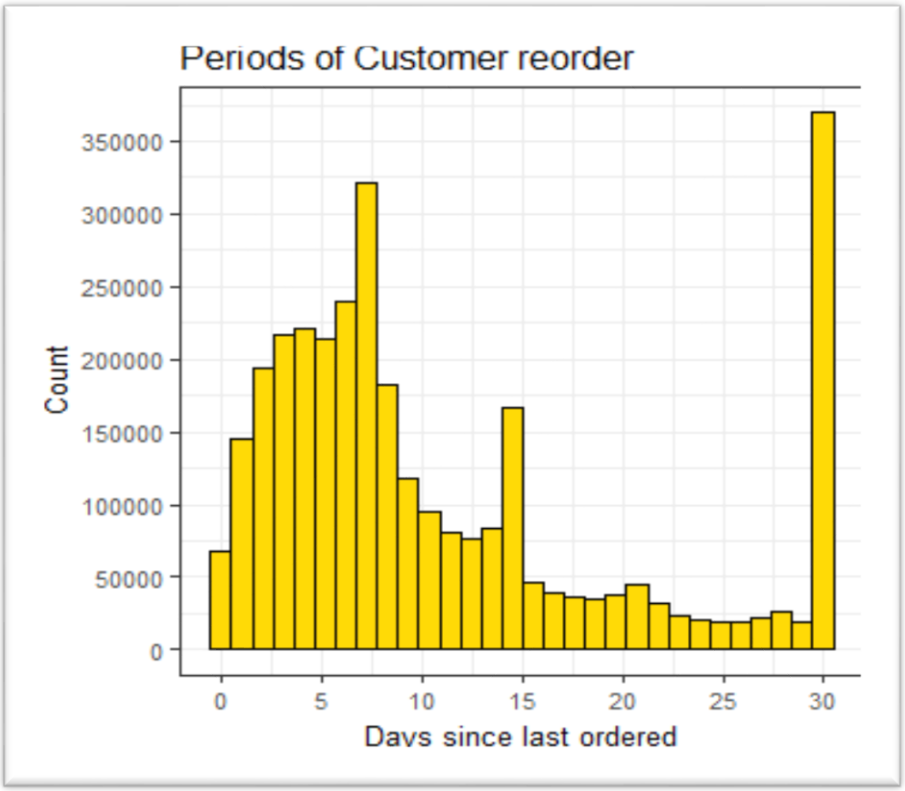

Figure 4 - Histogram of periods of Customer reorder

Majority of customers reorder from Instacart after 30 days which can be deemed as after a month. We can dig into the items ordered in this period and recognize if these are monthly groceries or similar patterns. We can also notice the next frequent orders come from after 7 days or a week and subsequent orders are after 15 days.

A pie chart of top 10 products ordered are shown below. Bananas followed by Organic Bananas and Organic strawberry are the most ordered products. In fact, organic products are preferred over non-organic is our study.

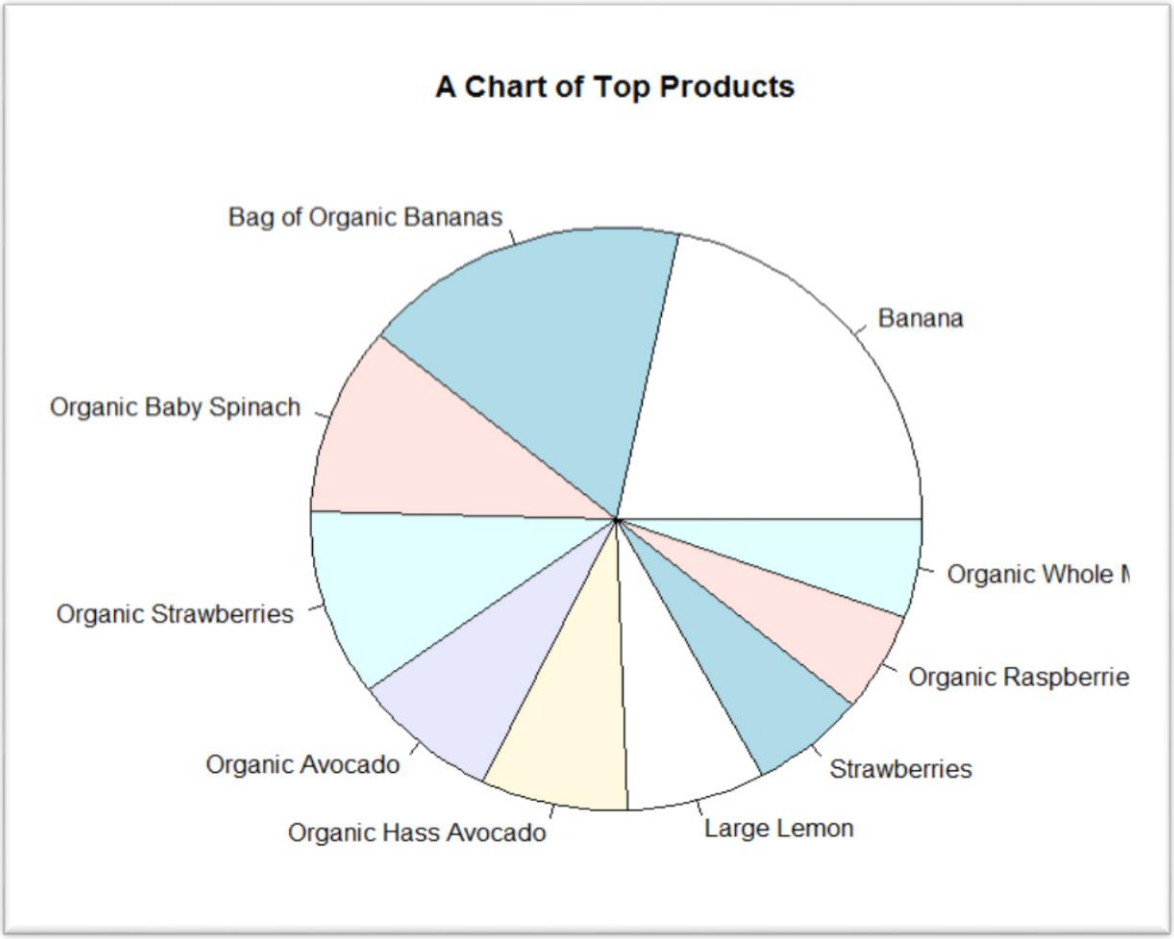

Figure 5 – Pie chart of top 10 products sold

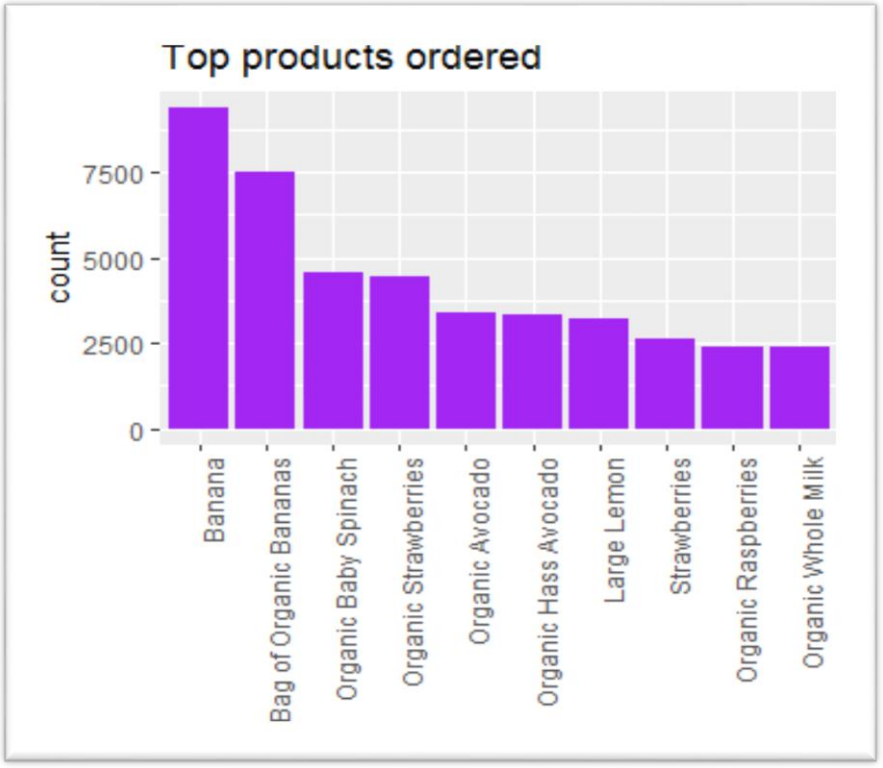

Figure 6 - Histogram of Top 10 ordered products

5. Determining missing values:

There are very few missing values in the prior dataset and none in train and test datasets.

In the below graph, we can see the spread of missing values in the order_prior dataset. The "days_since_prior_order" has significant missing values.

In the consistent plot, we can see the proportion of missing values with variables.

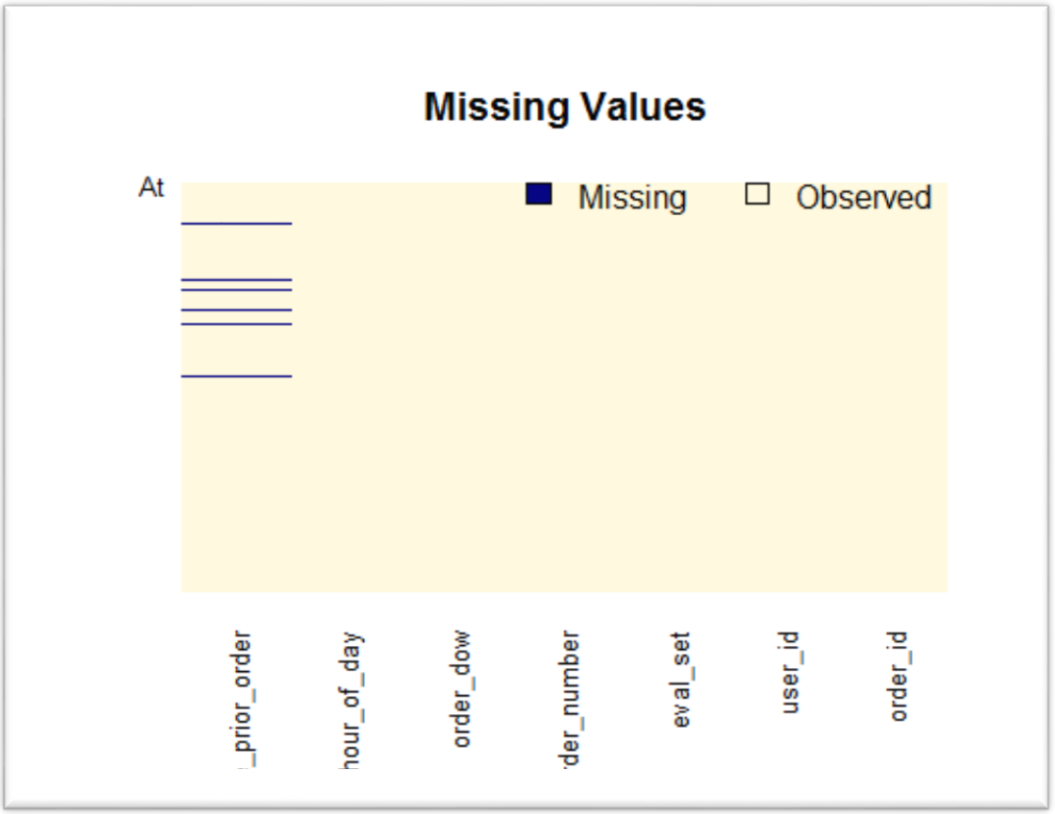

Figure 7 - Missing value frequency plot

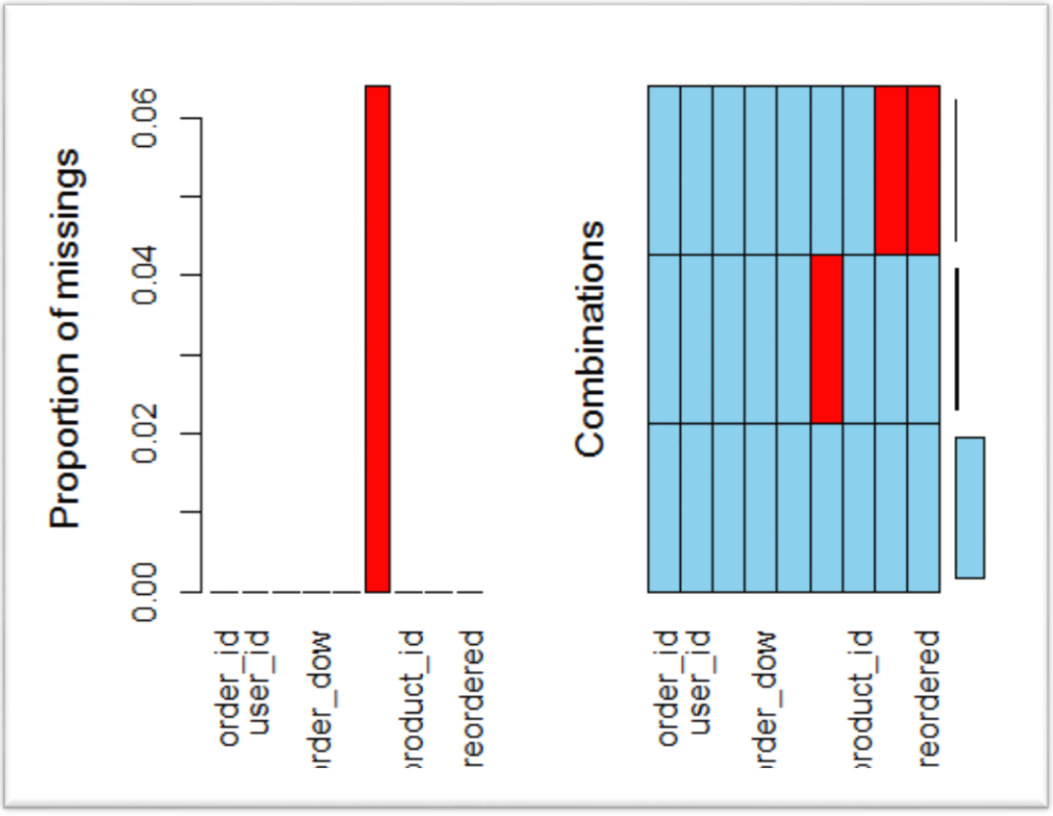

Figure 8 - Proportion of missing values

5.1 Dealing with missing values:

We will deal with missing values intuitively here.

The first order has NA values for “reordered” and “add_to_cart_order”. Replace it with zeros.

And “days_since_prior_order” have missing values as interpreted from the graphs. This means that the user has ordered for the first time. Thus, we impute with zeros once again.

6. Understanding how to work on the data

Aisles organized within departments:

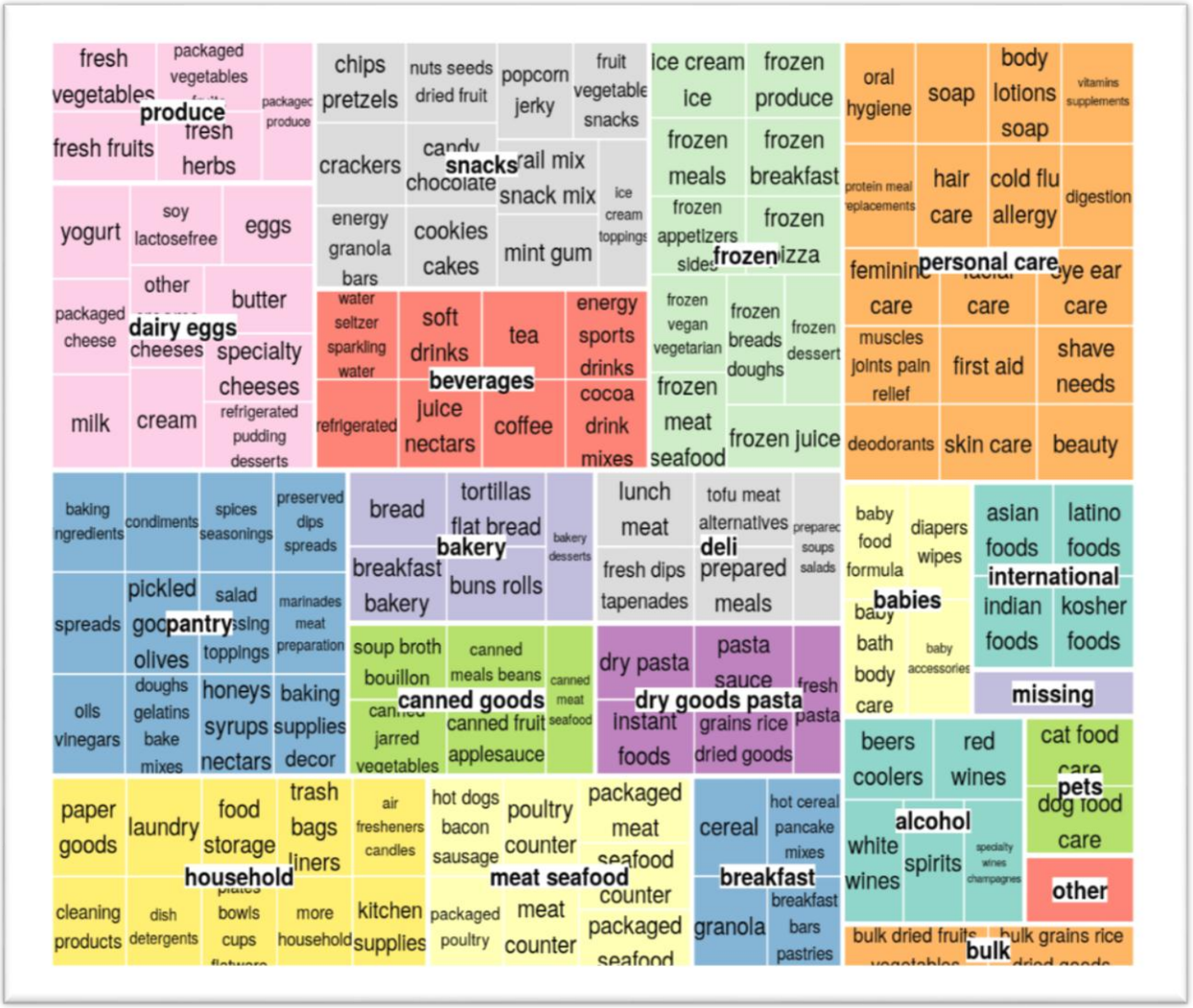

Figure 9 - Aisles within departments

Departments Distribution:

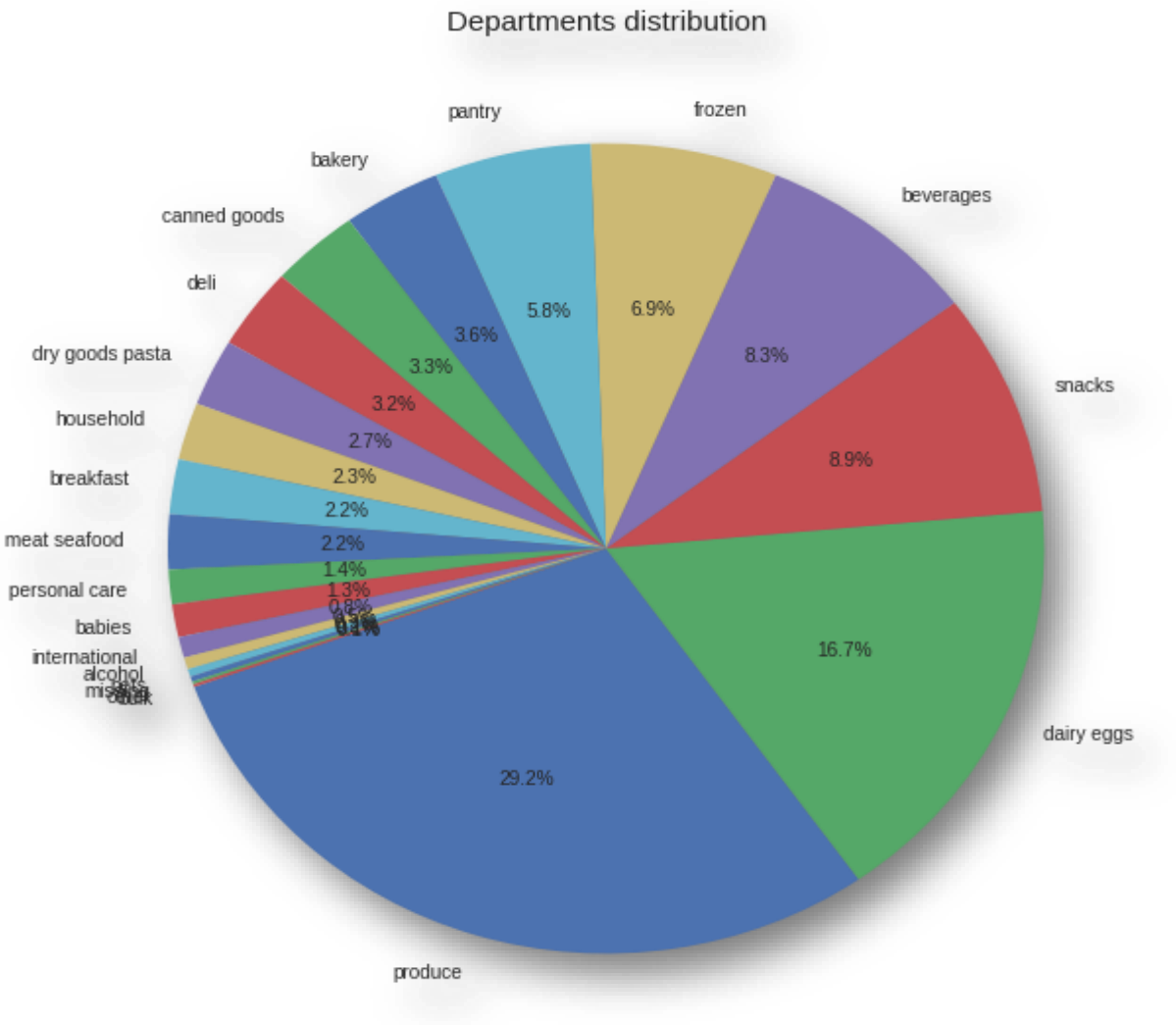

Figure 10 - Departments

The datasets have prior and train data. The prior indicates the previous orders of customers. We will segregate prior, train and test. Now we can combine the prior and train products with the orders.

The Aisle, Department and Products can be inner joined. We can further associate the goods details with our orders based on "product_id".

Learning the patterns on prior and train, we can predict products of the test set.

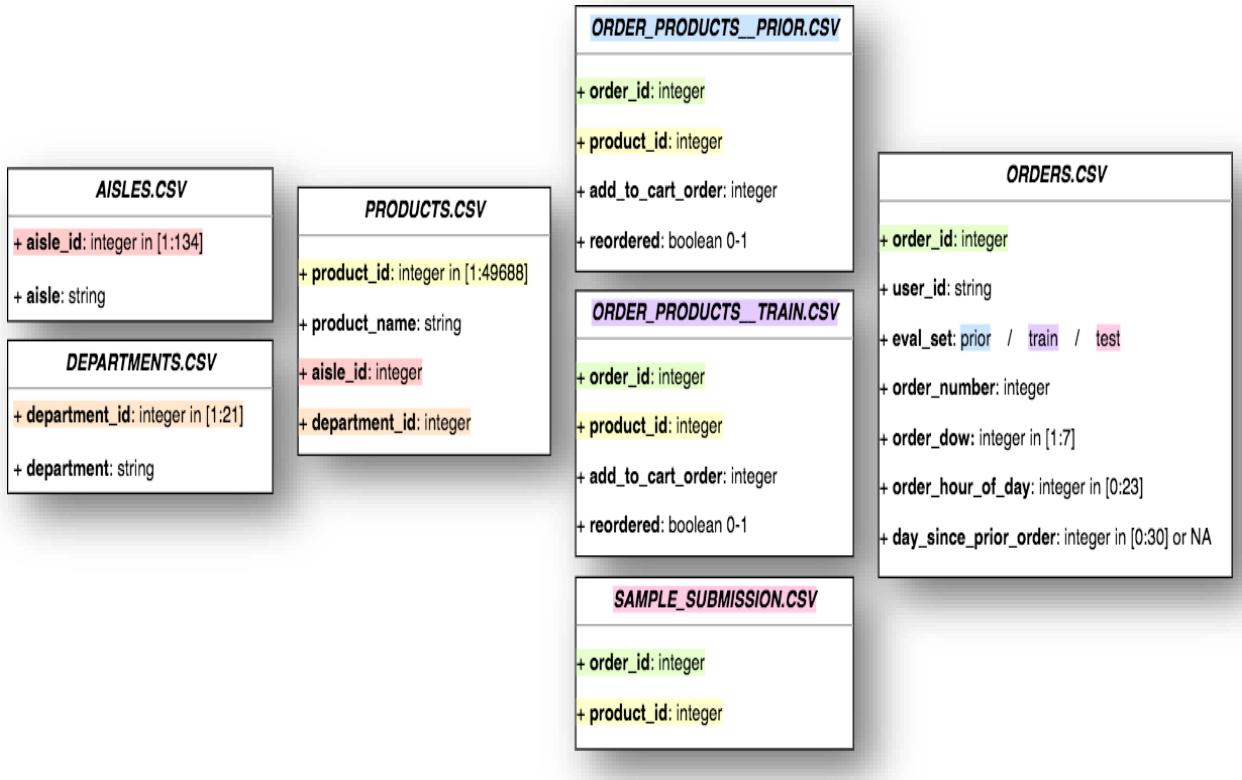

Figure 11 - File relations

7. Market Basket Analysis

7.1 Apriori

Steps:

a. Extract orders (from order_products_prior) corresponding to users (in test orders)
b. Combine products the users ordered
c. Convert this into rules using arules package in R
d. Sort by confidence
e. Now find all rules whose LHS matches the test dataset
f. Subset the suitable rules
g. Now extract the corresponding RHS

7.2 Bayesian Updating of a Prior

In Bayesian Updating of a Prior,
P(R|O) - Probability of reordering a product given a user's order history is proportional to P(O|R) x P(R).

P(R|O) ~ P(O|R) P(R)

Posterior ~ Bayes Factor x Prior

Where P(R) is the probability by product_id that it is reordered over all users.
P(O|R), for each user, is the probability of that a user will order a product given their reordering history.

We will consider the inputs as order_products_prior.csv and orders.csv. We will use order_product_prior, with a groupby 'product_id' to calculate P(R) for each product (our Prior in Bayes Theorem).

Let us now merge the test orders with order_products_prior.csv, for each test user, and produce P(O|R) - The probability that they will reorder any product in their prior orders (our likelihood in Bayes Theorem).

Now the table has these columns:
user_id, product_id, reorder_count, reorder_sum

P(O|R) = (reordered_sum+1)/(reordered_count+2)

The mean of the reordered_count for each user is round down and used for guessing the number of items N that a user will reorder.

The posterior is, for each product listed in the user's order history, the product of P(O|R)1 x P(O|R)2 x... x P(R).

For each test user, select the N most likely items from their posterior as the solution.

8. Conclusion:

I have chosen Bayesian Updating of a prior model because of the accuracy and practical convenience of use over Apriori and the product items for Instacart data have been predicted.

_____THANK YOU_____