**INSTACART MARKET BASKET ANALYSIS**

**WHICH PRODUCTS WILL AN INSTAKART CUSTOMER BUY AGAIN?**

In this competition, Instacart is challenging the Kaggle community to use this anonymized data on customer orders over time to predict which previously purchased products will be in a user’s next order.

The dataset for this competition is a relational set of files describing customers' orders over time. The goal of the competition 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, they provide between 4 and 100 of their orders, with the sequence of products purchased in each order. They also provide the week and hour of day the order was placed, and a relative measure of time between orders

Here

Each entity (customer, product, order, aisle, etc.) has an associated unique id.

It consist of six files

1. Aisles.csv

2. Departments.csv

3. Product.csv

4. Orders.csv

5. Order\_products\_\_train.csv

6. Order\_products\_\_prior.csv

The submission file should consist of order\_id and product\_id.

The orders.csv contain prior, train, test details.

Here I need to predict product id of test data.

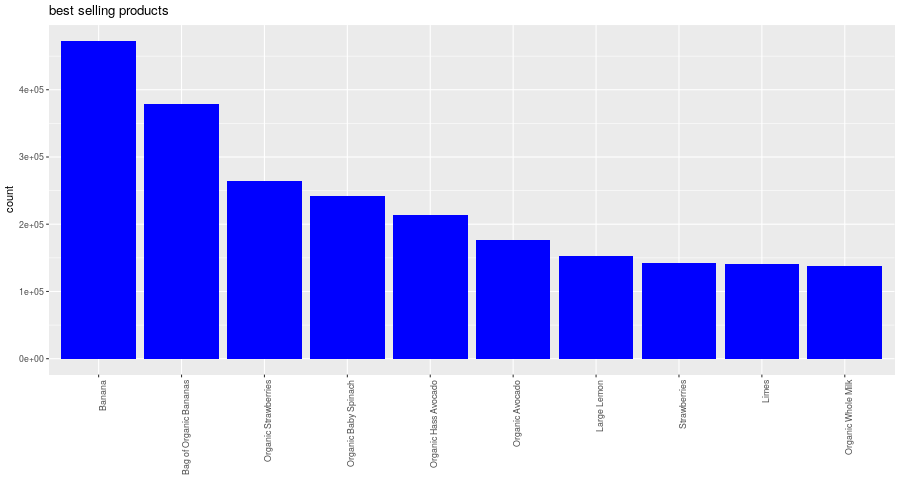
The problem statement is **classification problem.**

This is rules based recommendation problem. That is why I used decision tree algorithm that is **C50** algorithm for **classification**.

I got accuracy **69.2%**

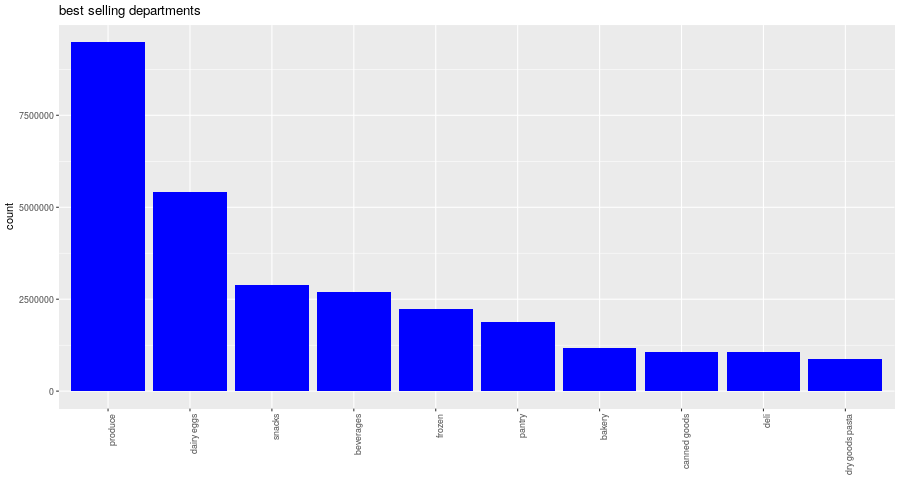
**Visualization of data.**

**Best selling products**

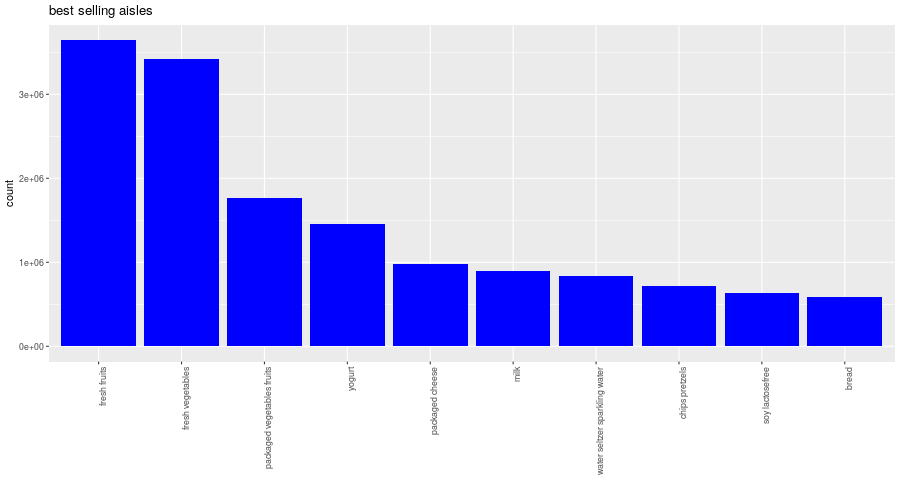


Looks like banana is sold many times.

Best-selling departments



Best selling aisles



Visualizing Unique counts

