Summary\_V1

Cherry

11/17/2019

## Chapter 1 - Introduction

Every time when you shop from meticulously planned grocery lists you will leave the marks of browsing or ordering. Instacart’s 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. But how they can know what kinds of food are exactly what you need?

Instacart released a dataset, “The Instacart Online Grocery Shopping Dataset 2017”, which contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, they provided their orders and the sequence of products in each orders. They also provided the time including the week and the hour of day the order was placed, and the relative time between orders. Using this data to test models, Instacart are enable to predict products that a user will buy again, try for the first time or add to cart next time.

Our project using this datasets which are sourced from Kaggle aims to predict which previously purchased products would be in a consumer’s next order. This report focuses on exploratory data analysis and model prediction of orders and products according to Instacart online transaction, the rest of this report contains five chapters.

## Chapter 2 - Data Descriptioin

### 2.1 Source of the Data

The source dataset of our analysis contains relational set of .csv files which all from Kaggle competition website. The dataset consists of information about 3.4 million grocery orders, distributed across 6 csv files.

### 2.2 Description of dataset

‘orders.csv’ gives a list of all orders and 1 row per order. (Including variables: order\_id, user\_id, eval\_set(prior, train and test), order\_number, oder\_dow(the day of week), order\_hour\_of\_day(the hour of day) and days\_since\_prior\_order.)

‘products.csv’ file contains the names of the products with their corresponding product\_id. (Including variables: product\_id, product\_name, aisle\_id and department\_id.)

‘order\_product\_prior.csv’ contains previous order contents for all customers. (Including variables: order\_id, product\_id, add\_to\_cart\_order(the sequence of products that the customers put in the cart) and reordered(1 represents the customer has a previous order that contains the product, 0 means not contain).)

‘order\_product\_trian.csv’ is structurally similar to ‘order\_product\_prior.csv’, but it is specially used for data training. (Including variables: order\_id, product\_id, add\_to\_cart\_order, reordered.)

‘departments.csv’ file shows the department information about different kinds of products. (Including variables: department\_id and department.)

‘aisles.csv’ is presenting the aisles information about each product. (Including variables: aisle\_id and aisle.)

## Chapter 3 - Exploratory Data Analysis

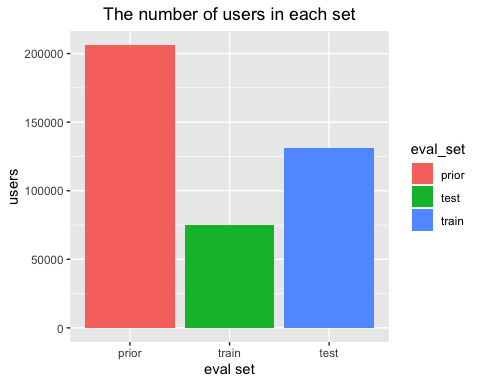
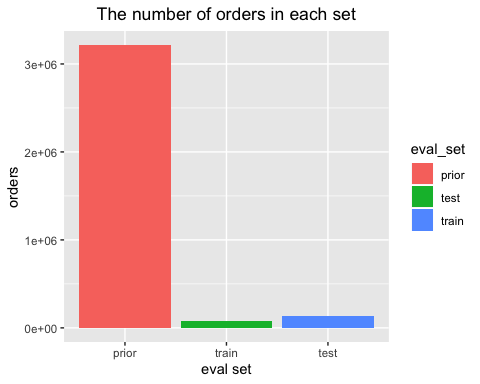
### 3.1 Orders

In the first part, we will explore ‘order.csv’ file. This data set records each order of each customer in detail, including the number of orders and, specific time of each order and the time interval between two orders. In this data set, all the data is divided into three sets, which are prior, train and test.

#### 3.1.1 Basic analysis

We found that there are missing value existed in days\_since\_prior\_order column. The missing value means those are the first order for each user. Here we replace these missing values with zero instead.

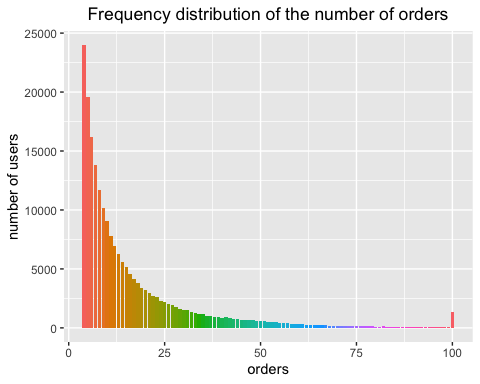
#### 3.1.2 How many orders and products of each set?



3214,784 orders belong to the prior set, and the rest orders which are the last order of each customer are seperated into train set and test set. The train set has 131,209 observations and the test dataset has 75,000 observations.

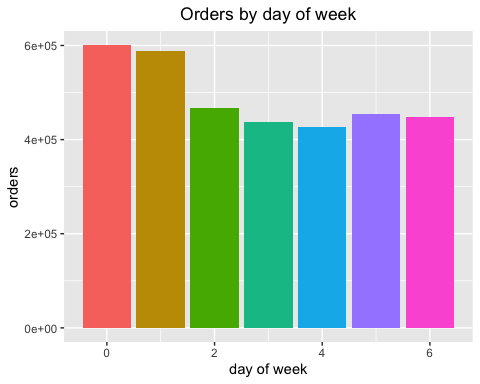
There are 206,209 customers in total. Out of which, the last purchase of 131,209 customers are given as train set and we need to predict for the rest 75,000 customers.

#### 3.1.3 How many orders are customers usually ordering?



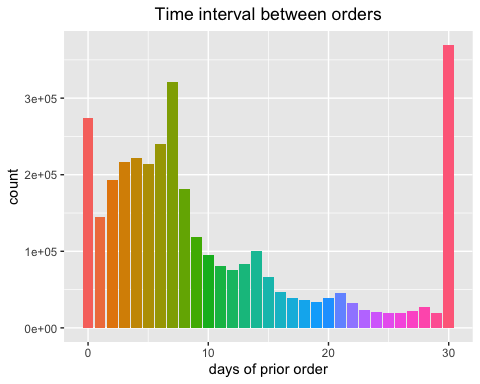
Customers usually have more than 4 orders, and the maximum number of orders are 100.

#### 3.1.4 When do customers usually place orders？



So majority of the orders are made during day time. The 10am hour is the most popular time to make orders, followed by a dip around lunch time and a pickup in the afternoon.Now let us combine the day of week and hour of day to see the distribution. It looks as though 0 represents Saturday and 1 represents Sunday. Wednesday is then the least popular day to make orders.

#### 3.1.5 How often do people order? (time interval between each order）



While the most popular relative time between orders is monthly (30 days), there are “local maxima” at weekly (7 days), biweekly (14 days), triweekly (21 days), and quadriweekly (28 days). Looks like customers order once in every week (check the peak at 7 days) or once in a month (peak at 30 days). We could also see smaller peaks at 14, 21 and 28 days (weekly intervals).

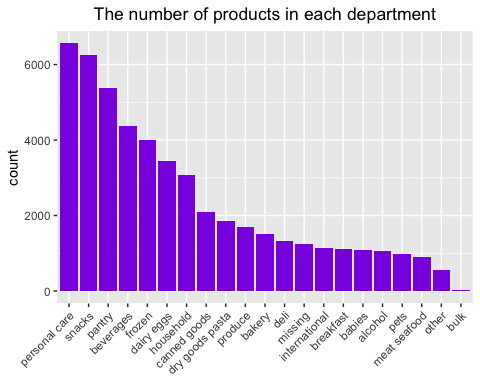
### 3.2 Products & Department and Aisles

In this section we combined ‘products.csv’, ‘aisles.csv’ and ‘departments.csv’ to get a new data frame. Next we will have a general understanding of the storage of these products, we will anlysis how many products in each department and aisle.

#### 3.2.1 Basic analysis

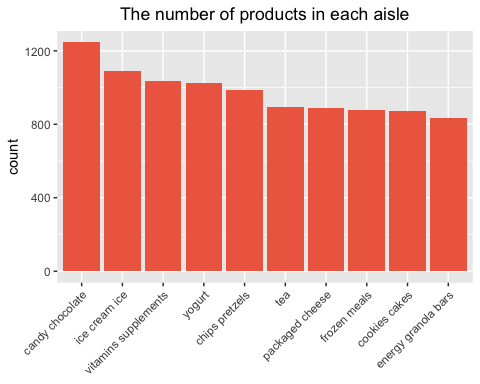
There are 134 aisles, 21 departments in the data and 49688 products in the data set. There are ’r sum\_fianlpro` complete rows in the data, hence no missing observation.

#### 3.2.2 How many products in each department?



The top five departments with the most products are personal care, snacks, pantry, beverages and frozen.

#### 3.2.4 How many products in each aisle?



This plot presents the top 10 aisles with highest number of products.

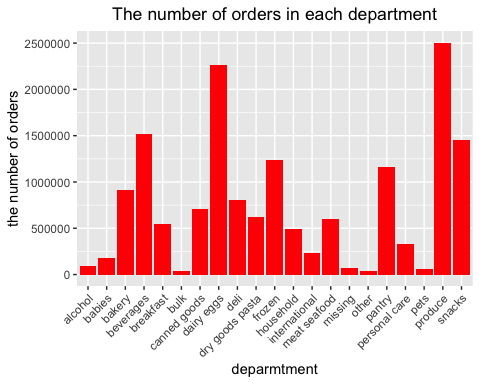
### 3.3 Orders & Department and Aisles

In this section we will explore the sales of different departments and aisles, including order frequency and reordered rate. Then, we are also interested in finding relationship between the sequence of each products putting in the cart by customers and the reordered chance of the product.

#### 3.3.1 Basic analysis

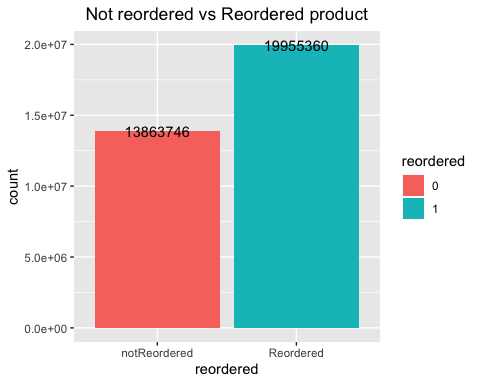
First, we vertically combine ‘order\_product\_prior.csv’ and ‘order\_product\_train.csv’ and create a new data frame called ‘all\_order’. Since there are no key variable between ‘product’ and ‘orders’, we have to merge these two datasets by merging them to ‘all\_order’. Eventually, ‘product\_order’ is the final data frame we created to do further analysis.

#### 3.3.2 What is the best selling department?



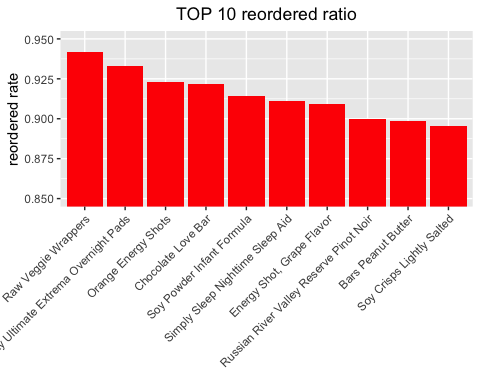
The most three popular departments are produce, dairy eggs and beverages.

#### 3.3.3 How many products are reordered?



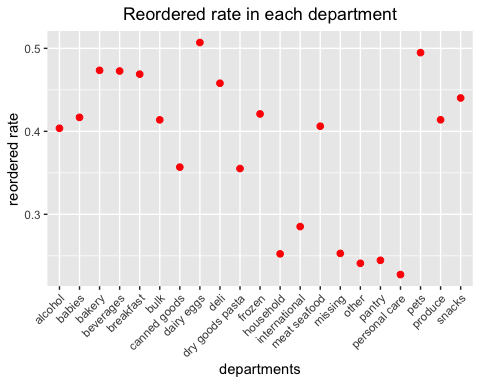
There are 19955360 products are reordered by customers, 13863746 products are never reordered.

#### 3.3.4 What is the most reordered product?



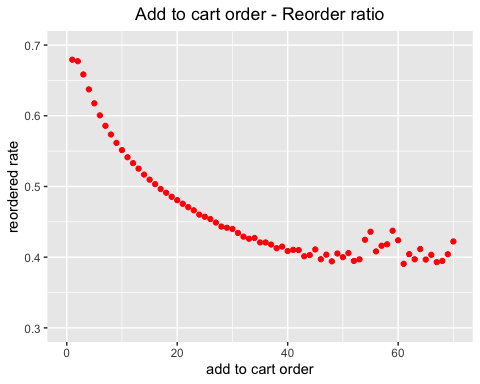
The top three products with highest reordered rate are Raw Veggie Wrappers, Serenity Ultimate Extrema Overnight Pads and Orange Energy Shots.

#### 3.3.5 Which department has the highest reorder ratio?



Personal care has lowest reorder ratio and dairy eggs have highest reorder ratio.

#### 3.3.6 Is there any relationship between the sequence of adding to cart and reordered chance？



##   
## Welch Two Sample t-test  
##   
## data: add\_to\_cart\_order by reordered  
## t = 744.44, df = 25825697, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.922225 1.932374  
## sample estimates:  
## mean in group 0 mean in group 1   
## 9.487599 7.560300

This graph shows that products placed initially in the cart are more likely to be reorderd than one placed later in the cart. We also did t-test to verify whether the sequence of products adding to cart are siginificantly different between reordered products and not reordered products. We can conclude from the results showing the p-value is smaller than 0.05 that the sequence of products adding to cart significantly influence whether the products are reordered.