Capstone Project 2 Milestone Report 1

Predicting the Products an Online Grocery Shopper Will Purchase Again

Yi Li

1. Introduction

Online grocery shopping is growing rapidly these years. According to the U.S. Online Grocery Survey 2020, released on May 6 2020, one half (52.0%) of all respondents had bought groceries online - more than double the shopper numbers from two years ago. The coronavirus pandemic is transforming consumers' needs and behaviors, and has encouraged more grocery shoppers to start buying or buying more online.

There are many grocery delivery apps in the market today such as Instacart, Shipt, Amazon prime now, and Walmart grocery delivery etc. Features that help customers' shopping experience more easy and efficient will make the app stand out from others. Correctly predicting customers' shopping behavior using machine learning, and incorporate it into the features of the apps will make their consumers' shopping experience more pleasant.

In this project, I am going to use a dataset from a Kaggle competition to predict the products a customer will buy again. This dataset is provided by Instacart. This dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, 4 to 100 of their orders are provided, with the sequence of products purchased in each order. The week and hour of day the order was placed are also provide, and a relative measure of time between orders.

Dataset: https: "The Instacart Online Grocery Shopping Dataset 2017", Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on <2020/05>

2. Exploratory Data Analysis

2.1 Basic Structure of the Datasets

The Instacart Online Grocery Shopping Dataset consists of 7 datasets, in which "order_products__train" is held of Kaggle for analyzing the result of competition, I will use the other 6 datasets for this project.

The structure and description of each dataset are as follow:

column	description	dtype
aisle_id	aisle identifier	integer in [1:134]
aisle	the name of the aisle	string

Table 1. Basic structure of aisles.csv

aisle_id	aisle
1	prepared soups salads
2	specialty cheeses
3	energy granola bars
4	instant foods
5	marinades meat preparation

Table 2. Head of aisles.csv

column	decription	dtype
department_id	department identifier	integer in [1:21]
department	the name of the department	string

Table 3. Basic structure of department.csv

department_id	department
1	frozen
2	other
3	bakery
4	produce
5	alcohol

Table 4. Head of department.csv

column	decription	dtype
product_id	product identifier	integer in [1:49688]
product_name	name of the product	string
aisle_id	aisle identifier	integer
department_id	department identifier	integer

Table 5. Basic structure of products.csv

product_id	product_name	aisle_id	department_id
1	Chocolate Sandwich Cookies	61	19
2	All-Seasons Salt	104	13
3	Robust Golden Unsweetened Oolong Tea	94	7
4	Smart Ones Classic Favorites Mini Rigatoni Wit...	38	1
5	Green Chile Anytime Sauce	5	13

Table 6. Head of products.csv

column	decription	dtype
order_id	order identifier	integer in [1: 3421083]
user_id	customer identifier	integer in [1: 206209]
eval_set	which evaluation set this order belongs in	category(prior/train/test)
order_number	the order sequence number for this user (1 = first, n = nth)	integer in [1:100]
order_dow	the day of the week the order was placed on	integer in [1:7]
order_hour_of_day	the hour of the day the order was placed on	integer in [0:23]
days_since_prior	days since the last order, capped at 30 (with NAs for order_number = 1)	float in [0:30] or NA

Table 7. Basic structure of orders.csv

order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
2539329	1	prior	1	2	8	NaN
2398795	1	prior	2	3	7	15.0
473747	1	prior	3	3	12	21.0
2254736	1	prior	4	4	7	29.0
431534	1	prior	5	4	15	28.0

Table 8. Head of orders.csv

column	decription	dtype
order_id	order identifier	integer
product_id	customer identifier	integer
add_to_cart_order	order in which each product was added to cart	integer
reordered	1 if this product has been ordered by this user in the past, 0 otherwise	integer(0/1)

Table 9. Basic structure of orders_products__prior.csv and order_products__train.csv

order_id	product_id	add_to_cart_order	reordered
1	49302	1	1
1	11109	2	1
1	10246	3	0
1	49683	4	0
1	43633	5	1

Table 10. Head of orders_products__prior.csv

order_id	product_id	add_to_cart_order	reordered
1	49302	1	1
1	11109	2	1
1	10246	3	0
1	49683	4	0
1	43633	5	1

Table 11. Head order_products__train.csv

After importing all the dataset as pandas DataFrames, the only DataFrame that has NaN in it is the “order” DataFrame. The “days_since_prior_order” value of each user_id’s first order is NaN. There is no other missing values in the DataFrames.

2.2 Exploratory Data Analysis and Statistical Inference

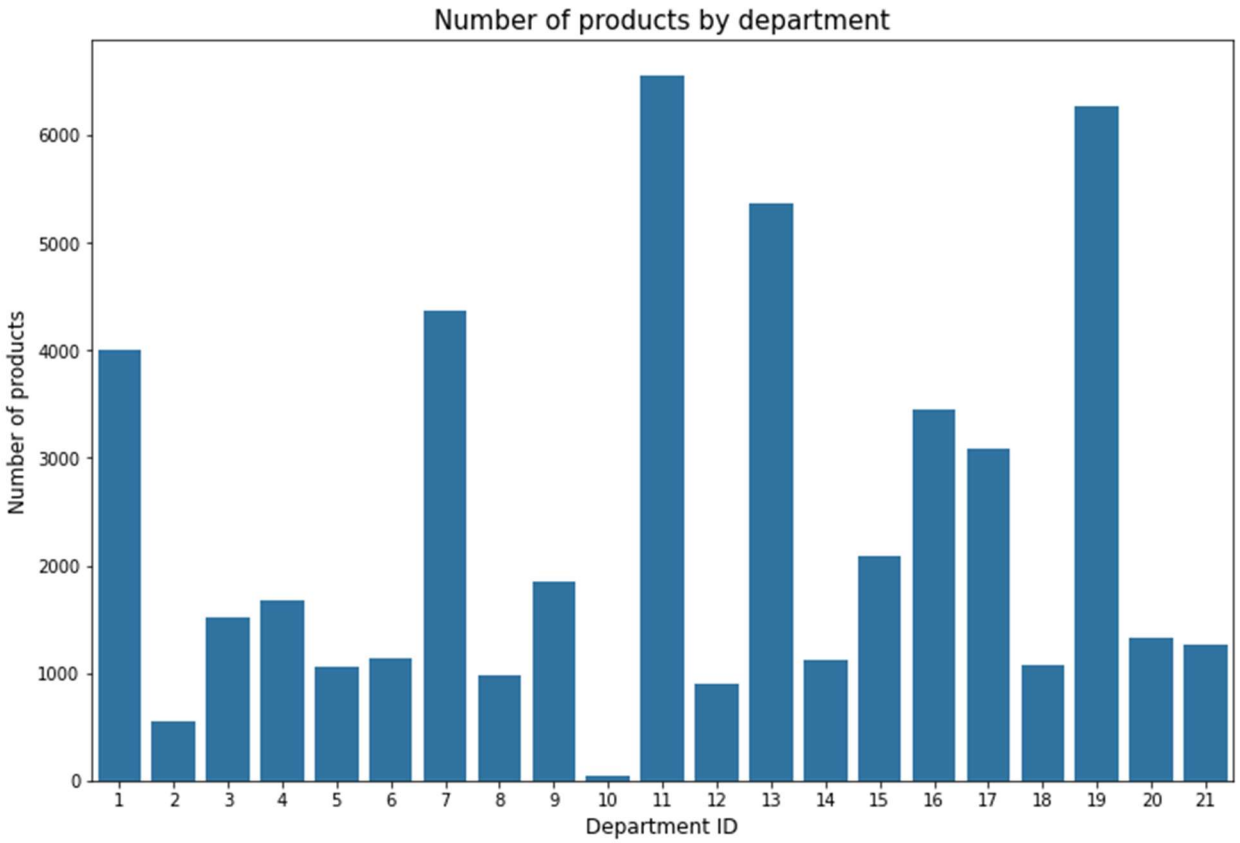

FIGURE 1. Number of products in each department.

By plotting the number of products in each department, we can see that department of personal care and department of snacks have the most types of products, and department of bulk has least types of products.

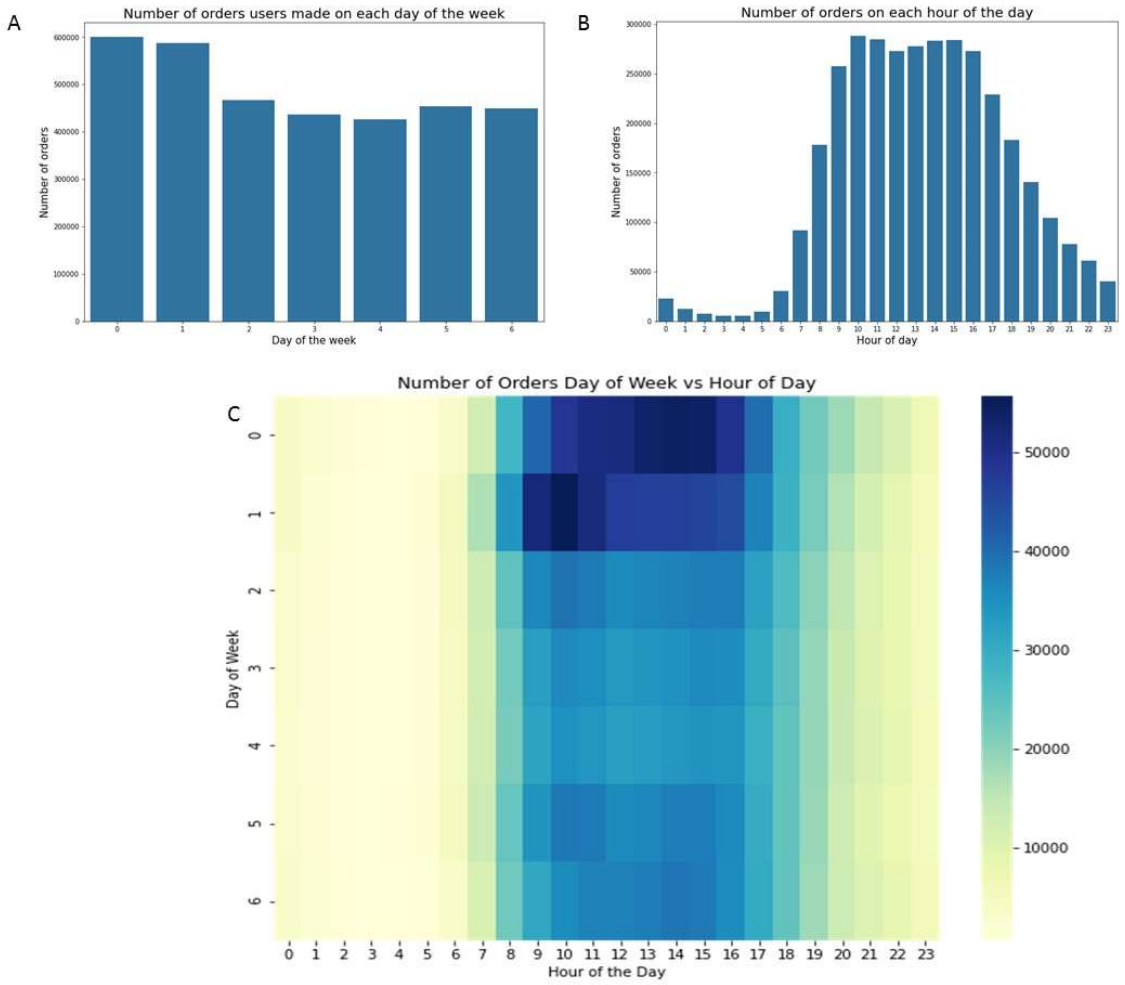

FIGURE 2. Number of orders by time

By plotting the number of orders placed on each day of a week and each hour of a day, we can see that between 9AM and 16PM on Saturday and Sunday is the most popular time to place orders.

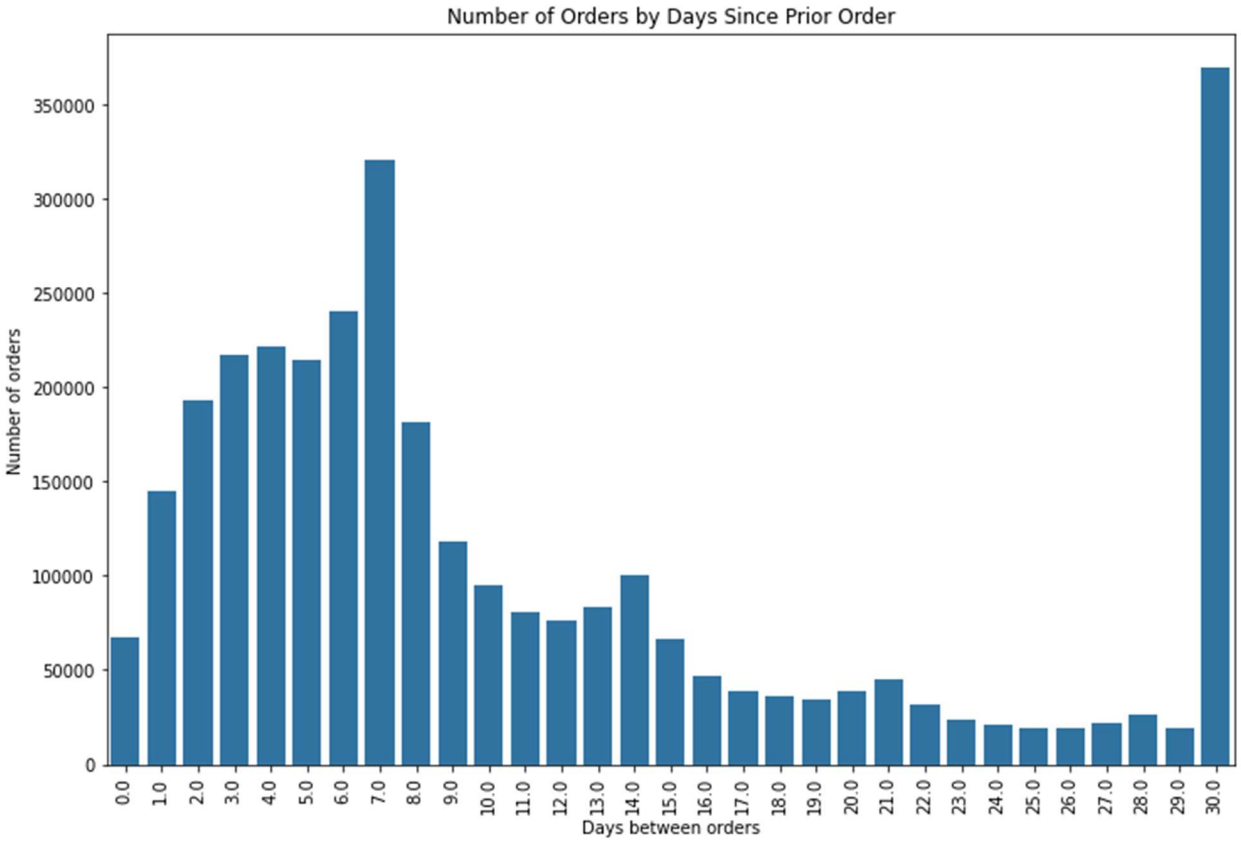

FIGURE 3. Number of orders by days since prior order

By plotting the number of orders by days since prior order, we can see that lots of users order once in every week (local peak at 7 days). We can also see smaller local peaks at 14, 21 and 28 days. There seems to be a cut off value of 30 days for days since prior order.

In prior orders, there are about 59.0% of products are reordered, and about 87.9% of ordered containing reordered products. In the last orders, there are about 59.9% of products are reordered, and about 93.4% of orders containing reordered products.

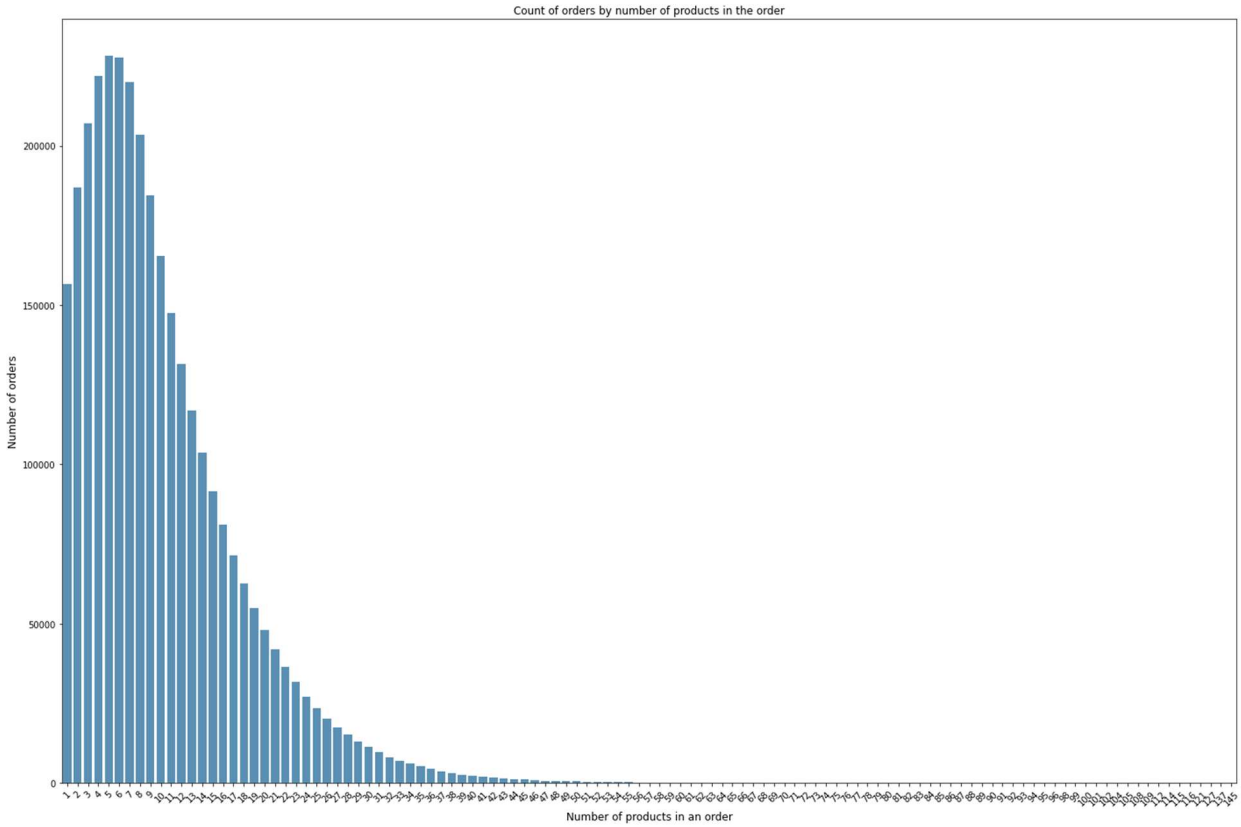

FIGURE 4. Number of orders by number of products in the order

The distribution of number of orders by number of products in the order is right-skewed, very few orders containing more than 50 products, the mode is 5, and the median is 8.

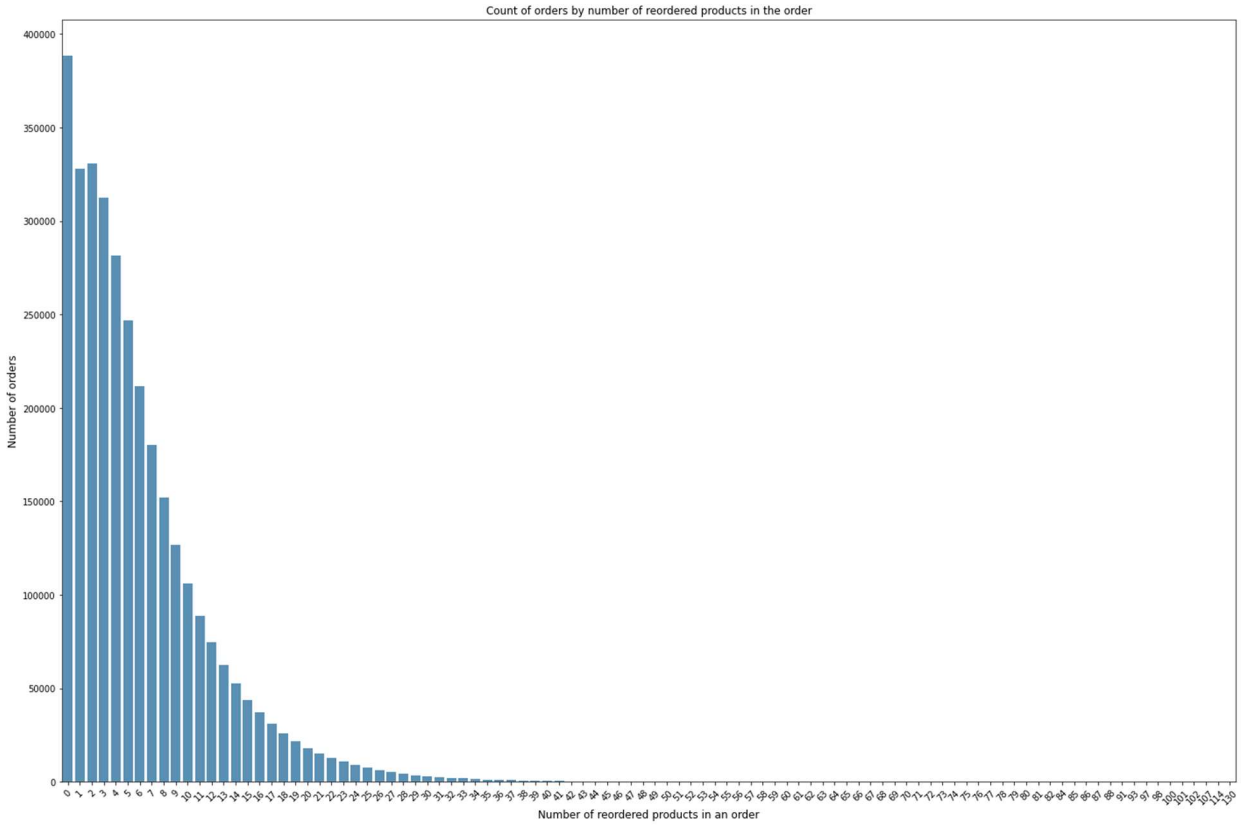

FIGURE 5. Number of orders by number of reordered products in the order

The distribution of number of orders by number of reordered products in the order is also right-skewed, the mode is 0 (no reordered product in the order), and the median is 4.

The orders and order_products_prior DataFrames were merged together into one DataFrame called prior_products, and One-way ANOVA was performed to test whether there is difference in mean number of reordered products in orders placed on different time or day. (H0: There is no difference in mean of number of reordered products in orders placed on different time or day. H1: There is difference in mean.) The p value for both tests is 0.0 (reject H0), this indicated that there is statistical significant difference in mean number of reordered products in orders placed on different time or day.

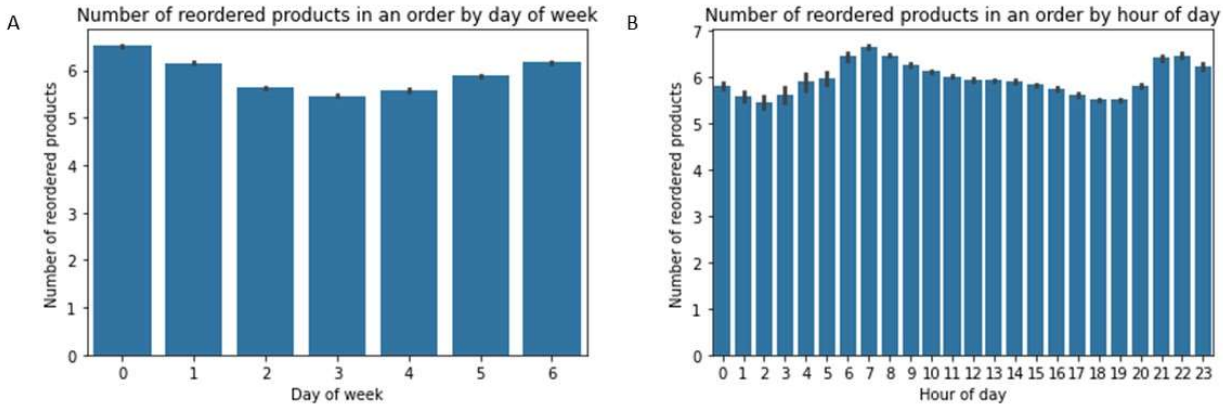

FIGURE 6. Bar plots present the difference in mean of number of reordered products in an order by day of week or hour of day.

REFERENCE

https://www.kaggle.com/sudalairajkumar/simple-exploration-notebook-instacart