Capstone Project 1 – Milestone Report

Topic: Instacart Market Basket Analysis

Background: Instacart is an online grocery application. It allows customers to order their groceries online and have someone buy and deliver their groceries in front of their doorsteps.

Problem: With a ton of customers' data available, Instacart is looking to predict customers' buying behavior based on each customer's order history, which includes: items previously purchased, time and day of purchase in order to improve its recommendation system to customers and increase its revenue per customer by retaining their loyalty.

Dataset

There are 5 datasets available here:

1. Aisle: Includes the aisle id and the aisle name

	aisle_id	aisle
0	1	prepared soups salads
1	2	specialty cheeses
2	3	energy granola bars
3	4	instant foods
4	5	marinades meat preparation

2. Departments: Includes the department id and the department name

	department_id	department
0	1	frozen
1	2	other
2	3	bakery
3	4	produce
4	5	alcohol

3. Order_products: Includes order id, product id, add to cart order, and if the customer reorders the product or not.

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

4. Orders: divides the data into three sections: prior, train and test data. All have the same columns, which include: order id, user id, evaluation set(prior,train, or test), order number, order day of week, order hour of day, and days since prior order.

	order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
0	2539329	1	prior	1	2	8	NaN
1	2398795	1	prior	2	3	7	15.0
2	473747	1	prior	3	3	12	21.0
3	2254736	1	prior	4	4	7	29.0
4	431534	1	prior	5	4	15	28.0

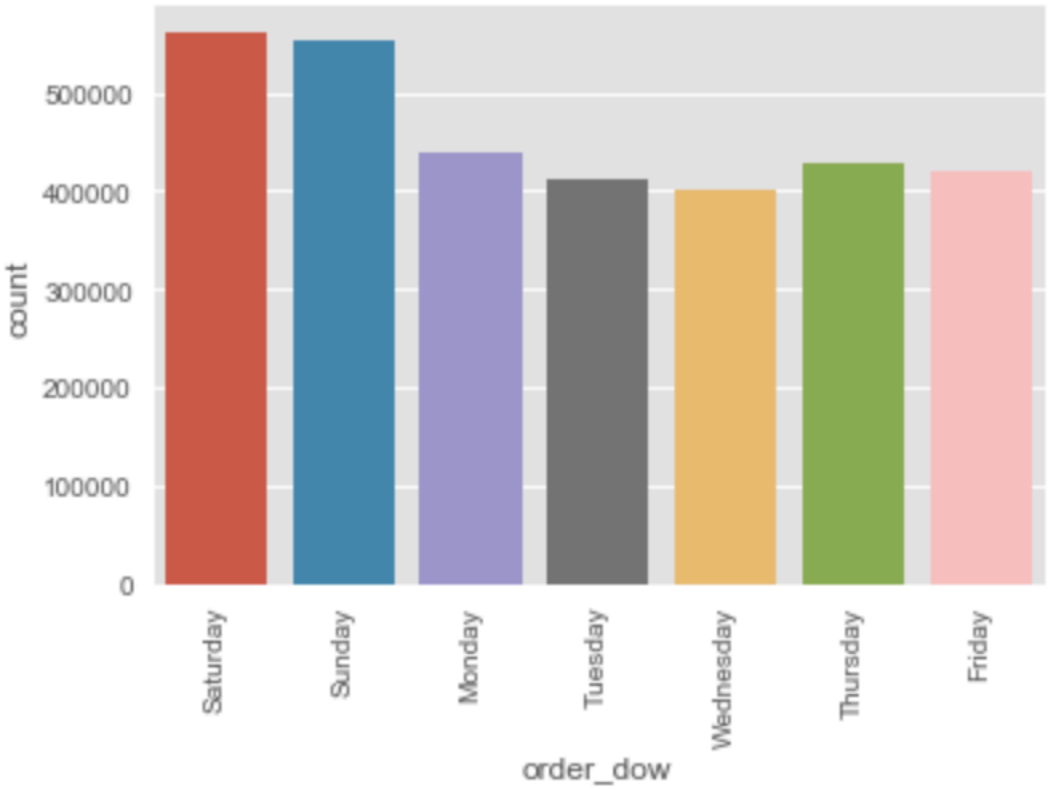
5. Products: Includes the product id, product name, aisle id, and department id

	product_id	product_name	aisle_id	department_id
0	1	Chocolate Sandwich Cookies	61	19
1	2	All-Seasons Salt	104	13
2	3	Robust Golden Unsweetened Oolong Tea	94	7
3	4	Smart Ones Classic Favorites Mini Rigatoni Wit...	38	1
4	5	Green Chile Anytime Sauce	5	13

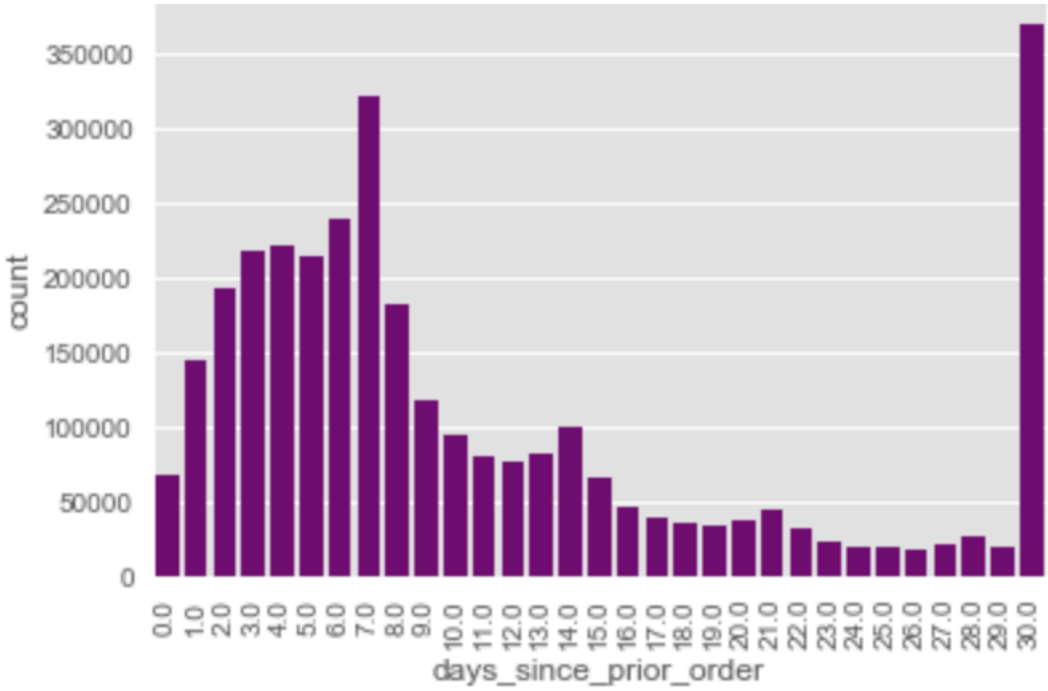
Data Wrangling Techniques

Since the data is taken from Kaggle, we can see that the data is pretty clean and there is not a lot of data cleaning that has to be done. Missing values are present only in the orders dataset and comprises only 6% of the total data in the orders data, allowing us to exclude them. Additionally, the "NaN" values here mean that the customers have no previous order history.

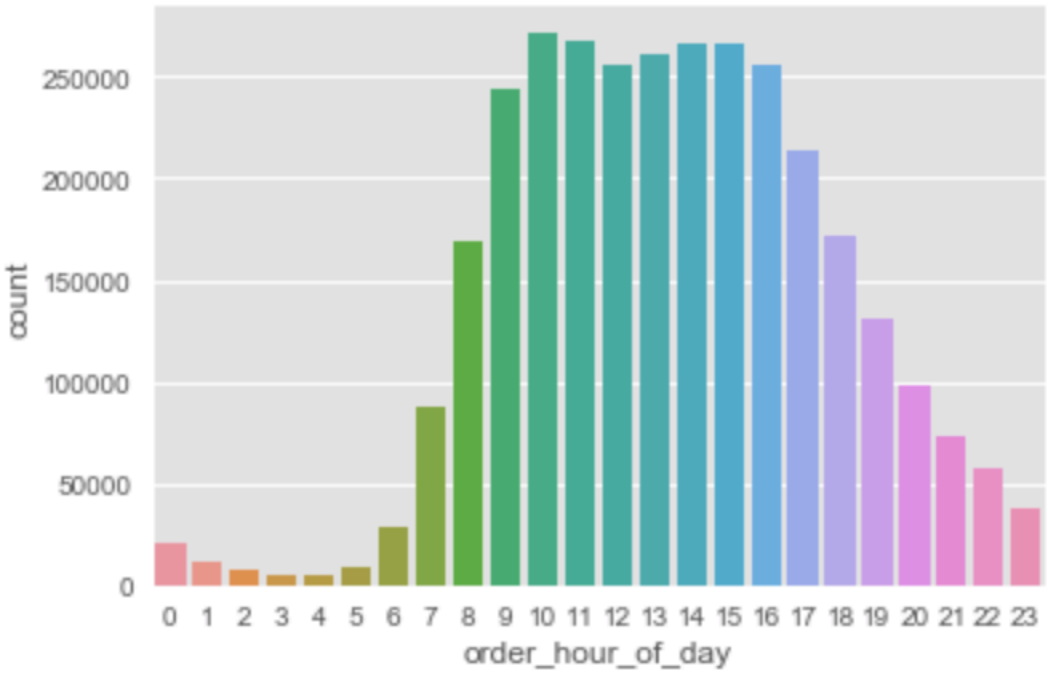
Data Exploration



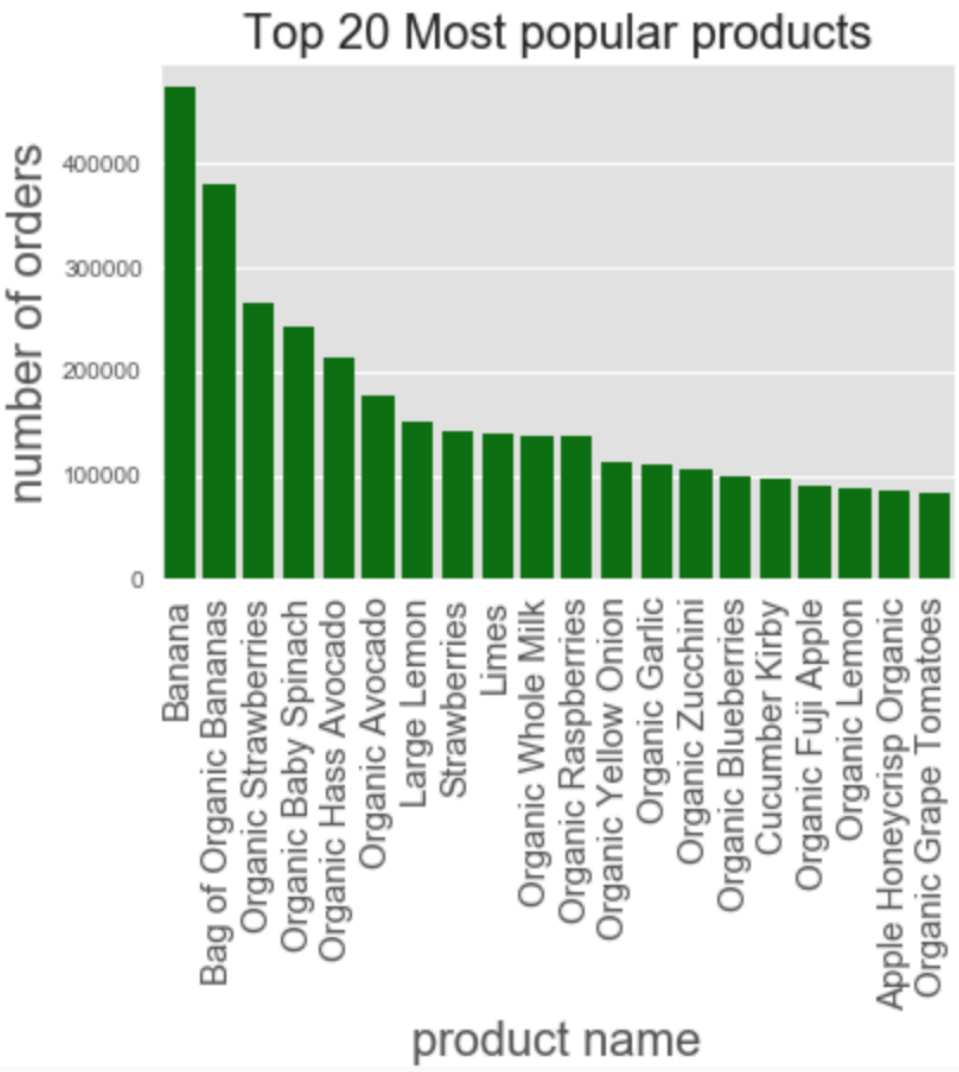
We can see that a lot of orders from customers come in during the weekend: Saturdays and Sundays.



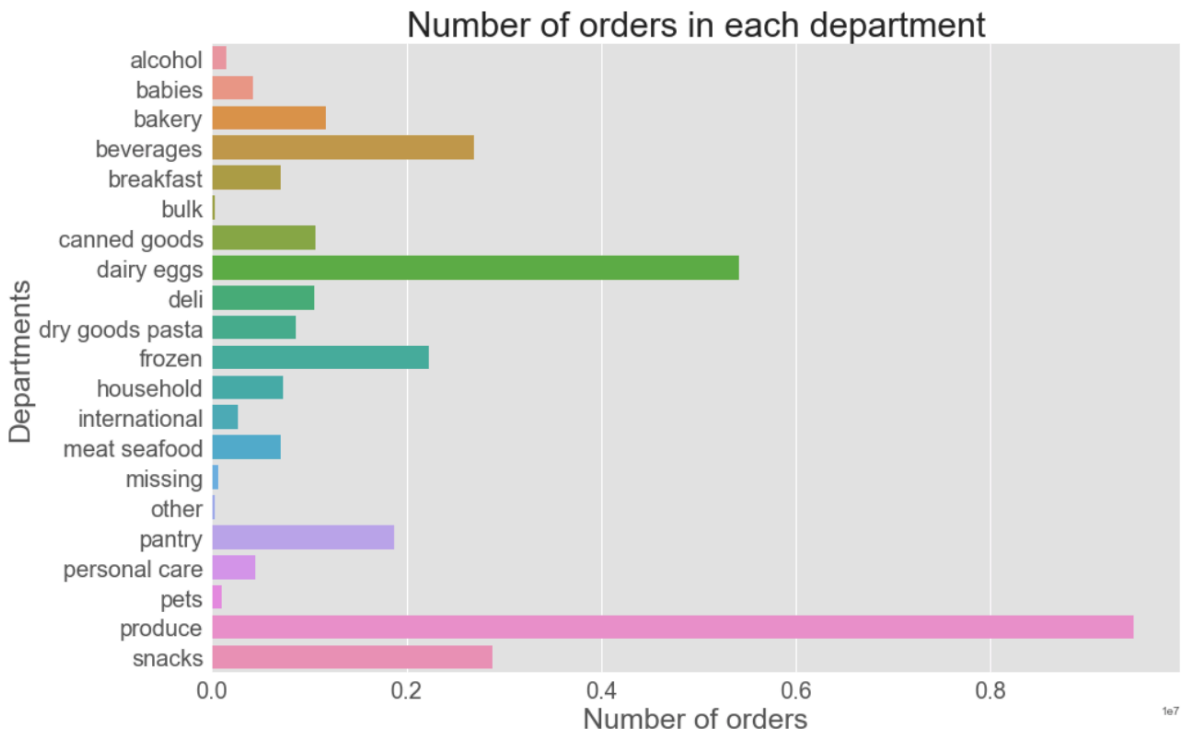
To figure out how long it takes one to re-order on average, we can see that most people reorder every 7 days or every 30 days. At first, when I was using boxplot to figure out if there is any outliers, people who order every 30 days would simply be excluded as they would be consider outliers. However, plotting the information on a bar graph actually shows that we could not simply exclude them.



Looking at the time of day that one usually purchases his or her groceries, we can see that most people order at the time range: 9am – 5pm, which is during working hours. This information can be very useful because Instacart could target its customer during those hours or remind the regular customers to reorder their groceries when they have not done so.



The top 20 most popular groceries products mostly comprise of fruits, such as bananas, strawberries, etc. By knowing this information, Instacart could increase its revenue by offering promotional discounts on other items not on the list or even do bundle the least popular products with any of the top-selling items.



The information above tells us that the least popular departments are: pets, international, and bulk. On the other hand, the most popular departments are produce, dairy eggs, and beverages. By knowing this, I was thinking that organizing the list of recommended items to include some of the least popular product could increase sales on the least popular products. By doing so, Instacart is able to increase its revenue per customer.