****Objective:****

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, 4 and 100 of their orders are given, with the sequence of products purchased in each order

Let us start by importing the necessary modules.

As we could see, orders.csv has all the information about the given order id like the user who has purchased the order, when was it purchased, days since prior order and so on.

The columns present in order\_products\_train and order\_products\_prior are same. Then what is the difference between these files.?

As mentioned earlier, in this dataset, 4 to 100 orders of a customer are given (we will look at this later) and we need to predict the products that will be re-ordered. So the last order of the user has been taken out and divided into train and test sets. All the prior order informations of the customer are present in order\_products\_prior file. We can also note that there is a column in orders.csv file called eval\_set which tells us as to which of the three datasets (prior, train or test) the given row goes to.

Order\_products\*csv file has more detailed information about the products that been bought in the given order along with the re-ordered status.

Let us first get the count of rows in each of the three sets.

So 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.

Now let us validate the claim that 4 to 100 orders of a customer are given.

So there are no orders less than 4 and is max capped at 100 as given in the data page.

Now let us see how the ordering habit changes with day of week.

Seems like 0 and 1 is Saturday and Sunday when the orders are high and low during Wednesday.

Now we shall see how the distribution is with respect to time of the day.

So majority of the orders are made during day time. Now let us combine the day of week and hour of day to see the distribution.

Seems Satuday evenings and Sunday mornings are the prime time for orders.

Now let us check the time interval between the orders.

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).

Since our objective is to figure out the re-orders, let us check out the re-order percentage in prior set and train set.

On an average, about 59% of the products in an order are re-ordered products.

****No re-ordered products:****

Now that we have seen 59% of the products are re-ordered, there will also be situations when none of the products are re-ordered. Let us check that now.

About 12% of the orders in prior set has no re-ordered items while in the train set it is 6.5%.

Now let us see the number of products bought in each order.

A right tailed distribution with the maximum value at 5.!

Before we explore the product details, let us look at the other three files as well.

Now let us merge these product details with the order\_prior details.

Wow. Most of them are organic products.! Also majority of them are fruits.

Now let us look at the important aisles.

The top two aisles are fresh fruits and fresh vegetables.!

****Department Distribution:****

Let us now check the department wise distribution

Produce is the largest department. Now let us check the reordered percentage of each department.

****Department wise reorder ratio:****

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

****Add to Cart - Reorder ratio:****

Let us now explore the relationship between how order of adding the product to the cart affects the reorder ratio.

****Looks like the products that are added to the cart initially are more likely to be reordered again compared to the ones added later.**** This makes sense to me as well since we tend to first order all the products we used to buy frequently and then look out for the new products available.

****Reorder ratio by Time based variables:****

Looks like reorder ratios are quite high during the early mornings compared to later half of the day.