Mentor: Srdjan Santic *By Anushree Srinivas*

Capstone Project: Instacart Market Basket Analysis

Inferential Statistics

For inferential statistics, I performed four different normality tests for each of the following 5 variables: order_dow, order_hour_of_day, days_since_prior_order,order_number,reordered. The normality tests that I conducted were Shapiro-Wilk test (**stats.shapiro**) (which is not accurate in this case since our sample size >5000), Anderson-Darling (**stats.anderson**), Kolmogorov-Smirnov test (**stats.kstest**) and finally the D'Agostino and Pearson (**stats.normaltest**) test.

To further check if any of the variables are correlated, I tried a correlation plot for the same five variables. I used **sns.heatmap()** and the **corr()** functions to get my corresponding correlation plot.

In the orders correlation plot, using the **orders.corr()** function and get the correlation plot for the orders data set. In the plot there's a small negative correlation (-0.36) between order_number and days_since_prior_order.

Moving on to the merged data set using the **op_prior_merged.corr()** function of order_products_prior, we can see a slight negative correlation(-0.13) reordered and add_to_cart_order. There's also a minor positive correlation (0.062) correlation between department_id and aisle_id.

Then I take a look at the merged_reorder dataset(**merged_reorders.corr()**). There are several small correlations here. In addition to the three correlations I've already mentioned;(1. between order_number and days_since_prior_order, 2. between reordered and add_to_cart_order), there are three other correlations that come to light here. There's a good positive correlation of 0.31 between order_number and reordered. Then we can see a slight positive correlation between add_to_cart_order and days_since_prior_order of 0.054. From this plot we can see that the correlation between reordered & days_since_prior_order and add_to_cart and reordered is the same (0.13).