Data Wrangling

Collecting data for this project was straight forward and was deflated from a tar.gz file format. Patoolib is a library that makes extracting data from an archive simple. It consists of seven methods to be used and is easily controlled through the os library from python. To further the ease of the extraction, I created two functions to minimize semi-repeated cells for creating dataframes with pandas. I considered using dynamically named variables in the dataframe creation but exploring the options it was not recommended. I initially decided to create functions to check for missing values but decided to have a more focused control on each dataframe. However, to check for missing data, one last for loop was used in concert with all of the files. The orders dataframe consisted of 206209 null values in the days since prior column. This can be deduced as a customer's initial purchase therefore, the rows will not be eliminated from the dataframe. Also, some values identifying the aisle and department in which the product came from contains about 1200 missing values. These missing values will not have a significant effect in the final goal however results from the frequency in which a product is sold is missing a considerable amount of information. It does not exceed 20% of the values so it can still be used in analysis.

The first concept I thought of to explore within this data was the items that are offered by instacart. In order to explore this efficiently, products, departments and aisles were merged together to contain a single accessible dataframe. During the creation of the dataframe for departments it threw a Future Warning Error. After further investigation it was caused by the disagreement of Numpy ndarrays and Python string comparisons. If any comparisons are required in the future, warnings will be imported to ignore these warnings..

Overall, the types in each column are consistent and do not require any conversions. There exists little extraneous data and a greater knowledge of domain can be collected during EDA. Some questions to be considered during EDA that may require additional cleaning includes the central tendencies for the amount of products, times and reorders, product complements, sequencing of products ordered in one purchase, customer habits and associations. Perhaps these associations can include probabilities of events. With the visualizations, outliers, such as infrequent large purchases, and top selected products can be identified and examined to minimize it's effects. The last consideration of wrangling the data involves increasing the amount of previous orders included in the training data set. Next step is EDA, where I plan to continue to munge data based on increasing discoveries and inferences where I believe more questions and answers will appear.