Data Wrangling

The Olist Data consists of 9 separate datasets:

1. Customer Dataset
2. Orders Dataset
3. Geolocation Dataset
4. Order Payment Dataset
5. Order Reviews Dataset
6. Order Item Dataset
7. Products Dataset
8. Sellers Dataset
9. Product Category Name Translation Dataset

In the customer dataset, there were 96096 unique customers (Customer\_Unique\_Id) with 3345 customers having repeat purchases. As there is a customer id created for each order placed, there are 99441 customer ids. There was no missing value in this data set.

In the orders dataset, we can link with customer dataset using the Customer\_Id as key. As there are different order status like Delivered, Cancelled, Shipped, Unavailable etc.. only Delivered has been considered as analysis is going to be done on orders that have been delivered(97% of the data). That reduced the data points from 99441 to 96478.

There were further 65 missing values in various rows which were removed from orders dataset to bring final number to 96455.

There were no missing values in Order Payments Data and also Order Items data. But in the Order Reviews data, there were many data points missing but that was because most customers had not given a review. As the number of missing values is large, all that can be done is ignore the reviews columns and only consider the product rating column.

Products data consists of weight and dimensions of product as well a product description in Portuguese that be converted to English by joining the data set with the poduct category English translation dataset. Initial data wrangling does not show any alarming outliers.