Exploratory Data Analysis

The initial Average Order Value (AOV) that was provided in the document was achieved by using the standard mean formula, (Summation of order_amount) / # of orders. This is a quick way to achieve an average value but it does not take into account outliers and anomalies that could be inside the dataset.

During my first look of the dataset, I visualized the total_items column using a boxplot which showed me that there were a few orders skewing our data which can be seen in Figure 1.

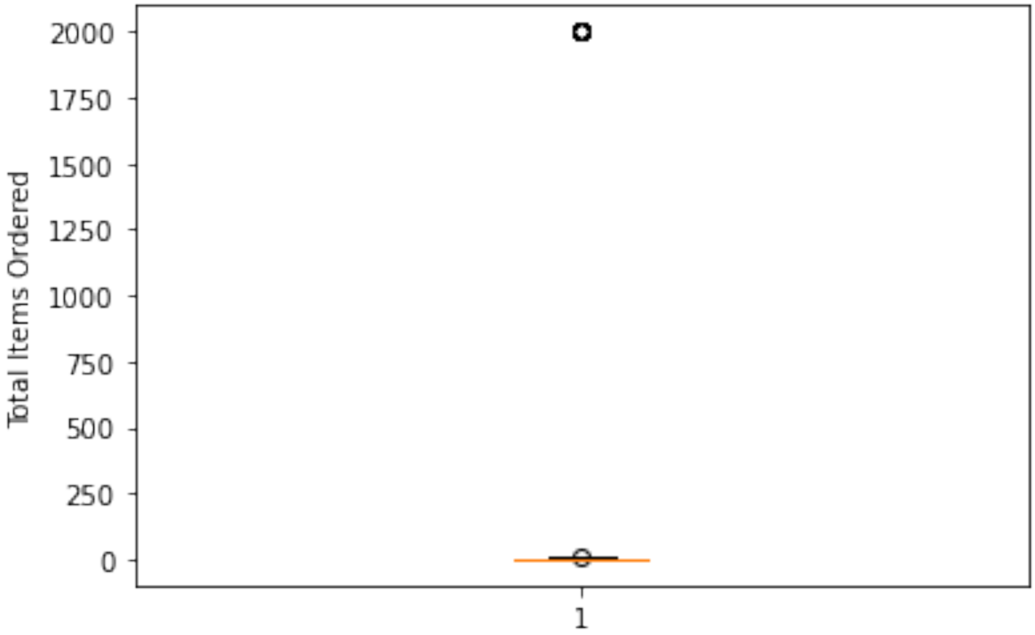

Figure 1.

We can see that there are extreme outliers in this dataset in terms of total_items, in total there were 17 orders with a total_items count of 2000 which were also all made by the same at the same shop. With orders like this at a shop, we might suspect that they could have been fraudulent, accidental, or potentially bulk orders from that shop. From my research, it wouldn't be justifiable to omit these orders due to the reasoning of fraudulent or accidental because the orders were placed at different dates throughout the month of 2017-03. Without a greater understanding of the customer and shops relationship it would be difficult to reason why such orders were made because there should be some degree suspicion around these orders.

After attempting to get a new AOV by using the average cost per item in each shop, we can see that there is another outlier in this area of data as well which can be seen in Figure 2. A shop has a per item cost of $25,725 whilst the average item cost for all shops is only $387.86.

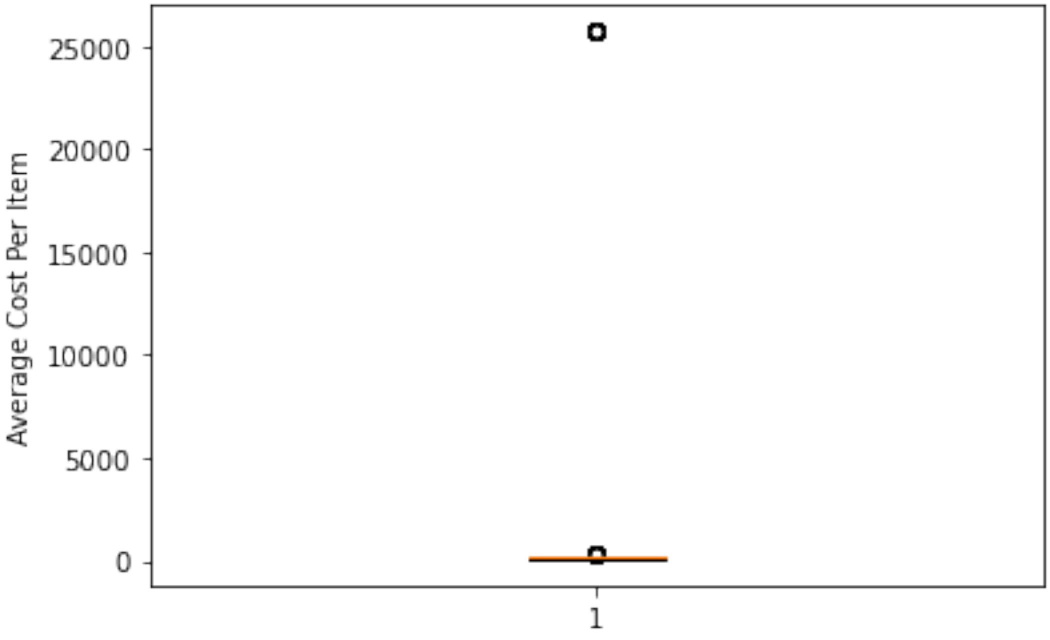

Figure 2.

Calculating a New Average Order Value

A naive new AOV that we can calculate is by removing orders which have a 'total_items' equal to 2000, this results in a new AOV of $754.09 per order which could be a more reasonable average.

But if we didn’t remove anything from the dataset, a new metric that we could use is the value that each **user** is spending at a **shop.** First we group the data by user_id and shop_id, then we find the mean for each of the **user_id - shop_id** relationships. The new Average Order Value that results from using the mean formula is **$937.19** which seems much more reasonable of an AOV, there are some pretty extreme outliers in the data, i.e. $25,725 cost per item in a store and a user purchasing 2000 items at a time.