I used four python libraries for my analysis. They include pandas, numpy, matplotlib and seaborn.

Data wrangling process

Order data set:

1. I dropped the eval\_set column as it was irrelevant to my analysis.
2. Renamed column order\_dow to order\_day\_of\_week.
3. Changed data type of the following columns: order\_id, user\_id, order\_number, order\_day\_of\_week, order\_hour\_of\_day and days\_since\_prior\_order to reduce memory usage.

Products data set:

1. Changed data type of the following columns: product\_id, aisle\_id, department\_id, prices, add\_to\_cart\_order ansd average\_price to reduce memory usage

Products Prior data set:

1. Changed data type of the following columns: order\_id, product\_id, add\_to\_cart\_order and reordered to reduce memory usage.

Customers data set:

1. Renamed columns Surnam to Surname, fam\_status to family\_status, n\_dependants to no\_of\_dependants for easy comprehension.
2. Dropped columns First Name and Surname due to data privacy concerns.
3. Changed data type of the following columns user\_id, Age, no\_of\_dependants and income.

Kindly refer to sheet ‘Wrangling steps’ of the Excel report for more details.