DATA QUALITY ASSESSMENT

1. CUSTOMER ADDRESS DATASET :

a. Columns before cleaning: ‘customer_id’, address’, 'postcode', 'state', 'country', ‘property_valuation'

b. There were no null values in any columns.

c. As “customer_id” column was just a serial number given to rows , it was unwanted for any insights , visuals or model building . Hence , I dropped the particular column.

d. ‘property_valuation’ column had rating ranging from 1-12 , I sorted the column in ascending order.

e. ‘state’ column had three main states namely - New South Wales', 'Victoria', ‘Queensland’. But it had short forms of the same like -'NSW', 'VIC', ‘QLD as new values . By checking unique values, I replaced the short form values with the existing state names and hence cleansed the state column in the dataset.

f. Columns after cleaning : 'address', 'postcode', 'state', 'country', 'property_valuation'

2. CUSTOMER DEMOGRAPHIC DATASET:

a. Columns before cleaning: ‘customer_id’,’first_name', 'last_name', 'gender', 'past_3_years_bike_related_purchases', 'DOB', 'job_title', 'job_industry_category', 'wealth_segment', 'deceased_indicator', 'owns_car', ‘tenure’,’default’.

b. Dropped unwanted serial numbering column which was - customer_id. Along with this , also drooped “default” which had unknown and irrelevant values .

c. Null values in last_name column was filled with 0 as few people do not have last names.
While , other columns like - DOB , job title , job industry category and tenure had many missing values.

d. After sorting DOB column in ascending order , I noticed that there are few wrong DOB entries like 1843-12-21 , where the person’s age should be 178 years which is not possible.

e. Gender column had wrong spelling entries like ‘Femal’ and usage of short forms like U , F and M. Hence I replaced these values with Female , Male and Undefined accordingly.

3. NEW CUSTOMER LISTS DATASET :

a. Null values in last_name column was filled with 0 as few people do not have last names.
While , other columns like - DOB , job title ,and job industry category had many missing values.

b. 21st column was a duplicate of rank column . Hence , I dropped the column.

c. Also , unknown column names with vague irrelevant values from 17th to 20th column .Hence , I dropped the column.

4. TRANSACTIONS :

a. Dropped unwanted serial numbering column which was - “transaction_id”.

b. Missing values in online_order , brand, product_line, product_class , product_size ,standard_cost , product_first_sold_date .

c. Sorted the transaction column in ascending order.