

Data Quality Assessment

Sprocket Central Pty Ltd

Drafted by Muhammad Farmal Khan

Transactions Data Assessment

Issue 1: Missing values

Columns	Number of Missing values
Online_order	360
Brand	197
Product_line	197
Product_class	197
Product_size	197
Standard_cost	197
Product_first_date_sold	197

Solution:

The 197 missing records have no values in the columns: product_line, product_class, product_size, standard_cost and product_first_sold_date with the product_id = 0. These records can be deemed and irrelevant for data analysis and can be removed.

The remaining missing values are in the online_order column. These can be assigned filled and assigned by a new category: N/A.

Issue 2: Incorrect data format Product_first_sold_date column

Solution: This column must be in the date format of DD/MM/YYYY similar to transaction date. To change the format to date the excel option of format can be changed from number to short date.

New Customer List Data Assessment

Issue 1: Missing values

Columns	Number of Missing values
Last_name	29
DOB	17
Job_title	106
Job_industry_category	165

Solution: The missing values can be filled with N/A.

Issue 2: Property_valuation column

This column contains 2 records with 7.00 which is in decimal format while the other records are in whole number format. These 2 records should also be converted to whole numbers i.e (property_value = 7).

Customer Demographics Assessment

Issue 1: Missing Values

Columns	Number of Missing values
Last_name	125
DOB	87
Job_title	506
default	302
Tenure	87

Solution: The last_name, DOB, and job_title can be filled with N/A. For the tenure column the missing values can be filled by the median or mean of the values in that column.

Issue 2: Incorrect values in DOB column

Solution: The date of birth values indicate 1 customer is more than 100 years old. This is obviously an error and to rectify it a new column 'Age' can be created through calculation and remove the DOB column. If an outlier exists in Age column, replace it with median value of the column.

Issue 3: default column is irrelevant

Solution: The default column contains gibberish values that are hard to read and understand and will provide no useful information in any sort of analysis. This entire column can be removed.

Issue 4: Inconsistent data format in gender column

Solution: It is best to follow in data format for the gender column with either:

(M and F) or (Male and Female)

One record also contains 'Femal' which should be corrected to F or Female depending upon the data format chosen.

Customer Address Assessment

Issue 1: Inconsistent data format in state column

Solution: The records containing 'New South Wales' and 'Victoria' should be replaced with 'NSW' and 'VIC' to keep the data format consistent.