Name:Noor Ul Huda
ID:FLX-S8-077

Subject: summary of data quality issues and mitigation strategies

Dear team,
I hope you are doing well.I've reviewed the four data sets provided by Sprocket Central Pty Ltd using key quality Framework Dimensions including completeness,Accuracy,consistency etc.Below is the summary of main issues I found along with steps to fix them.

1. CustomerDemographic.csv

Key issues identified:

- Several unrealistic date of births (e.g. DOB in 1800's)
- Missing values in fieldset as last_name, job_title and job_industry_category.
- Gender field containing multiple values (e.g. F,Female,U)
- Default colum contain corrupted or malicious text (e.g. 0/0,...//..)

Recommendations:

- Apply logical validation rules to remove unrealistic dates of birth.
- Identify and assign values where possible.
- Standardize gender values.
- Correct or remove default column.

2. Transactions.csv

Key issues identified:

- Product_first_sold is in excel serial format then in date format.
- Check online_order or unexpected or missing values.

Recommendations:

- Convert the serial date value to standard date format.
- Standardize boolean online order field.

3. NewCustomerList.csv

Key issues identified:

- Multiple blank columns appear in the data set.
- Missing values in various customer fields.

Recommendations:

- Remove or properly label unused columns.
- Scan all and clean missing values.

4. CustomerAddress.csv

Key issues identified:

- Both full state name and abbreviation appears.
- Certain customer id are missing when compared to CustomerDemographic.

Recommendations:

- Convert all states into uniform format.
- Check all customer ids across the table and add missing.

Once these data quality issues are addressed, the datasets will be in a suitable condition for further analysis as required in the project.

Kind regards,
Noor Ul Huda