To whom it may concern,

I hope this email finds you well.

Thank you for sending us your datasets for our assessment.

After careful assessment of the datasets provided to us, we have identified a few quality issues with the datasets. There are some attributes we look at when evaluating the quality of a dataset. Listed in this email are the issues that we came across, along with their mitigation methods.

Below is a summary statistics table from the three datasets received. Please inform us if the figures do not align with your understanding.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **No. of records** | **Distinct Customers IDs** | **Date Data Received** |
| Customer Demographic | 4,000 | 4,000 | 1/10/2020 |
| Customer Address | 3,999 | 3,999 | 1/10/2020 |
| Transaction Data | 20,000 | 3,494 | 1/10/2020 |

**Accuracy – Correct Values**

**Issue**: The values of the data do not correspond to reality. One of the errors was the birth year of a customer which was 1843.

**Mitigation**: Added an age column within the dataset to accurately tell what customers’ ages are.

**Recommendation**: Generally, the date of birth should be a valid date that falls within an interval of 01/01/1900 to 01/01/2010.

**Completeness – Data Fields with Values**

**Issue**: Additional “*customer\_ids*” rows were inconsistent among “*CustomerDemographic*”, “*customer\_address*” and “transactions” sheets.

**Mitigation**: Maintained all “*customer\_ids”* from 1 to 3500 for all sheets as this is the range of customer ids consistent across all rows, dropped/filtered out any rows outside this range.

**Recommendation**: Ensure tables are up to date (from the same period). For our model, only “*customer\_ids*” from 1 to 3500 will be used as they have complete dates (free from missing values).

**Note**: Data received may not be synchronised across all spreadsheets. The results of the analysis may be skewed because of incomplete data. To overcome this issue, spreadsheets sharing similar information should be frequently crosschecked and synchronised.

**Issue**: Blanks/Missing values/null values

**Mitigation**: Dropped all missing values for “*job\_title*”, ”*brand”,* and“*online\_order*”columns in “*transactions*” and “*CustomerDemographic*” sheets.

**Recommendation**: Dropdown options should be used instead. This will ensure that there will not be blanks as there is a dropdown option for all choices.

**Note**: Blanks are treated as incomplete data. Having to drop entire rows of data due to blanks in particular cells can further skew analysis.

**Consistency – Values Free from Contradiction**

**Issue**: “Gender” column in “*CustomerDemographic*” sheet and “state” column in “*customer\_address*” sheet. Input values referring to the same thing were different.

**Mitigation**: Replaced with values that match with other values.

**Recommendation**: Dropdown options can be used here as well. This will maintain consistency in the input values across all the rows.

**Currency – Values up to Date**

**Issue**: “*deceased\_indicator*” column in “*CustomerDemographic*” sheet.

**Mitigation**: All data with the value “Y” were dropped from the dataset.

**Recommendation**: As the data of customers who are deceased are not considered as current customers, these rows of customers should be filtered out/dropped.

**Note**: It could be challenging to check for deceased customers, but once information is received the customer’s information should be updated accordingly.

**Relevancy – Data Items with Value Meta-data**

**Issue:** Existence of meta-data in “*default*” column of “*CustomerDemographic*” sheet.

**Mitigation**: Dropped entire column of metadata.

**Recommendation**: Check for incomprehensible Meta-data and delete or format to make them comprehensible.

**Validity – Data Containing Allowable Values**

**Issue**: Inconsistent data type for the same attribute (e.g. numeric values for some fields and strings for others)

**Mitigation**: Format “*list\_price*” column to currency to maintain consistency

**Recommendation**: Set up columns in a format that readily receive inputs of the same format. This will aid in interpretation/analysis of results at a later stage.

**Uniqueness – Records that are Duplicated**

**Issue**: None detected

**Mitigation:** -

**Recommendation:** -

These are the summary of all data quality issues found in the first stage of the data quality analysis. Mitigation strategies are noted below every issue for effectively improving data quality in the future. By following these strategies, data quality can be improved which in turn improves quality of analysis. A high-quality data not only improves efficiency in business processes and data lifecycle processes, but it also aids in making well informed data-driven decisions.

If you have any questions regarding data quality issues, please do not hesitate to contact me. Thank you.

Kind regards,

Jeremy Chia