Dear Manager,

Sprocket Central Pty Ltd,

Thank you for providing us with the four datasets from Sprocket Central Pty Ltd.

The summary table below highlights key data quality issues we have discovered in the data cleaning process. Please let us know if you have any queries concerning the issues presented.

I have evaluated the datasets according to the data quality dimensions framework as follows:

* Accuracy – correct values
* Completeness – data fields with values
* Consistency – values free from contradiction
* Timeliness – values up to date
* Relevancy – data item with value meta-data
* Uniqueness – records that are not duplicated.
* Validity – data containing allowable values.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Completeness** | **Consistency** | **Timeliness** | **Relevancy** | **Uniqueness** | **Validity** |
| (Dataset 1)  Transactions | Missing Multiple Values | Missing (Customer ID, online order, brand, product line,  Product class, Product Size, Standard Cost and first sold date) | Formatting of list price, standard cost Inconsistent | - | Cancelled order Visible | - | First sold date is an invalid format |
| (Dataset 2)  New Customer | - | Missing (Last name, DOB, Job Title, Job Industry) | - | - | Unnamed columns are irrelevant | Duplicate Customers | - |
| (Dataset 3)  Customer Demographics | DOB values are not possible | Missing (Last name, Job Title, Job Industry, Tenure | Formatting of Gender Inconsistent | Deceased Customer out of date | Corrupted Default Column | - | Age not reliable |
| (Dataset 4)  Customer Address | - | - | Formatting of state inconsistent | - | - | - | - |

# Accuracy issues:

* In the customer demographics dataset, the DOB is inaccurate, e.g., a data entry indicated a customer was born in 1843. An age column could be derived to improve comprehensibility.

**Mitigation:** Filtering out all the outliers from the DOB section.

**Recommendation:** Create an age section along with the DOB section to allow easy checking for errors.

# Completeness issues:

* In the Transactions dataset, the customer IDs are inconsistent, with some customer IDs being null. Date is not in sync across all the dataset tables. e.g., Customers with Customer ID 3501-4999 are missing.

**Mitigation:** Only take customer ID from 1-3500.

**Recommendation:** Ensure tables are up to date. Only complete data can be used for analysis. If the data is not in-sync across all spreadsheets.

* In the customer demographics dataset and the new customer dataset, some records of last name, job title, job industry, tenure and DOB are missing.
* In dataset Transactions, some records of the online order, brand, product line, product class, product size, standard cost and first sold date are missing.

**Mitigation:** To fill these null values, there is sufficient date so an average of the groups can be taken.

**Recommendation:** Make these values mandatory to enter to proceed.

# Consistency issues:

* For Transactions dataset, the list price and standard had an inconsistent format.
* For Customer Demographics dataset, the gender had an inconsistent format.
* For Customer Address dataset, the state had an inconsistent format.

**Mitigation:**

* + The prices were formatted to 2 decimal places for the currency $.
  + Changed all Variants of Male and Female to one standard.
  + Replace the names of states to abbreviations to ensure consistency across the datasets.

**Recommendation:** Instead of using a variable text field as the data type, categorical data should be used to prevent multiple representations of the same value. Dropdown menus reduce discrepancies and human error in manual data entry by various staff members and enhance the reading and interpretation of the data.

# Relevancy issues:

* For Transactions Dataset, the order status showed cancelled orders.
* For New Customer Dataset, there were several hidden columns.
* For Customer Demographics Dataset, there was a default column with corrupted data.

**Mitigation:** Filtering out all the cancelled orders and dropping all the corrupted and hidden columns.

**Recommendation:** Remove all the Corrupted date from files.

# Uniqueness issues:

* For New Customer Dataset, some of the customers are duplicated, e.g., customers that have the first name as the same as other customers but no last name.

**Mitigation:** Merge the customers with the same first name.

**Recommendation:** Ensure Customers enter both first name as well as last name to correctly identify.

# Validity issues:

* For Transaction Dataset, the product sale date is an integer which may cause confusion.

**Mitigation:** Standardize all the dates and currency values.

**Recommendation:** Make sure all the data is in the correct format

Please let us know if you have comments or questions on the above as I would be happy to discuss to ensure that all assumptions applied align with Sprocket Central Pty Ltd.’s understanding.

Warm regards

Akshaj Kabthiyal

Data Intern, KPMG