Dear (Client Name),

We have received the three raw datasets from SP rocket central Pty Limited. As per the preliminary task, in the below-mentioned list, we have analysed the quality of the raw data and we found multiple quality issues that need to be addressed. Also, we have suggested recommendations to mitigate the quality issues and improve the effectiveness of the data.

1. **Accuracy issues:**

**Issue**: **DOB** was inaccurate for **“Customer Demographics** ” and missing an **age\_column**;

Missing a **profit\_column** for **“Transactions”**

Mitigation: Filter out outlier in **DOB**

Recommendation: Create an **age\_column**, allowingfore more comprehensible data and easierto check for errors. Create a **profit\_column** in **“Transactions”** to check accuracy of sales.

Creating additional columns for age and profit will allow for easier identifications of errors. The **profit\_column** will assist in future monetary analysis.

1. **Uniqueness.**

**Issue**: Some of the data values are outliers and can disrupt the whole dataset. For example, The customer ID “34” with the name of Jephthah Bachmann was born in 1843, meaning that he is 175 years old which is an error in the data in the Customer Demographic Table.

**Recommendation**: Remove the redundant data as it may skew the distribution of the dataset.

1. **Currency:**

**Issue**: People that are “Y “ in “**diseased\_indicator”** are not current customers for **“Customer Demographics** ”

**Mitigation:** Filter out customers checked why entices indicator

**Recommendation**: It can be difficult to check for deceased customers but once this information is received one should update data accordingly.

Deceased customers are not current customers, removing them from data will increase currency of data and will result in more accurate estimates in future analysis.

1. **Completeness:**

**Issue1**: Additional **“customer\_ids”** were inconsistent among **“Customer Demographics** ” **“Customer Address”** and “**Transactions”**

**Mitigation :** filter or customer IDs from 1 to 3500

**Recommendation :** Ensure tables are up to date (from the same time). For our model only customer IDs from 1 to 3500 will be used as they have complete data.

The data received may not be in syn across all spreadsheets with incomplete data the analysis results may be skewed. This is a “completeness “ issue to prevent future occurrences it's encouraged to cross check spreadsheets and sync data.

There are an additional number of entries in customer\_ids in the Transactions table than Customer Demographic and Customer Address Table. Hence, the skewed data cannot be used if there are any missing records.

**Recommendation:** We will only perform the analysis on the synced data of all the three customer tables across the customer\_ID.

**Issue2**: Multiple attributes like “**Online Order”, “Brand Name”, “Product Line”, “Product Class”, “ Product Size”, “Standard Cost”,** and “**product\_first\_sold\_date”** in the **“Transactions”** table had blank values. Also, In the **“Customer Demographics** ” , **“Job Title”, “Job Category”** and **“Tenure”** some of the records are missing.

**Recommendation:** As the percentage of missing values in the datasets is low as compared to the whole dataset we can go proceed by removing them.

1. **Relevancy:**

**Issue**: lack of relevancy or comprehensibility in **default column** for **“Customer Demographics** ” and **order\_status** for **“Transactions**”.

**Mitigation:** deleted metadata in **default column,**  filter out canceled **order\_status.**

**Recommendation:** check for incomprehensible metadata and delete or format to make comprehensible .

Canceled **order\_status** is irrelevant information for future analysis as it can skew data for example to the number of customers per annum will be an overestimate.

1. **Validity:**

**Issue:**  former of “**list\_price”**, “ **product\_sale\_date“** for **“Transactions**”.

**Mitigation:** format “ **product\_sale\_date“** to short date format, format “**list\_price” to**  currency.

**Recommendation:** Set up columns to that formats such as price and decimals are already in place when entering new data.

Allowable values will make data to be interpreted more easily formatting enterprise and allowing for either two or three decimals placed consistently will increase readability this will reflect positively on speed and accuracy of analysis for business decisions.

1. **Consistency**

**Issue**: In the “State” Column of the Customer Address Table multiple duplicate values were found such as “VIC” & Victoria, “NSW” & “New South Wales”. Also, the issue is in the “Gender” column of the Customer Demographic Dataset.

**Recommendation**: To use abbreviations of the states instead of full names for all the records to ensure consistency across addresses. For Gender Column, the records “U” can be imputed with the distribution of the dataset.

Please look into the above-mentioned quality issues along with the recommended changes to ensure the consistent quality of the dataset across all the tables. If all the suggestions are matched we can proceed with further analysis of the data to find some suitable insights for the company.

Regards,

Imen Khemaissia