Dear [Name of the Client],

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 analyzed 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.

The quality issues and their corresponding mitigations for the 3 datasets are mentioned below in order as they appear in the fields.

1. **Customer Demographics (Issues)**:

* Gender – This field suffers from the data consistency and duplicate value issue. There are multiple values for the same entry. For e.g., M, Male, male.

**Recommendation**: All the values can be abbreviated instead of using full names.

* DOB – This column suffers from Outlier issue. For customer\_id 34, the DOB is 1843, signifying that the customer is alive for 175 years. This might be an issue with the data collection.

**Recommendation**: We suggest this record be removed as there seems to be an error during the data collection for this record.

* Default – The values in this column are not understandable and hence they cannot be used for further analysis.

**Recommendation**: We can remove this column from the data to avoid redundant data being used for further analysis.

* Nulls – The following columns suffer from missing data entries, null values, and blanks. Columns: last\_name, DOB, job\_title, default, tenure

**Recommendation**: It is recommended to impute the missing values or remove the corresponding records depending upon the field of missing value. We can proceed with both the options depending upon the field.

1. **Customer Address (Issues)**:

* State – Similar to the gender field in customer\_demographics sheet, this field has multiple values signifying the same meaning. i.e., the field suffers from inconsistent data For e.g., VIC, Victoria, NSW, New South Wales.

**Recommendation**: It is advised to proceed with the abbreviations rather than complete words.

* Customer\_id – This field has absent data for customer\_id 3 and 10 missing i.e., there is no address recorded for these customer ids.

**Recommendation**: We can remove these entries for our further analysis, or if possible, we can try to re-capture the data values for these ids.

1. **Transactions (Issues)**:

* Customer\_id – This field contains an extra id (5034) which is not available in the demographics data. This is the data point for which no customer information like address or demographic information is available.

**Recommendation**: We can remove this data point from our further analysis.

* Standard\_cost – The field contains cost value including the “$” symbol. And few values with more than 2 digits after the decimal point. This is a case of data type mismatch and data inconsistency.

**Recommendation**: We can remove the $ symbol to bring all the values in an uniform numeric format and also keep the round format up to 2 digits after the decimal for all such values.

* Nulls – The following columns suffer from missing data entries, null values, and blanks. Columns: online order, brand, product line, product class, product size, standard cost, product\_first\_sold\_date

**Recommendation**: It is recommended to impute the missing values or remove the corresponding records depending upon the field of missing value. We can proceed with both the options depending upon the field.

Please investigate 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.

Warm Regards,

Prathmesh Lonkar