Dear Sir/Madam,

Thank you for providing us with the following 3 raw datasets from SP Rocket Central Pte Ltd.

* Transactions in the past 3 months
* Customer Demographic
* Customer Address

As per the preliminary task, we have analyzed the quality of the raw data and 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. Please kindly refer to the list below for more details.

1. **Missing Values**

**Observation:** In the *Transactions* data frame, columns such as “online order” and “brand” contain significant number of missing values. In the *Customer Demographic* data frame, columns such as “job title” and “job industry category” contain significant number of missing values.

**Recommendation:** If there are only a few missing values in the column, proceed to remove the entire row of data that contains the missing value. If there is a significant number of missing values in the column, replace the missing values with the mean or mode of the column, whichever that is more appropriate.

1. **Inconsistent Entries**

**Observation:** In the *Customer Demographic* data frame, the "gender" column contains multiple expressions with the same meaning, such as Female, Femal and F, which can all be grouped under Female. In the *Customer Address* data frame, the "state" column contains multiple expressions with the same meaning, such as NSW and New South Wales, which can all be grouped under NSW.

**Recommendation:** Group the expressions with the same meaning under one common expression. For instance, replace entries of New South Wales with NSW.

1. **Outlier Values**

**Observation:** In the *Customer Demographic* data frame, the "DOB" column contains an outlier, since a person born in 1843 will be 180 years old now, which is impossible.

**Recommendation:** Since the outlier makes no logical sense, remove the row containing it.

1. **Illogical Values**

**Observation:** In the *Transactions* data frame, the “standard cost” column contains 3 price values with more than 2 decimal places, which is impossible.

**Recommendation:** Since the price values make no logical sense, remove the rows containing them.

1. **Invalid Columns**

**Observation:** In the *Transactions* data frame, the “product first sold date” column contains integers instead of datetime. In the *Customer Demographic* data frame, the “default” column contains numerous gibberish values.

**Recommendations:** Since the values in the columns make no logical sense, please kindly check what data was originally intended to be stored in these columns. If it is still possible to retrieve these data, replace these columns with them. Otherwise, remove these columns.

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. Thank you very much and have a great week ahead!

Best Regards,

Li Mingyang