Dear Sprocket Central Pvt Ltd,

Upon reviewing the three datasets from SP Rocket Central Pty Limited, our preliminary analysis has highlighted several quality issues within the raw data.We have further given our comments about how we tackled the data quality issues and laid out a plan to move forward with the data cleaning.

|  |  |  |  |
| --- | --- | --- | --- |
| Worksheet name | Total Records | **Unique Customer ID**s | Data Quality Issue |
| Transaction | 20000 x 13 | 3494 | Completeness & Relevancy |
| New Customer List | 1000 x 23 | - | Completeness & Consistency |
| Customer Demographic | 4000 x 13 | 4000 | Completeness,Consistency & Relevancy |
| CustomerAddress | 3999 x 6 | 3999 | Consistency |

The table above brings attention to notable data quality issues, assessed against standard dimensions such as completeness, relevancy, and consistency. To prevent the recurrence of these issues, the following recommendations have been proposed:

1. **Transaction Worksheet : The dataset exhibits missing values across various columns like “brand name”, “Product Line”, “Product Class”, “ Product Size”, “Standard Cost”. The column “product\_first\_sold\_date” was converted into a date/time format.**
2. We observed blank values in the specified columns, impacting data completeness and potentially causing inaccuracies during modeling. To address this data quality concern, it is crucial to eliminate these blank values from the dataset..
3. We addressed the issue with the 'product\_first\_sold\_date' column, which was initially in an unintelligible integer format. This problem commonly arises when exporting data from third-party sources. To enhance interpretability, we converted this column into a date/time format, facilitating a clearer understanding of the data.
4. **NewCustomerList Worksheet : Missing values identified and also inconsistency in “gender” column data values .**
5. As mentioned above ,blank values were discovered in the sheet for the column “second\_name”,however, it is not an issue as we may only use first name instead ,therefore, it is not important,however, there were still blank values .These were followed by more blank and null values in columns “job\_title” and “job\_industry”
6. The column “gender” consisted of an irrelevant variable “U” which was discarded from the column. However, if you could provide more clarity on its meaning then we could manage it in other ways.

1. **CustomerDemographic Worksheet :** Several crucial columns such as 'Last Name,' 'DOB,' 'job\_title,' 'job\_industry\_category,' and 'tenure' have missing values. Inconsistency for gender values and an irrelevant field called “default”.
2. The 'gender' column contains multiple values and spelling variations (e.g., U, F, Femal, M), leading to ambiguity.Therefore, this was changed to the “M” for male and “F” for female.
3. Null values were removed from “job\_title” and “job\_industry\_category”.
4. The 'default' column lacks clarity in its interpretation.Hence, it was removed.
5. Additional Customer IDs: There are additional customer ids that are not present in ‘CustomerAddress’ and ‘Transactions dataset’.
6. **CustomerAddresses Worksheet :**
7. State Name Format Inconsistency:The 'state' column exhibits inconsistencies in name formatting.There are some states mentioned using abbreviations and other using full names which creates duplicacy.
8. Unclear Property Valuation Metric:The 'property\_valuation' metric lacks information about the unit of valuation (e.g., millions, billions)

Moving forward, the team will continue with the data cleaning and transformation process for the purpose of modelling. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.If you have any inquiries or require further clarification, please feel free to reach out.  
  
Regards,

Rajat Mishra

Junior Data Consultan