Data Quality Assessment - Update

Dear Team

Sprocket Central Pty Ltd,

This is Rajat Srivastava, from the KPMG Data & Analytics team. This email is in response to the preliminary data quality assessment on the datasets provided to us. First of all, thank you for providing us dataset to analysis and get great insights, it is extremely important for us to clean the data, or missing values involved. The importance of quality of data can never be underestimated.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | Number of Records | Unique Customer Ids | Date Data Received |
| Customer Demographics | 4000 | 4000 | 04.01.2023 |
| New Customer Data | 1000 | 1000 | 04.01.2023 |
| Customer Addresses | 3999 | 3999 | 04.01.2023 |
| Transaction | 20000 | 3494 | 04.01.2023 |

Please let us know if the figures are not aligned with your understanding.

I would like to mention the parameters on the basis of which data quality checks are done.

* Accuracy - Whether the values filled are correct
* Completeness - The dataset should not have null values in it
* Consistency - Values free from contradiction
* Currency - Values up to date
* Relevance - Data Items with value metadata
* Validity - Data containing allowable values
* Uniqueness - Duplicated data shouldn’t be involved.

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are mentioned below. Furthermore, recommendations have been provided to avoid the recurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

**Data Quality Assessment Results and Mitigation Measures:**

* Customer Demography - It had accuracy issues e.g. a data entry indicated a customer was born in 1843. An age column could be included to avoid such errors. The mean year was imputed here. The dataset also had completeness issues in last names, job titles, industry. Consistency issues in ‘gender’ like it had various entries involved. A corrupted ‘default’ column was there. Cross checking the spreadsheets and allowing only valid entries would help. Enforce a drop-down list for the user entering the data rather than a free text field in case of categorical entries like brands, states, gender etc.
* Customer Address - This dataset was complete, with just inconsistency issues in the ‘State’ feature, which was replaced by the team.
* New Customer - There were no accuracy issues in this but it also had a completeness issue. There were null values in last names, job titles, and industry. Some inconsistency was there like there were unnamed columns involved which would be harder to interpret in order to come up with marketing strategies later.
* Transaction Records - Incomplete data was in various columns such as order status, brand, product class, line, standard cost. We had to impute these with forward and backward crossmatching and filling but it would definitely affect the end results. Transaction data entries are advised to be filled with utmost sincerity. This would be affecting marketing strategies directly.

If the number of null-values are small, they are imputed using various statistical measures but if it’s so much that the column or feature itself isn’t providing any value to the analysis, it is better to drop them. These are some errors we come up with while investing in any dataset. It would be great if these small measures are taken while data entry, to get a high end result. There’s a detailed report attached regarding the assessment.

The team will continue with the data cleaning, standardization and transformation process

for the purpose of model analysis. 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.

Feel free to connect in case of any issues.

Kind Regards,

Rajat Srivastava

KPMG - Data & Analytics