Dear client,

I’m writing this email to confirm the 3 datasets from Sprocket Central Pty Ltd with you. Thank you for providing these datasets with us. After I reviewed the datasets, there have some issues I want to mentioned. The table below shows the statistical summary of the 3 datasets. Please take a look of the summary and let me know if the data are not match with your company’s record.

|  |  |  |
| --- | --- | --- |
|  | Number of record | Number of customer\_id |
| Customer Demographic | 4000 | 4000 |
| Customer Addresses | 3999 | 3999 |
| Transaction Data | 20000 | 3494 |

**Inconsistent data records**

Mitigation: The data records in each tables should be recorded from the same period. The customer number are not same among these 3 datasets.

Recommendation: The datasets may miss some data records. Therefore, please double check with your business record to ensure there have no missing data.

**Multiple columns contain empty values (such as brand, product\_line, job\_title and etc.)**

Mitigation: If the empty values only occupied a small portion of the dataset, then filter out the empty values. Otherwise, counted into the dataset.

Recommendation: Only a small number of rows contain empty values in transaction and customer demographic datasets. These minor record will not affect the further analysis, then those records have been removed from the original datasets.

**Inconsistent value for the same content(such as gender and state)**

Mitigation: Using the unified expression for the same value in the datasets.

Recommendation: Change those value into a same format. For example, change ‘F’, ‘Femal’ into ‘Female’. Using the same expression is better to possess the data.

**Inconsistent data type(such as product\_first\_sold\_date should be defined as date rather than number)**

Mitigation: Using appropriate data type to present data.

Recommendation: Change the data type into an appropriate data type. And there may contain more than one data type in one column, therefore adjust the data into a same data type(convert string into number, convert number into date).

**Inaccurate values (such as impossible birthday)**

Mitigation: Filtering out all the inaccurate and wrong values from the datasets.

Recommendation: Those error may recorded by mistakes, and the outlier could affect the analysis. Therefore, the outliers were removed from the datasets.

First, most of the data are correct and accurate except for some minor error such as the customer’s birthday which is impossible, I highlighted the error with yellow. And this data should not be include in the dataset.

Thank you again for providing the datasets with us again. Our team will continue to work on with the data cleansing to make sure the data quality is ready for the further analysis. And please let me know if you have any concerns about the summary and I’ll keep updating our progress with you. Thank you!

Best wishes,

Rebecca