Sprocket Central Pty Ltd

Dear Sir/madam

Re: Report on Data Quality Assessment

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding

|  |  |  |  |
| --- | --- | --- | --- |
| Table name | No. of records | Distinct Customer ids | Date data received |
| Customer  Demographic | 4000 | 4000 | 26-09-2021 |
| Customer  Address | 3999 | 3999 | 26-09-2021 |
| Transaction  Data | 20000 | 3494 | 26-09-2021 |

Looking at the datasets in relation to data validity, data uniqueness, data precision\accuracy, Reliability, timeliness and consistency the following are some of the data quality issues we came across;

* **There is inconsistency in some of commonly used attributes for instance, the gender column instead of having three possible values that is, male, female or U for Unisex, values such as femal and F and Mal came up. Also states names shows inconsistency, having VICTORIA same time VIC, New South Wales and NSW**

*Mitigation: please ensure that data entry for commonly used attributes remains consistent, therefore we will change the F and Femal values to Female and mal to male. We going to convert new South Wales to NSW and Victoria to VIC*

*Recommendation: Enforce a drop-down list for data entry rather than a free text field.*

* **The ‘default’ column provides irrelevant information; it contains different character forms which are not related and makes no sense.**

*Mitigation: therefore, we are going to drop the column for our training set.*

*Recommendation: While collecting data one should consider the relevance of the feature before including it.*

* **Data completeness is also an issue here, some of the most essential attributes have missing values. The missing features include**.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column Name | Last name | DOB | Job title | industry | default | Tenure | postcode | state | address | country | Property  valuation |
| Missing Values | 124 | 87 | 506 | 656 | 302 | 87 | 4 | 4 | 4 | 4 | 4 |

*Mitigation: We are going to fill some of the categorical data with mode and the numerical data with mean. Some will be dropped off*

*Recommendation: you should enforce data entry by applying NOT NULL constraint to ensure every information concerning a record is captured*

* **Date of birth for customer ID 34, year being 1843 is inaccurate figure. This wrong value interferes with data accuracy. This also affects data validity since age 175 is an invalid entry**. **Such age does not exist**

*Mitigation: We are going to drop the entry or rather provide us with the intended exact value, these inaccurate values will affect our model accuracy.*

*Recommendation: while capturing data you should ensure precision and accuracy is enforced in the entry tools*

* **From the task’s description, Transaction data should be for the last three months but after carefully looking into data, it covers data for twelve months, period between month of January 2017 to December 2017.**

*Mitigation: Therefore, you should clarify what months or period you wish us to look into or we will have to use the entire 12-month period dataset*

**● Inconsistent data type for the same attribute (‘the standard cost’ column values, some are numeric i.e. 45.87 while some are strings e.g. 1,167.68)**

*Mitigation: Convert selected records in characters to numeric. Remove non-numeric characters from string. The dollar sign included is going to be dropped and added to the column title name*

*Recommendation: Ensure that fact tables in the given database have constraints on data types. Having different data types for a given field make it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field*

* **There is data redundancy aspect. The customer ID has been used for both customer demographic table and customer address table**

*Mitigation: We are going to merge the two tables on customer ID basis to enhance uniqueness in our data exploration*

* **Additional customer\_ids in the ‘Transactions table’ but not in ‘Customer Master (Customer Demographic)**’

*Mitigation: Please ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for our model. This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records.*

*Recommendation: Pleases ensure regular updates across all datasets*

* **Data is not up to date; we still have the deceased data in our data**

*Mitigation: We are going to delete the two entries for the deceased customers*

*Recommendation: Make sure record update is done regularly.*

Moving forward, 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.

Regards

Bramuel Aswani