Dear sir/madam,

Thanks for providing the 3 datasets from Sprocket Central Pty Ltd. After having explored the data set there were a lot of data quality issues found in the 3 datasets where these issues have been summarized in tabular format which contains details regarding the data issues found in the respective columns of the 3 datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.NO.** | **DATA QUALITY ISSUE** | **SHEET NAME** | **COLUMN NAME** |
| 1. | Null Values | Transactions  Customer Demographic | online\_order  brand  product\_line  product\_class  product\_size  standard\_cost  product\_first\_sold\_date  last\_name  DOB  job\_title  tenure |
| 2. | Data Inconsistency | Customer Demographic  Customer Address | gender  state |
| 3. | Incorrect Data Format | Transactions  Customer Demographic  Customer Address | transaction\_id  product\_id  customer\_id  transaction\_date  list\_prices  standard\_cost  product\_first\_sold\_date  customer\_id  past\_3\_years\_bike\_related\_purchases  tenure  customer\_id  postcode  property\_valuation |
| 4. | Irrelevant Data | Customer Demographic | Default |
| 5. | Incorrect Data | Customer Demographic | DOB |

Details regarding the data quality issue:

1. There were a lot of **Null/Blank** values present in the respective columns of the 3 datasets and it is important to communicate with the client as to whether any data is there to be **filled** or else whether these blank values can be **eliminated**.

2. Data inconsistency issue was found in the gender column of the Customer Demographic dataset which contained more values such as **Male, M, Female, F, Femal, U** instead of either having **M** or **F** Or **Male or Female**. The state column of the Customer Address dataset contained values such as **NSW, QLD, VIC, New South Wales, Victoria** instead of either **NSW, QLD ,VIC** or **New South Wales, Queensland, Victoria**.

3. The columns of the three datasets were not present in the right Data format. A lot of columns that contained **numbers** as the data were in **Text and Genera**l format and columns containing **date** were also in the **General** format especially the **products\_first\_sold\_date column** present in the **Transactions** dataset was in the **General** format and was converted to **date** format in order to get the exact date values.

4. The **Default column** in the **Transactions** dataset contained some absurd values which made the data irrelevant resulting in an **Irrelevant** data quality issue. After having tried to convert the data in different formats the data remained absurd and therefore I suggest to remove the **Default** column from the **Transactions** dataset.

5. The **DOB** column in the **Customer Demographic** dataset contained one value of **DOB** which is **1843-12-21** as the data which was **Incorrect** making it absurd and irrelevant.

The data quality issues mentioned above were addressed by correcting the above and the 3 datasets were completely cleaned by keeping the data issues in mind and is ready for the next phase of the analysis process.

Looking forward to work with your company in order to further continue the analysis process regarding the datasets and recommend some strategies in order to drive your company in the right and positive direction.

Regards

Anantharaman H

Junior Consultant