Data quality issues.

Dear Manager,

Thanks for providing us the three datasets of *Sprocket Central Pty Ltd.* I have analysed these three data sets and identify the following table of data quality issues. I am going to highlight the data quality issues.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data quality matrices | Data accuracy | Data completeness. | Data consistency | Data currency | Data relevance | Data validity |
| * Customer Demographic | Inaccurate data in DOB column  Age column is missing | Job\_title column contains the blanks.  Customer\_id column data is missing, | Gender column the data is in inconsistent form | Deceased column filter\_out | Default column is completely irreverent from the given data. |  |
| * Customer address |  | Customer\_id is incomplete | State column contains the inconsistent data |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| New\_customer\_list |  | First\_name and last\_name columns ha missing values and also g4s index value missing |  |  |  | Values, property\_value and rank columns have formatted. |
| Transactions | Profit column is missing | Customer\_id column has missing data.  Online order column has blank values  Brand column has blank values |  |  | Cancelled status order value is filter out | list\_price column is formatted.  Product\_sold\_first\_date  Column is formatted. |

**Accuracy Issues:**

Accuracy issue in the DOB column of **customer\_demographic dataset**. I spotted the outlier in the DOB column of the dataset and another issue is the missing information of age column so, we can accurately spot the information regarding person to avoid such confusions.

In the **Transactions dataset,** the profit information is missing. I have created the profit column by minus the sales of the product from the standard rate to calculate how accurately the profit has been made.

**Completeness Issues:**

**In Customer Demographic dataset,** job\_title has the blank spaces and customer\_id has missing values and I have identified it using filter out function.

Similarly, In **Customer address dataset,** customer\_id column is incomplete.

Talking about the **New\_customer\_list and Transcations dataset,** have the First\_name and last\_name columns have missing values and Customer\_id column has missing data.

Online order column has blank values, Brand column has blank values respectively and these values have been identified by using filter out function.

**Data consistency:**

**In Customer Demographic dataset,** the data is in the inconsistent form particularly in the gender column. So, by using the find and replace method I fixed that. Similarly, the state column of the **customer address dataset,** also contains the inconsistent records and I fixed that by using the method used above.

**Data currency:**

There is the data currency (data quality) issue in the **Customer Demographic and Transactions dataset,** because there are some values in the status\_order\_value and decreased column which are not required/updated. So, to update the columns values I use the filter and filter out the un-necessary values.

**Data relevance:**

In the **Customer Demographic dataset**, the default column is completely irrelevant from the provided data. So, to make the whole data relevant I removed the targeted column from the dataset.

**Data validity:**

There are the formatting issues in the maniple columns of the different dataset. For most of the formatting issues related to the converting datatypes from one datatype to another suitable datatype. For example, Converting General to text data type. These sorts of issues were present in the Values, property\_value and rank columns of the **New\_customer\_list dataset** and list\_price, and Product\_sold\_first\_date columns of the **Transcation dataset**. I formatted the datasets’s columns to handle these sorts of data quality issues.

Here are some of the important data quality issues that I solved because if it persists in the dataset and I analyse the dataset by having these issues remain present in the dataset. So, it can lead us to the wrong results.