**SPROCKET CENTRAL PTY LTD DATA QUALITY EXPLORATION**

Dear client,

After careful examination of the dataset you provided, I discovered some data quality issues which can cause discrepancies and make our analysis of the dataset bias. Though we were given 3 sheets names to analyze but from the Excel file sent, we discovered an additional sheet, so we decided to analyze it as well since it contains records of your new customers. The general summary of the dataset is:

|  |  |
| --- | --- |
| **Sheet Name** | **Number of records** |
| Transactions | 20000 |
| NewCustomerList | 1000 |
| CustomerDemographic | 4000 |
| CustomerAddress | 3999 |

From the above sheets, I realized some columns have data quality issues. In the Tables below, I have highlighted those columns in each sheets and strategies to mitigate those issues:

**Transactions Sheet**

|  |  |  |
| --- | --- | --- |
| **Column Names** | **Data Quality Issues** | **Strategies to mitigate the issues** |
| online\_order |  | These columns contains 197 null values. I would advise you recheck |
| brand |  | your database for the value of those |
| product\_line | Incompleteness | blanks which may be as a result of |
| product\_class |  | oversight. But if they are not found, |
| product\_size |  | type “**not given**” in the blank rows |
| standard\_cost |  | Or they should be removed. |
| product\_first\_sold\_date | Incompleteness and Inconsistency | The null values should be removed and the format/datatype should be changed from General to Short Date. |
| list\_price | Inconsistency | The format should be changed from General to Currency. |

**NewCustomerList Sheet**

|  |  |  |
| --- | --- | --- |
| **Column Names** | **Data Quality Issues** | **Strategies to mitigate the issues** |
| last\_name | Incompleteness | The null values can removed or |
| job\_title |  | replaced with “**not given**”. |
| dob | Incompleteness and Inconsistency | The null values should be removed and the format should be changed from Custom to Short Date. |
| gender | Inaccuracy | U- Value is not well interpreted. |
| property\_valuation |  |  |
| postalcode | Inconsistency | The formats should be changed from |
| past\_3\_years\_bike  related\_purchases |  | Text to Numbers. |

**CustomerDemograhic Sheet**

|  |  |  |
| --- | --- | --- |
| **Column Names** | **Data Quality Issues** | **Strategies to mitigate the issues** |
| last\_name | Incompleteness | The null values can removed or |
| job\_title |  | replaced with “**not given**”. |
| dob | Incompleteness and | The null values should be removed and the format for dob, all should be changed to Short Date while format |
| tenure | Inconsistency | for tenure, it should be changed to Number. |
| gender | Inaccuracy | Spelling Error as “**Femal**” instead of Female and U- Value is not well interpreted. |
| deceased-indicator | Irrelevancy | Remove the Y value because we only  need the dataset of people alive and |
| default |  | removed default column from the dataset. |

**CustomerAddress Sheet**

|  |  |  |
| --- | --- | --- |
| **Column Names** | **Data Quality Issues** | **Strategies to mitigate the issues** |
| State | Inaccuracy | New South Wales and Victoria should be abbreviated to NSW and VIC so that it can align with the others. |

Please I would suggest you quickly look into these issues so you can make the needed corrections and we can move on to providing insights on the dataset to help your company grow because our client’s growth is our major priority. Hope to hear from you soon.

Kind regards,

Okinedo Blessing.