Dear Sir/Madam,

In response to your request, I have assessed the quality of the data you sent us and found some issues with the data, some of which may not be necessarily significant enough to affect future analysis while others would impact the effectiveness of any analysis to be carried out forthwith.

Fortunately, no duplicate records were found.

Some of the issues identified are:

1. Inaccurate data – the date of birth of Jephthah Bachman is stated as 1834 which would make him/her 189 years old today which is impossible. Kindly confirm his/her correct birth date.
2. Some entries also have misspellings such as Argiculture in the Job industry category. This would be mitigated by providing a drop-down list with all the job industry categories listed for easy choosing with a final option of others which would apply branching logic with another entry box for specifying the other job industry category.
3. Many records are incomplete as they contain blank entries. This would be mitigated and prevented by ensuring validation of the data entry tools to reject submission of records with blank entries. Other records within the transaction worksheet contain incomplete data where they lack product details invalidating the nature of these transactions.
4. Some data that I would have considered irrelevant was included in the document i.e., the default column in the customer demographic worksheet. It contains invalid data that comprises of a mixture of alphabetical and non-alphabetical characters with no clear pattern or trend. Others include the 4 columns to the left of the rank column. They are randomly changing hence provide no significance to the analysis of the data.
5. Some of the product\_ids in the transaction worksheet are invalid as different products seem to have the same product\_id. This can be prevented by validating the entry of product\_id to unique numbers or numbers that are automatically incremental.
6. Some of the data entries were inconsistent with the nature of the data to be entered. Some dates were saved as text instead of date format. This can be prevented by validating the entry form to only allow date format or include a calendar data input.
7. Also, some entries in the gender variable were inconsistent with some being male while others are M, some female while others are F or femal and an inapprehensible U which I assume is supposed to be Unspecified. This can be prevented by using drop down lists from which the respondent can choose from.

Below is a full breakdown of the issues identified and mitigations undertaken

## Transactions

|  |  |
| --- | --- |
| Problems Identified | Interventions taken |
| 360 records have blank records in the online\_order column | Replaced with Unspecified |
| 197 records seem to have made a transaction but without a listed product evidenced by no product\_id, no brand, no Product\_line, no product\_class, no product\_size, no product\_first\_sold\_date | Replaced with null |
| Different products seem to have similar product\_id hence not a unique identifier |  |
| The number format of the list\_price column was set to general | Number format changed to dollar currency |
| The number format of the product\_first\_sold\_date column was inconsistent with the type of data present | Number format changed to short date |
| The text case of online\_order, product\_class and product\_size columns was inconsistent | Text case changed to Proper case |

## NewCustomerList

|  |  |
| --- | --- |
| Problems Identified | Interventions taken |
| The column headers Rank and Value are inconsistent with the rest. They are in proper case | Changed to lower case |
| The column header postcode has been misspelt | Corrected to postal\_code |
| The column header DOB is inconsistent as it may have several meanings | Changed to date\_of\_birth |
| 29 records have missing data in the last name | Blanks replaced with null |
| 87 records have U as the entry in under the gender column which is erroneous | Changed to Unspecified |
| The number format of past\_3\_years\_bike\_related\_purchases column has been set as text but it contains numbers | Entries converted to numbers and format set as general |
| The number format of date\_of\_birth has been set to text | Number format changed to date |
| 17 records have missing data in the date\_of\_birth column | Blank records replaced with null |
| 106 records have missing data in the job\_title column | Blank records replaced with null |
| Agriculture in the job\_industry category is misspelt (argiculture) for 26 records | The spelling is corrected |
| The number format of property\_valuation column has been set as text but it contains numbers | Entries converted to numbers |
| The worksheet contains 4 columns to the left of the rank column that contain data that is constantly changing as they are dependent on a RANDBETWEEN function | Deleted the 4 columns as they are not significant for analysis |

## CustomerDemographic

|  |  |
| --- | --- |
| Problems Identified | Interventions (to be) taken |
| Jephthah Bachman’s date of birth is stated as 1834 meaning he is currently 189 years old which is impossible. | Enquire from the client about that specific customer. Cell filled with red in the meantime |
| 87 records including Jephthah Bachman’s gender has been input as U which is erroneous | Changed to Unspecified |
| 87 records are missing the date of birth | Blank records filled with null |
| 125 records are missing the last name | Blank records filled with null |
| Entries in the gender column are inconsistent with some being represented as M and F for male and female respectively while some are misspelt (femal) | Corrected to either Male or Female |
| 506 records have missing data in the job\_title column | Blank records filled with null |
| Agriculture in the job\_industry category is misspelt (argiculture) for 113 records | The spelling is corrected |
| The default column contains inconsistent data. | Deleted the column |
| The column header DOB is inconsistent as it may have several meanings | Changed to date\_of\_birth |

## CustomerAddress

|  |  |
| --- | --- |
| Problems Identified | Interventions taken |
| The column header postcode has been misspelt | Corrected to postal\_code |
| Under the state column, there is inconsistent data with some entries being entered in different formats i.e., New South Wales and NSW; Victoria and VIC | All records with New South Wales changed to NSW and Victoria to VIC for consistency |
| The entries in customer\_id are inconsistent having skipped 3, 10, 22 and 23; the last customer\_id is 4003 | Assuming that when sorted from largest to smallest, the specified customer\_ids were mistakenly skipped hence the order is correct, the customer\_ids were renumbered so as to have the last customer\_id being 3999 |

Kindly get back to me in case of any consultation regarding these issues or any other query.

Regards,

Ian Kigamba.