**Data Quality Assessment**

For *Sprocket Central Pty Ltd*

Dear *[Client point-of-contact]*,

Thank you for providing us with the 3 datasets. Below is the summary statistics of the provided data. 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 | *4,000* | *4,000* | *18-12-2020* |
| Customer Address | *3,999* | *3,999* | *18-12-2020* |
| Transaction Data | *20,000* | *3,494* | *18-12-2020* |

Notable data quality issues were encountered therefore the assessment has been split into two parts. The first one describes the general issues in detail – it helps to understand why certain changes are required and acts as a guideline if you decide to further review and polish the data with our team or on your own accord. This would help us to meet your preferences the best. The second part tells what necessary mitigations have been done to accurately perform feature engineering and data representation in further stages.

**Part 1**

To make things as brief and clear as possible, we have separated the explanations/suggestions with concrete issues. Each sheet has been assessed separately while considering overarching (global) issues too. The suggestions have been categorized and labeled with numbers corresponding to the explanations below.

**Reasons and advice:**

1. In terms of ***completeness***, CustomerDemographic and Transactions contain some empty cells. This can cause issues when a specific piece of data is queried and there is nothing to return. If such cells are referred in other tables, an error can occur. Therefore, it is suggested to leave no empty cells or, if there is no data, to denote it explicitly by inserting *“n/a”* like in other cells where data is not provided.
2. ***Accuracy*** could be improved in CustomerDemographic and Transactions as there exist some typos and junk values. This is to ensure data can be sorted and classified correctly, as well as to assure the correct values are returned when data is queried. Junk values take up space and are unnecessary. They do not provide any useful information. It is suggested to recheck the pointed-out cells for any typos and replace or remove any rows or columns with junk data.
3. ***Consistency*** should be kept within columns in CustomerDemographic, CustomerAddress and Transactions. It is best to keep a single name for values of specific category because it makes things easier to manage as data is in one place. Same format across the values within a column allows correctly comparing data when doing analysis. It is advised to pick one or the other naming for a certain class and to stick with only one cell format within a single column and every other column of the same type.
4. ***Validity*** should be checked for some data in CustomerDemographic, CustomerAddress and Transactions. Values must be within appropriate bounds and use a correct data type and format. Having incorrect type of values can cause errors and inconsistencies when data is handled, as well as provide falsifying information. It is best to look at each case separately and fix and approve the data reflects the reality and is formulated correctly.
5. ***Relevancy*** must be ensured in CustomerAddress and Transactions. Irrelevant data consumes extra space and requires extra management when a dataset is expanded. Having old data might not reflect the information correctly if it is not used anymore. It is suggested to get rid of columns or rows that are not required.

**Concrete issues:**

Customer Demographic

* Columns *last\_name*, *DOB, job\_title, default* and *tenure* have cells with ***missing values* (1)**. For example, the row with a value of *“4”* in *customer\_id* has no data in *last\_name* and *job\_title.*
* Columns *gender, DOB* and *default* have ***inaccuracies*** **(2)**.
  + One cell has a value of *“Femal”,* lacking *“e”* at the end.
  + The value of “*1843-12-21”* in *DOB* column is most likely mistyped.
  + The entire *default* column consists of junk data.
* Columns *gender,* *DOB*, *job\_title* and *deceased\_indicator* with *owns\_car* have ***inconsistencies*** **(3)**.
  + There are two values, “*M”* and *“F”,* which refer to *“Male”* and *“Female”* and therefore are inconsistent.
  + Row with *DOB* value of “*1843-12-21”* has a different cell format from the other values within that column.
  + The last word of the value *“Analog Circuit Design manager”* in *job\_title* is not capitalized which makes this value inconsistent with other ones.
  + *“Yes”/ “No”* and *“Y”/ “N”* are used to express the same thing between columns *deceased\_indicator* and *owns\_car.*
* Column named *past\_3\_years\_bike\_related\_purchases* is ***valid (4)***, but it could be improved (shortened and simplified) as it would make it easier to perform any queries to this name.

Customer Address

* Columns *address* and *state* have ***inconsistencies (3)***.
  + Some addresses use different notation for street numbers, for example, number 7 is sometimes written as *“7”*, sometimes as *“07”*. There should be no *“0”* in front of the number if the amount of digits in every street number is different. Otherwise, 0’s could be added for formatting and sorting reasons.
  + *“New South Wales”*/ *“Victoria”* and *“NSL”*/ *“VIC”* are used to denote the same thing. It is suggested to stick with the short version as it will be consistent with *“QLD”*.
* Some values in columns *customer\_id* and *address* are ***not valid (4)***.
  + When comparing IDs in CustomerAddress with IDs in CustomerDemographic, some of them are not included in one or the other dataset. For example, ID *“4003”* is only found in CustomerAddress.
  + Address numbers usually do not have more than 3 digits and they never have a value of 0.
* Column *country* might be ***redundant (5)*** for the dataset if all the customers use addresses in Australia.

Transactions

* Columns *online\_order, brand, product\_line, product\_class, product\_size, standard\_cost* and *product\_first\_sold\_date* contain some ***empty cells*** ***(1)***.
* Column *customer\_id* contains ***inaccurate (2)*** value of *“5034”*. No customer has this ID in CustomerDemographic.
* Columns *transaction\_date, list\_price* and *standard\_cost* have ***inconsistencies*** ***(3)***.
  + *transaction\_date* has different date notation than in other dataset sheets. DD/MM/YYYY should be changed to YYYY-MM-DD. It would also be wise to use a custom cell format rather than date format to keep it consistent with the other sheets.
  + *list\_price* has different price notation than in *standard\_cost.*
  + Some values in *standard\_cost* have different format than the majority, for example, *“312.7350159”* has more than 2 decimal values and does not have a dollar symbol in front.
* Columns *product\_id,* *product\_first\_sold\_date* and *standard\_cost* contains some values that are ***not valid*** ***(4)***.
  + It would be good to check whether *product\_id* denotes a unique item. There are some duplicate values for *product\_id* with different *brands* or *product\_lines*. One ID should refer to only one item, however, if there are same IDs but for different brands, a composite key should be used to uniquely identify the product.
  + Values (or their notation) in *product\_first\_sold\_date* should be changed to represent the date like in other columns of the same type. If the values do not refer to date, the column should be renamed.
* Some records might not be ***relevant (5)*** anymore as it was mentioned that the dataset should contain only the transactions within the last 3 months.

The following section provides some tips on how these issues can be avoided when gathering future information from customers.

**Recommendations:**

* Enforce a drop-down list for the user entering certain data, such as gender and state, rather than a free text field.
* Make certain fields, such as DOB and Job Title, compulsory or provide an option of “n/a” where information is not applicable.
* Use an external interface for address finding to allow users to pick the correct information in the address fields rather than manually typing them.
* When information is passed to the database, have it validated with regular expressions and approve its type and format. It is wise to have constrains on datatypes.
* Adapt a tracking system for the transaction data to be notified when a certain record can be removed.

**Part 2**

Some cleaning has been already done for the 3 datasets. Please note that only the changes that are necessary have been applied. We will adapt our analysis to the current state of the training sets – it is ready and we can move forward. The summary of the implemented mitigations is provided below:

* Rows with missing fields, where the total amount of such records is less than 1%, have been removed if they do not affect other records globally (e.g., this was applied in Transactions).
* Likewise, records with invalid values have been removed (this includes customers with IDs not found in the main dataset, CustomerDemographic).
* Values in columns *list\_price* and *standard\_cost* have been changed to numerical format with 2 decimal places.
* All the mentioned inaccuracies in CustomerDemographic have been fixed.
* Consistency problems have been minimized – values and their types in columns *gender, state, transaction\_date, list\_price* and *standard\_cost* are now correct.

Below is a short sum-up of the currently cleaned data which is ready for further analysis. We have also attached the cleaned document for you to review. Please inform us if anything else should be considered or if any changes should be reverted.

|  |  |  |
| --- | --- | --- |
| **Table name** | **No. of records** | **Distinct Customer IDs** |
| Customer Demographic | *3,999* | *3,999* |
| Customer Address | *3,996* | *3,996* |
| Transaction Data | *19,800* | *3,493* |

We look forward to hearing from you.

Kind regards,

*[Junior Consultant Name]*