Dear (Sprocket Central Pty Ltd),

Thank you for providing us with the 4 datasets from Sprocket Central Pty Ltd. Our team had gone through dataset and found some data quality issues. We would like to reach out to your team for further information and clarification required about the attributes of the dataset.

Here are the summary statistics of the 4 datasets. Please let us know if the figure is not aligned with your understanding:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table name** | **No. of records** | **No. of columns** | **No. of columns having missing values** | **Date data received** |
| Transaction | 20000 | 13 | 6 | 12/06/2020 |
| NewCustomerList | 1000 | 23 | 4 | 12/06/2020 |
| CustomerDemographic | 4000 | 13 | 6 | 12/06/2020 |
| CustomerAddress | 4000 | 6 | 0 | 12/06/2020 |

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

1. **Transaction Table analysis:**

* The dataset is having information for whole 2017 year not just 3 months.
* We are not getting the clear idea of values in column “product\_first\_sold\_date”.
* Summary Statistics of columns where records were missing:

|  |  |  |
| --- | --- | --- |
| Column name | Count | Percentage |
| online\_order | 360 | 1.8% |
| brand | 197 | 0.985 |
| product\_class | 197 | 0.985 |
| product\_size | 197 | 0.985 |
| standard\_cost | 197 | 0.985 |
| product\_first\_sold\_date | 197 | 0.985 |

* In the “standard\_cost” columns 99 percent of the data lies below 1610.90, except some rows values 1759.85 which counts to 1 percent.

Please refer to excel file ‘standard\_cost\_outliers.xlsx’ for the list of outliers

* Column “list\_price” and “standard\_cost” is showing positive relation.
* ***Recommendation:***
* Please recheck the “product\_first\_sold\_date” column or give us idea of what the column consists of.
* Provide us the process of constructing the database so we can deal with missing data.
* In “standard\_cost” column some values seem to be an outlier and need to be rechecked whether it is a genuine input, data is missing or a technical glitch.
* Give us more information about the “list\_price” column.

1. **NewCustomerList dataset analysis:**

* There are 5 unnamed columns.
* Summary Statistics of columns where records were missing:

|  |  |  |
| --- | --- | --- |
| Column name | Count | Percentage |
| last\_name | 29 | 2.9 |
| job\_title | 106 | 10.6 |
| DOB | 17 | 1.7 |
| job\_industry\_category | 165 | 16.5 |

* “gender” column is having label ‘U’.
* 5 percent of the data in “property\_valuation” column is having values equal to or less than 2.

Please refer to excel file ‘property\_valuation\_outlier.xlsx’ for the list of outliers

* Some of the unnamed columns, rank and “property\_valuation” columns are positively co-related.
* “post\_code” and “property\_valuation” columns are showing negative relation.
* ***Recommendation:***
* Label the unnamed columns and provide with additional information.
* In “property\_valuation” column some values seem to be an outlier and need to be rechecked whether it is a genuine input or a technical glitch and also provide us with more information about the column

1. **CustomerDemographic dataset analysis:**

* Column name “default” is having un recognizable values.
* Summary Statistics of columns where records were missing:

|  |  |  |
| --- | --- | --- |
| Column name | Count | Percentage |
| last\_name | 125 | 3.125 |
| job\_title | 506 | 12.650 |
| DOB | 87 | 2.175 |
| job\_industry\_category | 656 | 16.5 |
| default | 302 | 7.55 |
| tenure | 87 | 2.175 |

* Column “gender” is having different labels for the same categories.
* In “DOB” column, person name ‘Jephthah Bachmann’ is having birth year 1843

Customer Id of that person is 34.

* ***Recommendation:***
* Values in “default” column is either due to technical error or database software incompatibility issues (if using one). Need to contact IT department for software incompatibility issue.

1. **CustomerAddress dataset analysis:**

* The “state” column is having different labels for the same categories.
* Column “post\_code” and “property\_valuation” is showing negative correlation.
* **Lack of completeness of certain columns across the table**

Various columns have empty values in certain records. Summary statistics of records having empty values are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Table name | Column name | Count | Percentage |
| Transaction | online\_order | 360 | 1.8% |
| Transaction | brand | 197 | 0.985 |
| Transaction | product\_class | 197 | 0.985 |
| Transaction | product\_size | 197 | 0.985 |
| Transaction | standard\_cost | 197 | 0.985 |
| Transaction | product\_first\_sold\_date | 197 | 0.985 |
| NewCustomerList | last\_name | 29 | 2.9 |
| NewCustomerList | job\_title | 106 | 10.6 |
| NewCustomerList | DOB | 17 | 1.7 |
| NewCustomerList | job\_industry\_category | 165 | 16.5 |
| CustomerDemographic | last\_name | 125 | 3.125 |
| CustomerDemographic | job\_title | 506 | 12.650 |
| CustomerDemographic | DOB | 87 | 2.175 |
| CustomerDemographic | job\_industry\_category | 656 | 16.5 |
| CustomerDemographic | default | 302 | 7.55 |
| CustomerDemographic | tenure | 87 | 2.175 |

***Mitigation***: If only a small number of rows are empty, filter out the record entirely from the training set for prediction. Else, if it is a core field, impute based on distribution in the training dataset.

***Recommendation***: Recheck the process of constructing the database. Missing data having very less percentage (1% or 2%) can be removed from training dataset.

* **Inconsistent of values of certain columns across table**

(e.g. “standard\_cost” column in Transaction table, 99 percent of the data lies below 1610.90, except some rows values 1759.85 which counts to 1 percent)

***Mitigation***: Some values seem to be an outlier and need to be rechecked whether it is a genuine input, indicator of empty space or a technical glitch.

***Recommendation***: See whether all the values lies in accepted range.

Please refer to excel file ‘data\_outliers.xlsx’ for the list of outliers between tables.

* **Inconsistent in labels same categories of certain columns across table**

(e.g. gender column in CustomerDemographic table have ‘F’ and ‘Female’ for female category)

***Mitigation***: Remap the labels of the categorical columns.

* **Relevant columns showing strong correlation**

(e.g. Pearson correlation coefficient is 0.55 of

column “list\_price” and “standard\_cost” in Transaction table)

***Mitigation***:Check for derivatives or duplicated columns.

***Recommendation***: Plot graphs between columns to better visualise the output.

* **Unexplained columns**

Some columns were unnamed or the values were un explainable

(e.g. column ‘default’ in CustomerDemographic table)

***Mitigation***:Label the unnamed columns, check for unexplainable data and provide with additional information.

***Recommendation***:Could be either due to technical error or database software incompatibility issues (if using one). Need to contact IT department for software incompatibility issue.

* Invalid records of certain column

(e.g. Jephthah Bachmann birth year 1843 in DOB column in CustomerDemographic table)

***Mitigation***: Reach out to the customer for valid information. Or try to contact domain expert for guidance.

Moving forward, the team will continue with the data cleaning, standardisation and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Kind regards,

Prashant Lal