***Task 1 : Data Quality Assessment***

**Data Quality Assessment**

**Assessment of data quality and completeness in preparation for analysis.**

Dear Manager,

Thank you for providing the opportunity and the dataset from Sprocket Central Ltd. As we had assessed the data quality and completeness for analysis, and we found multiple quality issues that need to be addressed. Also, we have suggested recommendations to mitigate the quality issues and improve the effectiveness of the data.

**Data Quality dimensions:**

|  |  |
| --- | --- |
| Accuracy | If a customer’s age is 32, but the system says she’s 34, that information is inaccurate. |
| Completeness | You might make a customer’s middle name optional, but as long as you have the first and last name, the data is complete. |
| Consistency | If your human resources information systems say an employee doesn’t work there anymore, yet your payroll says he’s still receiving a check, that’s inconsistent. |
| Timeliness | You need financial information every quarter; if the data is ready when it’s supposed to be, it’s timely. |
| Uniqueness | Data duplication is a frequent occurrence. “Daniel A. Robertson” and “Dan A. Robertson” may well be the same person. |
| Validity | Many systems ask you to enter your birthday in a specific format, and if you don’t, it’s invalid. |
| Relevancy | Relevant data for Analysis. |

**Now when we have our dimensions, we found some insights that would be helpful for further and better analysis.**

**FINDINGS:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Accuracy | Completeness | Consistency | Timeliness | Relevancy | Uniqueness | Validity |
| Transactions |  | customer ID (3501-4999), online order, brand, product line/class/size, standard cost, first sold date.  **(BLANK VALUES)** | list price, standard cost |  | Filtered: cancelled orders |  | Product sold date. (Not in Date format),  List Price (Missing currency) |
| New Customer |  | last name, DOB, job title, job industry.  **(BLANK VALUES)** | Property Valuation. (some values are in decimals) |  | Irrelevant:  unnamed columns | Duplications:  customers |  |
| Customer Demographics | DOB | last name, job title, job industry, tenure  **(BLANK VALUES)** | Gender | Deceased customer | default column |  |  |
| Customer Address |  |  | State |  |  |  |  |

***Recommendation:***

1. ***Accuracy issues –***

ISSUE – We have Dob Column having inaccurate values.

* We need to filter out the outliers in DOB.
* Create an age column for more comprehensible data and to allow checking for errors.

1. ***Completeness* –**

ISSUE – We have Many columns with blank/Missing values.

* Weightage of missing values in datasets is low as compared to whole dataset we can go proceed by removing them.
* Only complete data will be used. If the data is not in-sync across all spreadsheets, data analysis with incomplete data may skew the results.

1. ***Consistency –***

***ISSUE –*** *We have**a* list price and standard costs were in inconsistent formats, *State column that contained abbreviation for same duplicated value, Gender column that contained abbreviation for same duplicated value for Male, Female.*

* Skewed data cannot be used if there are any missing records.
* By providing dropdown options, you can standardize data entry, reducing inconsistencies and human errors caused by manual input from different personnel.
* This approach enhances data interpretability and readability, making it easier to draw meaningful insights.
* When dealing with sensitive attributes like gender, it's essential to be respectful. Alongside traditional gender categories, include an "Others" option represented by 'U' for individuals who don't identify within the binary gender spectrum.

1. ***Timeliness –***

***ISSUE –*** *We have**a column which is unnamed, default column which has irrelevant values.*

* After filtering out the customers marked as deceased, it is essential to ensure the accuracy of this information.
* Timely updates will help maintain the data's reliability and prevent any potential errors or misinterpretations in future analyses.

1. ***Relevancy –***

***ISSUE –*** *We have**a column which is unnamed, default column which has irrelevant values.*

* To improve data comprehensibility, it is advisable to remove/reformat any unclear meta-data.
* By simplifying the data and focusing only on relevant information, the analysis will become more straightforward and effective in making meaningful conclusions.

1. ***Uniqueness –***

***ISSUE –*** *We have**duplicated customer id.*

* Ensure each customer record includes first name and last name, with a unique ID that is consistent.
* Ensure data uniqueness by removing duplicates, and validating data entry to maintain the integrity of the dataset. Uniqueness guarantees accurate analysis and prevents inconsistencies in the results.

1. ***Validity –***

***ISSUE*** *– We have*Product sold date column that is (Not in Date format), List Price column that has (Missing currency).

* For accurate data analysis with currency values and a proper datetime format, ensure all columns have consistent formats.
* Use standardized currency codes and validate data for correctness, handle missing values appropriately, and utilize data cleaning tools when necessary.
* Properly formatted data enhances the reliability and clarity of the analysis, supporting well-informed decision-making processes.

The above summary highlights the primary data quality issues identified during the initial analysis stage. As we proceed, the team will engage in stages 2 and 3, focusing on the extraction, transformation, and load processes for model analysis.

Please let us know if you have comments or questions on the above as I would be happy to discuss to ensure that all assumptions made align with Sprocket Central Ltd.'s understanding. Your input is valuable to maintain the accuracy and reliability of our analysis.

**Regards,**

Saksham Mishra.

Data Consultant*, KPMG*