|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | ACCURACY | Completeness | Consistency | Currency | Relevancy | Validity | Uniqueness |
| Dataset 1  (Customer Demographics) | Erroneous: DOB | Missing: last name, job title, job industry, tenure | Format: Gender | Out of date: deceased customer | Unusable column Default: deleted | Unreliable: Age |  |
| Dataset 2  (Customer Address) |  |  | Format: States |  |  |  |  |
| Dataset 3  (Transactions) | Profits (Price- Cost) column not there | Missing: Customer Id (3501-4999), online order, brand, product line/class/size, standard cost, first sold date | Format: list price and standard cost |  | Cancelled Order | Format: Product sold date |  |

Data Analyses: Data Quality Check Update

Dear Manager,

Thanks for providing the required datasets of Sprocket Central Ltd.

The table below provides a brief summary of some of the key data quality issues we have found in the data cleaning process. Please let us know if you have any query regarding the following issues.

In depth data quality issues

In the following I’ll discuss the in greater detail about the data quality issues that we have discovered and the strategies used to mitigate the same, along with the recommendations as to how to improve the accuracy of the data sources to avoid data quality issues in the future. This will result in the improvement in the accuracy of the dataset hence helping in the future business decisions.

We have evaluated the datasets according to the data quality framework as follows:

* Accuracy- correct values
* Completeness- data field with values
* Consistency- values free from contradiction
* Currency- values up to date
* Relevancy- data item with value meta-data
* Validity- data containing allowable values
* Uniqueness- duplicated values

***Accuracy***

Dataset 1 has an incorrect DOB value for e.g. a data entry indicates the customer to be born in 1843.

Mitigation: Removed the outlier value

Recommendation: An age column is more suitable for a more thorough data and for checking of possible errors.

***Completeness issues***

In dataset 3, the customer ids from 3501-4999 are missing while the same ids are there in the other two datasets.

Mitigation: I have included customer ids from 1-3500 only.

In dataset 1 some records of last name, job title, job industry, DOB and tenure are missing.

In dataset 3, some records of the online order, brand, product line, product class, product size, standard cost and first sold date are missing.

Mitigation: I have used forward/backward filling method to fill in the missing values. The average of the group is also used to fill in where appropriate.

Recommendation: Providing drop-down options for job title, online order and brand column. Convert the first sold date into a standard format. The use of pre-defined options will allow for more complete data.

***Consistency***

For dataset 1, the gender was in inconsistent formats.

For dataset 2, the state was in inconsistent formats.

For dataset 3, the list price and standard costs were in inconsistent formats.

Mitigation: Replaced variations of Men under the category of ‘Male’, and similarly all the variations of Women under ‘Female’. Similarly, the names of states replaced to abbreviations, e.g., ‘Victoria’ to ‘VIC’ to ensure consistency across the datasets.

Recommendation: Using dropdown options for categorical data like gender can reduce the inconsistencies to a large extent. It will improve interpretability and inconsistencies due to manual entries can be reduced.

### Timeliness issues

In dataset 1, a few customers were reported as deceased so they are not current customers.

Mitigation: Filtered out the deceased marked customers.

### Relevancy issues

For dataset 1, there was a default column with corrupted data.

For dataset 3, the order status showed cancelled orders.

Mitigation: Dropped the corrupted data columns, filtered out cancelled order status.

Recommendation: Cancelled order status may be ignored if it is not relevant to the analysis.

### Validity issues

For dataset 1, No age column.

For dataset 3, the product sale date is in integer format and not in Date Time format.

Mitigation: I standardised the product sale date and converted list price to currency format.

Recommendation: Ensure that all datasets are from the same time period, otherwise any duplicate or missing data records may skew the data analysis.

***Other Data Quality issues***

There were many missing datapoints across various columns.

Some of the data was out of sync having mismatch between datasets.

Inconsistent datatypes were used for same attributes.

Mitigation: I have filed the missing data with appropriate statistical models if the number of null values are small. If the null value is significant, I have dropped the feature from the datasets. The only exception I made was if the sample size is small and the datapoints are critical.

The above summarises the key data quality issues discovered through our first data analyses stage. The team will continue to extract, analyse, transform and load the dataset for the purpose of model analyses.

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 applied align with Sprocket Central Ltd.’s understanding.

Kind regards

Ashwani

Data Consultant, KPMG\*