**Re: Data analysis – Data cleaning update**

Dear Manager

Thank you for providing us with the four datasets from Sprocket Central Ltd.

The summary table below highlights key data quality issues we have discovered in the data cleaning process. Please let us know if you have any queries concerning the issues presented.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Completeness** | **Consistency** | **Timeliness** | **Relevancy** | **Uniqueness** | **Validity** |
| (Dataset 1)  Customer Demographics | Erroneous: DOB | Missing: last name, job title, job industry, tenure | Format: gender | Out of date: deceased customer | Corrupted:  default column deleted |  | Unreliable:  age |
| (Dataset 2)  Customer Address |  |  | Format: state |  |  |  |  |
| (Dataset 3)  Transactions | Missing:  profits (cost - price) | Missing:  customer ID (3501-4999), online order, brand, product line/class/size, standard cost, first sold date | Format:  list price, standard cost |  | Filtered: cancelled orders |  | Format:  product sold date |
| (Dataset 4)  New Customer |  | Missing: last name, DOB, job title, job industry |  |  | Irrelevant:  unnamed columns | Duplications:  customers |  |

In the following, I set out a more in-depth description of the data quality issues we have discovered and the strategies used to mitigate any data inconsistencies, along with recommendations and explanations as to how to improve the accuracy of the data sources to avoid data quality issues in the future. This will improve the accuracy of the data available to inform any future business decisions.

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

* Accuracy – correct values
* Completeness – data fields with values
* Consistency – values free from contradiction
* Timeliness – values up to date
* Relevancy – data item with value meta-data
* Uniqueness – records that are not duplicated
* Validity – data containing allowable values

### Accuracy issues

In dataset 1, the DOB is inaccurate, e.g. a data entry indicated a customer was born in 1843. An age column could be derived to improve comprehensibility.

Mitigation: I filtered out the outliers in DOB.

Recommendation: Create an age column for more comprehensible data and to allow checking for errors.

### Completeness issues

In dataset 3, the customer IDs are inconsistent among all datasets, with some customer IDs being null.

Mitigation: I included only customer ID from 1-3500.

Recommendation: Ensure tables are up to date. 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. To prevent future occurrences, it is advised to cross-check spreadsheets for completeness.

In dataset 1 and dataset 4, some records of last name, job title, job industry (as well as tenure and DOB, respectively) 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 filled in the missing null values using forward/backward filling methods or an average of the cluster group where appropriate.

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

### Consistency issues

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: I found and replaced variations of Men under the category of ‘Male’, and all variations of Women under ‘Female’. Similarly I replaced the names of states to abbreviations, e.g. ‘Victoria’ to ‘VIC’ to ensure consistency across the datasets. The prices were restricted to 2 decimal places for the currency $.

Recommendation: To avoid different representations of the same value, the data type should be categorical rather than a variable text field. Dropdown options minimise inconsistencies and human error in manual entries by different personnel and improves the data interpretability and readability. As gender is a protected characteristic, those identified as others may fall under the category of ‘U’.

### Timeliness issues

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

Mitigation: I filtered out the customers marked as deceased.

Recommendation: It may be difficult to verify this information, but where available this information should be updated as soon as possible.

### Relevancy issues

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

For dataset 3, the order status showed cancelled orders.

For dataset 4, there were several hidden columns.

Mitigation: I dropped the corrupted data columns, filtered out cancelled order status and dropped the hidden columns showing intermediate calculations.

Recommendation: Remove or reformat any incomprehensible meta-data to make it comprehensible. Cancelled order status may be ignored if it is not relevant to the analysis.

### Uniqueness issues

For dataset 4, some of the customers may be duplicated

Mitigation: Some records were filled in in a way which suggests that they relate to the same customer, e.g. same customer ID with the same first name but missing last name.

Recommendation: Ensure each customer record includes first name and last name, with a uniquely identifiable ID that is consistent.

### Validity issues

For dataset 1, there is currently no age column.

For dataset 3, the product sale date is an integer which may cause confusion.

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. The datasets have been merged to a single master dataset in a consistent format.

### Other data quality issues

There were many missing datapoints across various features/columns.

Some of the data was also out of sync, i.e. there was some mismatch between the datasets.

Inconsistent data types were used for the same attributes, e.g. integer for some fields and float for others which can introduce unintended bugs due to discrepancy in precision.

Mitigation: If the number of null-value is small, I have filled the records using appropriate statistical methods. Otherwise, if the number of null-value is significant, the records have been dropped from the master datasets. The only exception I made was if the sample size is small and the datapoints are critical. This achieved a standardisation of all fields to achieve constraints on the permitted data types.

The above summarises the key data quality issues discovered through the first, data quality analysis stage. Moving forward, the team will continue with the extraction, transformation and load process for the purposes of model analysis in stages 2 and 3.

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

Man Kit Yip

Data Consultant, KPMG\*

\*Disclaimer: This is a hypothetical scenario in the role of a data analysis consultant.