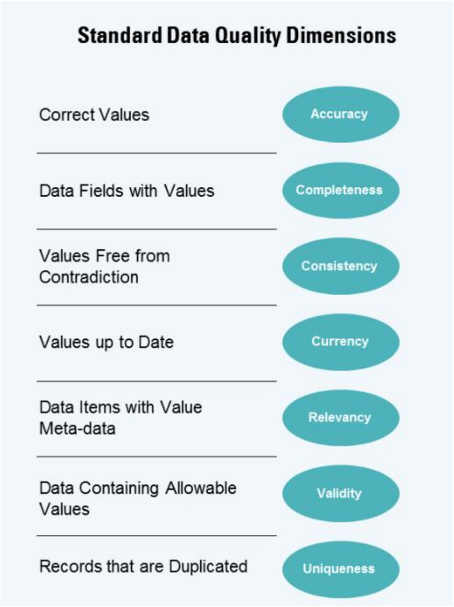
Re: Data analysis – Data cleaning update

Dear Manager

I have gone through the datasets provided by Sprocket Central Pty Ltd and here are some of the observations I would like to bring to your attention regarding the data quality.
Please find the summary as per the list of the Data Quality dimensions, we follow to evaluate a dataset.

Dataset	Accuracy	Completeness	Consistency	Currency/ Recency	relevancy	Validity	uniqueness
(Dataset 1) Transactions		Missing values in: online_order , brand, product_line, product_class, product_size, standard_cost, product_first_sold_date			Include: Quantity , Purpose unknown: cancelled orders	Format $: list_price and standard_cost , Incorrect format: product_first_sold_date	
(Dataset 2) Customer Demographics	Out of range: DOB	Missing values in: last_name, DOB, job_title, job_industry_category , default, tenure	Format: gender	Update: deceased_indicator	Invalid data: default, Include: Age Purpose unknown: deceased_indicator		
(Dataset 3) Customer Address			Format: state				
(Dataset 4) New Customer		Missing values in: last_name, DOB, job_title, job_industry_category			Irrelevant: unnamed columns, Include: customer_id		Include: customer_id

Below is a list of the Data Quality dimensions I followed to evaluate the dataset:



Standard Data Quality Dimensions

Correct Values	Accuracy
Data Fields with Values	Completeness
Values Free from Contradiction	Consistency
Values up to Date	Currency
Data Items with Value Meta-data	Relevancy
Data Containing Allowable Values	Validity
Records that are Duplicated	Uniqueness

Detailed description of data quality issues is given below. I have also included strategy to address data inconsistencies and recommendations to improve the accuracy of data and mitigate quality issues in future.

Accuracy

In Customer Demographics a data entry indicates a customer was born in 1843 this might be inaccurate.

Strategy employed:

I assumed that the year is entered incorrect and changed the year to 1943.

Recommendation:

Create a dropdown to select the birthday, this maintains the format and reduces error. Also, create an auto fill for age for cross verification.

Completeness

In Transactions dataset there are missing values in columns online_order, brand, product_line, product_class, product_size, standard_cost, product_first_sold_date.

In customer demographics there are missing values in columns last_name, DOB, job_title, job_industry_category, default, tenure.

The customer_id is inconsistent between Transaction's dataset, customer demographics and customer address dataset. Only customer_id from 1 to 3500 are consistent between these datasets.

In new customer dataset there are missing values in columns last_name, DOB, job_title, job_industry_category. Also, new customers dataset is not having customer_id.

Strategy employed:

Replaced most of the missing values in online_order with group mode of online_order when grouped by customer_id and rest to mode of online_order (True/ 1).

Similar methods are employed to treat missing values.

Recommendation:

Using drop down or predefined values will reduce manual error. It is suggested not to have any missing values in the dataset. Also, consistency has to be maintained across all column, this will reduce induction of bias into data.

Consistency

In Customer Demographics, there are variations in gender column.

In Customer Address, there are variations in state column.

In transactions, the list_price and standard_cost variables are in inconsistent formats.

Strategy employed:

Replaced variations in gender column such as F with Female and M with Male in gender column.

Similarly for state, I replaced the names of states New South Wales to NSW, Victoria to VIC.

For list_price and standard_cost variables I remove the symbol and convert the column to currency.

Recommendation:

Using drop down or predefined values will reduce manual error and improves the data interpretability and readability. Gender column has to be more inclusive i.e unidentified, transgender, others etc.

Currency/ Recency

In Customer Demographics dataset there is a deceased_indicator, the values in this column have to be verified for accuracy and updated regularly.

Strategy employed:

Deleted from Customer Demographics but stored the details of deceased customers into a separate dataset which can be used if required.

Recommendation:

It is sensitive to verify this information, if we wish to capture this information it has to be updated regularly.

Relevancy

In transactions dataset there are details regarding cancelled orders, purpose of this column is unknown.

In customer demographics dataset, default column has invalid data. Also, there is a deceased_indicator, purpose of this column is unknown.

In new customers dataset, there are unknown columns and there is no unique identifier such as customer_id for these customers.

Strategy employed:

Deleted from transactions but stored the details of cancelled orders into a separate dataset which can be used if required.

Deleted from customer demographics but stored the details of deceased customers into a separate dataset which can be used if required.

Deleted default column with invalid data from customer demographics dataset.

Deleted unnamed columns from new customers dataset assuming that the data from these columns is captured in other columns.

Recommendation:

Include quantity in transactions data to capture the quantity of products purchased.

Include autofill age column in customer demographics based on the birthday for cross-verification.

Include customer_id for new customers.

It is sensitive to verify deceased information, if we wish to capture this information it has to be updated regularly.

Suggested to check the data set for hidden columns and calculations to avoid loss of data.

Validity

In transactions, the list_price and standard_cost variables are in inconsistent formats. Also, product_first_sold_date is not in date format.

Strategy employed:

For list_price and standard_cost variables I remove $ symbol and converted column to currency.

Converted product_first_sold_date in transactions to date format.

Recommendation:

Using drop down or predefined values will reduce manual error and improves the data interpretability and readability. Gender column has to be more inclusive i.e unidentified, transgender, others etc.

It is important to maintain dates in all datasets in a standard format.

Uniqueness

So far, transactions_id for transaction dataset, customer_id for customer data set are unique meaning there are no duplicate entries.

In new customers dataset, there is no unique identifier such as customer_id for these customers.

Recommendation:

Suggest to restrict manual intervention to the unique identifiers such as transactions_id for transaction dataset, customer_id for customer data.

Include system generated unique value as customer_id for new customers.

I have merged the customer demographics dataset and customer address dataset into a single dataset.

For most of the cases missing value treatment is done as felt appropriate. But for entries where most of the data is missing it is deleted from the master dataset and saved into a separate dataset which will be used if the data is available.

In addition to above mentioned issues changes have been done to the data types of columns as needed. I have observed that transactions dataset has transaction details for old customers (customers with customer_id) and very few transactions by customers whose details are not available. Such entries are again stored into a separate dataset and deleted from the main dataset.

I request you to provide customer_id for the list of new customers and also transactions details for the same if available.

Please feel free to get back to me in case you have any queries concerning the issues presented.

Regards,

BRC

KPMG