Dear Tony Smith

Thankyou for providing us with three data sets from Sprocket central Pty Ltd. The summary table bellow highlights key quality issues that we discovered within the three data sets. Please let us know if you have any queries surrounding the issues presented.

- **Summary table,**

	Accuracy	**Completeness**	**Consistency**	**Currency**	**Relevancy**	**Validity**
Customer Demographic	>DOB inaccurate >Age Missing	>Job title: Blanks >Customer Id: Incomplete	>Gender: Inconsistency	>Decreased: Filtered out	**>**Default column: Delete	
Customer Address		>Customer Id: Incomplete	>States: inconsistency			
Transaction	>Profit : Missing	**>** Customer Id: Incomplete >Online Orders: Blanks >Brand:Blanks			**>**Cancelled Status Order: Filter out	**>**List price:Format >Product Sold Date: Format

Bellow are more description about data quality issue and migration used. Recommendation and explanations have also been included to avoid further data quality issues in the further. Following recommendation will improve accuracy of data used to influence business decisions of Sprocket Central Pty Ltd in the future.

Accuracy issues:

- **DOB was inaccurate for "Customer Demographic" and missing an age_column; missing a profit column for "Transactions"**

  *Mitigation: Filter out outlier in **DOB***

  *Recommendation : Create an **Age_column** allowing for more comprehensible data and easier to check for errors .Create a **Profit_column** in "**Transactions"** to check accuracy of sale.*

  Creating additional columns for age and profit will allow for easier identification of errors . The **Profit_Column** will assist in future monetary analysis.

Completeness:

- **Additional customer_ids were inconsistent among "Customer Demographic" ,"Customer addresses" and "Transactions"**

  *Mitigation: Filter all **Customer_ids from 1 to 3500.***
  *Recommendation : Ensure tables are up to date (from the same time period).for our model, only **customer_ids** from 1 to 3500 will be used as they have complete data*

  The data received may not be in sync across all spreadsheets, with incomplete data the analysis results may be skewed. This is a 'completeness' issue, to prevent further occurences it is encouraged to cross check spreadsheets and sync data.

- **Blanks in job_title for " Customer Demographic" in online_order and brand_column for "Transactions".**

  *Mitigation: Filter out 'blanks' for **Job-title, online_order and brand column.***
  *Recommendation: Simplify job titles to another category such as **industry_industry** or provide dropdown options for **job_title.** Provide dropdown options for **online_order and brand_column.***

  Blanks are treated as incomplete data and can skew further analysis results. The addition of dropdown options will allow to have more complete data and will result in more accurate analysis.

Consistency:

- **Inconsistency in gender for "Customer Demographic" and "Customer Address" respectively**

  *Mitigation: Filter all 'M' under category of "Male" filter all 'Femal' and 'F' for **Gender.** Filter all 'New South Wales' to "NSW" and 'Victoria' to "VIC" for **States.***

  *Recommendation: Create Dropdown options for 'Male' 'Femal' and 'U' in **gender**. Create dropdown options for all state abbreviations.*

  Dropdown options ,minimizes manual entry and human error. Allows for increase of consistency of terminology. Gender identity can be a sensitive topic, proceed with caution when creating options.

Currency:

- **People that are ‘Y’ in decreased indicator are not current customers for “ Customer Demographic’**

  *Mitigation: Filter out customers checked ‘Y’ in **Decreased_indicator.***
  Recommendation: Can be difficult to check for decreased customers ,but once this information is received one should update data accordingly.

  Decreased customers are not current customers removing them from the data will increase currency of data and will result in more accurate estimates in future analysis.

Relavancy:

- **Lack of relevancy or comprehensibility in default_column for “Customer Demographic” and order_status for “Transactions”**

  *Mitigation: Deleted Metadata in **default_colum.** Filter out ‘Cancelled’ **order_status**.*
  Recommendation: Check for incomprehensible Metadata and delete or format to make comprehensible .

  ‘Cancelled’ order_status is irrelevant information for future analysis, as it can skew data-for example total number of customer per annum will be an overestimate.

Validity:

- **Format of list_price, Product_sale_date for “Transactions”**

  *Mitigation: Format **Product_Sale_date** to short date format **list_price** to currency.*
  Recommendation: Set up columns so that formats such as price and decimals are already in place when entering new data.

  Allowable values will make data to be interpreted more easily. Formatting into Price and allowing for either 2 or 3 decimals placed consistently will increase readability. This will reflect positively on speed and accuracy for business decisions.

That summarises all data quality issues discovered through the first stage of the data quality analysis. The mitigation strategies suggested are simple and effective ways of improving data quality for future analysis. They will not only improve the analysis output that one can perform within the company but will increase the level of analysis that can be performed by KPMG and other hired analysis teams.

Please let us know if you have questions regarding mitigation or any data quality issues identified.

Kind regards,

Hrittik Saha.