Dear [Client],

Thank you for providing us with your three files of Sprocket Central Pty Ltd’s datasets related to transaction and customer information. In this email, we will evaluate data quality through issues that the datasets encountered and strategies to minimize the risk of affecting data analysis results in phase two, as follows:

* **Inconsistent values in the same column:**
* Issue: Gender in Customer Demographic table and State in Customer Address table include incompatible records such as *“F”, “Femal”, “M”* for Male/Female and *“Victoria”, “New South Wales”* for VIC/NSW. These errors reduce the accuracy of the data set, thereby reducing the reliability in statistical analysis and model distribution analysis.
* Mitigation: Regarding Gender variable, using replace function to change all inconsistent values into *“Male” or “Female”* options. Extended versions of customers’ state addresses are going to be replaced into abbreviated versions.
* **Invalid values for analysis:**
* Issue: Validity in data quality signifies that the value attributes are available for aligning with the specific domain or requirement. Therefore, a range of invalid values in customer’s age and deceased situation were still recorded, for example, a customer aged 122 or two customers with deceased status.
* Mitigation: All of the above values will be deleted from the dataset because it has no value in analysing customer behaviours to make valuable business decisions.
* **Incompleteness in a range of attributes:**
* Issue: The columns, such as job title, brand of purchasing or product information have empty values in certain records. When the dataset is not sufficient or complete enough, no meaningful inferences and decisions are going to be delivered.
* Mitigation: All rows including nulls will be deleted in its entirety to create the most complete set of data
* **Inaccuracies in determining data types of attributes:**
* Issue: Various columns in the Transactions table have the wrong types of values. In particular, they are text instead of currency in List price and Standard cost, text instead of the date in the product first sold date. Such inaccuracy makes the data’s attributes worthless and no insights exploited because of wrong values that fail to interpret results at the later stage .
* Mitigation: Changing wrong types of data into the right one

Our team fixed the errors to create the best value data set. You can open the edited version [here](KPMG_Modified_Dataset.xlsx). Please let us know if any changes are not aligned with your understanding.

Best regard

Van Duong Ngo