Dear Manager,

Thank you for providing us with these datasets from Sprocket Central Pty Ltd. According to the instruction given in Task 1, we reviewed the three datasets to assess their quality based on six dimension metrics and discovered notable data quality issues that affect the further analysis. Here are the quality assessment result and the methods used to mitigate the key quality issues.

	Customer Address	Customer Demographic	Transactions
Accuracy		• DOB inaccurate(1843-12-21)	• Profit column is recommend
Completeness	• Customer ID is incomplete	• Some data points in DOB and Job title are missing values. • Age column is missing.	• Some data points in online orders, brands, product line, product class, product size, and standard cost columns consists of missing value.
Consistency	• State column is inconsistent(Victoria, VIC, New South Wales, NSW)	• Gender column is not consistent(F, Female, Femal, Male,M)	
Currency		• Decreased indicator(Y is not representable)	
Relevancy		• Default column is omitted as it consists of irrelevant data points	
Validity			• Product first sold date consists of invalid data
Uniqueness			

Accuracy

- One of the data points (1843 -12-21) in the Date of Birth column from the Customer Demographic table is inaccurate; missing profit column in the Transactions table.
- **Solution**: Filter out outliers in DOB.
- **Recommendation**: Create an age column that will allow easier identification of errors regarding the birth date and create a profit column in Transactions to check accuracy. It will also help for future analysis.

Completeness

- Customer ID column in the Customer Address table is incomplete
- Additionally, customer IDs in Customer Demographic, Customer Address, and Transactions tables are inconsistent.
- Some data points in the last name, DOB, Job title, job industry category, and tenure columns of the Customer Demographic table consist of missing values. The job industry category consists of columns n/a value (no data point).
- Some data points in online orders, brands, product line, product class, product size, and standard cost columns from the Transaction Table consists of missing value.
- **Solution**: Filter out and remove the null values from the above-stated column, taking into consideration the number of rows to be removed.
- **Recommendation**: Fill in all the missing fields if it is in the core field (for example, in the last name column, since the first name is available it's better to keep the records and fill it). Change the job title to a simplified column like industry; add a drop-down option for brands as well on your website. Product class, product size, product line, standard cost and should be allocated the correct values by the system itself.

Consistency

- State column is inconsistent (Victoria, VIC, New South Wales, NSW) in the Customer Address table.
- Gender column is inconsistent (F, Female, Femal, Male, M) in the Customer Demographic table.
- **Solution**: Filter out outliers and replace them with consistent values.
- **Recommendation**: Avoid multiple representations.

Currency (Timeliness)

- Decreased indicator column in the Customer Demographic table is not current, i.e. people that are Y are not current customers.
- **Solution:** Filter out and remove the outliers.

Relevancy

- Default column in the Customer Demographic table consists of irrelevant data points, it is just random symbols which doesn't look correct.
- **Solution:** Remove the default column as it is not necessary for further analysis.

Validity

- Product first sold date consists of invalid data.
- **Solution:** converting the selected column in the string to Date data type.
- **Recommendations:** Ensure that the fact table in the given dataset has constraints on data types in order to make appropriate data transformations**.**

For better convenience, we have also attached the data quality issue after cleaning.

Best regards,

Yordanos Alemu (Virtual intern)