Hi **@Team Sprocket Central Pty Ltd.**,
I thank you on behalf of my team for providing us with the datasets. The following summary reflects the data received by our team, please reach out to us if there are any discrepancies (or) misunderstanding:

Dataset	Total No. of Records	Unique Customer IDs
Customer Demographics	4000	4000
Customer Address	3999	3999
Transactions	20000	3494

Upon receiving the datasets, we went ahead and analyzed its quality as per standard data quality framework procedure prescribed at KPMG AU. We have identified the underlying data quality problems in the data, we have made necessary changes and cleaned the data to account for further analysis. We hope that the necessary changes made can be further incorporated during the ETL process, prior to warehousing on your side. The following table summarizes the issues we came across during our initial analysis:

		Customer Demographics	Customer Address	Transactions
Accuracy	Correct Values.	● DOB ● job_industry_category		
Completeness	Data fields with values.	The given columns had NULL/NaN values. ● DOB ● job_title ● job_industry_category ● tenure	customer_id are not in sync with all three datasets.	The given columns had NULL/NaN values. ● online_order ● Brand ● product_line ● product_class ● Product_size ● standard_cost ● product_first_sold_date
Consistency	Values free from contradiction.	● gender	● states	
Currency	Values upto Date.	Deceased customers need to be excluded.		
Relevancy	Data Items with Value meta-data.	*default field can be excluded.		

Validity	Data containing allowable values.			product_first_sold_date - data type conversion to datetime format.
Uniqueness	Records that are Duplicated.	No duplicate values		

Detailed description:

1. **Accuracy Issues:** customer_id : 34 has a DOB : ‘1843-12-21’ making the customer over 176 years of age, making the value one of the DOB outliers in the customer demographics dataset. Few of the job_industry_category in the same data are misspelled.
   Mitigation: Replace the year of birth with suitable data (or) drop the rows if age is an important feature in further analysis.
2. **Completeness Issues:** Multiple columns in the *Customer Demographics, Transaction* dataset contain missing values. Also, the customer_id columns are not in sync with the other given datasets.
   Mitigation strategy: impute necessary datas with possible values (or) drop columns that have incomplete features.
3. **Consistency Issues:** gender column in customer demographics and states column in the customer address dataset needs to replace the typos and abbreviations with proper data.
4. **Currency Issues:** Deceased customers in the dataset should be filtered out.
5. **Relevancy Issues:** default feature in the customer demographics data should be filtered out as it is not relevant and the feature is not suitable for further analysis.
   Also the country column in the customer address should be dropped as all the customers data originate from Australia.
6. **Validity Issues:** In the transactions dataset product_first_sold_date needs to be in datetime datatype, few columns need to be reassigned with categorical datatype.
7. **There were no duplicates in the datasets provided to us.**

We hope that the initial findings were insightful, hoping to hear from your team soon on further analysis.

Regards,
Subhrajit Guchait.