Hi Sprocket Central Pty Ltd ,

myself , Yogesh Gavali from KPMG team. I am providing your esteemed company the details about the quality of the data which has been shared.

The data has been assessed carefully and the following mitigations are proposed. Please have a look and lets us know if there are any concerns .

Before proceeding further , the preliminary assessment of the datasets is as follows:

	Transactions Data	**Customers data**
Number of Observations	Transactions **:20000**	Customer Demographic : **4000** Customer Address : **3999**
Number of columns	Transactions :**13**	Customer Demographic : **13** Customer Address : **6**

Based on the preliminary assessment the Customer data can be merged into one by combining existing and new customer list as :

Customer Data:

i. The New dataset 'Customer' is created by merging 'CustomerDemographic' and 'CustomerAddress'.

Transaction Data :

i. We can see that customer_id has ranging from 1-3500. An outlier of customer_id of '5034'.

To handle null values the missing rows are imputed with 'Unknown' or the rows are removed as they contribute less than 1% of total data.

		Customers Data	*Transactions data*
Missing values	**Observed**	**1.** last_name 125 **2.** DOB 87 **3.** job_title 506 **4.** job_industry _category 656 **5.** tenure 87 **6.** address 4 **7.** postcode 4 **8.** state 4 **9.** country 4 **10.** property_valuation 4	**1.** online_order 360 brand **2.** 197 product_line 197 **3.** product_class 197 **4.** product_size 197 **5.** standard_cost 197 **6.** product_first_ sold_date 197 **7.**
	Mitigation	**1.** last_name Filled with 'Unknown' **2.** DOB converted to age, age_group	**1.** online_order Dropped null rows **2.** brand Filled with 'Unknown' **3.** product_line Filled with 'Unknown' **4.** product_class Filled with 'Unknown'

Quality Check :

		and imputed with mode value. 3. job_title Filled with 'Unknown' 4. job_industry_category Filled with 'Unknown' 5. tenure Filled with mean value 6. address Filled with 'Unknown' 7. postcode Dropped null rows 8. state Dropped null rows 9. country Filled with 'Unknown' 10. property_valuation Dropped null rows	5. product_size Filled with 'Unknown' 6. standard_cost Dropped null rows 7. product_first_sold_date Dropped null rows
Outlier values	**Observed**	1. DOB has outlier of '1843-1221'	1. customer_id has value of 5034
	Mitigation	1. This is row is removed.	1. This is row is removed.
Inconsistence values	**Observed**	1. State – (NSW,VIC,QLD, New South Wales, Victoria) 2. Gender- ('F', 'Femal', 'M', 'U')	
	Mitigation	1. Replaced 'M': 'Male', 'Femal' : 'Female', 'F' : "Female", 'U' : "Unknown" 2. Replaced 'New South Wales': 'NSW', ' 'Victoria': 'VIC'	
Inconsistence data type	**Observed**		1. product_first_sold_date is in float64 type
	Mitigation		1. product_first_sold_date is converted to datetime
Columns dropped	**Observed**	1. Default 2. deceased_indicator	
	Mitigation	1. These columns are dropped as they are used in calculating 'Rank' column.	

		2. 'default' Column contains metadata and does not provide any additional information 3. deceased_indicator is dropped as we have 2 rows in Y category , It doesn't provide much insight.	
Duplicated Values	**Observed**	**1.** No duplicate rows	**1.** No duplicate rows
	Mitigation		

Please provide the acknowledgement for the assumptions taken while this data quality assessment so we can proceed for the data cleaning and further analysis required for modelling.

Regards,

Yogesh Gavali

(KPMG team)