Data Quality Assessment

CustomerDemographic		
	Issue	Proposed action to be taken
Accuracy	n/a	
Completeness	columns/fields last_name, DOB, job_title, job_industry_category and tenure are incomplete (125, 87, 506, 656, 87 empty cells respectively). Also, it doesn't have address, postcode, state, country and property_valuation columns	Data should be provided where needed and if the needed data isn't available, then we need to agree to fill text empty cells with "n/a" or "unknown" and value empty cells with the mean for that column. Missing columns can be looked up and filled using the vlookup function.
Consistency	n/a	
Currency	n/a	
Relevancy	default field is not relevant has it is not consistent across the dataset	We need to get approval to delete the default column
Validity	gender column has some ambiguous categories and misspellings (Femal, F, M, U). Kindly specify what U stands for and whether it fits into the gender categories	We need approval to maintain (Male and Female) categories and also U category needs to be defined
Uniqueness	customer_ids are unique as no duplicates were found	

Transactions		
	Issue	Proposed action to be taken
Accuracy		
Completeness	online_order, brand, product_line, product_class, product_size, standard_cost, and product_first_order_date have empty cells (360, 197, 197, 197, 197, 197, 197 respectively).	Data should be provided where needed and if the needed data isn't available, then we need to agree to fill text empty cells with "n/a" or "unknown" and value empty cells with the mean for that column
Consistency	It was observed that brand, product_line, product_class, product_size, standard_cost, and product_first_order_date are empty across their respective rows	
Currency		
Relevancy		
Validity	product_first_sold_date does not match date format/type of our dataset.	We need to confirm if those are actual dates, and if they are, actual dates should be provided or converted to the right date format
Uniqueness		

CustomerAddress		
	Issue	Proposed action to be taken
Accuracy		
Completeness	Data is not complete as customers with ids (3,10,22 and 23) are not found in the list. This will make looking up certain details like address for customers difficult	A complete list with customers id needs to be provided
Consistency	Inconsistent values were found in the state attribute (New South Wales, NSW, Victoria, VIC)	We need to decide whether to go with the abbreviations or not. And if we choose to go with the abbreviations, then QLD has to be defined in full
Currency		
Relevancy		
Validity		
Uniqueness	in the address column, it was observed that six different customers with unique ids have same address in common with different post codes and states. These are the duplicate address (3 Talisman Place, 64 Macpherson Junction, 3 Mariners Cove Terrace and states QLD, NSW and VIC respectively)	

Cleaning Process

Transaction Table		
Attribute	Issue	Action taken
Date	I needed measurable date data	Created new features Month and weekday_name
New Customer List Table		
Attribute	Issue	Action taken
past_3_years_bike_related_purchases, postcode, property_valuation	Attributes past_3_years_bike_related_purchases, postcode, property_valuation were in text format when they are clearly values	Used the "Value" function to convert all three attributes to values
DOB	Not measurable	Created a new feature age to show age of customers
DOB	17 empty cell found	Repaced empty cells with N/A
Age	Value error	Used IFERROR function to make value error N/A
gender	U not a known gender category	Filter U out of the data set

Customer Demographic		
gender	Inconsistent values for the same attribute (Female, F, Femal, Male and M)	Replaced Femal and F with Female, while M to Male
DOB	Not measurable	Created a new feature age to show age of customers
default	Attribute appeared to be irrelevant to the dataset	Deleted the attribute from the dataset
Customer Address		
state	Inconsistent values for the same attribute (Victoria, VIC, NSW, New South Wales)	Replaced New South Wales with NSW and Victoria with VIC