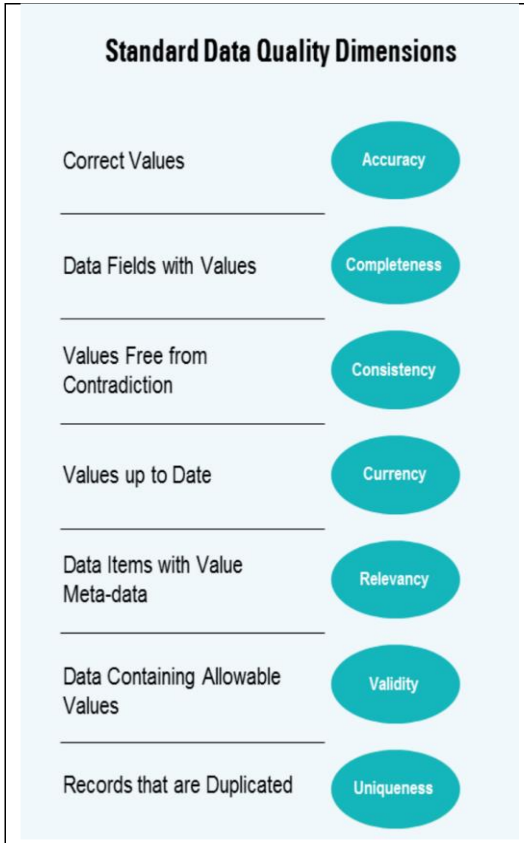
Task 1: Checking data first

Dear Manager,

Sprocket Central Pty Ltd,

After reviewing the datasets provided by Sprocket Central Pty Ltd, some data quality issues were encountered. The methods used to mitigate the inconsistent data are as follows:

Summary table

 Standard Data Quality Dimensions Correct Values Accuracy Data Fields with Values Completeness Values Free from Contradiction Consistency Values up to Date Currency Data Items with Value Meta-data Relevancy Data Containing Allowable Values Validity Records that are Duplicated Uniqueness		**Customer Demographic**	**New Customers List**	**Transactions**	**Customer Address**
	Accuracy	DOB: inaccurate			
	Completeness	DOB,Last Name, Job title, tenure: Blanks Customer_id: incompleteness	DOB, Last Name, Job title: Blanks job_industry_category: n/a	Online order, Brand, Product line, class, size; Product sold date: blanks	Customer_id: incompleteness
	Consistency	Gender: Inconsistent	DOB: Format	Standard Cost: Format	State: Inconsistent
	Currency	deceased_indicator:			
	Relevancy	Default:delete			
	Validity	DOB: 1843-12-21		List Price, Standard Cost, Product Sold Date: Format	State: format
	Uniqueness	No duplicate values	No duplicate values	No duplicate values	No duplicate values

Below is more in-depth description of data quality issues found and methods of mitigation used.

Accuracy issues:

- DOB was inaccurate for “Customer demographic”
- Mitigation: Filter out outlier in DOB

Completeness

- customer_id were inconsistent among “Customer Demographic”, “Customer Address”, and “Transactions”
- Mitigation: Filter all customer_ids from 1 to 4003
- Blanks in Online order, Brand, Product line, class, size; Product sold date in “Transactions” should be removed.

Consistency:

- Gender in “Customer Demographic” and State in “Customer Address” are all inconsistent.
- Mitigation: Filter all ‘M’ under category of ‘Male’, filter all ‘Femal’ and ‘F’ under ‘Female’ for gender. Filter all ‘New South Wales’ to ‘NSW’, and ‘Victoria’ to ‘VIC’ for States.
- DOB in “New Customers List”, and Standard Cost in “Transactions” need to be formatted (Mitigation).

Currency

- People that are ‘Y’ in deceased_indicator are not current customers for “Customer Demographic”
- Mitigation: Filter out customers checked ‘Y’ in deceased_indicator.

Relevancy

- Lack of relevancy in Default in “Customer Demographic” and order_status in “Transactions”
- Mitigation: Deleted metadata in Default column. Filter out “cancelled” order_status.

Validity

- Format of List Price, Standard Cost, Product Sold Date in “Transactions”
- Mitigation: format List Price, Standard Cost to currency, format Product Sold Date to date format,

Furthermore, some following recommendations have been suggested to avoid the reoccurrence of data quality issues.

- Data Validation for Gender column (in List), State (in List).
- Data Validation for Product Sold Date column (in Date)

Please let me know if you have any concerns surrounding the issues found.

Kind regards,
Thach.