Good day,

These are the data quality issues we found with the provided datasets.

* This customer\_id[5034] in the customer\_transaction table is not available in the customer\_demographic table, which means the customer details are not registered, but a transaction took place.
* The samples are not consistent. happen to find 'F', 'M', which are clear for female and male respectively. but 'U’ Is unclear where it belongs. sample distribution: {'F': 1, 'Male': 1872, 'Female': 2037, 'U': 88, 'Femal': 1, 'M': 1}
* {'1843-12-21': 1} we found a sample year with the following year, not valid, as the customer happens to be very old.
* Job\_title, there are more than 500 missing(nan) job titles in the dataset.
* Job\_industry\_category, there are a lot of missing(nan) values in the datasets, approximately ~656.
* Default, we found this column to not be useful, the data might have been encoded.
* Tenure, most of the values are missing(nan)
* Online\_order, there is some missing data in this column, sample distribution: {nan: 360, False: 9811, True: 9829}
* Brand, there are missing values in this column, approximately, 197.
* product\_line, there are missing values in this column, approximately, 197.
* Product\_class, we found missing values in this column. Sample distribution: {nan: 197, 'low': 2964, 'high': 3013, 'medium': 13826}
* Product\_size, we found missing values in this column. distribution: {nan: 197, 'small': 2837, 'large': 3976, 'medium': 12990}
* Standard\_cost, we found missing values of about 197.

Strategies to use to mitigate data quality issues:

* Data validation: by checking data so that it meets rules and criterias, like the range, format, and constraints within which the data must fall into. Example making sure input forms for collecting data have validation mechanisms. This will help to prevent getting wrong data.
* Data governance: Establish rules and policies for data quality and management to ensure that the data is consistent, accurate. This can involve tasks such as defining data quality metrics, setting data standards, and implementing data governance processes.
* Data cleansing: Clean the data to fix the data quality issues. This can involve tasks such as filling in missing data, correcting errors, standardising data, and removing duplicates.
* Data profiling: Analyse the data to identify data quality issues, such as missing data, duplicates, outliers, inconsistencies, and errors.
* Data privacy: Protect the privacy and security of the data to ensure that it is not compromised or misused. This can involve tasks such as encrypting sensitive data, implementing access controls, and complying with data privacy regulations.
* Data quality training: Train and educate data users and stakeholders on data quality best practices and procedures to ensure that they understand how to use and manage the data effectively. This can involve tasks such as providing data quality training, creating data quality guidelines, and conducting data quality awareness campaigns.
* Data monitoring: Continuously monitor the data to detect and correct data quality issues in real-time. This can involve tasks such as setting up data quality alerts, tracking data quality metrics, and performing regular data audits.

Sincerely,  
Kibambe.