Dear Sir or Madam,

Thank you for providing us with three datasets from Sprocket Central Pty Ltd. The summary table below highlights key quality issues that we discovered within the three datasets. Please let us know if you have any queries surrounding the issues presented.

We used Python Pandas package to identify the data quality issues, and making modifications.

**Summary Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Completeness** | **Consistency** | **Relevancy** | **Validity** |
| **Transactions** | 1. missing profit | 1. online order: null values 2. Brand: null values |  |  | 1. Product first sold date: format |
| **NewCustomerList** |  |  | 1. gender: inconsistency | 1. unnamed value columns: delete |  |
| **CustomerDemographic** | 1. missing age | with null values 1. last\_name 2. DOB 3. job\_title / category | 1. gender: inconsistency | 1. default: delete | 1. deceased\_indicator : drop "Y" |
| **CustomerAddress** |  |  | 1. state: inconsistency |  |  |

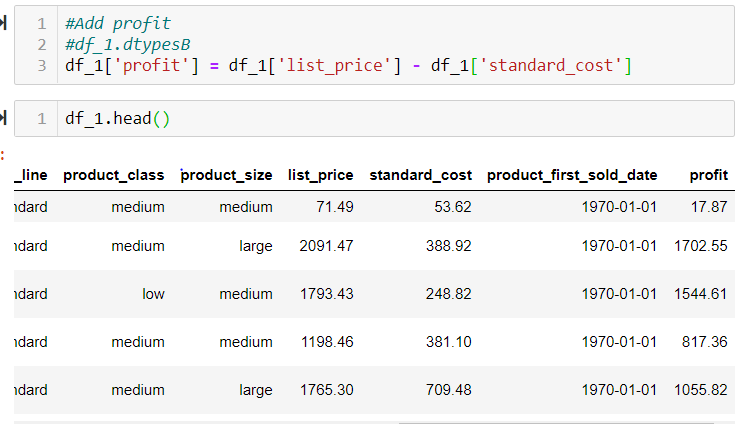
Below are more in-depth descriptions of data quality issues and methods of mitigation used. Recommendations and explanations have also been included to avoid further data quality.

**Accuracy Issues:**

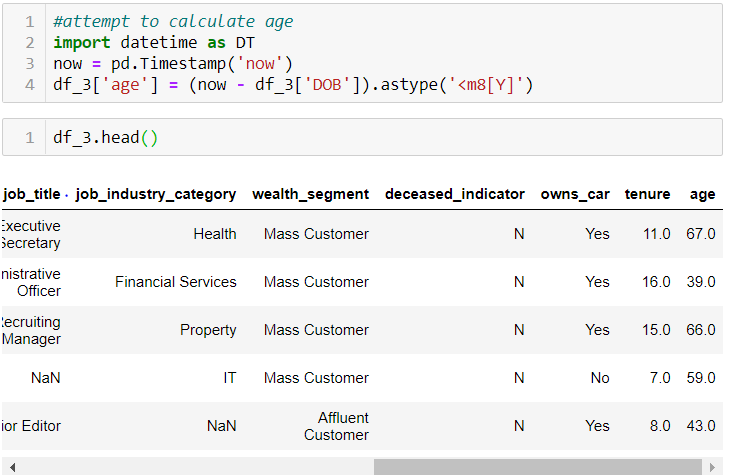
* In the Transactions table, we need a new column to record the sales profit.
* In the Customer Demographic table, we need a column for customer age.

**Accuracy Mitigations:**

* Added profit column by subtracting standard cost from list price

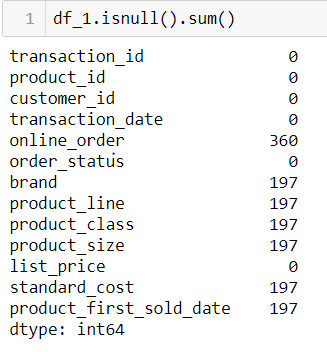


* Add age column by subtracting DOB from current date

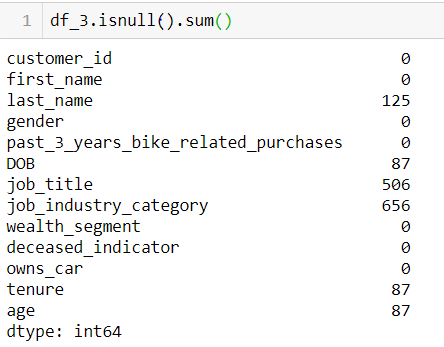


**Completeness Issues**

* In the Transactions table, null values exist in online order status, and brand / product relevant columns.



* In the customer demographic table, null values exist in last name, job title, category and tenure

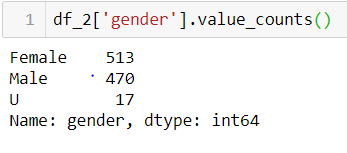


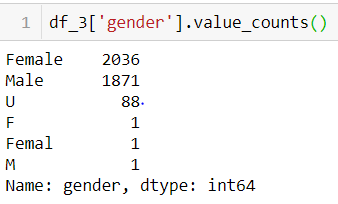
**Completeness Mitigations:**

* For transactions table, since we are difficult to find unified replacement value, we can choose to drop null values, or we need the update of data table to provide all the data information.
* For customer demographic table, we can either drop null values or update the table to ensure all information to be filled without null values.

**Consistency Issues:**

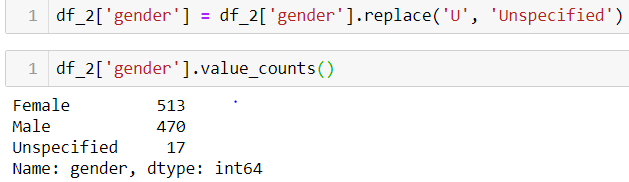
* Inconsistency in gender for NewCustomerList and customer demographic tables respectively. Details as shown in below screenshots.

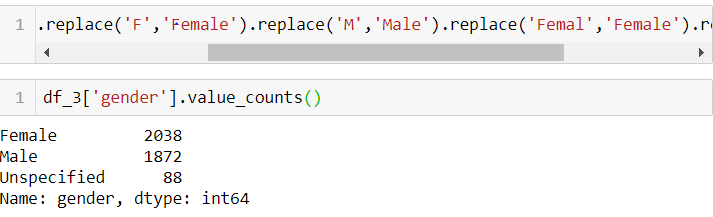




**Consistency mitigations:**

* Replace “U” with “Unspecified” for gender in New customer list. Also replace other inconsistent values in the customer demographic table.





**Relevancy Issues:**

* New customer list has 5 unnamed columns
* Customer demographic table has default column.

**Relevancy mitigations:**

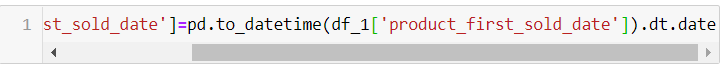
* Delete unnamed columns and the default column.

**Validity Issues:**

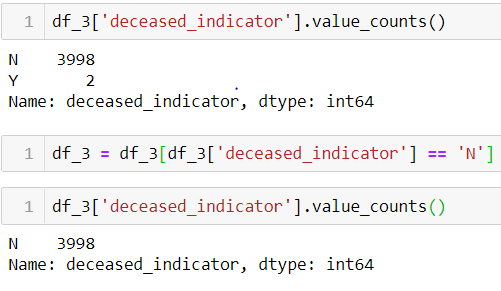
* In the transactions table, the product first sold date values has float datatype which needs to be converted to datetime.
* In the customer demographic table, the deceased indicator has 2 “Y” values.

**Validity mitigations:**

* Convert first sold date from float to datetime.



* Drop rows with “Y” in deceased indicators.



That summarises all data quality issues discovered through the first stage of the data quality analysis. The mitigation strategies suggested are simple and effective ways of improving data quality for future analysis.

Please let us know if you have questions regarding mitigation or any data quality issues identified.

Kind regards

Hang Liu