Dear Sprocket Central Pty Ltd,

Thank you for providing the datasets. The KPMG Data Analytics virtual experience team has reviewed the data sets and found several quality issues with respective recommendations to avoid recurrence of the issues.

Methods to address the data quality issues across all sheets are as follows:

Accuracy - The Date of Birth was incorrect and there is no age column for Customer Demographic dataset; Transactions does not have a profit column

* The incorrect date of birth is an outlier, filter it out as the age is too high.
* In the future, age outliers can be found with an addition of an age column.
* Adding a profit column can serve as a useful data feature.

Completeness - Some columns in the data sets have empty values: job title, brand of purchase, online order

* For columns with very few rows with missing data, filter them out from the training dataset when fitting models or predicting.
* For transactions dataset, less than 1% of rows have missing data. They should be removed.

Consistency - Data types are inconsistent; Gender values are inconsistent, States are inconsistent

* Python regular expressions can abbreviate equivalent values to create consistency across addresses
* Limit the choice of entering in data if one value means the same as another (e.g. ‘Femal’, ‘Female’, ‘F’ is all the same). This can be in the form of a drop down for users. Allow for ‘U’ or ‘unspecified’ out of gender identity respectfulness.
* Remove numeric characters or strings to keep a single data type for a feature.

Additional information - Customer Demographic does not have customer\_id but the column exists in the Transactions and Customer Address dataframes

* Cross check datasets to prevent a skew in analysis as this could just be missing data or that there is enough information to map the dataframes together.
* For now, if the data is consistent with the company’s interests, filter out the rows with the missing ids.

With these suggestions, the team will continue to clean the data for model training. The preliminary cleaning and analysis were done with Python’s Pandas library. Please let us know if the suggestions do not align with your understanding or if they raise any questions.

Best,

Brian Doan