KPMG Virtual Internship – Data Quality Assessment

Greetings,

This is Sai Chand Sunkara from KPMG virtual internship team. After reviewing the datasets provided in the data quality analysis phase, the final results in this phase are as given below,

There are some errors in the dataset, and to correct these errors the below points can be considered to improve the quality of data in each data set,

1. Transactions table

* Make 1st row as header of column
* Change datatypes of transaction\_id, product\_id, customer\_id to “int” and list\_price, standard\_cost to “float” and product\_first\_sold\_date to “datetime” format.
* There are few “NaN” values in online\_order (360), brand (197), product\_line (197), product\_class (197), product\_size (197), standard\_cost (197), product\_first\_sold\_date (197). Need to drop them as there are few or apply mean, median or mode.
* There are no duplicated values.

1. Customer demographics table

* The data given was separated, so merge NewCustomerList, CustomerDemographic and CustomerAddress tables according to customer\_id column to make the final customer demographic table.
* Change customer\_id, past\_3\_years\_bike\_related\_purchases, tenure, postcode, property\_valuation, rank, value to “int” and DOB to “datetime” datatypes
* There are “Nan” values in last\_name (154), DOB (104), job\_title (612), job\_industry\_category (821), tenure (87), address (4), post\_code (4), state, country (4), property\_valuation (4). And rank and value have 4000 “NaN” values which are generated as we merged the tables. Rank and value columns can be obtained by constructing a regression model and can be inserted. And rest all columns can be either treated with mean, median, mode or can be dropped if negligible.
* There are no duplicated values.

Regards,

Sai Chand Sunkara (KPMG virtual internship team).