RE: Sprocket Central – Data Quality Issues and Recommendations for Data Treatment

Dear Mr./Mrs. _____:

As per your request, our team had carefully assessed the dataset that you provided. Here is a summary of the data quality issues that might have negative impacts on the analysis going forward. Mitigation strategies are also recommended.

A Github repository was create to house the documents and results related to this work. Please refer to the Jupyter Notebook for the full data quality assessment.

- Module 1 Notebook: Link and Dashboard File: Link
- Project Github: Navyhoang/KPMG_Virtual_Internship_Challenge: Customers and transactions analysis (github.com)

Customer Demographic Data:

- Columns with missing values:
  - Last_name, job_title_, job_industry_category: 'nan' values can be filled with 'Unknown'.
  - Tenure: a conservative approach is to fill 'nan' with the minimal value of this column.
  - Age (calculated from DOB): filled with the mode value of the column.
- Mistyped data:
  - Gender: contains mistyped data such as 'F', 'M', 'U', 'Femal'. They can be simplified to "Female", "Male", or "Unknown".
- Irrelevant data:
  - Default: this column does not add any value to the analysis, it can be dropped.
  - Deceased_indicator: almost all values are 'N'. This column doesn't add much differentiating factor to the ML model in the next stage. It can be dropped.
- Invalidated/ contradicting data:
  - Age (calculated from DOB): max value is 177. This is out of the acceptable range since the world's oldest person record is 122. Values larger than 177 can be replaced by 122.

Customer Address Data:

- Mistyped data:
  - Address: 399 rows either have missing street numbers or numbers start with '0'. This is not a normal practice in real life. 0's can be stripped out. Those addresses with missing street numbers can still be kept for the analysis since the information from other columns can still be useful to answer other questions.
- Irrelevant data:
  - Country: all customers are from Australia. This column does not add any value to the analysis and can be dropped.

Transaction Data:

- Data type:
  - Product_first_sold_date: this column shows the number of days counting from Jan 1, 1900. The data can be converted to date_time data type.
- Columns with missing values:
  - Online_order: 360 rows with missing values. They can be dropped since they only make up ~1.8% of the dataset.
  - Brand, product_line, product_class, product_size, standard_cost, product_first_sold_date: 197 rows with missing values. They can also be dropped since they only make up <1% of the dataset.

Thank you for choosing KPMG for this exciting project work. We are happy to receive your feedbacks and iterate the analysis until it meets your expectations. Please let me know if you have any questions or concerns.

Best regards,
Thao Hoang

Data Analyst
Cell: (306)715-3789
Email: navy.hoang@mail.utoronto.ca