Name: Aravind Murali
College: Amrita University Coimbatore (Amrita School of Engineering)

A Report on -

VARIOUS DATA CLEANING AND ANALYSIS TECHNIQUES ON A MULTI-SERVICE BUSINESS DATASET

Aim: To generate a report on the various data cleaning and analysis processes involved while using a multi-service business dataset, with documentation on discrepancies or inaccuracies in the dataset (if any).

Procedure undertaken: The multi-service business dataset is provided in the form of an excel sheet (extension: .xlsx), with column names being 'Booking Type', 'Booking Date', 'status', 'Class Type', 'Instructor', 'Time slot', 'Duration (mins)', 'Price', 'Facility', 'Theme', 'service Name', 'subscription Type', 'Service Name', 'Servci Type', 'Customer Email' and 'Customer Phone'.

The dataset is loaded into front-end of Microsoft Power BI after the process of transformation of data in the backend interface called as Power Query Editor, where all the cleaning and preprocessing of data takes place.

Datatype name: 'Data Analyst_Assessment_Dataset'

The various data cleaning and transformation are done as follows -

- Since Booking ID, Customer ID and customer_Name should not contain duplicate values, 'remove duplicates' feature is used to remove any redundant values.
- The first row containing column names are promoted as headers using 'use first row as headers' feature.
- Since no column should contain null values, the columns 'Class Type' 'Instructor', 'Facility' & 'Theme' have their null values replaced with NotApplicable / NotSpecified / (N/A) using replace values feature.
- Since 'subscription type' column has empty values throughout, the column is removed completely, as it does not have information for any form of analysis.
- The columns 'customer Email' and 'customer Phone' have their null values replaced with (N/A).
- Empty spaces are removed, if any, using 'remove empty' feature of all columns.
- 'Time Slot' column datatype is changed from date/time to time.
- Column Quality, Column Profile and column Distribution features are used to evaluate columns and the information they contain; no empty values were found and all the information was given as needed.
- The 'customer phone' column is split into 'customer phone number' and 'country code' to separate country code from phone number. This is done using split columns using delimiter '-' (leftmost).

Conclusion: The cleaned dataset is evaluated once again and loaded to front end using 'close and Apply' feature