Group Name: Solo

Name: Armel Moumbe

Email : armel.moumbe@aivancity.education

Country : France

College: Aivancity school for Technology Business & Society

Specialization: Data Analyst

NB: This project is done by me alone due to not having any group members. Thank you for your time and understanding.

Problem description

XYZ company is collecting the data customer using google forms/survey monkey and they have floated n number of forms on the web.

The company wants to create a pipeline which will collect all the data of these google forms/survey monkey and visualize the data in the dashboard. The company wants clean data and if there is any data issue present in the data then it should be treated by this pipeline (duplicate data or junk data).

Data Cleansing and Transformation

The datasets were first loaded successfully on python and the cleaning and transformation was done there. As a reminder, we have three datasets, fitness analysis, fitness consumers and fitness trackers. Fitness analysis has 18 columns and 545 rows, fitness consumers has 22 columns and 30 rows and fitness trackers has 11 columns and 610 rows.

Here's a summary of the cleaning tasks performed:

Remove Duplicates: Identify and remove any duplicate rows in each dataset.

Handle Missing Values: Decide on a strategy to handle missing values (e.g., filling with a specific value, dropping rows, etc.).

Correct Data Types: Ensure all columns have the appropriate data types.

This is how the first approach went;

There was an issue converting the 'Reviews' column in the fitness trackers dataset to integers due to the presence of commas in the numbers. These values were then cleaned by removing the commas. Another issue then arises due to the presence of NaN values in the Reviews column, which could not be directly converted to integers. Opted for filling the missing Reviews values with 0, ensure again that there were no commas, and then converted the column to integers.

The second approach was to handle missing values by filling them using statistical methods such as mean, median, or mode. This technique was demonstrated using fitness analysis and fitness consumer datasets. Specifically:

1. **Impute missing numerical values** using mean and median.
2. **Impute missing categorical values** using mode

Another advanced approach can be to use model-based techniques, like KNN, to predict and fill in missing values.

The first approach made more sense since most of the data type in the datasets are string.

The Yaml file for the validation process will be used and shown in the next deliverable.

GitHub Repo link: https://github.com/m-armel/Data-glacier-Internship.git