Individual Challenge- Data Cleaning Expert

Nº62746 – Marcos Torres
Information Integration and Analytic Data Processing 2024/25
Master in {Data Science}
Faculdade de Ciências, Universidade de Lisboa

I. Challenge Overview

The objective of this challenge was to duplicate one of the datasets used in the project, and manipulate the records of said copy in order to create new types of significant data issues, which would then need to be dealt with using the appropriate procedures.

II. Dataset Selection and Initial Observations

As elaborated in the report of the actual project, several datasets were obtained through the Astroquery API, those being the TESS, ALMA, JWST and HAWKI datasets. For this challenge, the TESS dataset, corresponding to the Transiting Exoplanet Survey Satellite.
An initial overview of the dataset was obtained through the report generated by the yData-profiling package, revealing an extremely clean dataset. The only major issue being a small number of empty columns that could be easily removed, as they did not provide any relevant information for the objectives of the project.
While it's true that there are many high-quality datasets where not a lot of work is needed for them to be usable in projects, there are also just as many datasets with glaring quality issues.

III. Data Quality Issues

In order to lower the quality of the dataset, several python functions were created in order to introduce several types of data quality issues, those being:

- Misspellings;
- Missing Values;
- Misfielded Values;
- Duplicates;
- Wrong Data Type.

These quality issues were selected as they are some of the most common errors present in a dataset, usually caused by human error or poor quality sources.
While some of the functions were created with AI assistance, they all consist of selecting row index intervals defined previously, and changing the data in the specified columns in some capacity. Some care was taken to ensure that the resulting issues were realistic, and that they could be properly cleaned. The misspelling function was applied to the "provenance_name" column, corresponding to the pipeline that generated the observations, in this case being the SPOC (Science Processing Operations Center) pipeline. the missing values function and the wrong data type function were applied to the "s_ra" column, corresponding to the right ascension values of the observations, while the misfielded values function was applied to both the "em_min" and "em_max" columns, corresponding to the minimum and maximum electromagnetic range of the observations. The duplicates function was applied to a random row selected beforehand.
A yData report of the "dirty" dataset was also created for quick analysis.

IV. Data Cleanup Procedures

As for the cleanup procedures, the missing values and the values of the wrong data type were left for last, since if they were to be removed, they would affect the code used for cleanup.
Starting with the duplicates, their removal in this case was simple as the dataset contains several columns with unique IDs, which allowed them to be easily dropped. In the scenario that only specific columns had duplicated values, then it would depend on the type of information present in the column and how important it is to the task at hand.
Since the column with misspellings is supposed to have the same value in all rows, the rows with errors were easily corrected. Same with the columns that had misfielded values, these rows were cleaned easily based on the rest of the data present in the columns.
Finally, since the column affected by missing values and rows with the wrong data type corresponded to coordinates, there was no value that could be easily used as a replacement, and considering that this is already a very large dataset, dropping these rows would not be a significant loss of data.