Data Analysis - Master MLDM - 2023/2024
Project: Exoplanets dataset analysis

Project description

General objective: In this project, we will evaluate your ability to describe a dataset, and to extract the information from it using the techniques we have studied during this semester. Deliverables : source code that we can run on our machine (in Python) + report. You can use illustrations, graphs, tables etc...

Source code: The source file(s) should contain all the information needed to reproduce the experiments. The code should be commented, and well organized. This will be taken into account for your final grade.

Deadline : You will upload a first version of your code and report at the end of the session. After that you will have until 12/03/2023 23:59 to upload the final version of your project.

Report

Dataset description: In the dataset description section of the report, you must describe the dataset, its size, the variables, and their nature. Propose an overall analysis of the variables in the dataset: mean and total of the quantitative variables, proportion, variances of the variables of interest, etc. Do not hesitate to illustrate this part with numerous graphs.

Analysis: In the Analysis section of your report, you are free to provide any analysis you find interesting (e.g., find which planets are habitable). You can perform a wide range of method to extract information from the data. For example, you can try to train a linear regression model and see whether it fits some data information. Do at least one analysis, with full investigation of performance, optimal parameters, comparing different approaches, etc...

Additional informations

Dataset units: In the dataset, all the values are relative either to the Earth for planet variables or to the Sun for star variables.

Habitable zone distance: In order to consider an exoplanet to be "potentially habitable", many criterions are taken into account. One of them is "Habitable Zone Distance" which represents the range of orbits around a star within which a planetary surface can support liquid water given sufficient atmospheric pressure. We consider a planet to be in the Habitable Zone Distance if the orbital radius $R_{orbital}$ satisfies:

$$R_{orbital} \in [0.8 * \sqrt{L_{star}}, 2 * \sqrt{L_{star}}]$$

Total number of planets: Even though we only know of around 5,000 planets right now, we can estimate that there is roughly one planet for every star. Our galaxy has 100 billion stars, and so likely has around that many planets.