GDG Algiers

Noisy Data Cleaning

DevFest 2020

Description

Noisy data are data with a large amount of additional meaningless information in it called noise. This includes data corruption and the term is often used as a synonym for corrupt data. It also includes any data that a user system cannot understand and interpret correctly. The expected output of this challenge is a well presented notebook that illustrates all the steps of figuring out and cleaning the noisy data within the given dataset.

To pass the Challenge:

1. Create a notebook.
2. Load data.
3. Remove duplicate or irrelevant observations.
4. Fix structural errors (strange naming conventions, typos, or incorrect capitalization...)
5. Filter unwanted outliers.
6. Handle missing data.

Data guide

The given dataset contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. The data is stored as a CSV file that contains 32 columns. The dataset's size is up to 16MB.

Technical Requirements:

7. Use the given dataset.
8. The result of each treatment should appear in the notebook.
9. The submitted file should have the **.html** extension.
10. The notebook **must be well commented and presented**.