Pandas (Hands-on 2)

support@intellipaat.com

+91-7022374614

US: 1-800-216-8930 (Toll-Free)

Problem Statement

Large datasets need to be structured and free of outliers in order to be used for further analysis and model building in cases of supervised or unsupervised learning. Python pandas is one such python package that deals with manipulation of data with various operations. Use the modules and processes available in the python pandas module to come up with useful insights about the data.

Dataset Information

The dataset contains information about 260 cars that include horsepower, cubic inches, time to 60, brand, make year, weight, cylinders, mpg, etc.

Tasks To be performed:

1. Load the dataset into the program using the pandas module and perform the following operations.
    a. Print the first five rows of the dataset.
    b. Print the last five rows of the dataset.
    c. Print the column names from the dataset.
    d. Print the shape of the dataset.
    e. Print the information of the dataset with respect to each column in the data.
    f. Print the count of distinct values in each of the columns in the data.
    g. Check the dimensions and size of the data.
    h. Check for duplicate entries in each column.
2. Check for null values in the dataset, and if found any replace the null values with 0.
3. Describe the data based on parameters like mean, average, standard deviation, etc for each of the columns.
4. Calculate the **correlation coefficient** between each of the columns in the dataset.
5. Create separate dataframes with respect to each distinct value in the brand column, where the time-to-60 is less than 10.0.

    For example - Cars with brand name X will be in one dataframe, Cars with brand name Y will be in separate dataframe, and so on.

6. Create a separate data frame for cars whose **hp is greater than 200** and **time-to-60 is less than 15.0**
7. Create two separate data frames where in the first dataframe you sort the entire dataset in **descending order of hp**, and in the second dataframe, you sort the entire dataset in **ascending order with respect to time-to-60**.
8. Create a dataframe, in order of **top 10 heaviest cars** in the cars data. Repeat the same for **top 10 lightest cars**.
9. Create separate dataframes with sorted values with respect to **mpg in ascending and descending order**.
10. Split the dataset into two with respect to cars built **before 2000, and after 2000**.