Solved Case Study based on Open Datasets

UCI dataset is a collection of open datasets, available to the public for experimentation and research purposes. 'auto-mpg' is one such open dataset. It contains data related to fuel consumption by automobiles in a city. Consumption is measured in miles per gallon (mpg), hence the name of the dataset is auto-mpg. The data has 398 rows (also known as items or instances or objects) and nine columns (also known as attributes).

The attributes are:

mpg, cylinders, displacement, horsepower, weight, acceleration, model year, origin, car name.

Three attributes, cylinders, model year and origin have categorical values, car name is a string with a unique value for every row, while the remaining five attributes have numeric value.

The data has been downloaded from the UCI data repository available at http://archive.ics.uci.edu/ ml/machine-learning-databases/auto-mpg/.

Following are the exercises to analyse the data.

1) Load auto-mpg.data into a DataFrame autodf.

2) Give description of the generated DataFrame autodf.

3) Display the first 10 rows of the DataFrame autodf.

4) Find the attributes which have missing values. Handle the missing values using following two ways: i. Replace the missing values by a value before that. ii. Remove the rows having missing values from the original dataset

5) Print the details of the car which gave the maximum mileage.

6) Find the average displacement of the car given the number of cylinders.

7) What is the average number of cylinders in a car?

8) Determine the no. of cars with weight greater than the average weight