ANALYSIS OF CAR DATASET IN SPARK RDD

Description:

This data set consists of the various parameters related to cars being driven in United States, Europe and Japan. The table provided is in the .csv format.

Parameters:

1. **Make of the Car:** The table contains the brands of the cars being driven in the above mentioned regions. Example: Ford, Datsun, Chevrolet, Pontiac etc.

2. **Miles per Gallon (MPG):** The total miles that can be done by a particular car is given by MPG (Miles per Gallon). In the table provided it's in the range of 0 MPG to 46.6 MPG.

3. **Horsepower:** It is defined as the work done per unit time. The table contains data in HP for each model. It is in the range of 0HP-230HP.

4. **Displacement:** It is the volume of Air fuel mixture displaced by the piston in each stroke. In the given table it is in the range of 68cc-455cc.

5. **Weight:** The weight of each model is given in kilograms and is in the range of 1613kgs - 5140kgs.

6. **Acceleration:** Acceleration of each car model in the table is in the range of 8 miles/Sec2 to 24.8 miles/Sec2.

Problem and Solutions:

1. How many cars are developed by each country? What is the Minimum, Maximum and average number of a particular car Model developed by each country? Calculate by using schema of data frames.

    a. How many cars developed by per country?

    b. What is the Minimum, Maximum and average Numbers of a particular car Model developed by per country? Calculate by using schema of data frames.

2. Calculate the top 10 heavy cars from dataset by using sqlContext.

3. How many cars have 4 Cylinder engines and which have been developed in US? Load the dataset file from HDFS.

4. Find the best cars based on the criterions of:

    a. Acceleration more than 23.5

    b. Weight more than 3000

    c. And MPG more than 25.0