Exercise 5 – Detecting and Handling Skewness

In this exercise we'll join two datasets and try to detect and handle skewness.
The file UsedCarsFuel.csv contains the same data as in exercise 4, but the *fuel_type* column was replaced with the *fuel_code* column that contains a unique code for each fuel type (for example, "Diesel" gets code 1).
The file fuelTypes.csv contains for each fuel type its code.

Complete the following tasks:

1. Load both files into dataframes.
2. Perform an inner join between the datasets, based on the fuel_code column.
3. Open the Spark UI and investigate the time it took to perform the join. Can you detect a data skew?
4. Create a user defined function that adds a random character between 'a' and 'd' for each value in the fuel_code column. Write a function that explodes each row in the fuelTypes dataframe four times.
5. Run the join between the modified datasets and investigate the Spark UI.