Exercise 3 – Working with SparkSQL

In this exercise we'll read a csv file that contains data about used cars market in Pakistan. We'll use SparkSQL to investigate the data.

Open the UsedCars.csv file and investigate it's structure. Note at the columns names and values.
Complete the following tasks:

1. Load the csv file into a dataframe.
2. How many rows are there in the dataset?
3. The columns "currency", "description" and "item_condition" don't give any meaningful information. Drop them from the dataframe and save the result in a new dataframe.
4. How many cars run on petrol? Use the fluent sql api to show the number.
5. How many cars are from model date between 2000 and 2010?
6. It seems that the "brand" and "manufacturer" columns contain the same data. Validate that using query, and if so, drop the manufacturer column.
7. Count how many cars are there for each brand.
8. The "mileage_from_odometer" column represents the kilometrage of the car, but the values are strings. Write a user defined function that convert all values to integers by deleting the "km" and "," from each value. Create a new dataframe with the results in a column named "km" and drop the "mileage_from_odometer" column.
9. Find the model dates of all cars that traveled more than 100000km.
10. Write the dataframe to the disk, into a csv file called UsedCars2.csv.