Analysis of Earth Surface Temperature using Spark

Task 1:

Find the number of times where a city's average temperature on a day turned out to be higher than the city's average temperature throughout the dataset for a given country.

1. Load the dataset into data frame format using SQLcontext.read.csv with option headers is true.
2. Convert the data type of column Average Temperature from string to float, and drop any rows with null/nan values.
3. Filter out rows such that the city name is equal to the given input City name.
4. Group the data frame based on City and apply mean on the average Temperature.
5. Store data frame separately as df_avg.
6. Inner join the filtered data frame with df_avg on the condition that City value in both data frames is same.
7. Convert data frame to rdd apply map such that if average Temperature < avg (average Temperature) for a city store (City,1) into rdd else store (City,0).
8. Convert rdd to df and sort based on City, filter such that only City's that has value 1 are considered
9. Convert df to rdd and reduce by key to obtain aggregate for each City then apply Collect ().
10. Sort the final list obtained and print the city, value such that they are tab separated using a for loop

Task 2:

Find the number of times where a country's maximum average temperature on a date turned out to be higher than the worldwide land average temperature on the same date.

1. Load both the dataset into data frames format using SQLcontext.read.csv with option headers is true.
2. Convert data type of average Temperature column and Land Average Temperature column from string to float.
3. Drop rows with null/nan value in both data frames using na.drop('any')
4. Group by Country and date on the data frame of city.csv and apply max function on column average Temperature to obtain max temp for each day in a Country
5. Inner Join the resultant data frame with globa.csv data frame based on condition that dates are equal
6. Convert resultant df to rdd and apply map such that if max (average Temperature) > Land Average Temperature for a day, store (Country, 1) else store (Country ,0) into rdd
7. Reconvert the rdd to df format and apply sort based on Country column then filter the df so that it has only Country ,1 value in each row.
8. Covert df to rdd and reduce by key to obtain number of days where max (average Temperature) > Land Average Temperature followed by collect () and sorting of collected list
9. Print the Country, value in tab separated format