Homework 6

For this homework you will create a python notebook (`.ipynb` file). You'll need to use a pyspark kernel (for the 2nd part at least!). This file should then be uploaded to wolfware in the assignment link!

Goal

The purpose of this homework is to practice with MapReduce and basic pyspark coding. Remember, most homework assignments will have a part that pushes you beyond what was in the lectures! Learning to search for the right questions and browsing stackoverflow are really life skills that we should hone :) **Be sure to include markdown text describing what you are doing, even when not explicitly asked for!**

Part 1: Split data (8 pts)

Using the NFL box score data set (https://www4.stat.ncsu.edu/~online/datasets/scoresFull.csv):

- Split the data into separate .csv files based on the season. That is, you want to subset the data to obtain just one season and output that to a .csv file. You want to do this process for each season in the dataset.
- You can use a loop for this if you'd like!

Part 2: MapReduce Idea (no pyspark)

MapReduce part (16 pts)

- consider a variable to group on (like week) and a numeric target variable (like AQ1)
- write a mapping function to find the
  - sum of the target variable across the grouping variable
  - sum of the squared values of the target variable across the grouping variable
  - count or number of observations of the target variable in each group
- write a reduce function to combine the results across the season data sets

Summarizing bit (10 pts)

- take the final result and use it to construct the
  - mean at each level of the grouping variable
  - standard deviation at each level of the grouping variable (if the count for the group was larger than 1)
- create a function to put the MapReduce part and the final calculation part into an easy to use function, allowing you to change the grouping variable and target variable

Details/Hints

I'm not going to give as much structure on this one but I will give some requirements that should help you out.

- Once you've created the output data sets, read them back in and store each data set as an element of a list (which is iterable)
- You will need to write your own function to use with `map()`
    - Remember that this type of function **takes the data as an input and outputs key/value pairs (a dictionary for our purposes)**.

    - This mapping function should take in one of the data sets (one season of data), a grouping variable, and a target variable (I took these in as strings but you can do it however you'd like). This helped me but your mileage may vary.
- You should use the `map()` function to call the function written above across the iterable (list) of data sets.
    - As your mapping function takes in three arguments, you likely need to provide them all as iterables of the same length (I did).
        - The data sets should already be in an iterable (list or something similar)
        - You'll need to supply the grouping and target variables as iterables of the appropriate length (for instance, I did `["week"]*len(data_set_iterable)`)
- You should write a **reducer** function that takes in two dictionaries and combines them. Then use `functools.reduce()` (all of this very similar to how it was done in the notes!).
- You should write a function that takes the final result and returns the mean and standard deviation of the target variable at each level of the grouping variable (just don't return a standard deviation if the count is 1)
    - The mean is of course just the sum divided by the count
    - The sample standard deviation can be found using this formula

$$s = \sqrt{\frac{1}{Count - 1}(SumOfSquaredValues - Count * Mean^2)}$$

- Lastly, you should write a function that takes in the data (always the same iterable of our data sets), a target variable, and a grouping variable. The function should call your functions above to easily produce the final result with a single function call!

Part 3: Using pyspark (SQL)

This part should be very easy/short compared to the previous part! We'll use spark SQL functionality rather than writing our own MapReduce type code.) To do:

- Read in the full nfl data set into spark as a spark SQL style data frame
- Use spark SQL to find the mean and standard deviation for the AQ1, AQ2, AQ3, AQ4, AQFinal, HQ1, HQ2, HQ3, HQ4, and HFinal variables
- Repeat the previous item but return summaries at each level of the season variable.

Part 4: Using pyspark (pandas-on-spark)

Repeat part 3 but read the data into a pandas-on-spark data frame and use pandas-on-spark functionality to find the summaries! (This part should be short as well!)

Files to Submit

- Make sure all cells are run
- Click on File –> Save and Export Notebook

Upload the `.ipynb` file to the assignment link.