Web Data Analytics

Homework #2

Total Points: 50

Ground rules about study groups: You may form a study group with other fellow students. As a study group, you may discuss the assignment, exchange ideas, and clarify questions for each other. However, any form of plagiarism, e.g., copying each other’s answers, is strictly prohibited and will be prosecuted based on the academic integrity policies and procedures. If you participate in a study group, you must include the name of the group members in your submission.

You are free to consult any external resources for reference.

Study group members (if any):

**Instructions**

Create a new IPython Notebook, and use it to answer the following questions.

# Your file should be named in the following format:

# lastname\_firstname\_hw#

# For example, John Smith would name the file: smith\_john\_hw#.

# Question 1 (25 points)

In this question, you will be required to obtain the unemployment data for the following states: Indiana, Mississippi, and NewJersey (similar to what we did in class) and export them to individual files using python (This means there should be a separate file for each state). For each file, make the following changes:

* We only want four columns from the original data: series id, year, month, and value. Do not include any other columns from the original source data.
* Add a column called month\_year which holds a string that has the month year combination. For example, if a row has the month at 06 and the year at 2007, this column should have the following string: “06\_2007”
* We only want to keep the data from 2010 till 2015.

Hint: you can make use of a python method that is applied to a string called **.split(“,”).** Here is an example of what this will do, try to think about solving the problem above and thinking about the logical way you can use this method to assist you.

Lets say you have the variable “S”, which has the following string:

S = “apple,orange,pear,lemon,banana”

If you type:

Columns = S.split(“,”)

This will create a list, called Columns, where each element is the string separated by a comma. For example:

Columns[0] = “apple”

Columns[1] = “orange”

Columns[2] = “pear”

Columns[3] = “lemon”

Columns[4] = “banana”

Think about using “\t” instead of “,”.

**Question 2 (25 points)**

In this question, you will be required to obtain the unemployment data for the following states: New York, Mississippi, and New Jersey (similar to what we did in class) and export them to **one** file using python. Put all the data into one file, and add a column called “State” that has the state name.

* We only want the data from 2010-2015.
* We only want the data that correspond to the following months: June (06).

Please make sure to address the same issues associated with using one file (the ones we discussed in class – repeated headers and adding a state column for each row).