50 points total

MIS 3335 – Data Analysis Using Python

Homework 2 – Reading and Subsetting Data

The government of the United States makes many data sources available to the public. One such source is a summary of income tax filings for any given year. Files are produced that show filings for the country as a whole and for each state or other taxing entity within the US.

Accompanying this document are two additional files. The file **15zpnyagi.csv** is all the 2015 tax filing data for the state of New York. The second file, **15zpdoc.doc**, contains a description of the data. That includes an explanation of what all the column headers and codes mean. You will find this file to be very valuable.

Your assigned task is to use the contents of these files and your Python/pandas/Notebook skills to answer the following questions. ***Submit your notebook at the appropriate link in Blackboard***. Do not submit the data files.

5 pts

1. What is the total of “Salaries and wages amount” for all NY tax filers in 2015? [Answer: 514,823,973]

5 pts

2. What is the total for “Educator expenses amount” in Washingtonville, NY (zip code 10992)? [Answer: 60]
3. What is the mean number of dependents claimed by all New York tax filers in the adjusted gross income range of $50,000 to $75,000? Display your answer using a meaningful print statement and a formatted numeric answer.

   5 pts: create df with correct records

   A meaningful print statement would look something like this:

   ```
   The average number of dependents per return in t
   range is 4.58
   ```

   2.5 pts: calculation
   2.5 pts: meaningful output

   To format your answer to 2 decimal places, use the `format()` function inside the print function. For example, if you had a calculated answer of 4.578321490234 stored in the variable named `result`, you might use the following print statement:

   ```
   print('My answer is', format(result, '.2f'))
   ```

   2.5 pts: total num dependents
   2.5 pts: total num returns

   Note: You will want to use intermediate variables. You'll need to get the total number of dependents and the total number of returns to calculate this mean. See Table 2.2 on page 31 of P4E for series methods that might be of help.

4. Create data frames that contain only the following subsets of the original dataset:

   5 pts

   a. "df" - columns for zipcode, adjusted gross income category, number of exemptions, number of elderly returns, taxable Social Security benefits amount, and income tax amount.

   5 pts

   b. "df2" - the df columns plus the rows for only the lowest AGI category

   Display the first 10 rows for the `df` data frame and the last 8 rows of the `df2` data frame.

   2.5 pts 2.5 pts

10 pts: Overall style and format: comments, other documentation, follow standards