DSC 200 – Data Wrangling
Lab 6: Data Cleaning – Investigation, Matching and Formatting

Goals:

- Demonstrate ability to identify dirty data via code
- Demonstrate ability to use the **pandas** package to clean datasets

Instructions:

Part 1:

In part two of lab 5, you were required to choose 3 datasets for the exercises. In this part you are required to download your chosen datasets from that lab, write a Python function that merges the datasets into the single file which will be used for analysis later. Your python file should include a link to the files. Alternatively, you may include the downloaded file in your submission. I will be downloading each file and be testing your code using those files.

20 points

Part 2:

The files linked in the Canvas assignment from which this document is accessed have several data cleaning issues. Among the many issues the dataset has are repeated observations and features, and for some other categorical features, the inconsistent data. One of the files contains descriptive information about the data in the second file, while the other contains the actual data. More details about this file can be found at
https://archive.ics.uci.edu/ml/datasets/Adult

Write a Python function that cleans this data set. Make sure to check all the data cleaning issues we discussed in class and ensure that your script accepts the path to the file and creates a new file containing the cleaned-up data. This new file should be stored in the same directory as your script. Give the new file a meaningful name. The script should print the number of features and observations prior to and an after the cleaning task to the console.

30 points

What to submit:

Submit a single python script containing the two functions described in parts 1 and 2 above. If you choose to include the downloaded files from for part 1, submit a zipped file containing the python script and the data files. Implement a simple menu that asks the user to select one of the two functions. When selected, your scripts should perform the requested function.

Submission: Submit this work in Canvas via the linked assignment.