Project 1/31/2022

Initial Draft Due: 2/9/2022 5:29PM EST

Final Product Due: 2/14/2022 5:29PM EST

**Statement of Work**

* Create a reusable Jupyter notebook that’s purpose is to clean datasets and perform some initial exploratory data analysis (EDA)
* The notebook you create should have the flexibility to be applied to different datasets (don’t build it just to fit one dataset)
* Apply the notebook to a dataset of your choosing as an example of your process at work

**Project Goal**

* To create a useful, repeatable process/script that will help clean datasets
* Create a product that you can have for personal and professional use in the future
* Create a product that you can include in your personal Github repository to show to potential employers and assist in the interview process

**Project Requirements**

* Research and find a dataset that needs cleaning processes applied to it
* Include an introduction section that gives a verbal explanation of the data set, where you found it (site the source), what the data describes/the purpose of the data, what you could potentially use it for, and a brief overview of the contents/purpose of your process/notebook
* Provide a brief list/description of the different skills/tools/methods that you learned in this program and applied in the notebook (make this list as long as you want but make it concise)
* At the end of your process, summarize the difference between original and clean datasets and how you improved the dataset
* Your process needs to implement the pandas library, and the result of process should be a pandas dataframe
* All sub-processes created to clean data should be implemented via functions (one sub-process per function)
* Visualize your dataset in at least 2 different ways
* Prove that your dataset is either normally distributed, right-skewed, or left-skewed
* Calculate the 3 different measures of central tendency we discussed in class (mean, median, mode)
* Bonus: implement some or all of your functions via organized modules that you import from different file(s) (organized meaning all stats function in one file and all data cleaning processes in another)
* Bonus: provide a summary of how you think you could use the data to create predictions, calculations, or whatever future plans you potentially think you could implement (these don’t need to be too detailed for now; this is just to get you thinking – you can add on later)