**Data Analyst Assignment**

The purpose of this assignment is to assess your analytical and presentation skills.

You may use the language/software of your choice. This is open book/open notes, but work must be all your own. Aim to spend about 4-5 hours on this assignment.

The dataset comes from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets.html), It is called the Census Income Data Set, and it is originally pulled from US Census data. The data is in the file adult.csv, and descriptions for the attributes are listed in the file adult\_attributes.txt. Please provide the following 2 files when returning the assignment.

1. A file containing your results/analysis for Parts 1, 2, and 3 (i.e., Jupyter Notebook, Python file, R file, Excel, etc.).
2. A presentation containing answers for Part 2 and your model analysis for Part 3. *You should be prepared to walk through your analysis in a presentation format*.

**Part 1: Manipulating data**

Using the language of your choice, answer/complete the following questions about the data.

1. Add column headers to your data set based on what is described in adult\_attributes.txt – this must be conducted programmatically.
2. Remove dashes in the native country names and replace with spaces.

**Part 2: Data Visualization and Analysis**

Provide a profile of the data. This profile should include:

1. Information about missing data (i.e., how much is missing and how are these missing variables represented?). What considerations should we make about missing values?
2. Overall explanatory statistics about the data (e.g., quantiles, mode, mean, for numeric; histogram for categorical). Where is the value in this information?
3. At least two visualizations that show something interesting about this data set, and explain the meaning or implications of these visualizations.

**(Optional Bonus) Part 3: Building a model**

Develop a model to predict the probability of an individual making more than 50k per year. Use as many variables as you would like, and feel free to compare methods to yield a better performance.

Provide a written explanation that summarizes your methods and results. Can include numbers/tables if desired. Be sure to address the following questions in your summary:

1. What type of model(s) did you use and why?
2. How did you select the variables to be included in the model?
3. Did you use any methods for model selection and/or validation? If you are using multiple methods how did you compare the performance of your models?
4. What is the output of your model and how does it help address the question?
5. What would you do if you had more time?

**Part 2: Data Visualization and Analysis**

Provide a profile of the data. This profile should include:

1. Information about missing data (i.e., how much is missing and how are these missing variables represented?). What considerations should we make about missing values?
   1. **Missing numerical data is represented in numbers as ‘0’ in *capital-*gain and *capital-loss*, and as ‘?’ for string fields in *workclass* and *occupation* and *native-country***
2. Overall explanatory statistics about the data (e.g., quantiles, mode, mean, for numeric; histogram for categorical). Where is the value in this information?
3. At least two visualizations that show something interesting about this data set and explain the meaning or implications of these visualizations.