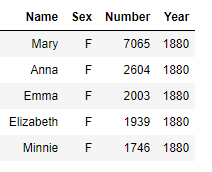
**Data Scientist Data Challenge for the Analytics Team of the Office of Digital**

Thank you for your interest in completing the data challenge for the data scientist position at the Office of Digital. In this challenge, you will be working with a publicly available dataset provided by the Social Security Administration of the US government. The data describe the yearly occurrences of each name on Social Security card applications. The data are from a 100 percent sample of Social Security card applications after 1880 through 2017. Each record in the dataset has the fields, "Name”, “Sex”, “Number", and “Year”, where “Name” is a 2 to 15 character name, “Sex” is M (male) or F (female), "Number" is the number of occurrences of the name, and “Year” is the year of the application. The first few rows of the dataset (after cleaning) look like this:



Below are questions that can be answered with the dataset. For each question, provide the answer and the code required to generate it. For the code, you can provide a single script or a Jupyter notebook with commented code. Be sure to document any data cleaning needed.

**Most common names**

1. What was the most common male name in 1989?
2. What was the most common female name in 1989?
3. Write code to output a list of the most common male names by year.
4. Write code to output a list of the most common female names by year.

**Name diversity**

1. How many unique names are there in the entire dataset?
2. How many unique male names were there in 1989?
3. How many unique female names were there in 1989?
4. How many unique names (male and female) were there in 1989?
5. Why is the number of unique names in 1989 (problem #6) smaller than the sum of unique male names (problem #7) and unique female names (problem #8) for that year?
6. Write code to output a list of names that show up in both the male and female groups in 1989. How many names show up in both groups for that year?
7. Show how the answer to problem #10 helps explain the answer to problem #9.

**Unisex names**

1. Provide a list of the most common unisex names through 2017. Describe and justify your criteria for a name to be considered unisex. What other criteria could you use for a name to be unisex?

**Machine learning exercise**

Design and create a machine learning model that classifies a name as either male or female based on the characteristics of the name. Write up a report with screenshots in a text document or Jupyter notebook that describes the following:

* the input data and its quality
* any data exploration and any interesting insights
* any preprocessing or cleaning of the data that was needed, including how you chose to label the data (i.e., how you labeled a name male or female as the ground truth for the model)
* the features used by your model and why you chose them
* your selection of machine learning algorithm and why it was chosen
* the accuracy of your model and any relevant metrics that describe model performance
* any model parameter tuning used to improve performance
* any relevant discussion, model interpretation, future steps, or further exploration needed

In addition, prepare a short slideshow presentation (~10 min.) of an overview of the machine learning exercise. You will present this to the analytics team.

There are no restrictions on the classification model that you use; however, be sure to keep in mind your ability to adequately explain and interpret the results.

Good luck!