CS 620–Introduction to Data Science, HW2, Fall 2022

The starter code and the data for this assignment are available here:
https://www.cs.odu.edu/~sampath/courses/f22/cs620/files/hw/hw2-starter.zip

A. (40 pts) Pandas basics**.** Please use the starter code (hw2-a.py) provided.

Let df be a pandas DataFrame constructed with the following;

```
data = np.array([1, 7, 3, 6, 2, 8, 5, 9, 4]).reshape(3, 3)
df = pd.DataFrame(data, index=['One', 'Two', 'Three'], columns=['a', 'b', 'c'])
```

Generate the following outputs.

i. [6, 8]

ii. [1 2 4] (hint: use numpy.diag)

iii. (10 pts) Display the following subset given any value of column ‘a’ is less than 6.

```
       a  b  c
One    1  7  3
Three  5  9  4
```

iv. (10 pts) Display the following result using “applymap” and “lambda”

```
       a   b  c
One    2   8  4
Two    7   3  9
Three  6  10  5
```

v. (10 pts) Display the following result using “apply” and “lambda”

```
One     7
Two     8
Three   9
```

B. (60 pts) Please use the starter code (hw2-b.py) provided. The “yob-names” directory (from social security administration, https://www.ssa.gov/oact/babynames/limits.html) contains number of text files, each contains the year of birth (yob) with rows of data, where each row has a sequence of columns, separated by a commas. For example, the first few rows of yob1880.txt are,

```
Mary,F,7065
Anna,F,2604
Emma,F,2003
Elizabeth,F,1939
```

We can interpret this data as, “In the year 1880, 7065 female babies were born named Mary; in the year 1880, 2604 female babies were born named Anna" and so on.

i. (30 pts) Write a python code to generate a single .csv file “yob-names.csv” that contains the data in the following format, where the file contains all the data merged from the yob files.

```
year,name,sex,frequency
1880,Anna,F,2604
1880,Emma,F,2003
1880,Elizabeth,F,1939
```

ii. (30 pts) Generate the results for the following queries based on the file “yob-names.csv” generated above. You **CANNOT USE** any regular python loops to answer the following questions (-5pts). Instead, use DataFrame techniques learned in pandas.

a) (5 pts) What is the most popular boys name in year 1980?
b) (10 pts) How many girls were born between 1990 and 1992?
c) (15 pts) Estimate the number of female Benjamin’s alive today (year 2022) who were born on or after 1950. For this particular query, use the given “cdc-life-expectancy.csv” file to generate this result. We can interpret the data from this file as, “The average life expectancy of U.S. babies born in each year, for Males and Females” and so on.

What to turn in:

Each file must exactly follow the naming convention: **Lastname-hw2.zip** should contain following 2 files.

HW2-a.py
HW2-b.py