CS 620–Introduction to Data Science, HW2

The template and data for this assignment are available here:
https://www.cs.odu.edu/~sampath/courses/f19/cs620/files/hw/wrangling.zip

A. (40 pts) Please use the code skeleton (HW2-a) provided.
   i. (10 pts) Write a python function to merge 2 sorted lists. The resultant list should be in sorted order.
   ii. (10 pts) Write a python function to calculate the summary statistics of an array of numbers: max, min, mean, standard deviation, median, 75 percentiles, 25 percentiles. You can write your own functions or use any existing modules from math, statistics or numpy.
   iii. (20 pts) Re-scaling is the process of linearly transform from one range to another. Write a python function that accepts a list of numbers in any range, then scales the numbers to [0, 1] using MinMax algorithm (given in the skeleton code).

B. (60 pts) Please use the code skeleton (HW2-b) 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 segment 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.
      a) (5 pts) What is the most popular boys name in year 1980?
      b) (10 pts) How many girls were born between 1990 and 2000?
      c) (15 pts) Estimate the number of female Benjamin's alive today (year 2019) 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