Stat 426 – Fall 2020

Midterm 1

Part 1

The Social Security Baby Names data is a popular and fun dataset. Here is a description of the data from the social security popular names website:
(See https://www.ssa.gov/oact/babynames/background.html)

> All names are from Social Security card applications for births that occurred in the United States after 1879. Note that many people born before 1937 never applied for a Social Security card, so their names are not included in our data. For others who did apply, our records may not show the place of birth, and again their names are not included in our data.
>
> All data are from a 100% sample of our records on Social Security card applications as of March 2019.
>
> *Data qualifications*:
>
> People using our data on popular names are urged to explicitly acknowledge the following qualifications.
>
> 1. Names are restricted to cases where the year of birth, sex, State of birth (50 States and District of Columbia) are on record, and where the given name is at least 2 characters long.
> 2. Name data are tabulated from the "First Name" field of the Social Security Card Application. Hyphens and spaces are removed, thus Julie-Anne, Julie Anne, and Julieanne will be counted as a single entry.
> 3. Name data are not edited. For example, the sex associated with a name may be incorrect. Entries such as "Unknown" and "Baby" are not removed from the lists.
> 4. Different spellings of similar names are not combined. For example, the names Caitlin, Caitlyn, Kaitlin, Kaitlyn, Kaitlynn, Katelyn, and Katelynn are considered separate names and each has its own rank.
> 5. When two different names are tied with the same frequency for a given year of birth, we break the tie by assigning rank in alphabetical order.
> 6. Some names are applied to both males and females (for example, Micah). Our rankings are done by sex, so that a name such as Micah will have a different rank for males as compared to females. When you seek the popularity of a specific name (see "Popularity of a Name"), you can specify the sex. If you do not specify the sex, we provide rankings for the more popular name-sex combination.
> 7. To safeguard privacy, we exclude from our tabulated lists of names those that would indicate, or would allow the ability to determine, names with fewer than 5 occurrences in any geographic area. If a name has less than 5 occurrences for a year of birth in any state, the sum of the state counts for that year will be less than the national count.

You can get the data directly from the Social Security website:

https://www.ssa.gov/oact/babynames/limits.html

The "National data" is a .zip file which, when unzipped, contains a .txt file for each year from 1880 to 2019. I have also included the .zip file in this repository and code to combine all the files into one.

The first and last three rows of the data :

name	gender	n	year
Mary	F	7065	1880
Anna	F	2605	1880
Emma	F	2003	1880
	...		
Zyking	M	5	2019
Zyn	M	5	2019
Zyran	M	5	2019

1. Add a column to the data that shows the proportion of people of that gender with that name that were born in that year. This column can be computed by dividing each row's n by the total number of applicants in that year for that gender.

   You can find the number of social security applicants by gender and year at the following url:

   https://www.ssa.gov/oact/babynames/numberUSbirths.html

   For full points, use the pandas function `read_html` to read the table into python and do not use loops in this part.

   The first and last three rows of the data with the new column should be:

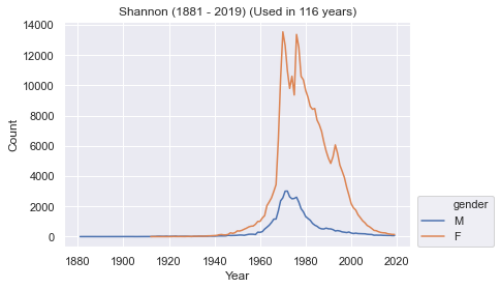
name	gender	n	year	prop
Mary	F	7065	1880	0.072383
Anna	F	2605	1880	0.026679
Emma	F	2003	1880	0.020521
		...		
Zyking	M	5	2019	0.000003
Zyn	M	5	2019	0.000003
Zyran	M	5	2019	0.000003

2. Write a pandas function that plots the frequency of a specific name over time by gender. You can assume that your function will only work with the DataFrame you made in problem (1) (meaning that the DataFrame does not need to be an argument of the function). For full points, your function should:

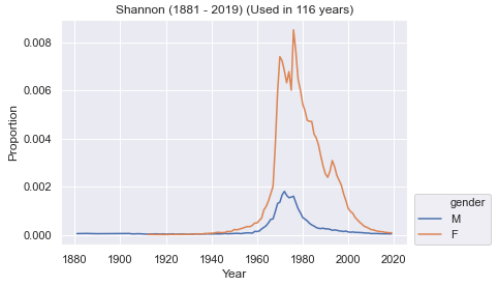
   - have an option to plot *either* the proportion *or* the count with appropriate axis labels, and legend, using the proportion as the default
   - have the plot title be the name that was plotted along with the first and last years the name appears in the data AND the number of years the name is in the data

   For example:

   `plotName("Shannon", var="n")`



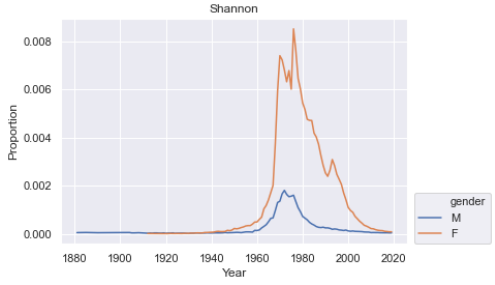
   `plotName("Shannon", var="prop")`



   At the very least your function should plot the proportion and use the plotted name as the title

   For example:

   `plotName("Shannon")`



3. Use the function you wrote in problem (2) to plot:

   (a) Your own name
   (if your own name is not in the dataset, plot the name "Eleanor")

   (b) The name "Semaj"

4. Answer the following questions about the data:

   (a) How many unique names are represented in the data?

   (b) What year has the most unique names?

   (c) What year has the fewest unique names?

   (d) What are the five most popular names (averaged over all years)

      i. overall?

      ii. for females?

      iii. for males?

   (e) What are the five most popular names in the year that you were born
   (please specify)
   If you would prefer to keep the year that you were born private, answer the question for the year 1998

      i. overall?

      ii. for females?

      iii. for males?

5. The goal of this problem is to find the top gender neutral names. Let's define a gender neutral name as a name that has approximately the same number of boys as the number of girls with this name. One way of doing this is to find the smaller of the ratios F/M and M/F, where F is the number of females with the name and M is the number of males with the name. If females with the name greatly outnumber males, then F/M will be large, but M/F will be small. If the sexes are about equal, then both ratios will be near one. The smaller ratio will never be greater than one, so the most balanced names are those with the smaller of the ratios near one.

   What are the top five most gender-neutral names? Restrict your analysis to names in which there are at least 10,000 of each gender. Do not include "Unknown" and "Baby" in your list of top five gender neutral names.

6. Using the function you wrote in problem (2), plot the #1 most general neutral name and the #10 most gender neutral name (these can be two separate plots).

7. For each year and gender, find the proportion of all names that are represented in the top 1000 names. (For a sanity check see https://www.ssa.gov/oact/babynames/limits.html which shows this percentage by year and gender for 2010 – 2019). Plot this proportion by year, with a separate line for males, females, and the total.

Part 2

Make a pandas DataFrame of significant earthquakes from 2001-2020 from the following Wikipedia pages:

`https://en.wikipedia.org/wiki/List_of_earthquakes_2001%E2%80%932010`

`https://en.wikipedia.org/wiki/List_of_earthquakes_2011%E2%80%932020`

Your DataFrame should have 693 rows. You should also have columns for the date, time, latitude, longitude, number of fatalities, and magnitude. Note that some entries in the "fatalities" column list the number of missing as well (for example: 70 dead 30 missing). The format for this columns is always the same (# dead # missing). Your "fatalities" columns should be the sum of the number dead and the number missing. Further note that number greater than 999 will have a comma (1,000).

8. For this problem, report both the month and the appropriate average:
   What month had:
   (a) the highest average number of earthquakes?
   (b) the lowest average number of earthquakes?
   (c) the highest average number of fatalities?
   (d) the lowest average number of fatalities?

9. For this problem, report both the hour and the appropriate average:
   What hour of the day had:
   (a) the highest average number of earthquakes?
   (b) the lowest average number of earthquakes?
   (c) the highest average number of fatalities?
   (d) the lowest average number of fatalities?

10. Round the magnitude to the nearest integer.
    (a) How many earthquakes of each rounded magnitude are in the data?
    (b) What is the total number of fatalities for each rounded magnitude?
    (c) What is the mean number of fatalities for each rounded magnitude?

11. Make a plot the earthquake longitude versus latitude. Change the size of the markers based on the square- or cube-root of the number of fatalities.