**Analyst Position Coding challenge – with Answers**

Using any programming language(s) you like, answer as many of the following questions as you can. It's okay if you don't do all the questions. If you get stuck, write pseudo-code or explain what you could do or would do.

Examples of languages in order of preference:  
R, python, SQL, Matlab/octave, Stata, SAS, perl, java, Julia, C, C++, javascript (node.js), objective C/Swift

***Don't use a spreadsheet in Excel or calculate by hand. Real datasets are too large for this to be practical.***

**About the data:**

The data come from CDC Wonder’s online open access data portal. The data contains state-level annual fertility information from the United States.

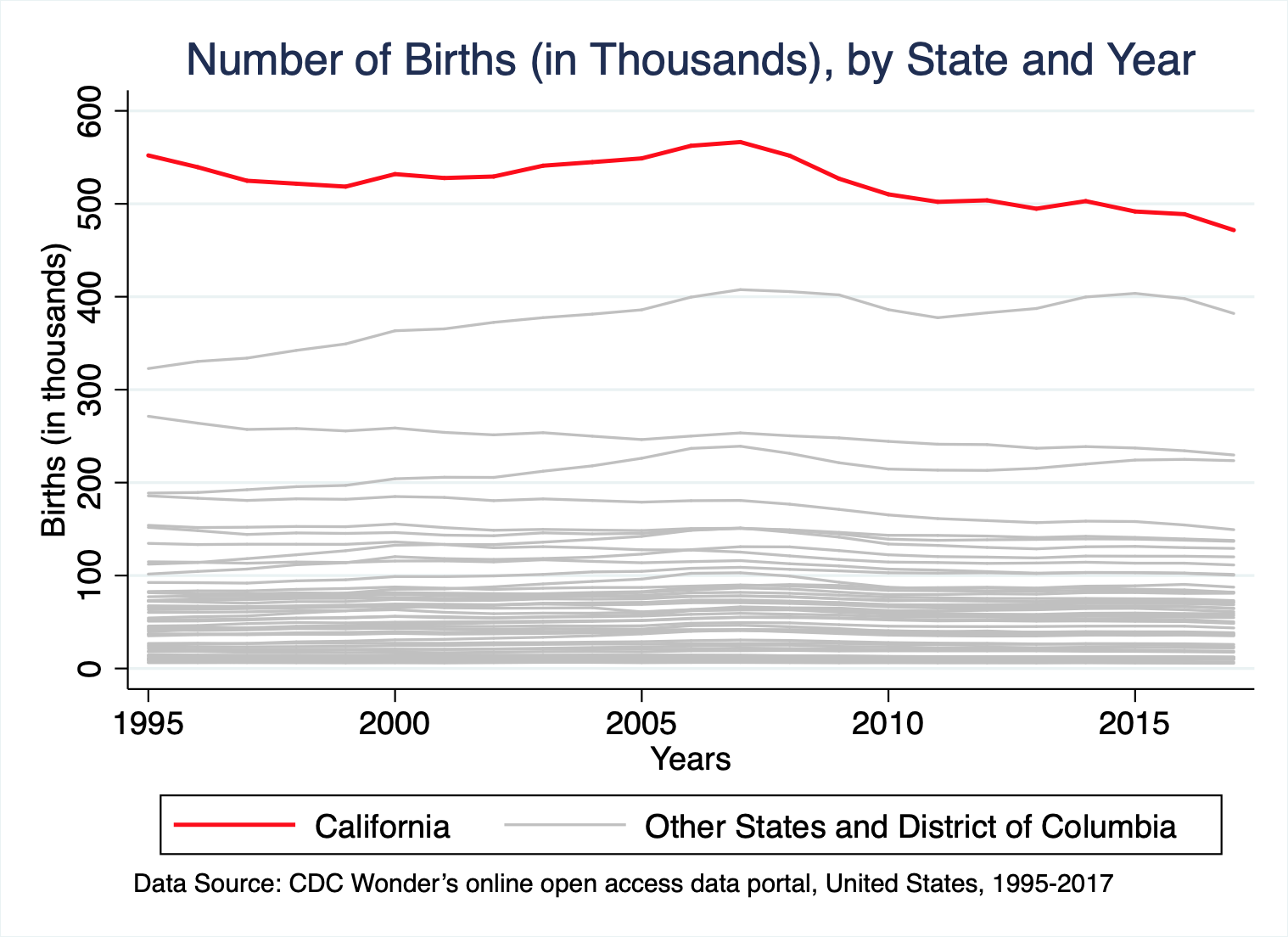
**Coding challenge questions**

1. Read in the data. *Bonus points for using the map package to avoid loops of repetitive code*
2. Dig into the data a bit! Explore the variables, examine missing data. Summarize the data as you see fit. Explain some basic data cleaning steps you took.

I followed these main steps to clean the data :

* 1. I identified and removed observations that had missing values for all variables except “notes”
  2. I made sure the state codes were the same for each state name in each of the 3 raw source files
  3. I looked at frequencies and number of observations with missing values for categorical variables
  4. I looked at mean, median, standard deviation, and minimum and maximum values for continuous variables; I also counted the number of observations with missing values
  5. After understanding why there still were missing values, I removed the observations with the “Total” note for each state and overall, since they will not be used in further analyses

1. Create a time series plot of the total number of births. *Bonus points for highlighting the state with the highest average number of births in red*



1. Estimate a regression model looking at how state and year impact birth rates.
   1. Did the birth rate increase or decrease over time?

The birth rate mostly decreased over time, as shown by the fact most coefficients associated with years are negative when compared to the omitted year (i.e., 2003). In addition, we can reject the null hypothesis that coefficients for all years are jointly equal to 0.

* 1. Interpret the intercept term

The intercept corresponds to the predicted birth rate for Alabama in 2003, since both statecode=1 (for Alabama) and year=2003 were omitted from the regression.

*10X bonus points for doing the challenge in RMarkdown and sharing your results in an HTML file*

*100X bonus points for doing the challenge and pushing your work to your Github repo*

*1000X points for presenting the results in a Shiny app*