R Practice

Bio 103

About this assignment

This assignment is a chance for you to practice data analysis skills in R a bit more. Because the goal is for you to learn data analysis, if you demonstrate your learning here we won't care whether you didn't demonstrate learning earlier. Thus, the grade on this assignment can replace the grade on a previous data analysis assignment. Make sure to learn this material before the next data analysis assignment so that one will be easy and you will do well on it too.

R Setup

- Make a new R project. If you are using a computer that does not belong to you make this project on your flash drive.

Load your data

- You will be working with data on populations from different countries around the world. Check out http://www.gapminder.org/ for information.
- You will work with a file called population_data.tsv

Your data should look like the following (except a lot more)

```
##       country year      pop continent lifeExp gdpPercap
## 1 Afghanistan 1952  8425333      Asia  28.801  779.4453
## 2 Afghanistan 1957  9240934      Asia  30.332  820.8530
## 3 Afghanistan 1962 10267083      Asia  31.997  853.1007
## 4 Afghanistan 1967 11537966      Asia  34.020  836.1971
## 5 Afghanistan 1972 13079460      Asia  36.088  739.9811
## 6 Afghanistan 1977 14880372      Asia  38.438  786.1134
```

- I named the variable holding my data "pops". You can name yours anything that is informative to you about the data. Click on the variable name to see the dataset. You should have a population size for each country for each year (along with some other information)

Plot your data

- Graph the change in population size for each country over time (hint: the population size is dependent on the year).

Hopefully you can guess (based on your plot) that if you fit a model to the data it won't fit very well. But maybe that's because there's so much variation between countries. There might be a pattern for each country individually.

- Filter out just the data for the United States using the following.

```
library(dplyr)  # don't forget to install.packages(dplyr) first
US <- filter(pops, country == "United States")
```

This command assigns to the variable `US`, every row of data from `pops` where the country column is equal to "United States".

- Plot the change in the US population over time.
- Fit a model to the US data (pop as a function of year) and plot this line.
- Fix your x and y axis labels
- Save your plot as a pdf and save your R script
- Repeat the previous analysis for another country to determine whether there is a pattern for each country even if this pattern is obscured when an analysis is done across all countries (if you are working with a friend make sure you pick a different country so we know you are doing your own work).