Lab 0

Rebecca Barter (based on materials by Yuval Benjamini, Jessica Li, Adam Bloniarz, and Ryan Giordano)

Due 9/1/2017

Note: This lab is not representative of the labs that you will receive in this class. Future labs will be significantly more difficult and open-ended.

This lab is not worth any grades; you can simply not complete the lab if you really don't want to, but you do need to submit *something* on GitHub (even if it is a blank `lab0.Rmd` and `lab0.pdf` file). This lab is an opportunity to make sure that you know how to submit your assignments, and for you to learn a little bit of Git/GitHub and R/tidyverse.

Install Git

1. Install Git on your system (https://git-scm.com/book/en/v2/Getting-Started-Installing-Git).
2. Sign up for GitHub (https://github.com/).
3. Go to https://education.github.com/ and sign up for the student pack to get unlimited private repositories.
4. Locally on your machine, clone my stat215a repository: `git clone https://github.com/rlbarter/stat215a`
5. On the GitHub website, log in and create a **private** remote repository called *stat215a*. Add me (*rlbarter*) as a collaborator for this repository (check out settings on the repo website).
6. Back in the terminal, set the origin of your local repository to be the remote repository that you just made. Change USERNAME below to your username. This tells git which remote repository to push your changes to when you `git push` (`git remote set-url origin https://github.com/USERNAME/stat215a.git`)
7. Inside this repository, you will find a file called *info.txt*. Edit this file to reflect your own information.
8. Check git status `git status`
9. Add (`git add info.txt`) and commit (`git commit -m "Updated info.txt with my own information"`) your edited *info.txt* file
10. Push your changes to your copy of the remote repository (`git push` or sometimes `git push remote origin`)
11. Check that info.txt has been updated in your remote github repository by navigating to https://github.com/USERNAME/stat215a (change USERNAME to your username)

Install R and RStudio

Install R from CRAN (https://cran.r-project.org/) and RStudio from RStudio (https://www.rstudio.com/products/RStudio/).

Install the tidyverse package in R

In the RStudio console, install the tidyverse package

```
# you only ever have to run the following once:
install.packages("tidyverse")
```

The best resource at the moment for learning the tidyverse is the book R for Data Science (http://r4ds.had.co.nz/) by Garrett Grolemund and Hadley Wickham. I also find the tidyverse website (https://www.tidyverse.org/) very helpful, but it probably not the place to start learning.

The tidyverse is actually a bundle of packages:

- `ggplot2` for visualization
- `dplyr` for data manipulation (SQL-style)
- `tidyr` for reshaping data (wide-form to long-form and vice verse)
- `readr` for loading data from a variety of formats
- `purr` for performing functional programming operations (e.g. maps to replace for-loops)
- `tibble` a more flexible alternative to data frames

The most important packages are `ggplot2` and `dplyr`, so if you decide to learn anything, learn these!

Other useful packages include:

- `lubridate` for dealing with dates
- `forcats` for dealing with factors

When writing code, you should follow the Google Style Guide (https://google.github.io/styleguide/Rguide.xml)

Analysis Instructions

Write up a report conducting the following analyses using R Markdown (if you prefer markdown) or R Sweave (if you prefer raw LaTeX). Note that both R Markdown and R Sweave can both can handle LaTeX equations contained within $ (inline) or $$ (new line) symbols.

1. Clone my STAT-215A-Fall-2017 repo (`git clone https://github.com/rlbarter/STAT-215A-Fall-2017`) to get the class materials and data for this week. These will live in the `week1/` folder.
2. If you have already done this, you can instead just pull any changes from the STAT-215A-Fall-2017 github repo (`git pull`).

Loading the data

1. Load USArrests in R (`data("USArrests")`)
2. Load the `statecoord.txt` into R from the `data/` folder in your cloned repository

Manipulating the data

1. Merge the two datasets together into a single data frame (using the `join()` functions from `dplyr`. Type `?dplyr::full_join`). Check that this worked correctly

Visualizing the data

1. Plot "Murder" vs "Assault" using the `geom_point()` function. What do you see?
2. Plot "Rape" vs urban population. There should be an outlier. Mark the outlier with a different colour.
3. Re-make these plots with the state names instead of the points (use `geom_text()`). Do you notice anything interesting?
4. Challenge exercise: Plot a map of the US colouring each state by its "Murder" rate. Check out `geom_polygon()`

Regression

You can fit a linear regression using the `lm()` function (or manually if you'd prefer!).

1. Fit a linear regression of urban population on "Rape".
2. Plot predicted values versus the residuals. Do you see any trends?
3. Replot "Rape" vs urban pop and draw a blue line with the predicted responses.
4. Now refit without the outlier and add a red line on the same plot.
5. Compare the lines. Are the linear responses a good description of the data?
6. Make a publishable graph. Add a header (`ggtitle`), axis labels (`xlab` and `ylab`) and customize the legend (`scale_color_manual`).

Submit the lab

Complete Lab 0 (within a folder called `lab0/`) and add, commit and push your changes to the stat215a github repository.

The `lab0/` folder (a sub-folder of `stat215a/`) should have the following structure:

```
lab0/
  data/
  lab0.Rmd
  lab0.pdf
  lab0_blind.Rmd
  lab0_blind.pdf
  R/
```