16:954:597:01 **Homework 5** Due: Wednesday, March 6th, 2019

- **NO** late submission will be accepted, except under special circumstances.
- Homework must be done individually and not in groups. Discussion of problems with others is permitted (and encouraged!), but you must write your own work in your own words.
- Submit your answers (via Canvas) as a single RMarkdown file that can be run on anyone's machine (i.e., that doesn't refer to your local files or directories). Your file name should have the following format: `lastname-NetID-week05.Rmd`. Make sure that your Rmarkdown file has yourself as author and has `output:html_document`.
- Be sure to include detailed explanatory text and remarks of what you are doing—don't just show a lot of `R` code and computer generated output. Use commands from the `tidyverse` and pipes whenever you can.

1. Download the texts of *Alice's Adventures in Wonderland* and *Great Expectations*, using the `gutenbergr` package:

   ```
   library(gutenbergr)
   books <- gutenberg_download(gutenberg_id = c(11, 1400),
   meta_fields = "title")
   ```

   (a) Find the 10 most common non-stop-words in *Great Expectations*. Create a world cloud of them.

   (b) Find the 10 most common bigrams in *Great Expectations* that do not include stop words.

   (c) Plot the sentiment for the two books.

2. Download the total federal R&D spending by agency/department here: `https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-02-12`.

   (a) Reformat this data with a separate variable called `rd_budget_frac` for the `rd_budget` as a fraction of `total_outlays` and two additional variables with upper and lower confidence bounds for `rd_budget_frac` from the linear model

   ```
   rd_budget_frac ~ department + year.
   ```

   Hint: the (95%) upper and lower confidence bounds for `rd_budget_frac` can be calculated using

   ```
   predict(lm, data, interval = "confidence"),
   ```

   or they can be calculated directly from `.fitted` and `.se.fit` from the `augment()` function in the `broom` package via

   $$\texttt{.fitted} \pm 1.96 \times \texttt{.se.fit}.$$

(b) Create four plots showing `rd_budget_frac` (along with the upper and lower confidence bounds from (a)) as a function of `year` for NASA, NSF, DHS, and DOD. (Be sure your figure looks polished.) Comment on any patterns you find.

3. Table 16 of the file UN_MigrantStockByOriginAndDestination_2015.xlsx (in the Week 5 Canvas folder) shows migration from one country to another in 2015. By eliminating the rows and columns that don't correspond to countries and then converting to a tidy dataset using `gather()` (i.e., a data frame with three columns: one for country of origin, one for country of destination, and the third indicating the number of people who immigrated from one country to the other), find

   (a) The top five countries from which people migrate to Canada.

   (b) The top five countries to which people migrate from Canada.

   (c) The top 10 migration pairs of countries.

   Your analysis can assume that the Excel file is in the same directory as your `RMarkdown` file.

   Hint 1: in `read_excel()`, you can name the sheet you want with the sheet option.
   Hint 2: The "country" codes for the regions and other non-countries start with 9. Try using regular expressions to eliminate these rows.