- **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-week03.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` whenever you can.

1. Read `R` for Data Science, Chapters 16, 13, and 14.1–14.3. Do the problems as you go along (you don't need to hand them in).

2. We would like to create a data frame like the `babynames` data frame, but for baseball players.

   (a) Use the `Master` data frame in the `Lahman` package to create a tibble with similar variables as the `babynames` data frame (i.e., `year`, `name`, and `n`), and ordered in the same way. You will need to use the `summarize()` function to get the counts of each name's use (according to the `nameFirst` variable). For year, use the year of birth.

   (b) In the `Master` dataframe, is the variable `birthYear` consistent with the year in `birthDate`? Use a function in the `lubridate` package to extract the year, and then use pipes and the `table()` function to see how often the first equals the second.

   (c) Create a data frame of players showing just the `playerID`, `first name`, `last name`, `given name`, and `career total` (meaning, summed over all years and all stints) of games (that is, the `G` variable) according to the `Fielding` data frame.

   (d) Using `mutate()` and `str_c()`, add a variable to your data frame in (c) for full name by combining the first name and last name with a space between them.

   (e) Use the data frames you've created to determine the 5 most popular first names in baseball among players who played at least 500 games. That is, first use the data frame from (c) to determine the set of players who played at least 500 games. Then create a data frame similar to (a), expect here, only consider the players who played at least 500 games. From them, determine the 5 most popular first names across all years. Plot them over time with lines in a single plot. Be sure to make the plot look nice by using a title and changing the axis labels if necessary.

3. We would like to answer some questions about the NYC Restaurant Inspections dataset.

(a) Using the `violation_description` variable, how many of the violations were issued for having a rodent related issue? To do this, search for the words "rodent", "mice", and "rat" using the appropriate command from the `stringr` library and tally their sum. Are restaurant owners more likely to get cited for rats or mice? Are there any cases of restaurants that are cited for both rats and mice? (Remember that the same restaurant can have multiple violations).

(b) Do the same thing as above, but look for violations involving insects. That is, search for "roach" and "flies". Which violation is more common? Are there any cases of restaurants that are cited for both roaches and flies?

(c) Are there any restaurants that have been cited for both rodents and insects?

(d) Using the `boro` variable, rank the five boroughs according to the total number of rodent and/or insect violations.