- **NO** late submission will be accepted, except under special circumstances.
- Midterm must be done individually. Discussion of problems with others is strictly prohibited.
- 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_midterm.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. Load the Philadelphia parking violations dataset from `https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-12-03` and answer the following questions.
   (a) What is the *most* common violation and its average fine? What are the *least* common violations and their respective fines?
   (b) For the most common violation, create a plot in `ggplot` of the number of violations for each *month* of 2017. What conclusions can you make?
   (c) The zip code for downtown Philadelphia is 19103. Are you more likely to get a ticket here than elsewhere? To visualize this, create a graph of the number of violations for each zip code.
2. The link `https://www.cnn.com/2020/03/03/health/us-coronavirus-cases-state-by-state/index.html` contains a breakdown of US cases and deaths of COVID-19 by state.
   (a) Using the node `.zn-body__paragraph` to extract the HTML content, create a data frame which lists the number of cases and deaths for each state. In particular, create a data frame with three columns, called `state`, `cases`, and `deaths` for all 51 states (including the District of Columbia). For example, you will need to extract numeric values of 1 and 2, respectively, from the text 'including one death' and 'including two deaths'. Be sure to make your code reproducible, as this is a live link that is being updated frequently.
   (b) Show the table in your `RMarkdown` file by using the function `knitr::kable()`.
3. Taking advantage of the `rvest` package, turn the table at `http://www.nature.com/articles/ng.3097/tables/3` into an `R` data frame.
   (a) Be sure to delete the rows "Genes with previous literature support (GRAIL)" and "New genes without previous evidence" (and don't do it by using the row number).
   (b) Be sure to convert the $p$-value column to numbers.

(c) The last column is shown in 3 rows in the journal, but most likely as one string in your table. Use regular expressions to insert semicolons (i.e., ";") between each of the original lines. For example,

"PI3K cascade (REACTOME, P = 6.2 x 10-13); Chronic myeloid leukemia (KEGG, P = 1.6 x 10-12); Response to fibroblast growth factor stimulus (GO, P = 5.4 x 10-11)"

(d) Show the table in your `RMarkdown` file by using the function `knitr::kable()`.