MSCS 264: Homework #11

Due Tues May 1 at 11:59 PM

You should submit a knitted pdf file on Moodle, but be sure to show all of your R code, in addition to your output, plots, and written responses.

Web scraping

1. Read in the table of data found at https://en.wikipedia.org/wiki/List_of_United_States_cities_by_crime_rate and create a plot showing violent crime rate (total violent crime / population) vs. property crime rate (total property crime / population). Identify outlier cities by using a plotting command such as:

```
ggplot(crimes4, aes(x = VCrate, y = PCrate, label = City)) +
    geom_point() +
    geom_text(data=subset(crimes4, VCrate > .004), check_overlap = TRUE,
              size = 2.5, nudge_y = 0.001)
```

Hints:

- after reading in the table using `html_table()`, create a data frame with just the columns you want, using a command such as: `crimes3 <- as.data.frame(crimes2)[,c(LIST OF COLUMN NUMBERS)]`. Otherwise, R gets confused since it appears as if several columns all have the same column name.
- then, turn `crimes3` into a tibble with `as.tibble(crimes3)` and do necessary tidying: get rid of unneeded rows, parse columns into proper format, etc.

2. As we did in class, use the `rvest` package to pull off data from imdb's top grossing films released in 2017 at https://www.imdb.com/search/title?year=2017&title_type=feature&sort=boxoffice_gross_us,desc. Create a tibble that contains the title, gross, imdbscore, and metascore for the top 50 films. Then generate a scatterplot of one of the ratings vs. gross, labelling outliers as in Question 1 with the title of the movie.

3. 5 points if you push your Rmd file with HW11 solutions along with the knitted pdf file to your MSCS264-HW11 repository in your GitHub account. So that I can check, make your repository private (good practice when doing HW), but add me (username = proback) as a collaborator under Settings > Collaborators.

Factors

Read Chapter 15 on factors and attempt the following problems:

4. In the `nycflights13` data, just consider flights to O'Hare (dest=="ORD"), and summarize the mean arrival delay by carrier (actually use the entire name of the carrier after merging carrier names into `flights`). Then use `geom_point` to plot mean arrival delay vs. carrier - first without reordering carrier names, and second after reordering carrier names by mean arrival delay.

5. Again considering only flights to O'Hare, create a new factor variable which differentiates national carriers (American and United) from regional carriers (all other which fly to O'Hare). Then create a violin plot comparing arrival delays for all flights to O'Hare from those two groups (you might want to exclude arrival delays over a certain level).