

MSCS 264: Homework #13

Due Tues Nov 20 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) vs. property crime rate (total property crime). Identify outlier cities (those with “extreme” values for VCrate and/or PCrate) by feeding a data set of outliers into `geom_label_repel()`.

```
url <- "http://en.wikipedia.org/wiki/List_of_United_States_cities_by_crime_rate"

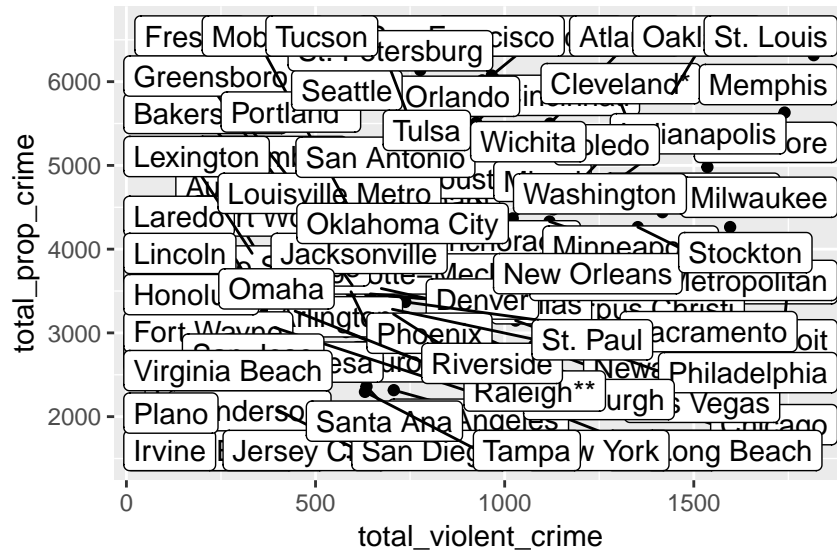
data <- read_html(url)
tables <- html_nodes(data, css = "table")
crimes <- html_table(tables, header = TRUE, fill = TRUE)[[2]]

crimes2 <- as.data.frame(crimes)[c(1:13)]

crimes3 <- as.tibble(crimes2)

crimes4 <- crimes3 %>%
  rename(`total_violent_crime` = `Violent Crime`,
         `murder/manslaughter` = `Violent Crime.1`,
         `rape` = `Violent Crime.2`, `robbery` = `Violent Crime.3`,
         `agg_assault` = `Violent Crime.4`, `total_prop_crime` = `Property Crime`,
         `burglary` = `Property Crime.1`, `larceny` = `Property Crime.2`,
         `gta` = `Property Crime.3`) %>%
  slice(2:n()) %>%
  mutate(`total_violent_crime` = as.numeric(`total_violent_crime`),
         `murder/manslaughter` = as.numeric(`murder/manslaughter`),
         `rape` = as.numeric(`rape`), `robbery` = as.numeric(`robbery`),
         `agg_assault` = as.numeric(`agg_assault`), `total_prop_crime` = as.numeric(`total_prop_crime`),
         `burglary` = as.numeric(`burglary`), `larceny` = as.numeric(`larceny`),
         `gta` = as.numeric(`gta`), Population = parse_number(Population))

ggplot(crimes4, aes(x = total_violent_crime, y = total_prop_crime)) +
  geom_jitter() +
  geom_label_repel(label = crimes4$City)
```



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.

Answer: St. Louis, Oakland, Memphis, Irvine, and Plano have extreme values of VCrate and PCrate.

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.

```
url <- "https://www.imdb.com/search/title?year=2017&title_type=feature&sort=boxoffice_gross_us,desc"

movies <- read_html(url)
movies2 <- html_nodes(movies, '.ratings-metascore , strong , .ghost~ .text-muted+ span , .list-item-h')
movies3 <- html_text(movies2)
movies4 <- c()

for(i in 7:206) {
  movies4[i-6] <- movies3[i]
}

movie_table4 <- tibble(title = character(), gross = double(), imdb_score = double(), metascore = double())

for(i in 1:49) {
  movie_table4[i,1] <- movies4[1+4*i]
  movie_table4[i,3] <- movies4[2+4*i]
  movie_table4[i,4] <- movies4[3+4*i]
  movie_table4[i,2] <- movies4[4+4*i]
}

movie_table4 <- movie_table4 %>%
  mutate(gross_mil = parse_number(gross),
```

```

    imdb_score = parse_number(imdb_score),
    metascore = parse_number(metascore))%>%
select(title, gross_mil, imdb_score, metascore)

ggplot(movie_table4, aes(x = metascore, y = gross_mil))+
  geom_point()+
  geom_label_repel(label = movie_table4$title)

```

3. 5 points if you push your Rmd file with HW13 solutions along with the knitted pdf file to your MSCS264-HW13 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.

```

ohare <- flights %>%
  filter(dest == "ORD") %>%
  group_by(carrier)%>%
  summarise(avg_delay = mean(arr_delay, na.rm = TRUE))%>%
  left_join(airlines)

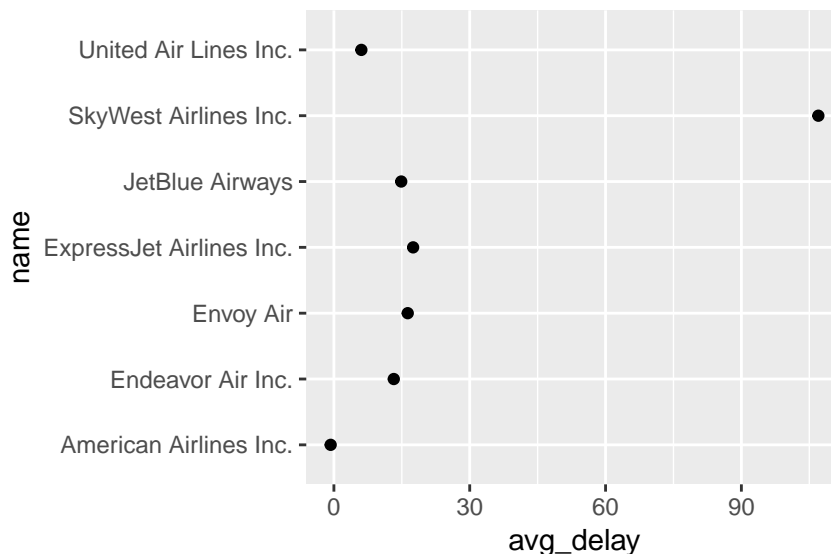
```

```
## Joining, by = "carrier"
```

```

ggplot(ohare, aes(x = avg_delay, y = name))+
  geom_point()

```

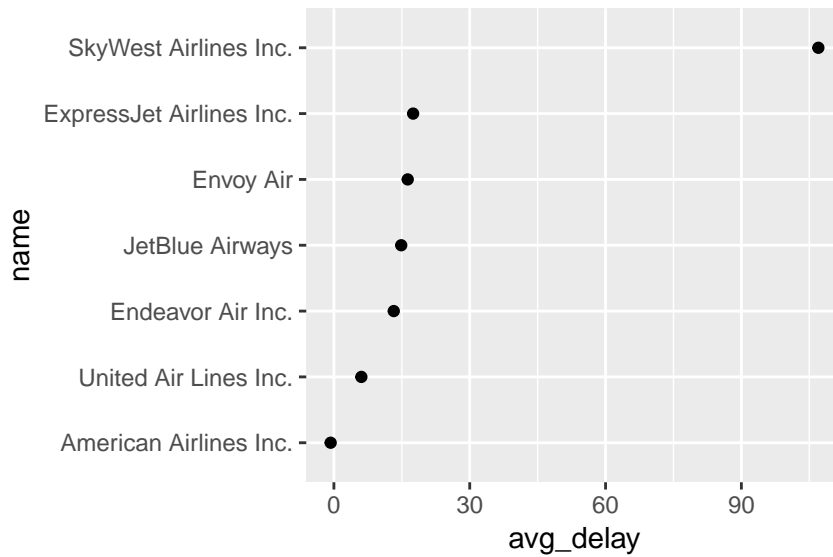


```

ohare2 <- ohare %>%
  mutate(name = fct_reorder(name, avg_delay))

ggplot(ohare2, aes(x = avg_delay, y = name))+
  geom_point()

```



5. Again considering only flights to O'Hare, create a new factor variable which differentiates national carriers (American and United) from regional carriers (all others 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).

```
ohare3 <- flights %>%
  filter(dest == "ORD") %>%
  group_by(carrier)%>%
  summarise(avg_delay = mean(arr_delay, na.rm = TRUE))%>%
  left_join(airlines)%>%
  mutate(airline_type = ifelse(name == "United Air Lines Inc." | name == "American Airlines Inc.", "national", "regional"),
         avg_delay = ifelse(avg_delay < 0, 0, avg_delay))%>%
  filter(avg_delay < 30)
```

```
## Joining, by = "carrier"
```

```
ggplot(ohare3, aes(x = airline_type, y = avg_delay))+
  geom_violin()
```

