Assignment 1

Nicole Blackburn

2025-09-16

This assignment is to be submitted individually. Turn in this assignment as an HTML or PDF file to ELMS. Make sure to include the R Markdown or Quarto file that was used to generate it. You should include the questions in your solutions. You may use the qmd file of the assignment provided to insert your answers.

Git and GitHub

1) Provide the link to the GitHub repo that you used to practice git from Week 1. It should have:

- Your name on the README file.
- At least one commit with your name, with a description of what you did in that commit.

Click here to see my practice Github Repo from week 1.

Reading Data

Download both the Angell.dta (Stata data format) dataset and the Angell.txt dataset from this website: https://stats.idre.ucla.edu/stata/examples/ara/applied-regression-analysis-by-fox-data-files/

2) Read in the .dta version and store in an object called `angell_stata`.

```
library(haven)
angell_stata <- read_dta("angell.dta")
```

3) Read in the .txt version and store it in an object called `angell_txt`.

```r
angell_txt <- read.table(file = "angell.txt")
```

4) What are the differences between `angell_stata` and `angell_txt`? Are there differences in the classes of the individual columns?

```r
angell_stata
```

```
# A tibble: 43 x 5
   city       morint ethhet geomob region
   <chr>       <dbl>  <dbl>  <dbl> <chr>
 1 Rochester    19     20.6   15   E
 2 Syracuse     17     15.6   20.2 E
 3 Worcester    16.4   22.1   13.6 E
 4 Erie         16.2   14     14.8 E
 5 Milwaukee    15.8   17.4   17.6 MW
 6 Bridgeport   15.3   27.9   17.5 E
 7 Buffalo      15.2   22.3   14.7 E
 8 Dayton       14.3   23.7   23.8 MW
 9 Reading      14.2   10.6   19.4 E
10 Des_Moines   14.1   12.7   31.9 MW
# i 33 more rows
```

```r
summary(angell_stata)
```

```
     city               morint          ethhet          geomob
 Length:43          Min.   : 4.20   Min.   :10.60   Min.   :12.10
 Class :character   1st Qu.: 8.70   1st Qu.:16.90   1st Qu.:19.45
 Mode  :character   Median :11.10   Median :23.70   Median :25.90
                    Mean   :11.20   Mean   :31.37   Mean   :27.60
                    3rd Qu.:13.95   3rd Qu.:39.00   3rd Qu.:34.80
                    Max.   :19.00   Max.   :84.50   Max.   :49.80
   region
 Length:43
 Class :character
 Mode  :character
```

```
angell_txt
```

```
                V1   V2   V3   V4 V5
1        Rochester 19.0 20.6 15.0  E
2         Syracuse 17.0 15.6 20.2  E
3        Worcester 16.4 22.1 13.6  E
4             Erie 16.2 14.0 14.8  E
5        Milwaukee 15.8 17.4 17.6 MW
6       Bridgeport 15.3 27.9 17.5  E
7          Buffalo 15.2 22.3 14.7  E
8           Dayton 14.3 23.7 23.8 MW
9          Reading 14.2 10.6 19.4  E
10      Des_Moines 14.1 12.7 31.9 MW
11       Cleveland 14.0 39.7 18.6 MW
12          Denver 13.9 13.0 34.5  W
13          Peoria 13.8 10.7 35.1 MW
14         Wichita 13.6 11.9 42.7 MW
15         Trenton 13.0 32.5 15.8  E
16    Grand_Rapids 12.8 15.7 24.2 MW
17           Toledo 12.7 19.2 21.6 MW
18       San_Diego 12.5 15.9 49.8  W
19       Baltimore 12.0 45.8 12.1  E
20      South_Bend 11.8 17.9 27.4 MW
21           Akron 11.3 20.4 22.1 MW
22         Detroit 11.1 38.3 19.5 MW
23          Tacoma 10.9 17.8 31.2  W
24           Flint  9.8 19.3 32.2 MW
25         Spokane  9.6 12.3 38.9  W
26         Seattle  9.0 23.9 34.2  W
27    Indianapolis  8.8 29.2 23.1 MW
28        Columbus  8.0 27.4 25.0 MW
29 Portland_Oregon  7.2 16.4 35.8  W
30        Richmond 10.4 65.3 24.9  S
31         Houston 10.2 49.0 36.1  S
32      Fort_Worth 10.2 30.5 36.8  S
33   Oklahoma_City  9.7 20.7 47.2  S
34     Chattanooga  9.3 57.7 27.2  S
35       Nashville  8.6 57.4 25.4  S
36      Birmingham  8.2 83.1 25.9  S
37          Dallas  8.0 36.8 37.8  S
38      Louisville  7.7 31.5 19.4  S
39    Jacksonville  6.0 73.7 27.7  S
```

```
40         Memphis  5.4 84.5 26.7  S
41           Tulsa  5.3 23.8 44.9  S
42           Miami  5.1 50.2 41.8  S
43         Atlanta  4.2 70.6 32.6  S
```

```
summary(angell_txt)
```

```
      V1                  V2              V3              V4
 Length:43          Min.   : 4.20   Min.   :10.60   Min.   :12.10
 Class :character   1st Qu.: 8.70   1st Qu.:16.90   1st Qu.:19.45
 Mode  :character   Median :11.10   Median :23.70   Median :25.90
                    Mean   :11.20   Mean   :31.37   Mean   :27.60
                    3rd Qu.:13.95   3rd Qu.:39.00   3rd Qu.:34.80
                    Max.   :19.00   Max.   :84.50   Max.   :49.80
      V5
 Length:43
 Class :character
 Mode  :character
```

- Angell_stata loads as a tibble and angell_txt loads as a data frame. Angell_txt columns were given names. The individual column classes are the same. City and Region are both (character or strings) classes. Moral Integration, Ethnic Heterogeneity, and Graphic Mobility are all (double or number) classes.

5) Make any updates necessary so that `angell_txt` is the same as `angell_stata`.

```
angell_txt <- angell_txt |>
  rename(city = V1, morint = V2, ethhet = V3, geomob = V4, region = V5)
```

6) Describe the Ethnic Heterogeneity variable. Use descriptive statistics such as mean, median, standard deviation, etc. How does it differ by region?

```
angell_stata |>
  group_by(region) |>
  summarise(mean = mean(ethhet),
            median = median(ethhet),
            sd = sd(ethhet),
            IQR = IQR(ethhet))
```

```
# A tibble: 4 x 5
  region  mean median    sd   IQR
  <chr>  <dbl>  <dbl> <dbl> <dbl>
1 E       23.5   22.1 10.8  12.3
2 MW      21.7   19.2  9.08 10.4
3 S       52.5   53.8 21.4  36.4
4 W       16.5   16.1  4.16  3.72
```

- The angell dataset contains data on the moral integration of American cities. The Ethnic Heterogeneity variable comes from percentages of nonwhite and foreign-born white residents. A higher value for ethnic heterogeneity can be interpreted as having more nonwhite and foreign-born white residents within the city. Each city is assigned a region in the United States: Northeast (E), Midwest (MW), Southeast (S), or West (W).
- When grouping by region, we can see the cities in the Southeast has the highest average heterogeneity with a mean value of 52.49%, with Northeast, Midwest, and West having mean values at 23.49%, 21.68%, and 16.55% respectively.
- IQR and sd measure the spread of the data. The Southeast region has the greatest spread in the data, with 36.45 and 21.44 for sd and IQR respectively, revealing that the cities located in the Southeast region differ from one another in terms of ethnic heterogeneity: some with a lower percentage and some with a greater percentage. In contrast, the West region has the lowest spread in the data, with 4.16 and 3.72 for sd and IQR respectively, revealing that the cities within the West region are more similar to one another in regard to ethnic heterogeneity.
- When comparing the Northeast and Midwest regions, values for mean, median, sd, and IQR look similar. Ethnic heterogeneity on average in the Northeast is 23.49%, while the Midwest region is 21.68%.

Describing Data

R comes also with many built-in datasets. The "MASS" package, for example, comes with the "Boston" dataset.

7) Install the "MASS" package, load the package. Then, load the Boston dataset.

```
df <- MASS::Boston
```

8) What is the type of the Boston object?

```
typeof(Boston)
```

```
[1] "list"
```

- Boston is a list.

9) What is the class of the Boston object?

```
attributes(Boston)
```

```
$names
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
 [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"

$class
[1] "data.frame"

$row.names
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
 [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
 [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
 [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
 [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
 [91]  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
[109] 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
[127] 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
[145] 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162
[163] 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
[181] 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198
[199] 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216
[217] 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234
[235] 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252
[253] 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270
[271] 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288
[289] 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306
[307] 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324
[325] 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342
[343] 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360
[361] 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378
[379] 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396
[397] 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414
[415] 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432
[433] 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450
[451] 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468
[469] 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486
[487] 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504
[505] 505 506
```

- Boston is a data frame.

10) How many of the suburbs in the Boston data set bound the Charles river?

```
table(df$chas)
```

```
  0   1
471  35
```

- There are 35 suburbs in the Boston data set that bound the Charles River.

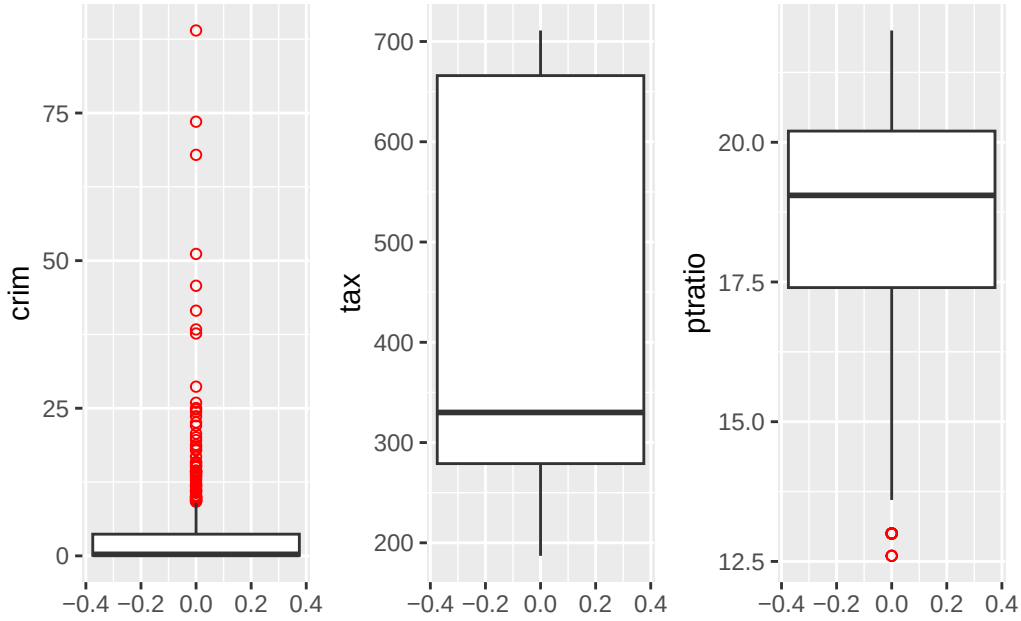
11) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each variable.

```
summary(df[c("crim", "tax", "ptratio")])
```

```
      crim                tax           ptratio
 Min.   : 0.00632   Min.   :187.0   Min.   :12.60
 1st Qu.: 0.08205   1st Qu.:279.0   1st Qu.:17.40
 Median : 0.25651   Median :330.0   Median :19.05
 Mean   : 3.61352   Mean   :408.2   Mean   :18.46
 3rd Qu.: 3.67708   3rd Qu.:666.0   3rd Qu.:20.20
 Max.   :88.97620   Max.   :711.0   Max.   :22.00
```

```
p1 <- ggplot(Boston, aes(y = crim)) +
        geom_boxplot(outlier.color = "red", outlier.shape = 1)
p2 <- ggplot(Boston, aes(y = tax)) +
         geom_boxplot(outlier.color = "red", outlier.shape = 1)
p3 <- ggplot(Boston, aes(y = ptratio)) +
         geom_boxplot(outlier.color = "red", outlier.shape = 1)

p1 + p2 + p3
```



- The vast majority of suburbs have low crime rates with an average per capita crime rate of 3.61%. Furthermore, 75% of all suburbs have a crime rate of 3.67% or less. Visualized by the boxplot, we see that some suburbs do have particularly high crime rates. Outliers are identified in red with the highest crime rate for one suburb being 88.98%.
- Full-value property-tax rates per $10,000 are visualized in the middle boxplot. The spread of tax rates is very wide, the lowest rate being 187 and the highest rate being 711; however, all of the property-tax rate values fall within the whiskers, meaning no outliers are present. Compared to the mean value of 408.2, the towns with higher tax rates are not particularly high.
- In the pupil-teacher ratio, the values range from 12.6 to 22.0. With a mean value of 18.46 and a median of 19.05, we see that some lower values are present. On the boxplot, two outliers are identified in red on the lower side, indicating there are some particularly low values for pupil-teacher ratio but no particularly high values.

12) Describe the distribution of pupil-teacher ratio among the towns in this data set that have a per capita crime rate larger than 1. How does it differ from towns that have a per capita crime rate smaller than 1?

```
crime_high <- df |>
  filter(df$crim > 1)

crime_low <- df |>
  filter(df$crim < 1)
```

```
summary(crime_high$ptratio)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.70   20.20   20.20   19.29   20.20   21.20
```

```
summary(crime_low$ptratio)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.60   16.80   18.30   18.02   19.20   22.00
```

- The distribution of pupil-teacher ratio among the towns in this data set that have a per capita crime rate larger than 1 ranges from 14.7 to 21.2. For towns with a per capita crime rate less than 1, the pupil-teacher ratio ranges from 12.6 to 22.0. Although towns with a per crime rate of less than 1 have a higher maximum pupil-teacher ratio, their mean and median pupil-teacher ratio values are lower than towns with a per capita crime rate greater than 1. Additionally, the minimum pupil-teacher rate is 12.6 which is found in a town with a per capita crime rate of less than 1.

Writing Functions

13) Write a function that calculates 95% confidence intervals for a point estimate. The function should be called `my_CI`. When called with `my_CI(2, 0.2)`, the function should print out "The 95% CI upper bound of point estimate 2 with standard error 0.2 is 2.392. The lower bound is 1.608."

*Note: The function should take a point estimate and its standard error as arguments. You may use the formula for 95% CI: point estimate +/- 1.96*standard error.*

Hint: Pasting text in R can be done with: `paste()` *and* `paste0()`

```
library(purrr)
my_CI <- function(point_est, se) {
  lower <- point_est - 1.96*se
  upper <- point_est + 1.96*se

  cat(paste0("The 95% CI upper bound of point estimate ", point_est,
         " with standard error ", se, " is ", upper,".\n",
         " The lower bound is ", lower,."))

}
```

```r
# call function
my_CI(2, 0.2)
```

```
The 95% CI upper bound of point estimate 2 with standard error 0.2 is 2.392.
 The lower bound is 1.608.
```

14) Create a new function called `my_CI2` that does that same thing as the `my_CI` function but outputs a vector of length 2 with the lower and upper bound of the confidence interval instead of printing out the text. Use this to find the 95% confidence interval for a point estimate of 0 and standard error 0.4.

```r
my_CI2 <- function(point_est, se) {
  lower <- point_est - 1.96*se
  upper <- point_est + 1.96*se
  return(c(lower, upper))

}

# 95% CI for point estimate 0 and standard error 0.4
my_CI2(0,0.4)
```

```
[1] -0.784  0.784
```

- The 95% confidence interval for a point estimate of 0 and standard error 0.4 is (-0.784, 0.784).

15) Update the `my_CI2` function to take any confidence level instead of only 95%. Call the new function `my_CI3`. You should add an argument to your function for confidence level.

Hint: Use the `qnorm` *function to find the appropriate z-value. For example, for a 95% confidence interval, using* `qnorm(0.975)` *gives approximately 1.96.*

```r
my_CI3 <- function(point_est, se, conf_lvl) {
  z_score <- qnorm(1-(1-conf_lvl)/2)
  lower <- point_est - z_score*se
  upper <- point_est + z_score*se
  return(c(lower, upper))

}

# test
my_CI3(0, 1, 0.95)
```

```
[1] -1.959964  1.959964
```

16) Without hardcoding any numbers in the code, find a 99% confidence interval for Ethnic Heterogeneity in the Angell dataset. Find the standard error by dividing the standard deviation by the square root of the sample size.

```
CI_eth <- function(point_est, se, conf_lvl) {
  z_score <- qnorm(1-(1-conf_lvl)/2)
  lower <- point_est - z_score*se
  upper <- point_est + z_score*se
  return(c(lower, upper))

}

CI_eth(mean(angell_stata$ethhet),
       sd(angell_stata$ethhet)/(length(angell_stata$ethhet))^1/2,
       0.99)
```

```
[1] 30.76074 31.98345
```

- The 99% confidence interval for Ethnic Heterogeneity in the Angell Dataset is (30.76074, 31.98345).

17) Write a function that you can `apply` to the Angell dataset to get 95% confidence intervals. The function should take one argument: a vector. Use if-else statements to output NA and avoid error messages if the column in the data frame is not numeric or logical.

```
angell_CI <- function(x) {
  if(is.numeric(x) || is.logical(x)){
    x <- na.omit(x)
    z_score <- qnorm(0.975)

    lower <- mean(x) - z_score*sd(x)/sqrt(length(x))
    upper <- mean(x) + z_score*sd(x)/sqrt(length(x))

  return(c(lower, upper))
  } else{
    return(c(NA))
  }
}

# test
lapply(angell_stata, angell_CI)
```

```
$city
[1] NA

$morint
[1] 10.13242 12.26758

$ethhet
[1] 25.27127 37.47292

$geomob
[1] 24.67187 30.52347

$region
[1] NA
```