Homework01

Emmenta Janneh

2024-01-18

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

R Basic

1. Load these data into R, directly using the URL (i.e., don't download it first to load it into R). Call the resulting data frame cdi.

```
cdi <- read.csv("https://dcgerard.github.io/stat_415_615/data/cdi.csv")
head(cdi)
```

```
##   id      county state area     pop percent_18_34 percent_65 physicians  beds
## 1  1 Los_Angeles    CA 4060 8863164          32.1        9.7      23677 27700
## 2  2        Cook    IL  946 5105067          29.2       12.4      15153 21550
## 3  3      Harris    TX 1729 2818199          31.3        7.1       7553 12449
## 4  4   San_Diego    CA 4205 2498016          33.5       10.9       5905  6179
## 5  5      Orange    CA  790 2410556          32.6        9.2       6062  6369
## 6  6       Kings    NY   71 2300664          28.3       12.4       4861  8942
##   crimes high_school bachelors poverty unemployment capita_income total_income
## 1 688936        70.0      22.3    11.6          8.0         20786       184230
## 2 436936        73.4      22.8    11.1          7.2         21729       110928
## 3 253526        74.9      25.4    12.5          5.7         19517        55003
## 4 173821        81.9      25.3     8.1          6.1         19588        48931
## 5 144524        81.2      27.8     5.2          4.8         24400        58818
## 6 680966        63.7      16.6    19.5          9.5         16803        38658
##   region
## 1      W
```

```
## 2     NC
## 3      S
## 4      W
## 5      W
## 6     NE
```

2. reate a new variable called log_capita_income, which is the log-transformed capita_income. Make sure this new variable is present in the cdi data frame.

```
cdi <- mutate(cdi, log_capita_income = log(cdi$capita_income))
head(cdi)
```

```
##   id      county state area     pop percent_18_34 percent_65 physicians  beds
## 1  1 Los_Angeles    CA 4060 8863164          32.1        9.7      23677 27700
## 2  2        Cook    IL  946 5105067          29.2       12.4      15153 21550
## 3  3      Harris    TX 1729 2818199          31.3        7.1       7553 12449
## 4  4   San_Diego    CA 4205 2498016          33.5       10.9       5905  6179
## 5  5      Orange    CA  790 2410556          32.6        9.2       6062  6369
## 6  6       Kings    NY   71 2300664          28.3       12.4       4861  8942
##   crimes high_school bachelors poverty unemployment capita_income total_income
## 1 688936        70.0      22.3    11.6          8.0         20786       184230
## 2 436936        73.4      22.8    11.1          7.2         21729       110928
## 3 253526        74.9      25.4    12.5          5.7         19517        55003
## 4 173821        81.9      25.3     8.1          6.1         19588        48931
## 5 144524        81.2      27.8     5.2          4.8         24400        58818
## 6 680966        63.7      16.6    19.5          9.5         16803        38658
##   region log_capita_income
## 1      W          9.942035
## 2     NC          9.986403
## 3      S          9.879041
## 4      W          9.882672
## 5      W         10.102338
## 6     NE          9.729313
```

3. Use R to calculate the mean and standard deviation of area.

```
area_mean <- mean(cdi$area)
area_sd <- sd(cdi$area)

area_mean
```

```
## [1] 1041.411
```

```
area_sd
```

```
## [1] 1549.922
```

4. Rename the pop variable to population. Make sure the cdi data frame has been modified.

```
cdi <- rename(cdi, population = pop)
head(cdi)
```

```
##   id      county state area population percent_18_34 percent_65 physicians
## 1  1 Los_Angeles    CA 4060    8863164          32.1        9.7      23677
## 2  2        Cook    IL  946    5105067          29.2       12.4      15153
## 3  3      Harris    TX 1729    2818199          31.3        7.1       7553
## 4  4   San_Diego    CA 4205    2498016          33.5       10.9       5905
## 5  5      Orange    CA  790    2410556          32.6        9.2       6062
## 6  6       Kings    NY   71    2300664          28.3       12.4       4861
##    beds crimes high_school bachelors poverty unemployment capita_income
## 1 27700 688936        70.0      22.3    11.6          8.0         20786
## 2 21550 436936        73.4      22.8    11.1          7.2         21729
## 3 12449 253526        74.9      25.4    12.5          5.7         19517
## 4  6179 173821        81.9      25.3     8.1          6.1         19588
## 5  6369 144524        81.2      27.8     5.2          4.8         24400
## 6  8942 680966        63.7      16.6    19.5          9.5         16803
##   total_income region log_capita_income
## 1       184230      W          9.942035
## 2       110928     NC          9.986403
## 3        55003      S          9.879041
## 4        48931      W          9.882672
## 5        58818      W         10.102338
## 6        38658     NE          9.729313
```

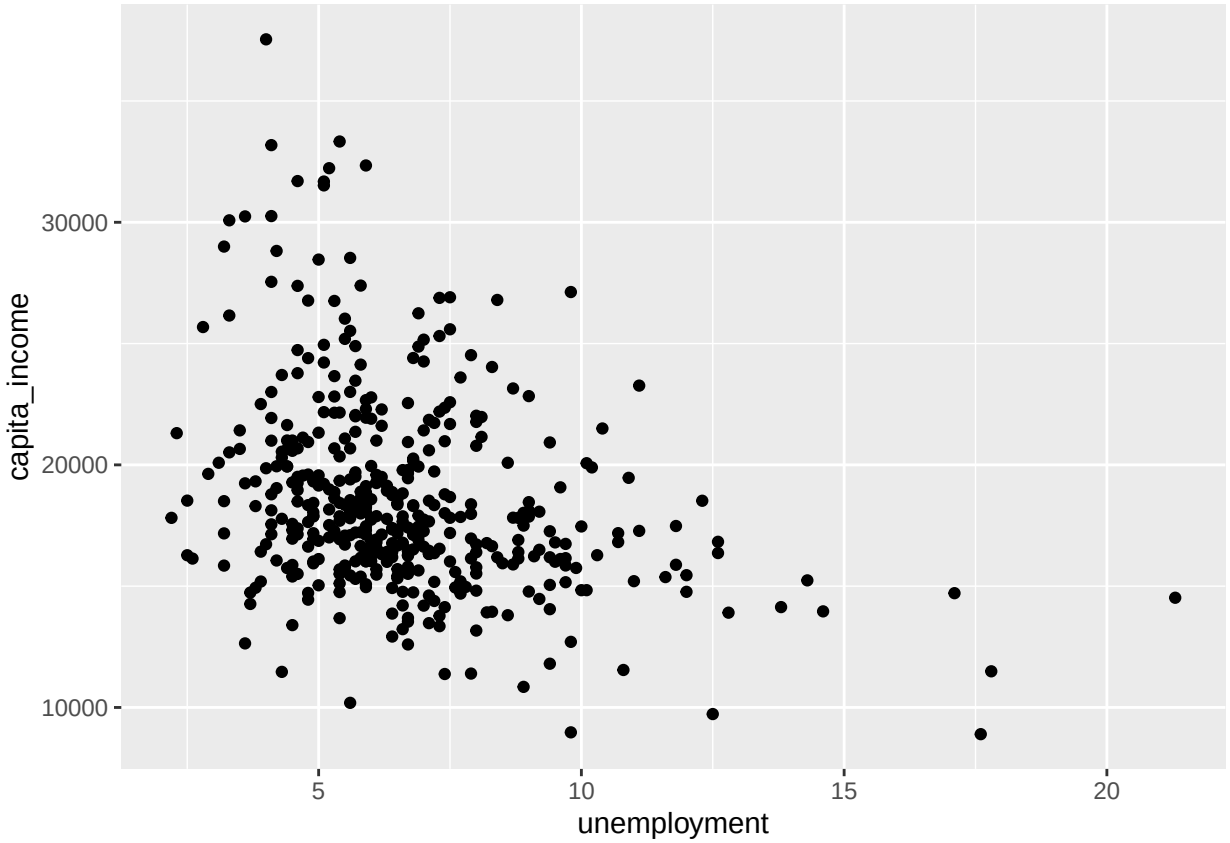
5. Use filter() to print out just the rows from Delaware.

```
filter(cdi, county == "Delaware")
```

```
##    id   county state area population percent_18_34 percent_65 physicians beds
## 1  83 Delaware    PA  184     547651          27.6       15.5       1374 1588
## 2 371 Delaware    IN  393     119659          32.9       12.7        217  494
##   crimes high_school bachelors poverty unemployment capita_income total_income
## 1  18924        81.4      24.8     5.0          5.3         23658        12956
## 2   1064        74.5      16.5    10.3          6.1         15697         1878
##   region log_capita_income
## 1     NE         10.071457
## 2     NC          9.661225
```

6. Use an appropriate plot (via {ggplot2}) to explore the relationship between capita_income and unemployment. Describe the relationship in a couple of sentences.

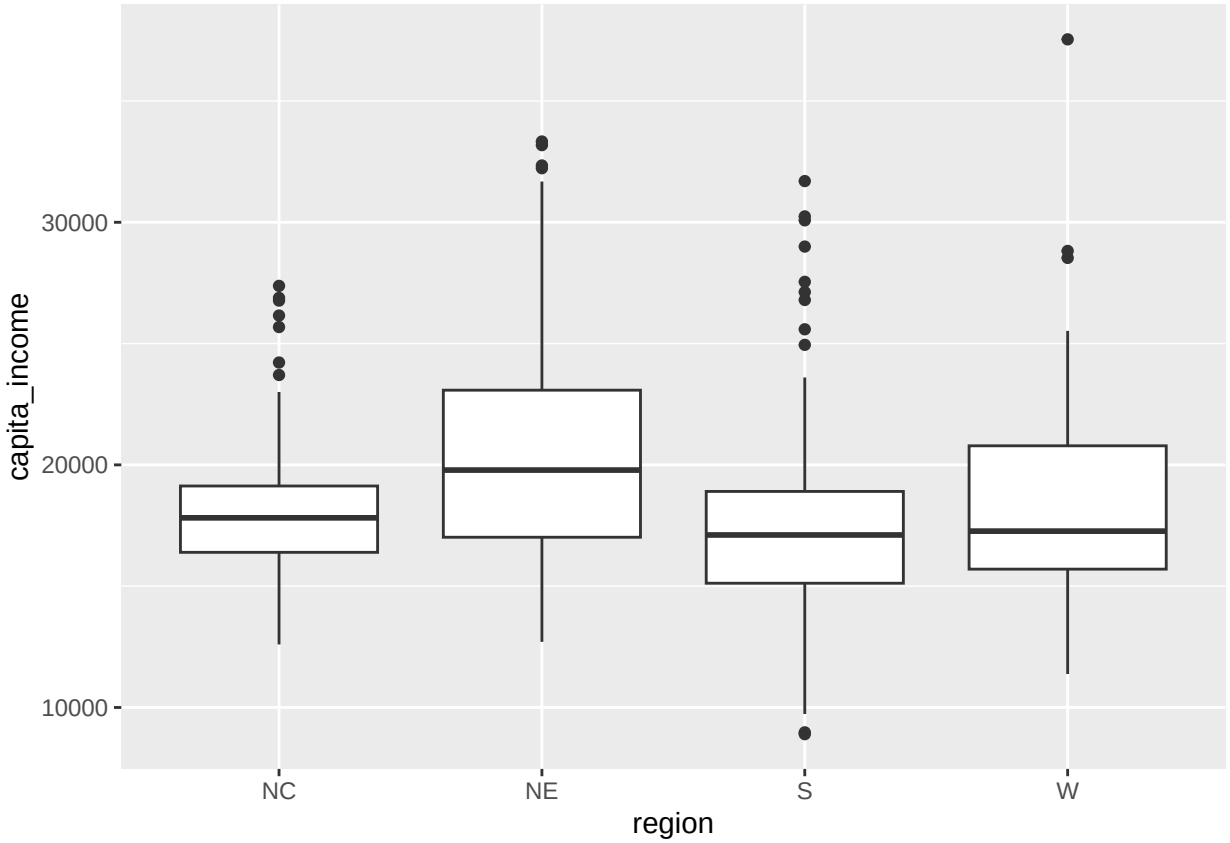
```
ggplot(cdi, aes(x = unemployment, y = capita_income)) +
  geom_point()
```



The scatterplot shows counties with less unemployment rate has much higher capital_income than counties with higher unemployment rate. Although counties with lower unemployment also have lower capita_income, this can be explained by other influencing factors other that unemployment rate.

7. Use an appropriate plot (via {ggplot2}) to explore the relationship between capita_income and region

```
ggplot(cdi, aes(x = region, y = capita_income)) +
  geom_boxplot()
```



Miscillaneous

```
x <- seq(1, 100, by = 3)
```

1. The sum of the log of the values of x

```
sum(log(x))
```

```
## [1] 122.594
```

2. The log of the sum of the values of x

```
log(sum(x))
```

```
## [1] 7.448334
```

Are these values the same? *NO, the vales are not the same, question one is looking for the addition of the log of the values of x, while question 2 is looking for the log of the sum of all x values.