Answer Sketch and Rubric for Lab 02

431 Staff and Professor Love

Last Edited 2020-09-11 17:03:08

```
library(tidyverse)
```

The Data for Lab 02

Lab 02 uses data from the `midwest` data set, which is part of the `ggplot2` package (which is part of the `tidyverse`) so by loading the `tidyverse` package, we will have direct access to the `midwest` data by typing `midwest`. The `midwest` data describe demographic information for 437 counties in the midwestern United States. You might use `?midwest` to obtain a little bit of additional information about these data, and/or use `View(midwest)` to get a look at a spreadsheet-style view of the data. We will focus on just four variables in Lab 02 taken from this data set:

- `county` = the name of the county
- `state` = the name of the state (each county is contained in a single state)
- `percollege` = the percentage of adult residents of the county who have completed a college degree
- `inmetro` = an indicator variable, which takes the value 1 if the county is contained in a metropolitan area, and 0 if it is not

Question 1

> Write a piece of R code that counts the number of observations (counties) in the data set within each state. Your result should also specify the states which are included in these data. Hint: The `count` function and the pipe `%>%` should be a big part of your code.

The `midwest` data contain information on a total of 437 counties, in 5 states. The breakdown by state of the number of counties available follows:

```
midwest %>%
    count(state)
```

```
# A tibble: 5 x 2
  state     n
  <chr> <int>
1 IL      102
2 IN       92
3 MI       83
4 OH       88
5 WI       72
```

So, for example, there are **88** counties from the state of Ohio in the data set, and the states represented also include Illinois, Indiana, Michigan and Wisconsin.

Additional Comments

1. It's important to realize that each row in the data set represents a single county.

- You can verify this by looking at the data:
  - perhaps with the command `View(midwest)`. This will bring up a spreadsheet of the data in a new R Studio window.
  - **or** perhaps by typing the name of the data set `midwest` into the Console in R Studio to get a listing.

The listing of the data set (which is a tibble) displays the first few variables for the first ten rows of the data set, like this.

```
midwest
```

```
# A tibble: 437 x 28
     PID county state  area poptotal popdensity popwhite popblack
   <int> <chr>  <chr> <dbl>    <int>      <dbl>    <int>    <int>
 1   561 ADAMS  IL    0.052    66090      1271.    63917     1702
 2   562 ALEXA~ IL    0.014    10626       759      7054     3496
 3   563 BOND   IL    0.022    14991       681.    14477      429
 4   564 BOONE  IL    0.017    30806      1812.    29344      127
 5   565 BROWN  IL    0.018     5836       324.     5264      547
 6   566 BUREAU IL    0.05     35688       714.    35157       50
 7   567 CALHO~ IL    0.017     5322       313.     5298        1
 8   568 CARRO~ IL    0.027    16805       622.    16519      111
 9   569 CASS   IL    0.024    13437       560.    13384       16
10   570 CHAMP~ IL    0.058   173025      2983.   146506    16559
# ... with 427 more rows, and 20 more variables: popamerindian <int>,
#   popasian <int>, popother <int>, percwhite <dbl>, percblack <dbl>,
#   percamerindan <dbl>, percasian <dbl>, percother <dbl>,
#   popadults <int>, perchsd <dbl>, percollege <dbl>, percprof <dbl>,
#   poppovertyknown <int>, percpovertyknown <dbl>,
#   percbelowpoverty <dbl>, percchildbelowpovert <dbl>,
#   percadultpoverty <dbl>, percelderlypoverty <dbl>, inmetro <int>,
#   category <chr>
```

Note that there are a total of 10 rows shown and an additional 427 rows hidden, so there are 437 rows in the data set, in total.

2. You can observe a help file for the `midwest` data by typing `?midwest` into the Console in R Studio, although there isn't much information in this case.

3. To obtain the number of counties (rows) in each state, I need only to obtain a count of the number of rows associated with each state. That's the command I used above, with `midwest %>% count(state)`.

4. There are some other things I used R to count in this case, for example:

   - To count the number of rows (counties) in the data as a whole, I used the command `nrow(midwest)`. The result is 437.
   - If I'd wanted to count the number of columns (variables) I'd have used `ncol(midwest)`. The result is 28.
   - I can get a count of both the rows and the columns by either listing the tibble with `midwest` or by capturing its dimensions (the size of the rectangle of data) with:

```
dim(midwest)
```

```
[1] 437  28
```

5. I did some (slightly) more sophisticated counting to understand:

- The number of (different) states in the data, using `n_distinct(midwest %>% select(state))`, which yields 5.
- The number of counties in the state of Ohio, using `nrow(midwest %>% filter(state == "OH"))`, which yields 88.

Question 2

> Use the `filter` and `select` functions in R to obtain a result which specifies the `percollege` and `inmetro` status of Cuyahoga County in the state of Ohio.

Here's the approach we used

```
midwest %>%
    filter(state == "OH") %>%
    filter(county == "CUYAHOGA") %>%
    select(state, county, percollege, inmetro)
```

```
# A tibble: 1 x 4
  state county   percollege inmetro
  <chr> <chr>         <dbl>   <int>
1 OH    CUYAHOGA       25.1       1
```

This displays the tibble, but restricted to the county of CUYAHOGA in the state of OH (Ohio) and to four of the available variables: the state, the county, the `percollege` value and the `inmetro` value.

So we conclude that 25.1% of the adult residents of Cuyahoga County have completed a college degree, and that Cuyahoga County is identified as being part of a metropolitan area.

Comments

1. Did we need the `filter(state == "OH")` line in our code? What happens if we leave this out?

```
midwest %>%
#    filter(state == "OH") %>%
    filter(county == "CUYAHOGA") %>%
    select(state, county, percollege, inmetro)
```

```
# A tibble: 1 x 4
  state county   percollege inmetro
  <chr> <chr>         <dbl>   <int>
1 OH    CUYAHOGA       25.1       1
```

Looks like there is only one county in the data set with the name CUYAHOGA, so we're OK.

2. What if we tried this approach (not specifying the state) with ADAMS county?

```
midwest %>%
  filter(county == "ADAMS") %>%
  select(state, county, percollege, inmetro)
```

```
# A tibble: 4 x 4
  state county percollege inmetro
  <chr> <chr>       <dbl>   <int>
1 IL    ADAMS       19.6        0
2 IN    ADAMS       16.1        1
3 OH    ADAMS        8.74       0
```

```
4 WI    ADAMS       12.4        0
```

3. How many unique county names are seen in the 437 counties in the `midwest` data?

```
n_distinct(midwest %>% select(county))
```

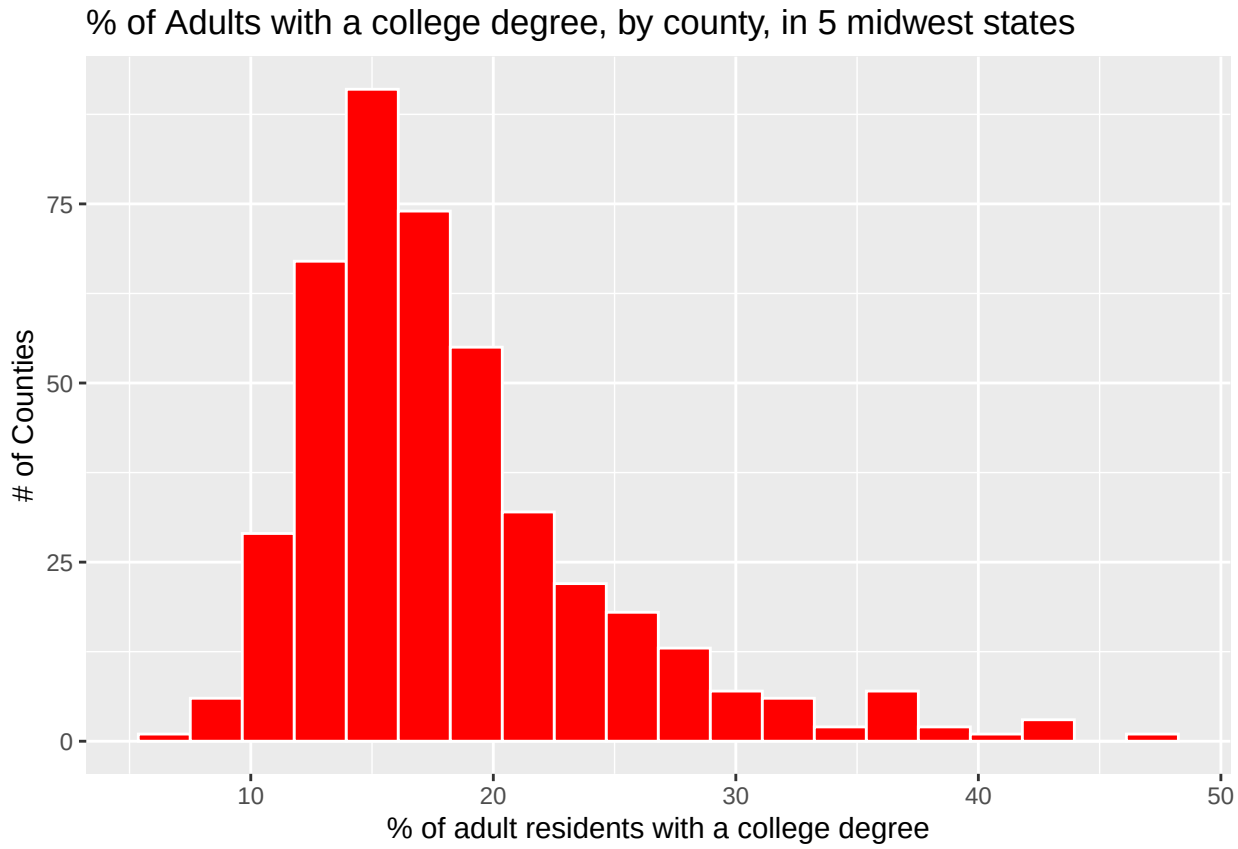
```
[1] 320
```

Question 3

> Use the tools we've been learning in the `ggplot2` package to build a histogram of the `percollege` results across all 437 counties represented in the data. Create appropriate (that is to say, meaningful) titles for each axis and for the graph as a whole (don't simply use the default choices.) We encourage you to use something you find more attractive than the default gray fill in the histogram.

Here is a simple and reasonable histogram for the `percollege` data, mostly using the template but filling in appropriate axis labels and a title, and using a red fill for the bars. 20 bins seems to work pretty well here.

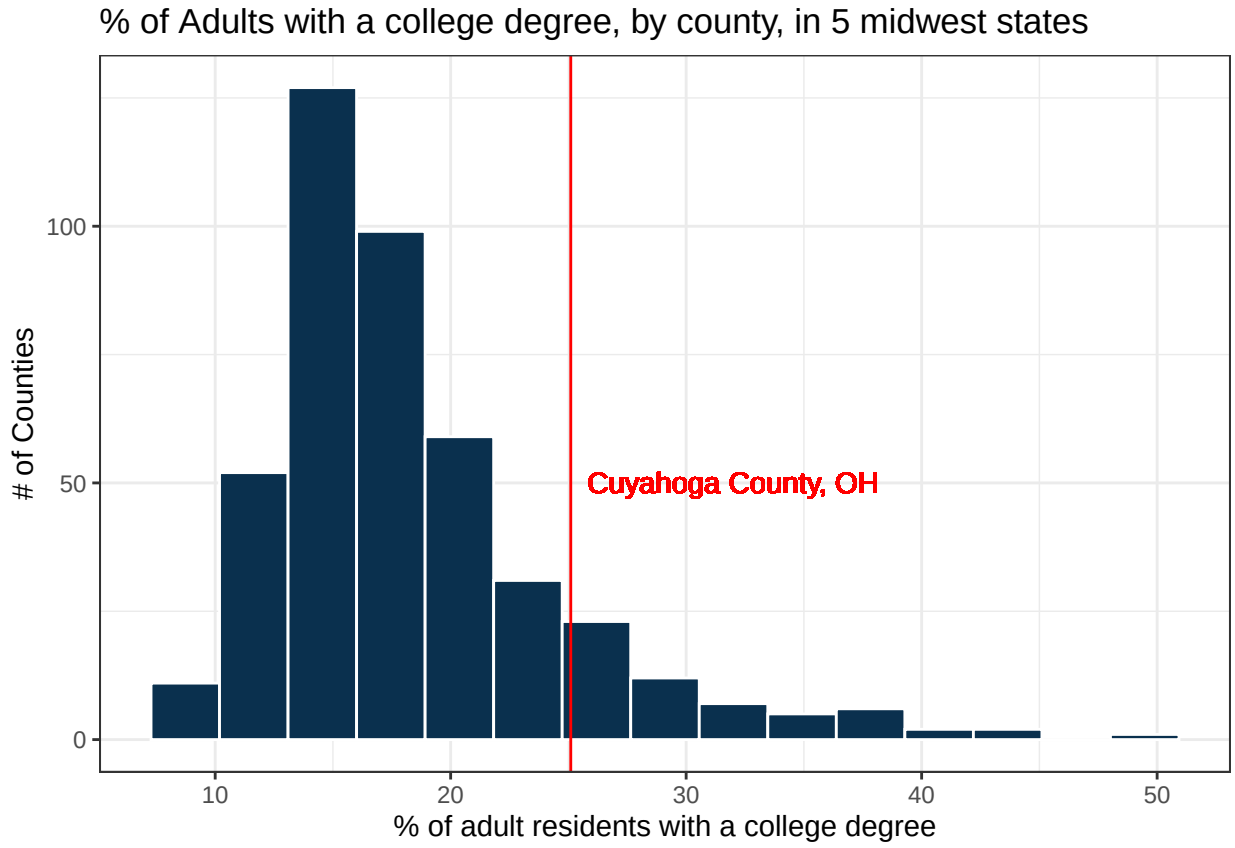
```
ggplot(midwest, aes(x = percollege)) +
    geom_histogram(bins = 20, fill = "red", col = "white") +
    labs(title = "% of Adults with a college degree, by county, in 5 midwest states",
         y = "# of Counties",
         x = "% of adult residents with a college degree")
```



Comment

The setup below uses `theme_bw()` to specify a revised theme, reduces the number of bins a bit, and creates the fill with the official blue color of CWRU[1]. Anticipating question 4, I also added a red vertical line to the plot showing the value of Cuyahoga County, Ohio, and a text annotation to indicate what the line means.

```
ggplot(midwest, aes(x = percollege)) +
    geom_histogram(bins = 15, fill = "#0a304e", col = "white") +
    geom_vline(xintercept = 25.1, col = "red") +
    geom_text(x = 32, y = 50, col = "red", label = "Cuyahoga County, OH") +
    theme_bw() +
    labs(title = "% of Adults with a college degree, by county, in 5 midwest states",
         y = "# of Counties",
         x = "% of adult residents with a college degree")
```



Question 4

> Based on your results in Questions 2 and 3, write a short description (2-3 sentences) of Cuyahoga County's position relative to the full distribution of counties in terms of `percollege`.

Cuyahoga County's `percollege` rate was 25.1%, which looks to be above the median level for the counties included in the data. There are certainly more counties with rates below Cuyahoga's than above it.

[1] CWRU's color guide is available at https://case.edu/umc/our-brand/visual-guidelines/color.

Comment

We could, if we like, be more precise, and perhaps identify the **ranking** of Cuyahoga County within the data set. We know there are 437 counties in all. How many have a *higher* value of `percollege` than Cuyahoga County?

```
midwest %>% count(percollege > 25.1)
```

```
# A tibble: 2 x 2
  `percollege > 25.1`     n
  <lgl>               <int>
1 FALSE                 386
2 TRUE                   51
```

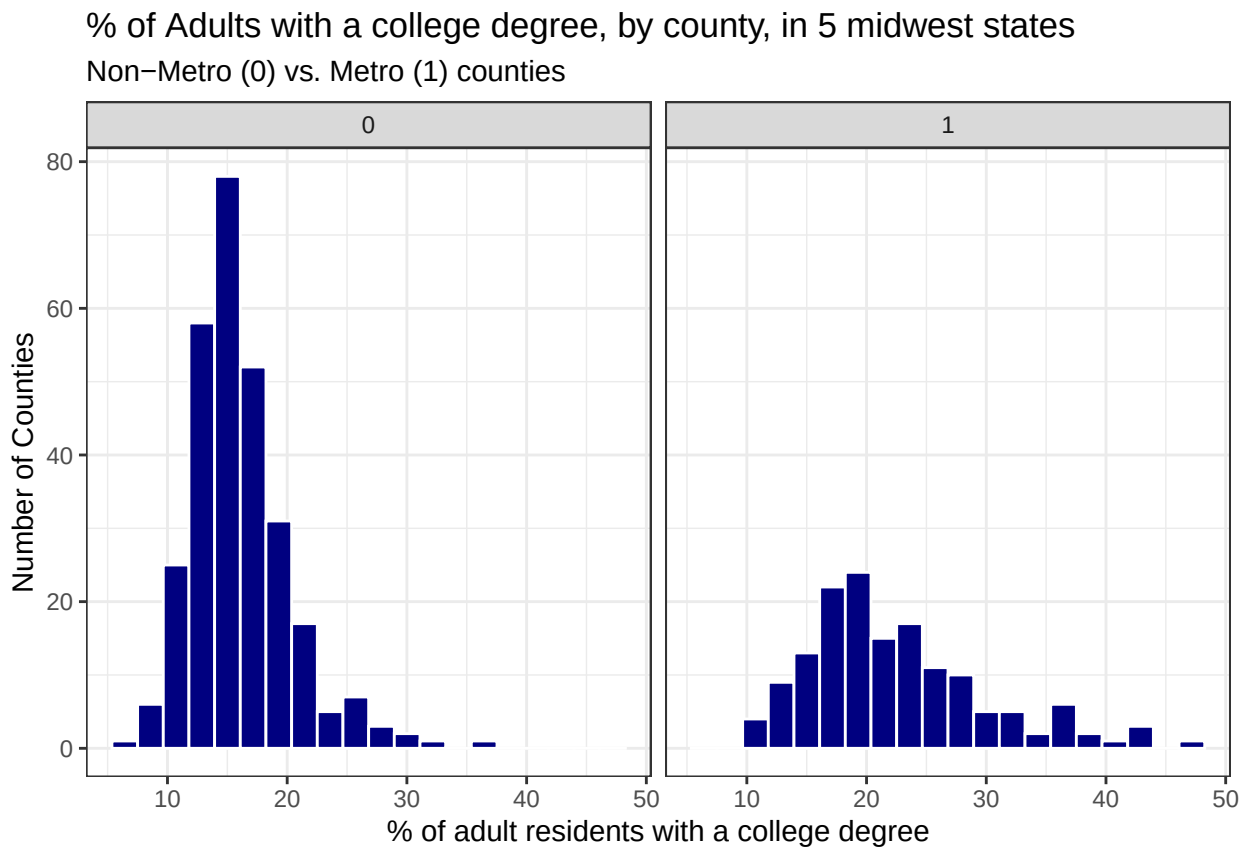
51 of the 437 counties rate above Cuyahoga on this measure, so that's 11.7% of those midwest counties. Cuyahoga County ranks 52nd among counties in midwestern states on this measure.

Question 5

> Use `ggplot2` to build a single plot (a pair of histograms after faceting would be one approach, or perhaps a comparison boxplot) which nicely compares the `percollege` distribution for counties within metropolitan areas to counties outside of metropolitan areas. Again, make an effort to build and incorporate useful titles and labels so that the resulting plot stands on its own, rather than just accepting all of the defaults that appear.

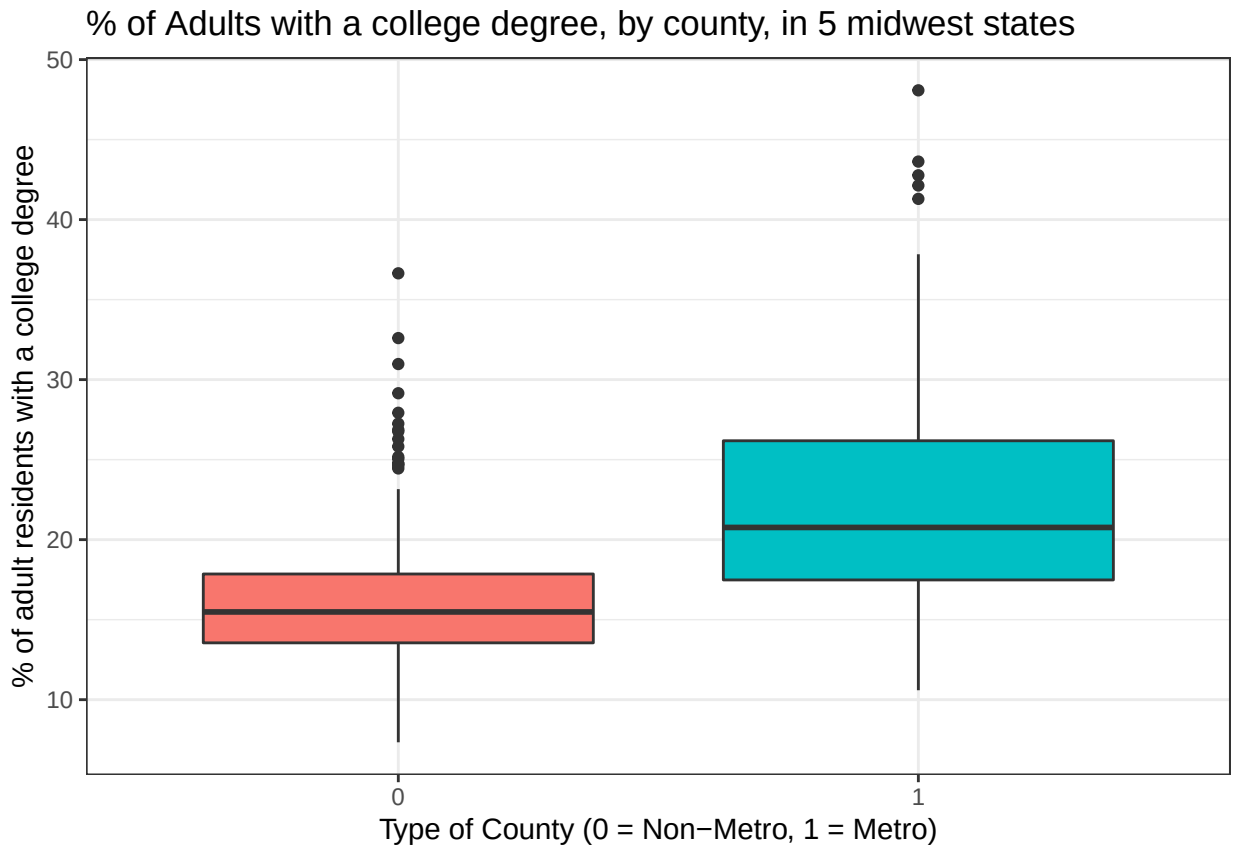
Here is a reasonable result.

```
ggplot(midwest, aes(x = percollege)) +
    geom_histogram(bins = 20, col = "white", fill = "navy") +
    facet_wrap(~ inmetro) +
    theme_bw() +
    labs(x = "% of adult residents with a college degree",
         y = "Number of Counties",
         title = "% of Adults with a college degree, by county, in 5 midwest states",
         subtitle = "Non-Metro (0) vs. Metro (1) counties")
```



Another common approach, that shows a bit less of the data, would be a comparison boxplot. Note that it's important to get R to treat the 0-1 information in `inmetro` as categorical here, and we do this by telling R to see it as a **factor**.

```
ggplot(midwest, aes(x = factor(inmetro), y = percollege,
                    fill = factor(inmetro))) +
    geom_boxplot() +
    guides(fill = FALSE) +
    theme_bw() +
    labs(y = "% of adult residents with a college degree",
         x = "Type of County (0 = Non-Metro, 1 = Metro)",
         title = "% of Adults with a college degree, by county, in 5 midwest states")
```

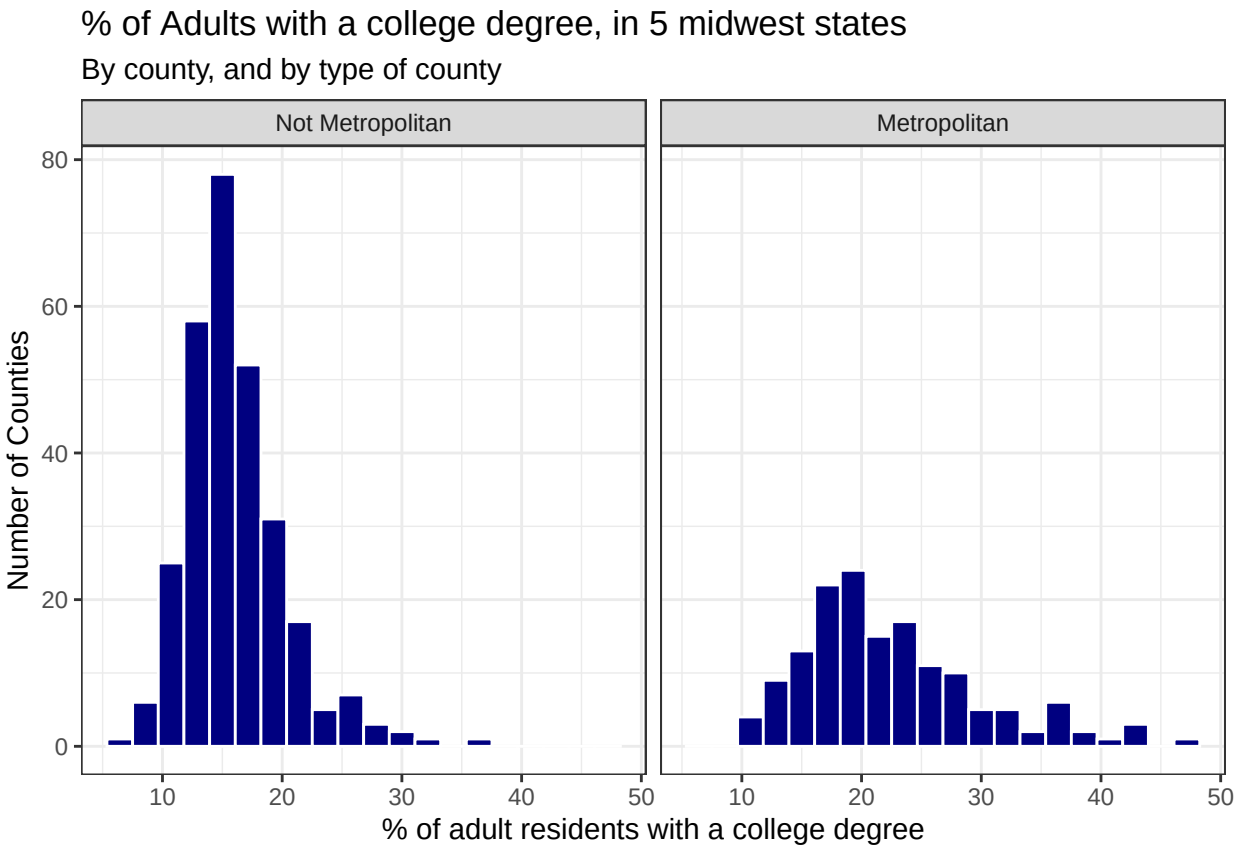


Comments

1. I think this version of the faceted set of histograms plot is slightly better. What has changed?

```
midwest2 <- midwest %>%
  mutate(inmetro_f = fct_recode(factor(midwest$inmetro),
                                "Metropolitan" = "1", "Not Metropolitan" = "0"))

ggplot(midwest2, aes(x = percollege)) +
    geom_histogram(bins = 20, col = "white", fill = "navy") +
    facet_wrap(~ inmetro_f) +
    theme_bw() +
    labs(x = "% of adult residents with a college degree",
         y = "Number of Counties",
         title = "% of Adults with a college degree, in 5 midwest states",
         subtitle = "By county, and by type of county")
```



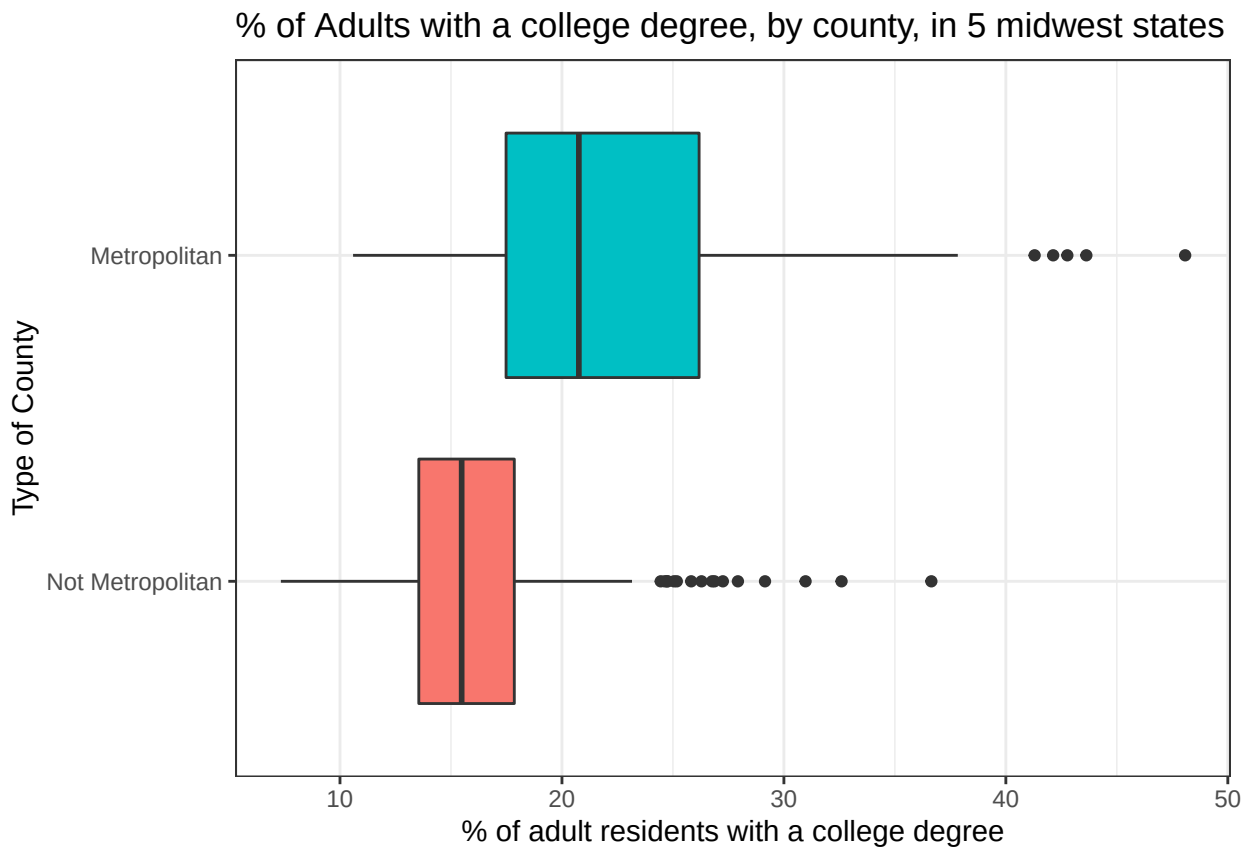
By pre-specifying the `inmetro` variable as a factor, and giving it meaningful names for its two levels (Metropolitan and Non-Metropolitan, instead of 1 and 0) we improve the plot.

2. Could we do the same sort of thing to improve the comparison boxplot? And what if we wanted a horizontal boxplot instead of a vertical one?

Sure. Note the use of `coord_flip()` to flip the Y and X coordinates and axes.

```
midwest2 <- midwest %>%
  mutate(inmetro_f = fct_recode(factor(midwest$inmetro),
                                "Metropolitan" = "1", "Not Metropolitan" = "0"))

ggplot(midwest2, aes(x = inmetro_f, y = percollege,
                     fill = inmetro_f)) +
    geom_boxplot() +
    guides(fill = FALSE) +
    theme_bw() +
    coord_flip() +
    labs(y = "% of adult residents with a college degree",
         x = "Type of County",
         title = "% of Adults with a college degree, by county, in 5 midwest states")
```



Question 6

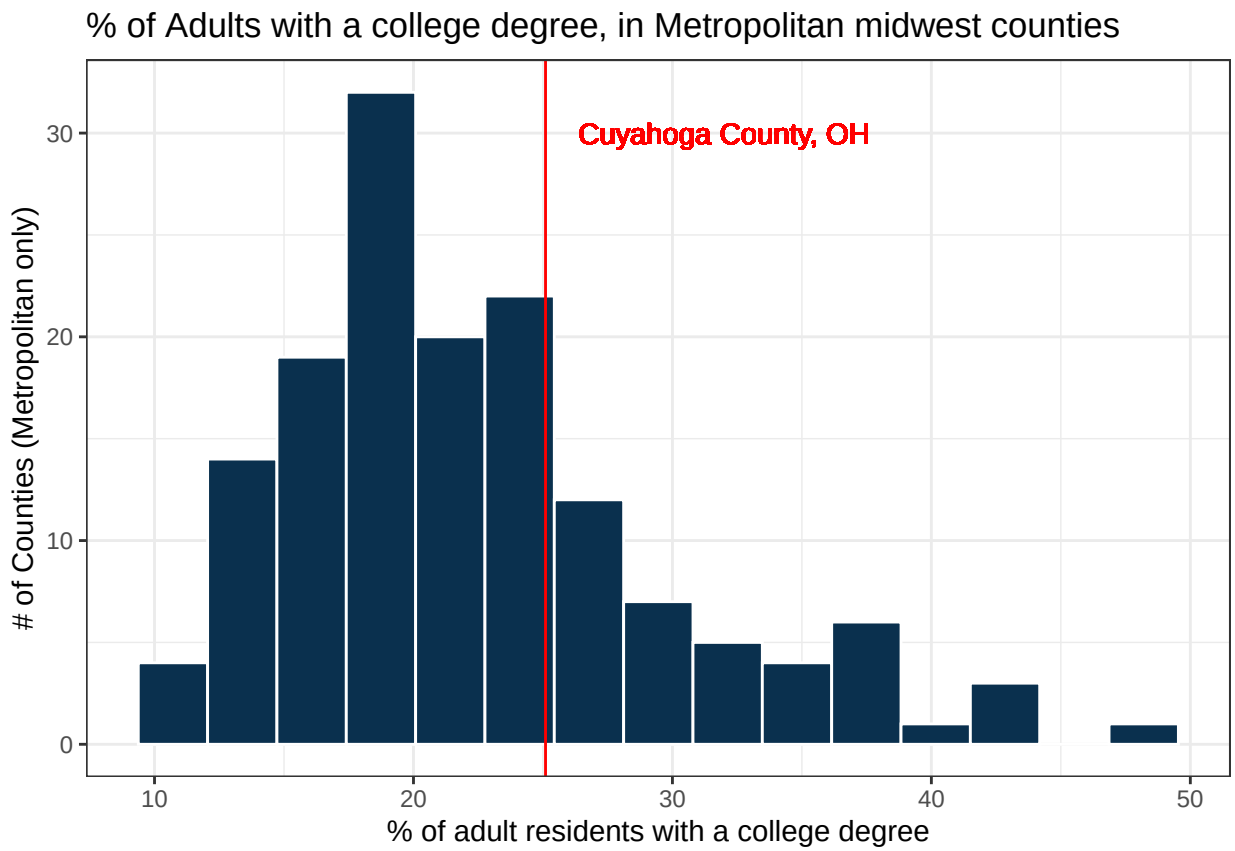
> Write a short description of where Cuyahoga County falls within the plot you built in Question 5. the position of Cuyahoga County in terms of `percollege` relative to the other counties within its `inmetro` category. Two sentences should be sufficient here.

Within the metropolitan counties, Cuyahoga's position is still above the median county, but a bit closer to the center of the group on `percollege` than it was when we looked at the whole data set.

Comments

The plot below is restricted to those counties in the "Metropolitan" group.

```
ggplot(filter(midwest, inmetro == 1), aes(x = percollege)) +
    geom_histogram(bins = 15, fill = "#0a304e", col = "white") +
    geom_vline(xintercept = 25.1, col = "red") +
    geom_text(x = 32, y = 30, col = "red", label = "Cuyahoga County, OH") +
    theme_bw() +
    labs(title = "% of Adults with a college degree, in Metropolitan midwest counties",
         y = "# of Counties (Metropolitan only)",
         x = "% of adult residents with a college degree")
```



It turns out that Cuyahoga County ranks 41st among the 150 metropolitan counties.

```
# number of Metropolitan counties
nrow(midwest %>% filter(inmetro == 1))
```

```
[1] 150
```

```
# number with percollege higher than Cuyahoga
midwest %>% filter(inmetro == 1) %>%
    count(percollege > 25.1)
```

```
# A tibble: 2 x 2
  `percollege > 25.1`     n
  <lgl>               <int>
1 FALSE                 110
2 TRUE                   40
```

40 of the 150 metropolitan counties in these five states rate above Cuyahoga County on this measure, so that's 26.7%. Compare this to Cuyahoga's ranking behind only 11.7% of all counties (metropolitan and non-Metropolitan) in those same five states.

Question 7

> By now, we'd like you to have read the Introduction and Chapters 1-3 of David Spiegelhalter's *The Art of Statistics*. What is the most important question you have after reading this material? Please cite the book appropriately when describing what Dr. Spiegelhalter has brought to your attention. Remember that a question ends with a question mark.

We don't write sketches for essay questions.

Grading

Lab 02 will be graded on a 0-100 scale.

For questions 1-6, students will receive up to 15 points, as follows:

- 10 points for a reasonable effort to build a good response that is more than slightly incomplete or incorrect.
- 12 points for a good effort that is either correct but not well written or the reverse, but has at most modest problems.
- 15 points for a complete, correct, well-written response.

Well-written responses use complete English sentences and show all required R code to achieve the desired result.

For question 7, a well-phrased question that is motivated by something from the required reading in Spiegelhalter will be awarded up to 10 points. Most students should score 8-10 points.

Later Labs will include more substantial essay questions and require deeper grading.