Class after Midterm_Exam

Fentaw Abitew

2022-10-14

AMERICAN UNIVERSITY

STAT 412/612

#Midterm 1

Be sure to provide all r code that produces the requested plots/output tibbles or data frames. Use ggplot coding to produce graphs and plots as demonstrated in class. Submit all results in an Rmardown file and a word file or an Rmarkdown file and a pdf.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
```

```
midwest
```

```
## # A tibble: 437 x 28
##      PID county    state  area poptotal popden~1 popwh~2 popbl~3 popam~4 popas~5
##    <int> <chr>     <chr> <dbl>    <int>    <dbl>   <int>   <int>   <int>   <int>
##  1   561 ADAMS     IL    0.052    66090    1271.   63917    1702      98     249
##  2   562 ALEXANDER IL    0.014    10626     759     7054    3496      19      48
##  3   563 BOND      IL    0.022    14991     681.   14477     429      35      16
##  4   564 BOONE     IL    0.017    30806    1812.   29344     127      46     150
##  5   565 BROWN     IL    0.018     5836     324.    5264     547      14       5
##  6   566 BUREAU    IL    0.05     35688     714.   35157      50      65     195
##  7   567 CALHOUN   IL    0.017     5322     313.    5298       1       8      15
##  8   568 CARROLL   IL    0.027    16805     622.   16519     111      30      61
##  9   569 CASS      IL    0.024    13437     560.   13384      16       8      23
## 10   570 CHAMPAIGN IL    0.058   173025    2983.  146506   16559     331    8033
## # ... with 427 more rows, 18 more variables: 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>, and abbreviated
## #   variable names 1: popdensity, 2: popwhite, 3: popblack, ...
```

```
view(midwest)

#V  =  pir2h
# SA  =  2(pirh  +  pir2)
```

For some of the problems, you have to determine what variables to use.Problems 1 - 10 are for undergradu

1. Using the midwest data frame produce a data table that shows output for the Ohio (OH) only. Produce correct output by using two methods. First use the piping method and then use the assignment method.

```
midwest %>% filter(midwest$state== "OH") ->oh_df # piping first then assignment

oh_df
```

```
## # A tibble: 88 x 28
##      PID county    state  area poptotal popden~1 popwh~2 popbl~3 popam~4 popas~5
##    <int> <chr>     <chr> <dbl>    <int>    <dbl>   <int>   <int>   <int>   <int>
##  1  2009 ADAMS     OH    0.035    25371     725.   25212      47      67      30
##  2  2010 ALLEN     OH    0.024   109755    4573.   96177   12313     202     572
##  3  2011 ASHLAND   OH    0.025    47507    1900.   46686     460      49     271
##  4  2012 ASHTABULA OH    0.041    99821    2435.   95465    3138     196     350
##  5  2013 ATHENS    OH    0.03     59549    1985.   56163    1678     167    1374
##  6  2014 AUGLAIZE  OH    0.024    44585    1858.   44225      66      50     177
##  7  2015 BELMONT   OH    0.031    71074    2293.   69520    1308      81     129
##  8  2016 BROWN     OH    0.028    34966    1249.   34487     406      28      30
##  9  2017 BUTLER    OH    0.028   291479   10410.  274892   13134     379    2659
## 10  2018 CARROLL   OH    0.024    26521    1105.   26254     135      65      29
## # ... with 78 more rows, 18 more variables: 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>, and abbreviated
## #   variable names 1: popdensity, 2: popwhite, 3: popblack, ...
```

2. Using the midwest data frame, produce a data table that shows white population that is greater than 50,000 but less than 90,000 for the state of Indiana (IN)

```
indiana_whitepop_less90<-filter(midwest, state=="IN" & popwhite> 50000 & popwhite<90000)
indiana_whitepop_less90
```

```
## # A tibble: 10 x 28
##      PID county      state  area popto~1 popde~2 popwh~3 popbl~4 popam~5 popas~6
##    <int> <chr>       <chr> <dbl>   <int>   <dbl>   <int>   <int>   <int>   <int>
##  1   665 BARTHOLOMEW IN    0.022   63657   2894.   61774    1005      97     610
##  2   672 CLARK       IN    0.022   87777   3990.   82289    4703     192     356
##  3   684 FLOYD       IN    0.009   64404   7156    61415    2642      92     175
##  4   689 GRANT       IN    0.024   74169   3090.   67817    5047     298     373
##  5   694 HENDRICKS   IN    0.024   75717   3155.   74519     685     157     275
##  6   696 HOWARD      IN    0.016   80827   5052.   75420    4398     226     457
##  7   703 JOHNSON     IN    0.018   88109   4895.   86455     845     139     534
##  8   705 KOSCIUSKO   IN    0.032   65294   2040.   64058     309     118     322
##  9   717 MORGAN      IN    0.024   55920   2330    55635       9     137      91
## 10   751 WAYNE       IN    0.024   71951   2998.   67532    3795     153     296
## # ... with 18 more variables: 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>, and abbreviated variable names 1: poptotal, 2: popdensity,
## #   3: popwhite, 4: popblack, 5: popamerindian, 6: popasian
```

```
#We can filter the popwhite column only if needed; I also added the county  for fullness
indiana_whitepop_less90%>%
  select(popwhite, state, county)
```

```
## # A tibble: 10 x 3
##    popwhite state county
##       <int> <chr> <chr>
##  1    61774 IN    BARTHOLOMEW
##  2    82289 IN    CLARK
##  3    61415 IN    FLOYD
##  4    67817 IN    GRANT
##  5    74519 IN    HENDRICKS
##  6    75420 IN    HOWARD
##  7    86455 IN    JOHNSON
##  8    64058 IN    KOSCIUSKO
##  9    55635 IN    MORGAN
## 10    67532 IN    WAYNE
```

3. Using the midwest data , produce a data frame (20 observations) that shows only the variables state, county, poptotal , popamerindian, percamerindian for the state of Indiana. Also your data frame should show popamerindian in descending order.Which county in Indiana has the highest number of Native Americans?

```
indian_data <- select(midwest,state,county,poptotal,percamerindan, popamerindian)%>%
  filter(state=="IN")

arrange(indian_data, desc(popamerindian)) # Marion has the highest popamerindian(Native Americans)
```

```
## # A tibble: 92 x 5
##    state county     poptotal percamerindan popamerindian
##    <chr> <chr>         <int>         <dbl>         <int>
##  1 IN    MARION       797159         0.213          1698
##  2 IN    ALLEN        300836         0.297           892
##  3 IN    LAKE         475594         0.182           865
##  4 IN    ST JOSEPH    247052         0.342           846
##  5 IN    MIAMI         36897         1.55            571
##  6 IN    ELKHART      156198         0.290           453
##  7 IN    TIPPECANOE   130598         0.245           320
##  8 IN    MADISON      130669         0.229           299
##  9 IN    GRANT         74169         0.402           298
## 10 IN    VIGO         106107         0.280           297
## # ... with 82 more rows
```

4. Using the midwest data and dplyr functions, create a data frame for only the state of Michigan (MI) showing those counties that have a known poverty population that is greater than 10,000 and a percentage of professionals that is greater than 10 percent. Only select variables that you need for the data frame, Your output should only have four variables and six (rows) / observations.

```
filter(midwest, state== "MI", poppovertyknown > 10000 & percprof >10) %>%
  select(state, county, poppovertyknown, percprof) ->michgan_poverty
```

```
michgan_poverty
```

```
## # A tibble: 6 x 4
##   state county    poppovertyknown percprof
##   <chr> <chr>               <int>    <dbl>
## 1 MI    INGHAM             261491     12.9
## 2 MI    ISABELLA            48498     10.0
## 3 MI    KALAMAZOO          212670     10.9
## 4 MI    MIDLAND             74135     11.2
## 5 MI    OAKLAND           1070844     11.2
## 6 MI    WASHTENAW          261261     20.8
```

5. Using the midwest data and dplyr commands and functions, write r code that will show the mean of the poverty population for the counties of each state.

```
midwest %>%
  group_by(county, state) %>%
  # Don't be alarmed here, I added state just to see the county and state together side by side and
  #since the question ask for the counties, I am NOT trying to group by using both variables. If I want
  # and by default r summarise has grouped output by the first variable only.
  summarize(mean_poverty= mean(poppovertyknown, na.rm=TRUE)) ->poverty_groupedby_county
```

```
## `summarise()` has grouped output by 'county'. You can override using the
## `.groups` argument.
```

```
poverty_groupedby_county
```

```
## # A tibble: 437 x 3
## # Groups:   county [320]
##    county    state mean_poverty
##    <chr>     <chr>        <dbl>
##  1 ADAMS     IL           63628
##  2 ADAMS     IN           30490
##  3 ADAMS     OH           25028
##  4 ADAMS     WI           14534
##  5 ALCONA    MI           10040
##  6 ALEXANDER IL           10529
##  7 ALGER     MI            8452
##  8 ALLEGAN   MI           88882
##  9 ALLEN     IN          296184
## 10 ALLEN     OH          104543
## # ... with 427 more rows
```

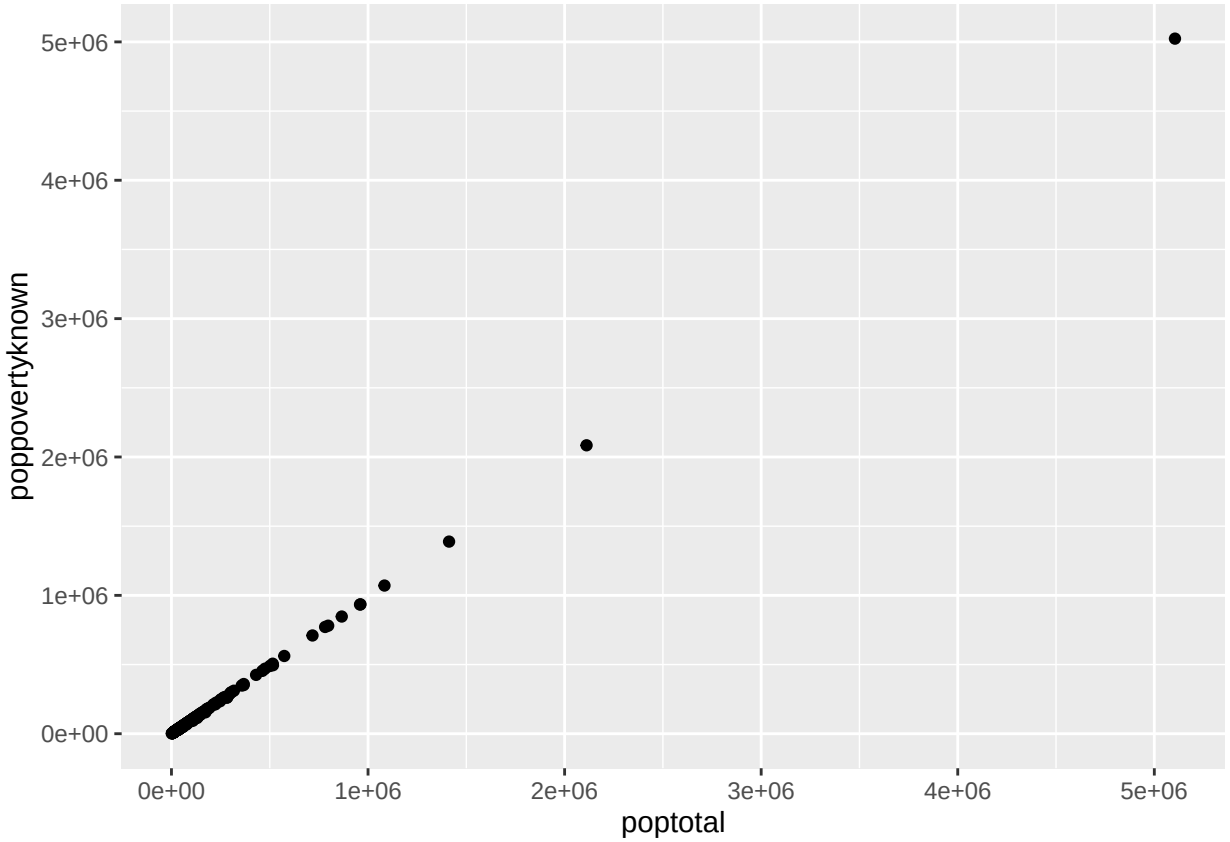
```
# Not asked to do, but kind of be curious and order to see the poorest county
arrange(poverty_groupedby_county, desc(mean_poverty))
```

```
## # A tibble: 437 x 3
## # Groups:   county [320]
##    county    state mean_poverty
##    <chr>     <chr>        <dbl>
##  1 COOK      IL         5023523
##  2 WAYNE     MI         2084529
##  3 CUYAHOGA  OH         1388547
##  4 OAKLAND   MI         1070844
##  5 FRANKLIN  OH          935142
##  6 MILWAUKEE WI          933532
##  7 HAMILTON  OH          846909
```

```
##  8 MARION    IN          780649
##  9 DU PAGE   IL          771641
## 10 MACOMB    MI          710217
## # ... with 427 more rows
```

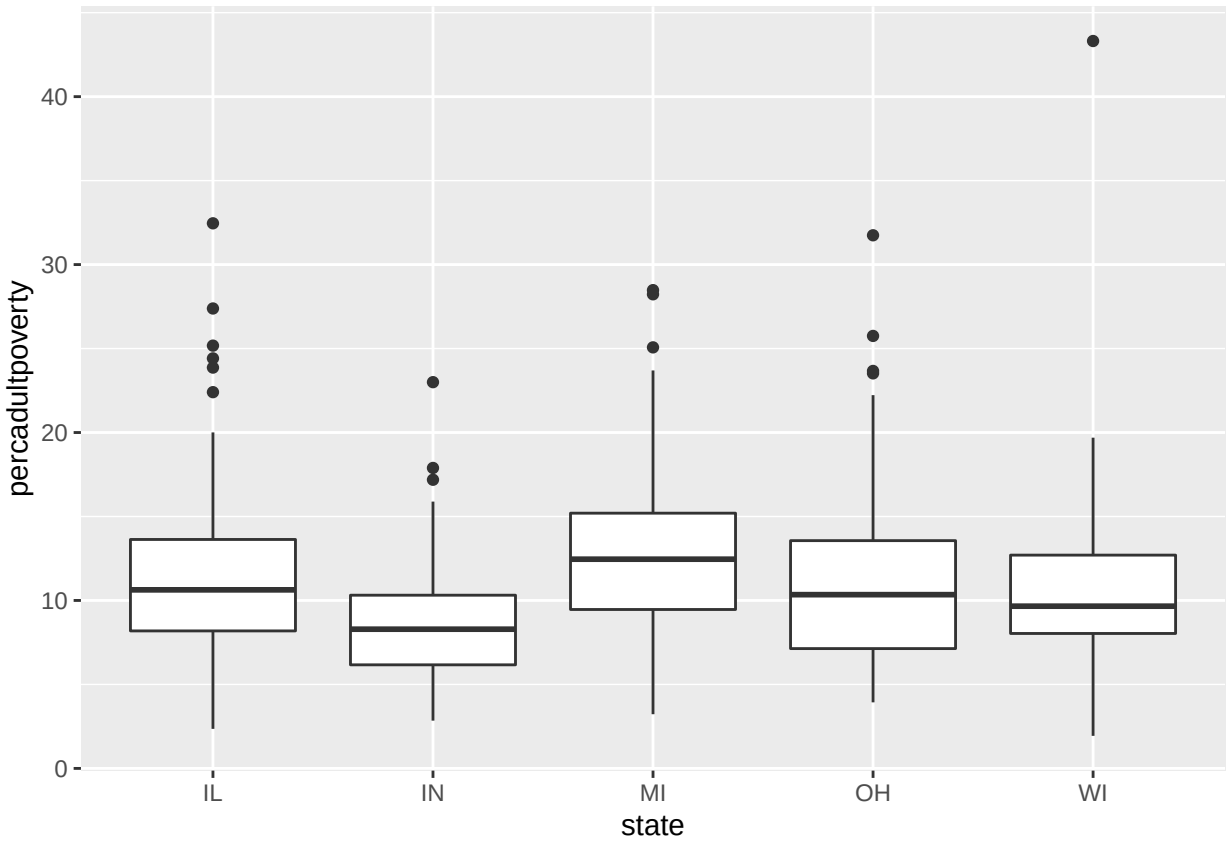
6. Using the midwest data, produce a scatter plot showing a relationship between the variables poppovertyknown and poptotal (Let poptotal = x and poppovertyknown = y).

```
ggplot(data=midwest, aes(x=poptotal, y=poppovertyknown)) +
geom_point()
```



7. Using the midwest data, write r code that will produce the following side by side boxplots.

```
ggplot(midwest, aes(x= state, y=percadultpoverty))+
geom_boxplot()
```



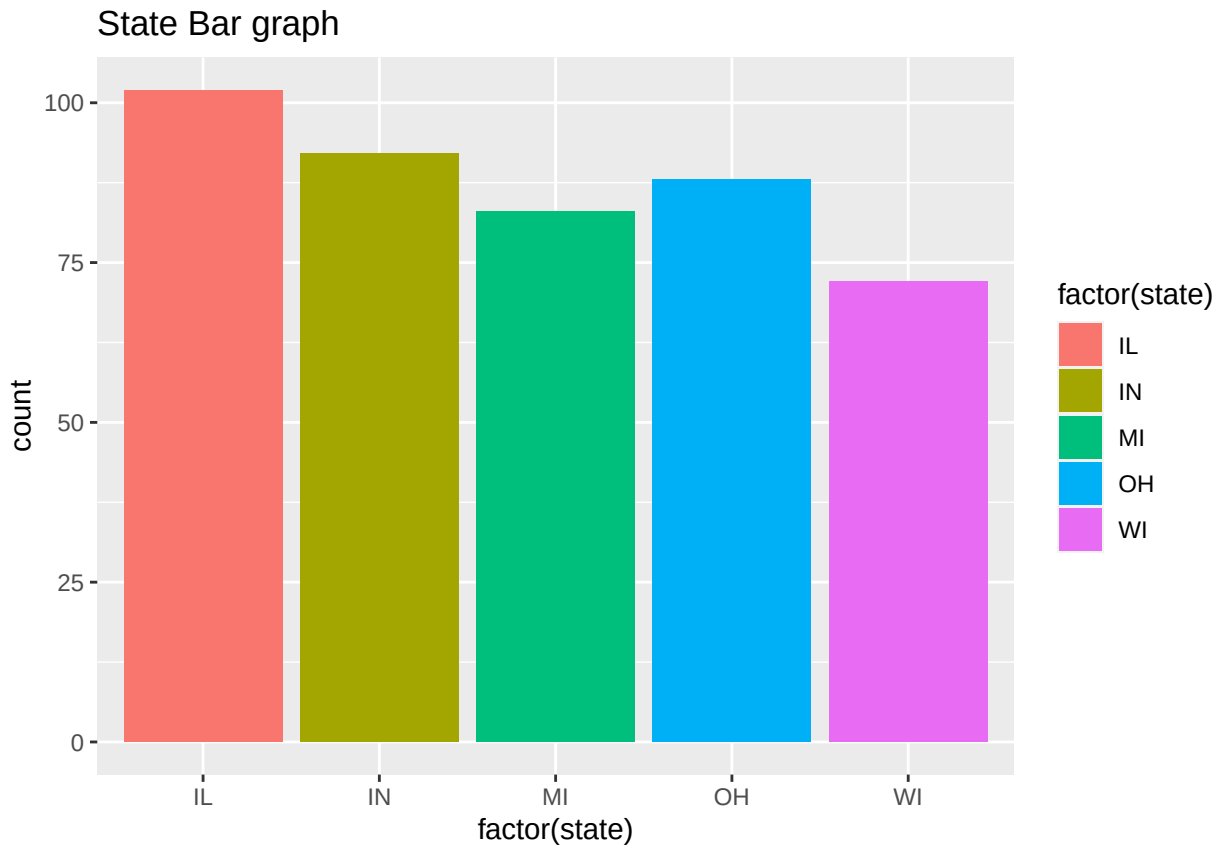
8. Using the midwest data, write r code that will produce a facet plot that shows scatter plots (red data points) with respect to the levels for the variable state. Also add code that will generate regression lines through your scatter plots that feature x = percollege and y = percprof. Title your facet plot "College/Professional Work Scatter Plots"

```
# Creating the scatter plot
Red_scatterplot <-ggplot(data=midwest, aes(x=percollege, y=percprof))+

  #Not asked but I added x and y lab
   xlab("Percentage of college") +
   ylab("Percentage of Professional") +
  ggtitle("College/Professional Work Scatter Plots")+
  geom_point(color='red') +
  geom_smooth(se = FALSE, method = lm)

# Facete wrap with respect to the levels for the variable state
facetplot_scatterplot <-Red_scatterplot + facet_wrap(~ state)
facetplot_scatterplot
```

```
## `geom_smooth()` using formula 'y ~ x'
```

9. Using the midwest data frame, create a bar graph that shows the different counts for each state in the data set. Your bars should have different colors. Which state has the highest count?

```
ggplot(data=midwest, aes(factor(state), fill = factor(state))) +
ggtitle("State Bar graph")+ # Not asked to do
geom_bar()
```




```
# State of Illinois(IL) has the highest count.
```

10. The formula used to find the volume of a cylinder is V = pi times r squared and the formula to find the Surface Area of a cylinder is A = 2(pi times r times h + pi times r squared) Using the formal notation and process for writing a function, as demonstrated in class, to write a function that will calculate the Volume and the Surface Area of a given cylinder. Test your function by calculating answers for r = 5 and h = 10.

```
cylinder_area = function(r,h)
{ area=(2*(pi*r*h + pi * r^2))
return(area)
}

cylinder_volume= function(r)
{
volume=pi*r^2
return(volume)
}
```

```
cylinder_area(5,10)
```

```
## [1] 471.2389
```

```
cylinder_volume(5)
```

```
## [1] 78.53982
```

```
Questions 11 and 12 are for graduate students (612) only
```

11. A partial data frame to be generated from the midwest data frame is given below. Write r code

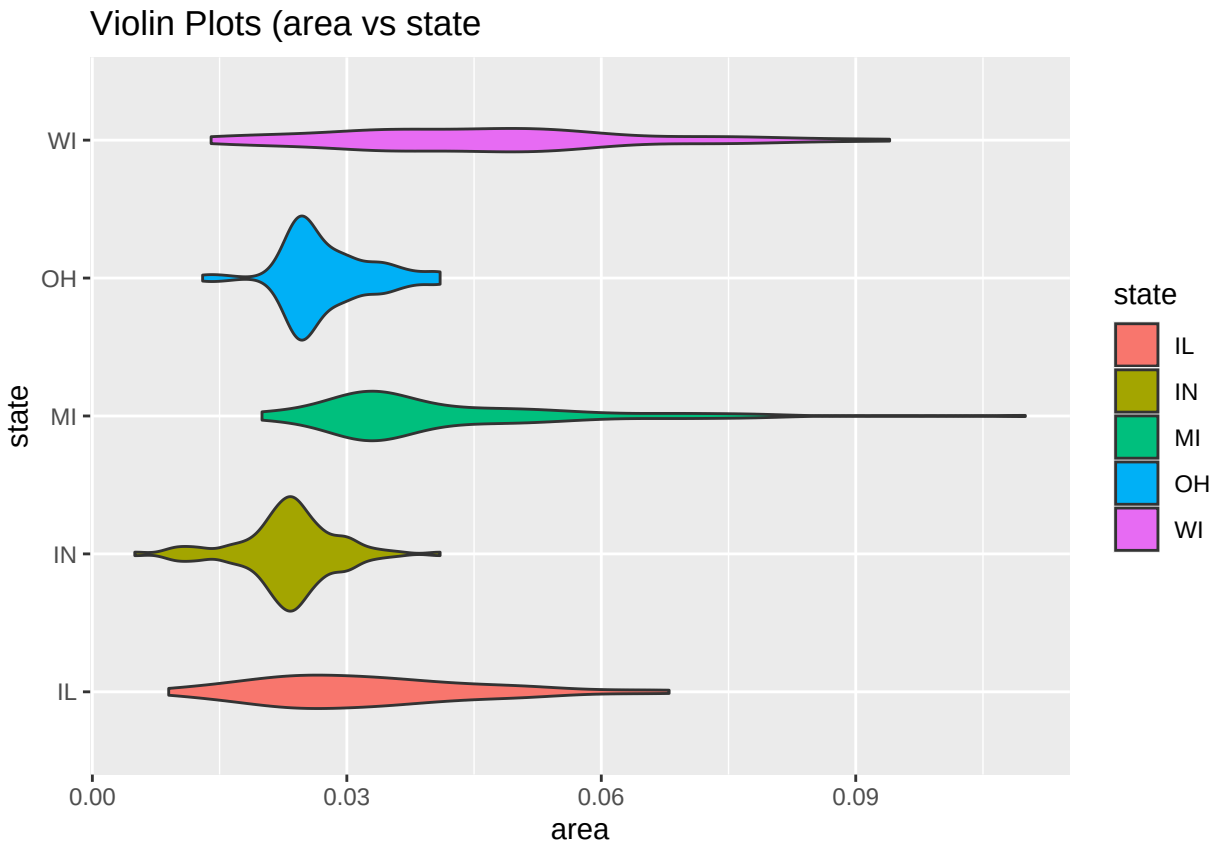
and apply dplyr functions that will produce an additional 20 rows to the 5 rows shown. A tibble: 72 x 6 state county poptotal popadults Ratio Percent 1 Wisconsin ADAMS 15682 11378 0.726 72.6 2 Wisconsin ASHLAND 16307 10262 0.629 62.9 3 Wisconsin BARRON 40750 26198 0.643 64.3 4 Wisconsin BAYFIELD 14008 9418 0.672 67.2 5 Wisconsin BROWN 194594 120575 0.620 62.0

```
midwest %>%
  select(state, county, poptotal, popadults) %>%
  filter(state=="WI") %>%
  mutate(ratio = popadults/poptotal,
         percent = ratio *100) -> wi_data
head(wi_data,25)
```

```
## # A tibble: 25 x 6
##    state county   poptotal popadults ratio percent
##    <chr> <chr>       <int>     <int> <dbl>   <dbl>
##  1 WI    ADAMS       15682     11378 0.726    72.6
##  2 WI    ASHLAND     16307     10262 0.629    62.9
##  3 WI    BARRON      40750     26198 0.643    64.3
##  4 WI    BAYFIELD    14008      9418 0.672    67.2
##  5 WI    BROWN      194594    120575 0.620    62.0
##  6 WI    BUFFALO     13584      8918 0.657    65.7
##  7 WI    BURNETT     13084      9045 0.691    69.1
##  8 WI    CALUMET     34291     20940 0.611    61.1
##  9 WI    CHIPPEWA    52360     33195 0.634    63.4
## 10 WI    CLARK       31647     19702 0.623    62.3
## # ... with 15 more rows
```

12. Use ggplot coding to produce the side by side plots shown below. (Hint: use the categorical variable state and the quantitative variable area of the midwest data table.

```
ggplot(midwest, aes(x=area, y=state, fill = state))+
  geom_violin() +
  ggtitle("Violin Plots (area vs state")
```


Violin Plots (area vs state
WI
OH
MI
IN
IL
state
0.00
0.03
0.06
0.09
area
state
IL
IN
MI
OH
WI