Test 1

Daniel Bernal

October 6, 2021

```
# Midterm Exam - Daniel Bernal - Fall 2021
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.0.5

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'purrr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## Warning: package 'stringr' was built under R version 4.0.5

## Warning: package 'forcats' was built under R version 4.0.5

## -- 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 popdensity popwhite popblack popamerindian
##    <int> <chr>      <chr> <dbl>    <int>      <dbl>    <int>    <int>         <int>
##  1   561 ADAMS      IL    0.052    66090      1271.    63917     1702            98
##  2   562 ALEXANDER  IL    0.014    10626       759      7054     3496            19
##  3   563 BOND       IL    0.022    14991       681.    14477      429            35
##  4   564 BOONE      IL    0.017    30806      1812.    29344      127            46
##  5   565 BROWN      IL    0.018     5836       324.     5264      547            14
##  6   566 BUREAU     IL    0.05     35688       714.    35157       50            65
##  7   567 CALHOUN    IL    0.017     5322       313.     5298        1             8
##  8   568 CARROLL    IL    0.027    16805       622.    16519      111            30
##  9   569 CASS       IL    0.024    13437       560.    13384       16             8
## 10   570 CHAMPAIGN  IL    0.058   173025      2983.   146506    16559           331
## # ... with 427 more rows, and 19 more variables: 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>

# 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.

# Pipping Method
midwest%>%
  filter(state == "OH")
```

```
## # A tibble: 88 x 28
##      PID county    state  area poptotal popdensity popwhite popblack popamerindian
##    <int> <chr>     <chr> <dbl>    <int>      <dbl>    <int>    <int>         <int>
##  1  2009 ADAMS     OH    0.035    25371       725.    25212       47            67
##  2  2010 ALLEN     OH    0.024   109755      4573.    96177    12313           202
##  3  2011 ASHLAND   OH    0.025    47507      1900.    46686      460            49
##  4  2012 ASHTABULA OH    0.041    99821      2435.    95465     3138           196
##  5  2013 ATHENS    OH    0.03     59549      1985.    56163     1678           167
##  6  2014 AUGLAIZE  OH    0.024    44585      1858.    44225       66            50
##  7  2015 BELMONT   OH    0.031    71074      2293.    69520     1308            81
##  8  2016 BROWN     OH    0.028    34966      1249.    34487      406            28
##  9  2017 BUTLER    OH    0.028   291479     10410.   274892    13134           379
## 10  2018 CARROLL   OH    0.024    26521      1105.    26254      135            65
## # ... with 78 more rows, and 19 more variables: 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>
```

```
# Assignment Method
oh = filter(midwest, state == "OH")
oh
```

```
## # A tibble: 88 x 28
##      PID county    state  area poptotal popdensity popwhite popblack popamerindian
##    <int> <chr>     <chr> <dbl>    <int>      <dbl>    <int>    <int>         <int>
##  1  2009 ADAMS     OH    0.035    25371       725.    25212       47            67
##  2  2010 ALLEN     OH    0.024   109755      4573.    96177    12313           202
##  3  2011 ASHLAND   OH    0.025    47507      1900.    46686      460            49
##  4  2012 ASHTABULA OH    0.041    99821      2435.    95465     3138           196
##  5  2013 ATHENS    OH    0.03     59549      1985.    56163     1678           167
##  6  2014 AUGLAIZE  OH    0.024    44585      1858.    44225       66            50
##  7  2015 BELMONT   OH    0.031    71074      2293.    69520     1308            81
##  8  2016 BROWN     OH    0.028    34966      1249.    34487      406           28
##  9  2017 BUTLER    OH    0.028   291479     10410.   274892    13134          379
## 10  2018 CARROLL   OH    0.024    26521      1105.    26254      135           65
## # ... with 78 more rows, and 19 more variables: 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>
```

```
# 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)

midwest%>%
  filter(state=="IN", popwhite > 50000, popwhite < 90000)
```

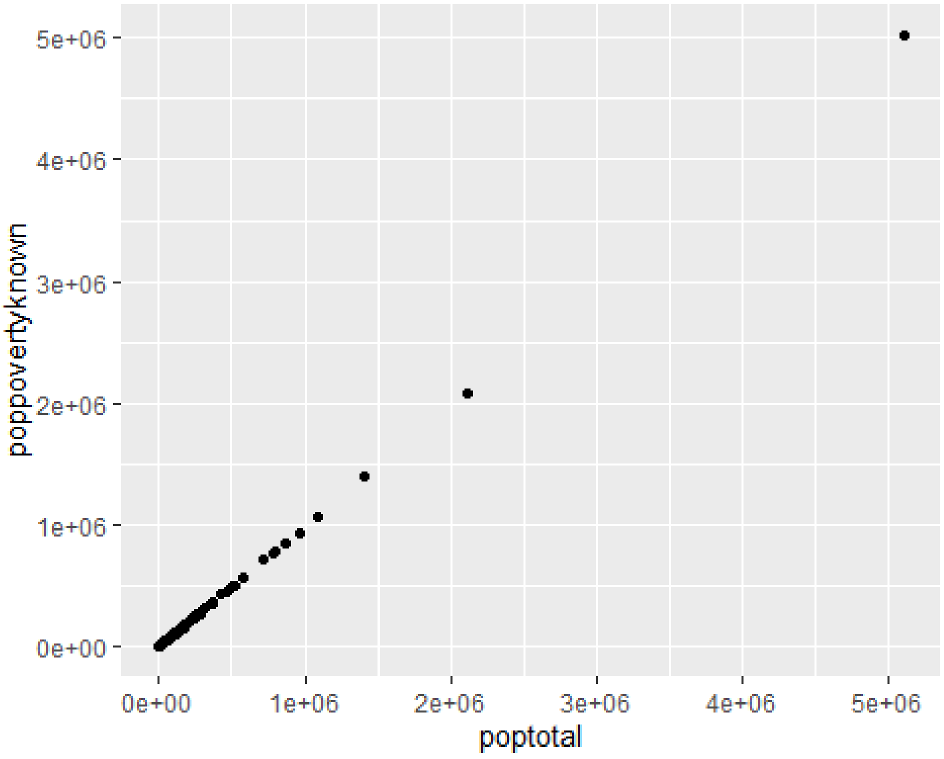
```
## # A tibble: 10 x 28
##      PID county      state  area poptotal popdensity popwhite popblack popamerindian
##    <int> <chr>       <chr> <dbl>    <int>      <dbl>    <int>    <int>         <int>
##  1   665 BARTHOLOMEW IN    0.022    63657      2894.    61774     1005            97
##  2   672 CLARK       IN    0.022    87777      3990.    82289     4703           192
##  3   684 FLOYD       IN    0.009    64404      7156     61415     2642            92
##  4   689 GRANT       IN    0.024    74169      3090.    67817     5047           298
##  5   694 HENDRICKS   IN    0.024    75717      3155.    74519      685           157
##  6   696 HOWARD      IN    0.016    80827      5052.    75420     4398           226
##  7   703 JOHNSON     IN    0.018    88109      4895.    86455      845           139
##  8   705 KOSCIUSKO   IN    0.032    65294      2040.    64058      309           118
##  9   717 MORGAN      IN    0.024    55920      2330     55635        9           137
## 10   751 WAYNE       IN    0.024    71951      2998.    67532     3795           153
## # ... with 19 more variables: 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>
```

```
# 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?

midwest%>%
  select(state, county, poptotal, popamerindian, percamerindan)%>%
  filter(state == "IN")%>%
  arrange(desc(popamerindian))%>%
  print(n=20)
```

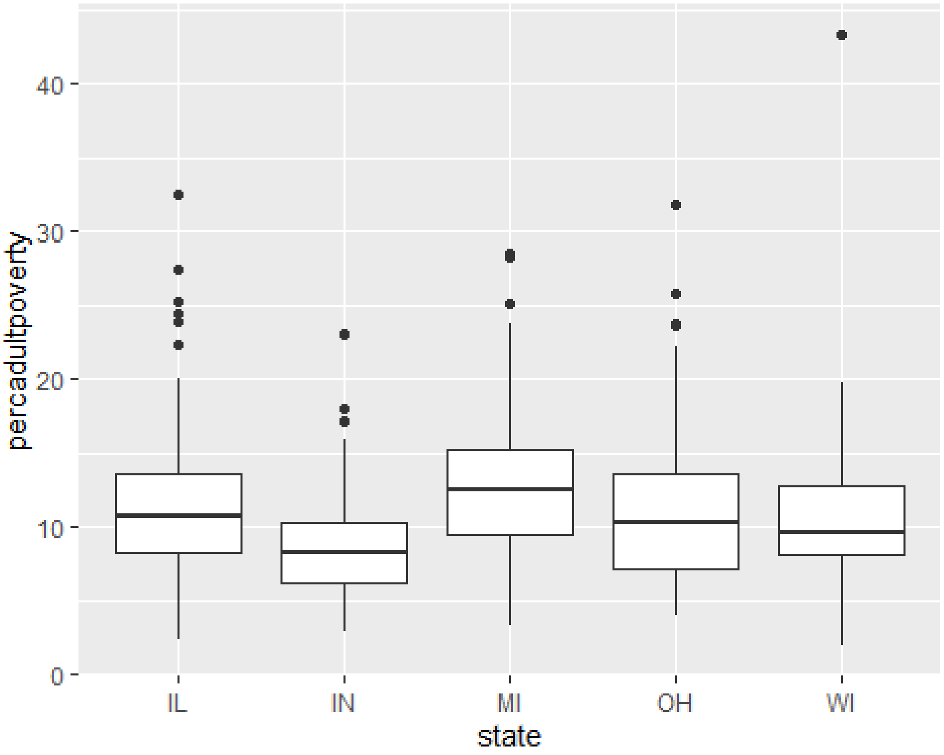
```
## # A tibble: 92 x 5
##    state county      poptotal popamerindian percamerindan
##    <chr> <chr>          <int>         <int>         <dbl>
##  1 IN    MARION        797159          1698         0.213
##  2 IN    ALLEN         300836           892         0.297
##  3 IN    LAKE          475594           865         0.182
##  4 IN    ST JOSEPH     247052           846         0.342
##  5 IN    MIAMI          36897           571         1.55
##  6 IN    ELKHART       156198           453         0.290
##  7 IN    TIPPECANOE    130598           320         0.245
##  8 IN    MADISON       130669           299         0.229
##  9 IN    GRANT          74169           298         0.402
## 10 IN    VIGO          106107           297         0.280
## 11 IN    VANDERBURGH   165058           284         0.172
## 12 IN    DELAWARE      119659           274         0.229
## 13 IN    LA PORTE      107066           259         0.242
## 14 IN    WABASH         35069           259         0.739
## 15 IN    PORTER        128932           243         0.188
## 16 IN    HOWARD         80827           226         0.280
## 17 IN    MONROE       108978         216        0.198
## 18 IN    CLARK        87777         192        0.219
## 19 IN    HAMILTON     108936         163        0.150
## 20 IN    HENDRICKS     75717          157        0.207
## # ... with 72 more rows

# The Marion county is the county with the highest level of native americans

# 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.

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

## # 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%>%
  select(state, county, poppovertyknown)%>%
  group_by(state)%>%
  summarise(meanpov = mean(poppovertyknown))
```

```
## # A tibble: 5 x 2
##   state meanpov
##   <chr>   <dbl>
## 1 IL    109253.
## 2 IN     58396.
## 3 MI    109362.
## 4 OH    120163.
## 5 WI     66029.
```

```
# 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)+
  geom_point(mapping = aes(x=poptotal, y=poppovertyknown))
```

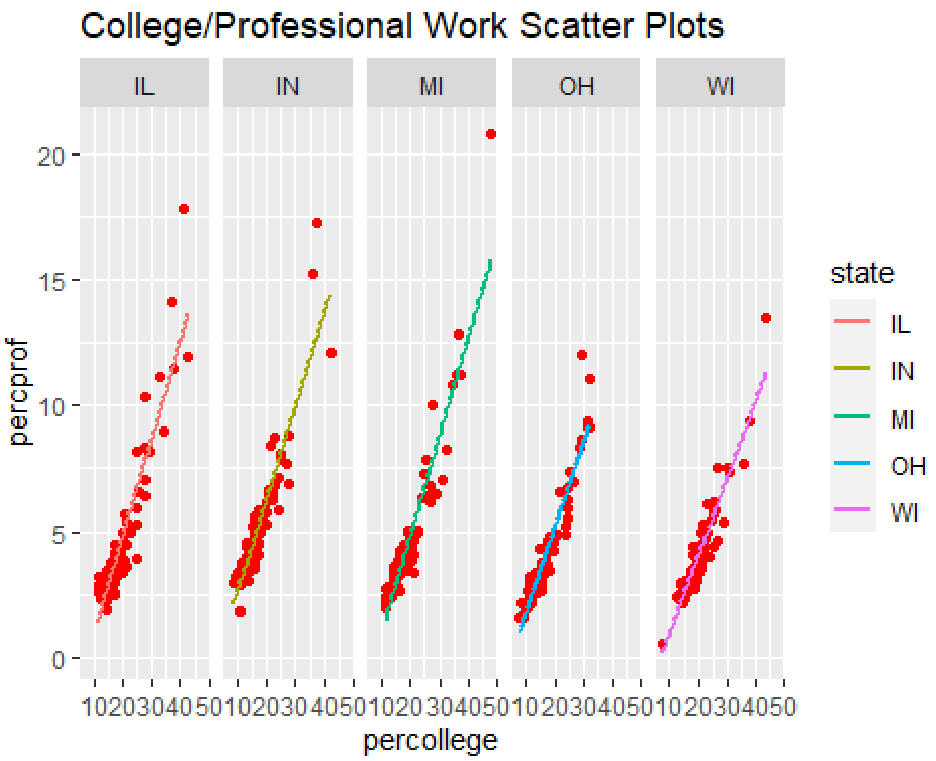


```
# 7
# Using the midwest data, write r code that will produce the following
# side by side boxplots.

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

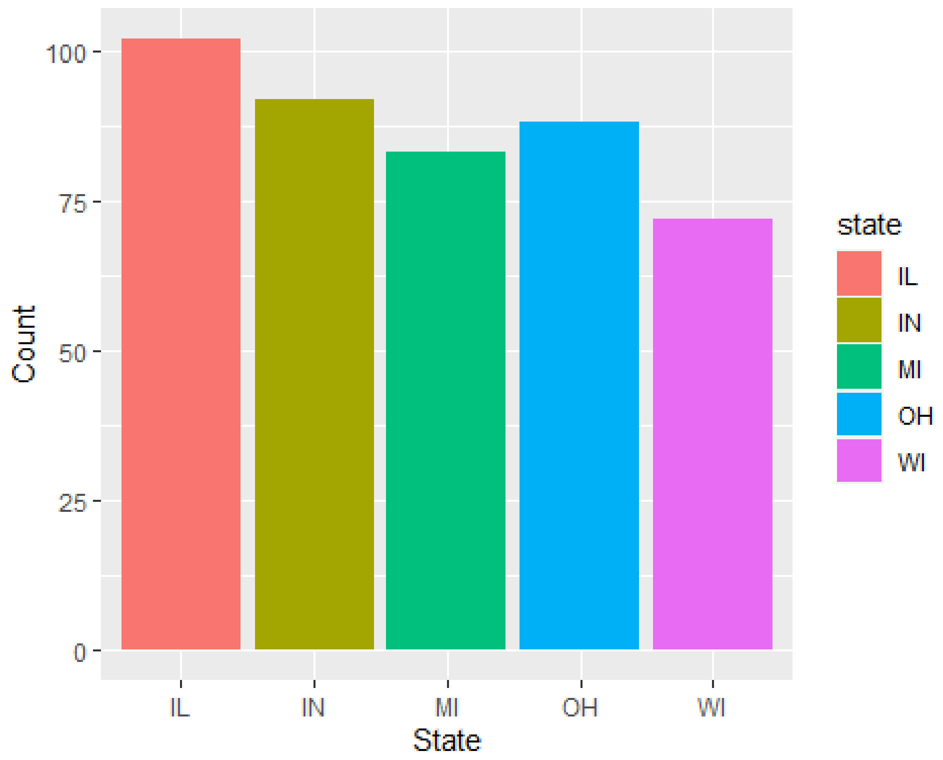


```
# 8
# Using the midwest data, write r code that will produce a facet plot
# that shows scatter plots (red data points) with respect to the levels
# for the variable state. Also add code that will generate regression
# lines through your scatter plots that feature x = percollege and y = percprof.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

ggplot(data=midwest)+
  geom_point(mapping = aes(x=percollege, y=percprof, color = state), color = "Red") +
  geom_smooth(method = lm, mapping = aes(x=percollege, y=percprof, color = state), se=F)+
  ggtitle("College/Professional Work Scatter Plots")+
  facet_grid(~state)
```

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

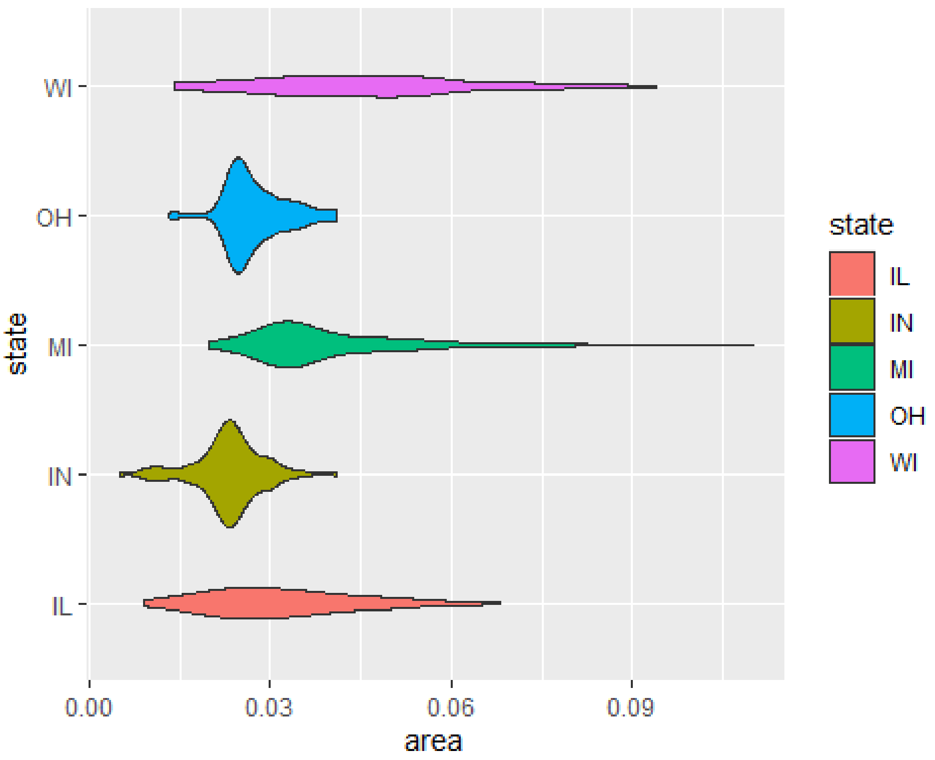


```
# 9
# Using the midwest data frame, create a bar graph that shows the
# different counts for each state in the data set. Your bars should
# have different colors.  Which state has the highest count?

ggplot(data=midwest)+
  geom_bar(mapping = aes(x=state, y=frequency(state),fill = state), stat = "identity")+
  ylab("Count")+
  xlab("State")
```



```
# The state of Illinois has the highest count out of all of the states considered in this dataset

#  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.

volume = function(r,h)
{pi*(r)**2*h
  return(pi*(r**2)*h)}

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

```
volume(5,10)

## [1] 785.3982

area(5,10)

## [1] 471.2389

# 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.

midwest%>%
  select(state, county, poptotal, popadults)%>%
  mutate(Ratio = popadults/poptotal, Percent = Ratio*100)%>%
  filter(state=="WI")%>%
  mutate(state=recode(state,"WI"="Wisconsin"))%>%
  print(n=25)

## # A tibble: 72 x 6
##    state     county    poptotal popadults Ratio Percent
##    <chr>     <chr>        <int>     <int> <dbl>   <dbl>
##  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
##  6 Wisconsin BUFFALO      13584      8918 0.657    65.7
##  7 Wisconsin BURNETT      13084      9045 0.691    69.1
##  8 Wisconsin CALUMET      34291     20940 0.611    61.1
##  9 Wisconsin CHIPPEWA     52360     33195 0.634    63.4
## 10 Wisconsin CLARK        31647     19702 0.623    62.3
## 11 Wisconsin COLUMBIA     45088     29637 0.657    65.7
## 12 Wisconsin CRAWFORD     15940     10169 0.638    63.8
## 13 Wisconsin DANE         367085  225973 0.616    61.6
## 14 Wisconsin DODGE         76559   49694 0.649    64.9
## 15 Wisconsin DOOR          25690   17369 0.676    67.6
## 16 Wisconsin DOUGLAS       41758   27060 0.648    64.8
## 17 Wisconsin DUNN          35909   19755 0.550    55.0
## 18 Wisconsin EAU CLAIRE    85183   49336 0.579    57.9
## 19 Wisconsin FLORENCE       4590    3057 0.666    66.6
## 20 Wisconsin FOND DU LAC   90083   56764 0.630    63.0
## 21 Wisconsin FOREST         8776    5608 0.639    63.9
## 22 Wisconsin GRANT         49264   29160 0.592    59.2
## 23 Wisconsin GREEN         30339   19708 0.650    65.0
## 24 Wisconsin GREEN LAKE    18651   12453 0.668    66.8
## 25 Wisconsin IOWA          20150   12747 0.633    63.3
## # ... with 47 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(data=midwest)+
  geom_violin(mapping = aes(x=area, y=state, fill=state))
```


WI
OH
MI
IN
IL
state
0.00
0.03
0.06
0.09
area
state
IL
IN
MI
OH
WI