AMERICAN UNIVERSITY

DATA 412/612

**TEST 2 dplyr**

*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 a*n *Rmardown file and a word file or a*n *Rmarkdown file and html file.*

library(tidyverse) library(dplyr) library(ggplot2)

midwest View(midwest)

For some of the problems, you have to determine what variables to use.

Problems 1 – 6 are for undergraduate (412) and graduate (612) students.

Graduate students are to also code for problems 11 and 12.

# **1**

# Using the **midwest** data table generate output for Ohio (OH) only. Produce

correct output by using two methods. First use the piping method and then use the assignment method.

# **2**

#Using the **midwest** data table, produce output that shows the

# white population that is greater than 50,000 but less than 90,000 for

# the state of Indiana (IN)

# **3**

# Using the **midwest** data table , produce output of (20 observations)

# that shows only the variables **state, county**, **poptotal** ,

# **popamerindian**, **percamerindian** for the state of Indiana. Also your data

# table table should show **popamerindian** in descending order.

# Which county in Indiana has the highest number of Native Americans?

# **4**

# Using the **midwest** data and dplyr functions, create a data table 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.

# **5**

# Using the **midwest** data table and dplyr commands and functions, write r code

# that will show the mean of the poverty population for the counties of each state.

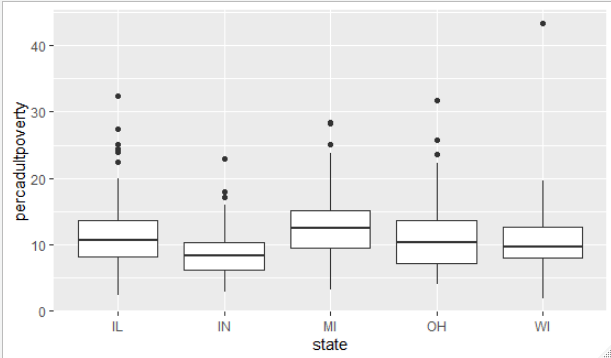
# **6**

**a)** Using the **midwest** data table, Use and show R coding to determine which State (**state**) has the largest mean population total (**poptotal**). **b)** Also code to determine how many counties this State has.

# **7**

# Using the **midwest** data table, write r code that will produce the following

# side by side boxplots.



# **8**

# Using the **midwest** data table, 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"

# **9**

# Using the **midwest** data table, 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?

# **10**

# a) Use R code to confirm that **midwest** is a tibble.

# b) Use two R coding methods to subset the column variable **poptotal** from

# the **midwest** data table.

# c) Use and show R coding to find the mean **area** for each **state** of the **midwest** data

table.

# 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

*<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

# **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.)

