BANA 7025
Data Wrangling with R
Homework #3

Deadline to submit via Canvas: 11:59PM, Sunday, Nov 8th, 2020

Submission expectations

You will submit an RMD script containing all code and a HTML report including explanation and results.

Reminder about code, output, and explanations

NOTE: Code or output submitted without any explanations or rationale will receive zero points. You must show code and explain what your code/output mean. However, you do not need an endless number of sentences for each question. Your responses should answer each question—no more, no less.

Homework Questions to Answer

You will import into RStudio the *acs_2015_county_data_revised.csv* file from the Week 4 folder you downloaded for today's class to complete this homework set.

Understand Context

0. Read the data dictionary included in this week's folder that you downloaded from the course website. *(Note: There is nothing to document on your homework submission—this step zero is a friendly reminder to read the documentation.)*

Importing Data and Data Cleaning

1. Import the data set using a *Tidyverse* function and NOT with a Base R function. How many rows and columns are in the data set?
2. Do any data types need changed? Show any code to change variable types and show code/output for a `glimpse()` command after you're finished.
3. Are there any missing values? How will you handle missing values? Will you impute a missing value with, for example, a mean or median value for the entire column, or will you remove the entire observation? Give a rationale for your decision and show any code/output to handle missing values.
4. Use the `summary()` function to examine any unusual values. Are there any? If so, how will you handle these unusual values? Show any code/output to handle unusual values.
   Notes:

- For the sake of time, you do not need to create any visualizations or other statistical summaries for every variable—the summary function will suffice for this homework).
- You should read the data dictionary for this homework to understand the context behind each variable.

Data Manipulation and Insights

5. How many counties have more women than men?
6. How many counties have an unemployment rate lower than 10%?
7. What are the top 10 counties with the highest mean commute? Show the census ID, county name, state, and the *mean_commute* in your final answer (sorted by *mean_commute*).
   Notes:
   - *Use the variable* mean_commute *to answer this question.*
   - *Leverage the* `dplyr::top_n()` function. Read the documentation for this function.
8. Create a new variable that calculates the percentage of women for each county and then find the top 10 counties with the lowest percentages. Show the census ID, county name, state, and the percentage in your final answer (sorted by ascending percentage).
9. Create a new variable that calculates the sum of all race percentage variables (these columns are the *"hispanic"*, *"white"*, *"black"*, *"native"*, *"asian"*, and *"pacific"* variables).
   a. What are the top 10 counties with the lowest sum of these race percentage variables?
   b. Which state, on average, has the lowest sum of these race percentage variables?
   c. Do any counties have a sum greater than 100%?
   d. How many states have a sum that equals exactly to 100%?
10. Using the *carpool* variable,
    a. Use the `dplyr::min_rank()` function to create a new variable called *carpool_rank* where the highest ranked county (rank = 1) is the county with the highest *carpool* value. Read the documentation carefully for the ranking function.
    b. Find the 10 highest ranked counties for carpooling. Show the census ID, county name, state, carpool value, and carpool_rank in your final answer.
    c. Find the 10 lowest ranked counties for carpooling. Show the same variables in your final answer.
    d. On average, what state is the best ranked for carpooling?
    e. What are the top 5 states for carpooling?