Homework Assignment: Week 3

Assignment Due: Sunday, January 2, 2022 by Midnight CT

Submit electronic solutions to drop box in BrightSpace

This homework, problems 1-9, will be using the file census.csv that is loaded in BrightSpace. The data was obtained from <https://www2.census.gov/programs-surveys/popest/datasets/2010-2016/national/totals/nst-est2016-alldata.csv>

For each question, you must turn in the R code as well as the answer. You may turn in one code file with the answers to each question (in fact, that is encouraged and probably easiest).

1. Write the code lines to complete a file investigation of census.csv. Describe the variables and observations in the file (in general, not one by one). Note that you may need to do a little research on the US Census web sites to find the definition of the variables.

The census data file has 2759 observations with 11 columns/variables. The first two variables (Name, LSAD) are Factor variables, while the other nine variables are all int variables. At first glance, this makes sense because the Name variable contains the city or county and state of the observation, and LSAD (Legal/Statistical Area Description) describes the entity observed. The Name variable has 2757 levels, and LSAD has 4 levels. The rest of the variables in the dataset are numerical metrics that tell us specific census data for each observation.

2. Determine if there any na’s in the data. Also, find the mean, median and standard deviation of the Census 2010 population variable. Compare the mean and median and make a preliminary assessment of the shape of the distribution.

There are no NAs in the data. The mean of the Census 2010 population variable is 239,150.9, the median is 62,544, and the standard deviation is 785,448.7Because the mean and median are relatively far apart, this leads me to believe there are some outliers in the data that are skewing the data. This is because the mean is much more susceptible to outliers than the median.

3. Sort the data frame by the Census 2010 population variable and determine the states with the smallest and largest populations. List the three with the smallest population and the three with largest population.

Smallest Population

* Kenedy County, TX (416)
* McPherson County, NE (539)
* Banner County, NE (690)

Largest Population

* New York-Northern New Jersey-Long Island, NY-NJ-PA (18897109)
* Los Angeles-Long Beach-Santa Ana, CA (12828837)
* New York-White Plains-Wane, NY-NJ (11576251)

4. Make a barplot, in base R, of the estimated 2016 population. Use the help menu in R for the barplot command to understand and use the **names.arg** and **las** arguments. Use these to improve the visualization of your bar plot.

See r code file.

5. Prepare a histogram in base R of the birth rate in 2016. Vary the breaks until you find a histogram that represents the data. Explain your decision. Add appropriate labels and title to the histogram.

I decided to use 30 breaks for my histogram. This is because I found that this allowed me to view enough specific detail about the data to determine the distribution. With any value less than 30, the data was collected together and did not allow me to see the full picture.

6. Prepare a boxplot in base R of the birth rate in 2016. Add appropriate labels and title to this histogram. Are there any extreme values? If so, identify the state(s) with the extreme values. Would the histogram alone have identified all of the extreme values? Explain.

There does appear to be at least 1 extreme value, in my opinion. There are a handful of observations that lay outside the outer quartiles, but not by a large enough margin. There is 1 observation that has a birth rate of 17.02922. Compared with the rest of the data, I would consider this an outlier here. The boxplot alone does not identify all extreme values, though. There could be other observations that should be considered outliers for other reasons. For example, if a particular observation has a relatively high birth rate when compared with geographically surrounding areas, this could be an outlier. But, this would not be evident in our boxplot. In this specific case, the observation that is an outlier is Utah.

7 Investigate the correlation between number of births and deaths for the years 2010 through 2016. You can specify the columns you want to us within brackets. For example, if your data is in columns 3 through 11 the code would be: cor(dataframe[, 3:11]) . Also, use the pair function to make a graphic of the correlations. Can you conclude anything about the population growth in the United States from this information?

It appears that each year has a high correlation between deaths and births (~99%). From this information, I can conclude that the US population is close to remaining relatively constant but probably increasing slightly. Each year with a high number of deaths will result in a high number of births. Since the general world population is always increasing slightly, I can conclude that the same holds true for the US. This data supports my conclusion.

8. Develop a theme following the ggplot2 theme code to visualize the rate of birth data comparing the years 2010 to 2013 to 2016. Compare the data using histograms, scatterplots, and boxplots, all under the same theme. At least one type of plot must use facets. What can you conclude about the birth rate in the United States over this time period?

From comparing the birth rates over the years, I can see that the birth rates remain very constant from year to year. They seem to only move by a few decimals. I’m not really sure what I expected before running this comparison, but it’s interesting to see. It’s interesting to see how constant the rate remains across states. In addition, it even looks like the most recent birth (2016) seems to be the highest among the other years. I’d be curious to see the latest years since 2016 to see if this trend continued.

9. Prepare a map using the Maps file example. Color the states (using the fill command) by the estimated population in 2016. Note: The population data must be numeric, and the states must be lower case.

See r file.

10. Finally, something new. Create a word cloud from a recent document you created (a longer email or a Word document). Save the file as a text file (.txt) and then create a word cloud. Does the word cloud convey the key points of the document? Explain.

Yes, I believe the word cloud does. I decided to use a word document I recently put together for work. If some of my coworkers were to take a look at the word cloud, they would under the gist of the topic I was talking about.