10.3 Final Project Step 3

Colin Michael

2021-11-07

setwd("/Users/colinmichael/Desktop/Data Science/DSC520")  
  
states\_df <- read.csv("StatesData.csv")  
  
head(states\_df)

## State CollegeGradRate BingeDrinkingRate ObesityRate LifeExpectancy  
## 1 Alabama 25.46833 12.2 39.0 74.9  
## 2 Alaska 29.55121 20.0 31.9 77.9  
## 3 Arizona 29.46681 15.0 30.9 79.2  
## 4 Arkansas 23.02779 15.2 36.4 75.4  
## 5 California 33.92596 16.7 30.3 81.0  
## 6 Colorado 40.91234 18.1 24.2 80.0

1.) I combined four different Excel CSV sources manually by copying and pasting the data. I altered the percentages to whole numbers for consistency. For example, 25% is shown as 25. I then used the read.csv function to load into a data frame.

2.) The final data set is a CSV excel file with the following columns: state, average college graduation rate, average obesity rate, average binge drinking rate, and average life expectancy.

3.) Future steps would include additional data fields to pull in related to state demographic data.

4.) I could potentially look at this data by grouping the states into regions.

5.) I plan to mostly pivot at the state level and run a regression.

6.) I could summarize the data by ordering states by different values and seeing the correlation to average life expectancy.

7.) I think scatterplots, bar charts, and histograms will all be useful. They will help show correlation and identify trends in the data.

8.) I think running multiple regressions will be very useful. I think finding out correlation coefficients will help me draw analysis and be able to help predict future life expectancy outcomes.

9.) Future steps mostly invovle adding more columns, creating quality data visualizations, and running regressions that are statstically significant.