Colin Michael

DSC 520: Statistics for Data Science

Final Project Step 1

**Introduction**:

The topic of health is something I am passionate about studying because it has a tremendous impact on people’s quality of lives. Health effects everyone in the world and is crucial towards allowing people to pursue their ambitions. Health is a very broad term but was summed up by the World Health Organization in 1948 as “a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.” (MedicalNewsToday.com)

The great thing about studying health is that there is a plethora of easily accessible data on the internet regarding health. For this research project, I will narrow my analysis down to health behaviors and outcomes at the state-level in the United States of America. I am curious to see how health behaviors and outcomes differ across each state and region. I believe habits and behaviors vastly differ across the United States of America and analyzing health at this scope will illuminate some interesting insights.

**Research Questions:**

1. What is the average life expectancy in each state?
2. What is the average B.M.I. in each state, and how is this distributed?
3. What percent of each state’s’ population consumes alcohol?
4. What percent of each state’s population smokes tobacco?
5. What are the average education levels across each state?
6. What is the average income level in each state?
7. What percent of each state’s population have health insurance?
8. What is the availability of hospitals/health care providers in each state?

**Approach/How my approach answers the problem**

My approach towards answering the problem of health outcomes across each state will be data driven. I want to use empirical data to study the relationship between various inputs and different health outcomes. The inputs, or behaviors, will be treated as independent variables. Some ideas of independent variables include income, education level, alcohol consumption, literacy rates, access to health insurance, age, and gender. Health outcomes will be treated as dependent variables. My health outcomes will be measured by life expectancy, childhood mortality, and suicide rates.

**Data**

1. Life Expectancy by State 2021: Pulled from World Population Review, data in CSV format. <https://worldpopulationreview.com/state-rankings/life-expectancy-by-state>
2. College Completion rates by State 2015-2019: Pulled from USDA.gov, data in .XLSX format. <https://data.ers.usda.gov/reports.aspx?ID=17829>
3. Adult Obesity rates by State 2020: Pulled from CDC.gov, data in .CSV format

[**https://www.cdc.gov/obesity/data/prevalence-maps.html**](https://www.cdc.gov/obesity/data/prevalence-maps.html)

1. Binge drinking rates by State 2019: Pulled from Statista.com, data in .XLS format

<https://www.statista.com/statistics/378966/us-binge-drinking-rate-adults-by-state/>

**Required Packages/Plots**

I predict that I may use ggplot2 to graph my data, tidyverse to manipulate the data, dplyr to summarize the data, and RMarkdown to produce a deliverable.

I believe I will use scatterplots, histograms, and bar charts to view trends and illustrate my data. I may have to use cbind and rbind to manipulate my data tables. I also anticipate using lm() to create models and do regression analysis with summary() and predict().

**Questions for Future Steps**

I have never worked with machine learning before. I admittedly do not know very much about it. I believe the chapters on machine learning will be highly informative for me and might influence my final project. I predict that machine learning will make my analysis more scalable and more easily maintained and updated in the future.