DSC520 Final Project 11.3

Colin Michael

2021-11-18

The aim of my research project is to use data to analyze and predict future health outcomes. I am extremely passionate about health because it is something that affects the quality of every person's life and happiness. People must have a certain degree of health to pursue their ambitions and make a positive impact on the world. 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) Health is something that is universally sought after and beneficial.

Health is a fascinating subject matter because there is a vast pool of data available and many different ways to examine health. Massive amounts of data can be overwhelming and may cause research projects to lose direction. I will focus my project on health behaviors and outcomes at the state-level in the United States of America. I believe this is a large enough pool of data to draw quality insights, but will not be overwhelming for my analysis. I am curious to see how different health behaviors are across the United States.

I created my own data set by collecting data from the following public websites: worldpopulationreview.com, usda.gov, cdc.gov, and statista.com. The dependent variable I am researching, or the health outcome I am studying, is life expectancy. This is not a perfect measure of a person's health, but it is an easily accessible piece of data that provides a useful overview on a person's health. At the very least, people tend to want to live longer lives.

The independent variables in my data set, or the health behavioral inputs, are college graduation rate, binge drinking rate, and obesity rate. I believe these three variables all impact health in different ways. Higher education allows people to better understand how the human body works and what foods are beneficial towards a healthy lifestyle. Binge drinking involves consuming large amounts of alcohol in a short period of time. Large consumption of alcohol negatively impacts the liver and also hurts people psychologically. Obesity is also shown to hurt people's health and negatively impact people's bodies.

```r
setwd("/Users/colinmichael/Desktop/Data Science/DSC520")
states_df <- read.csv("StatesData.csv")
head(states_df)
```

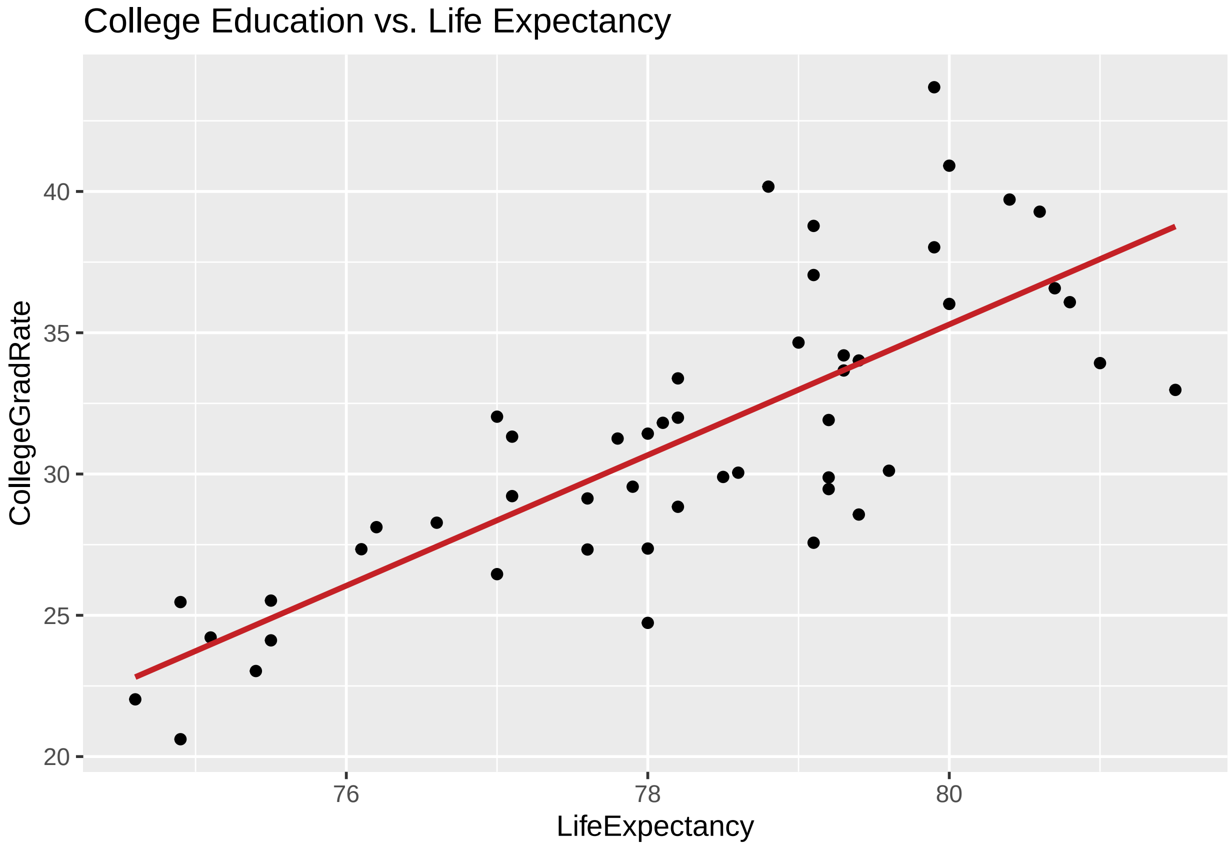
```
##        State CollegeGradRate BingeDrinkingRate ObesityRate LifeExpectancy
## 1    Alabama        25.46833              12.2        39.0           74.9
## 2     Alaska        29.55121              20.0        31.9           77.9
## 3    Arizona        29.46681              15.0        30.9           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
```

```r
library(ggplot2)
ggplot(states_df, aes(x=LifeExpectancy, y=CollegeGradRate)) + geom_point() + stat_smooth(method = "lm",
        col = "#C42126",
        se = FALSE,
        size = 1) + labs(title = "College Education vs. Life Expectancy")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



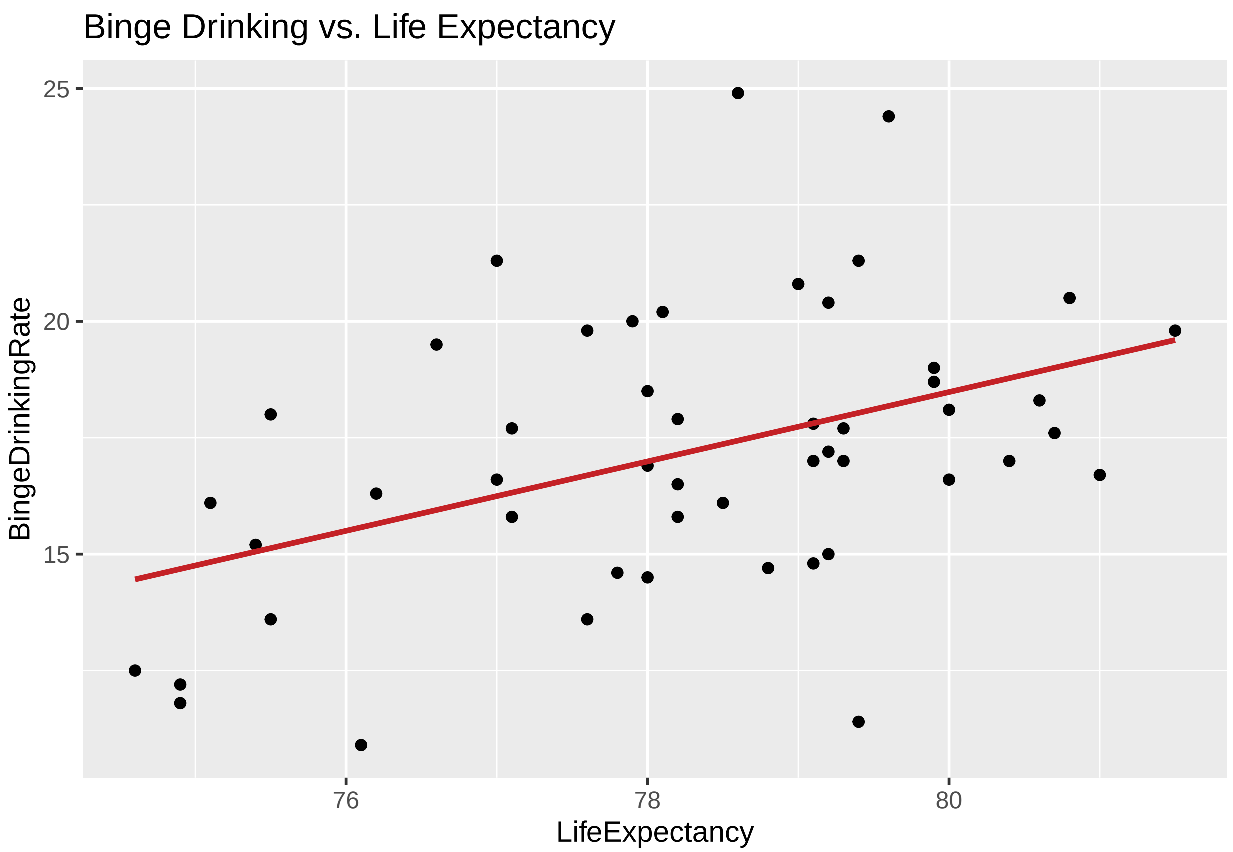
The graph shows a strong positive correlation between college education and life expectancy. This supports my hypothesis that more education leads to better health.

```
library(ggplot2)
ggplot(states_df, aes(x=LifeExpectancy, y=BingeDrinkingRate)) + geom_point() + stat_smooth(method = "lm
        col = "#C42126",
        se = FALSE,
        size = 1) + labs(title = "Binge Drinking vs. Life Expectancy")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



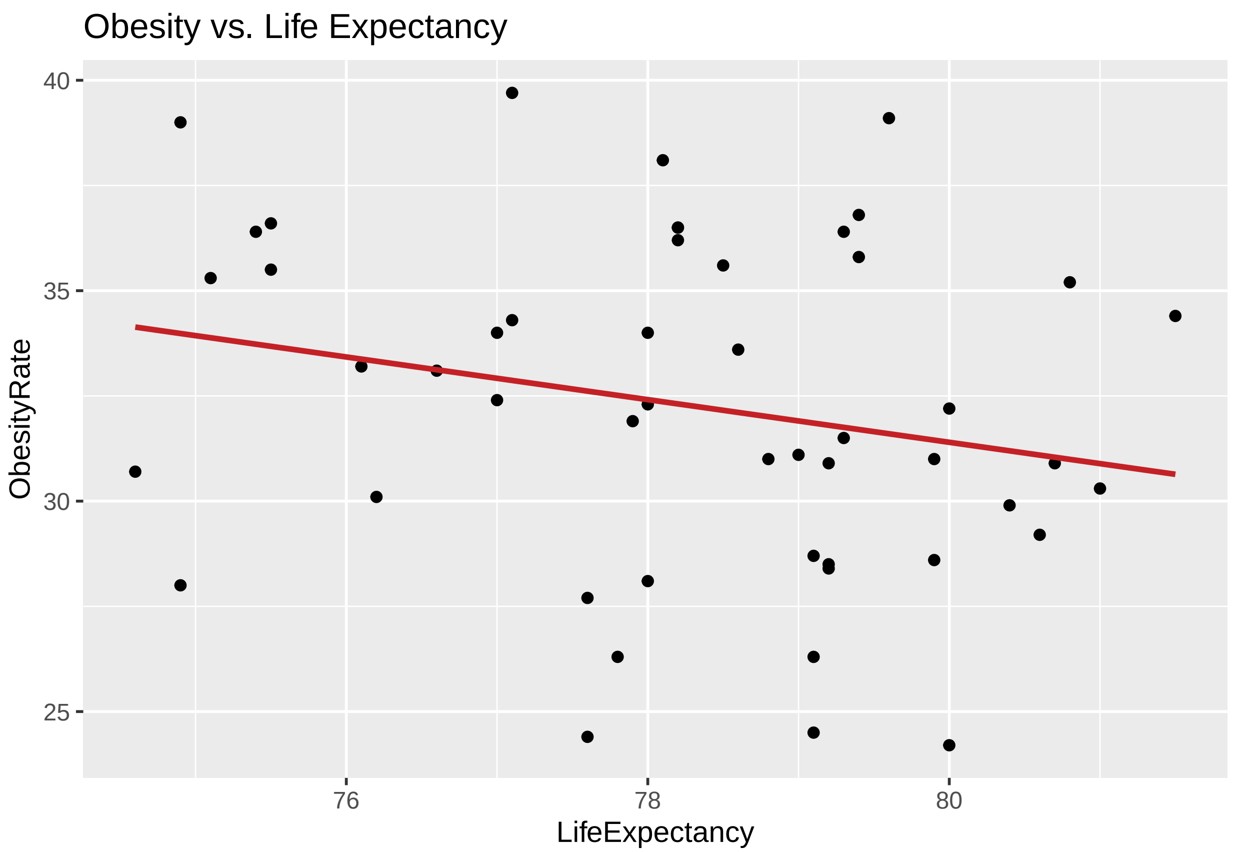
The graph shows a slight positive correlation between binge drinking and life expectancy. This shocks me. I thought there would be a negative correlation between binge drinking and life expectancy.

```
library(ggplot2)
ggplot(states_df, aes(x=LifeExpectancy, y=ObesityRate)) + geom_point() + stat_smooth(method = "lm",
        col = "#C42126",
        se = FALSE,
        size = 1) + labs(title = "Obesity vs. Life Expectancy")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



The graph shows a slight negative correlation between obesity rates and life expectancy. This supports my hypothesis that higher obescity negatively impacts life and health.

```
cor(states_df$LifeExpectancy, states_df$CollegeGradRate, use = "complete.obs")
```

```
## [1] 0.7814938
```

```
cor(states_df$LifeExpectancy, states_df$BingeDrinkingRate, use = "complete.obs")
```

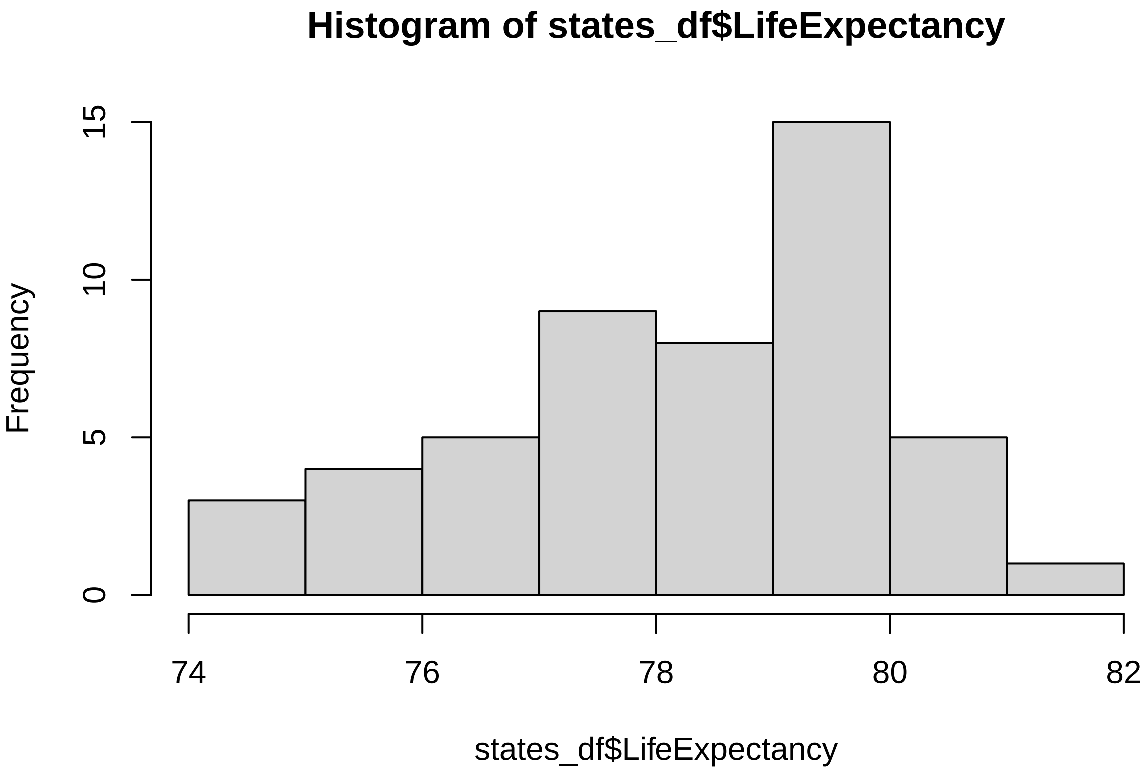
```
## [1] 0.4361845
```

```
cor(states_df$LifeExpectancy, states_df$ObesityRate, use = "complete.obs")
```

```
## [1] -0.2240112
```

The correlation between Life Expectancy is the strongest with College Graduation Rate at .78. Mild correlation is present with Binge Drinking at .42. There is a slight negative correlation with Obesity at -.22.

```
hist(states_df$LifeExpectancy)
```



Life Expectancy is normally distributed, which is an important condition for the regression I want to run.

```
states_lm<-lm(states_df$LifeExpectancy ~ states_df$CollegeGradRate + states_df$BingeDrinkingRate + stat

summary(states_lm)
```

```
##
## Call:
## lm(formula = states_df$LifeExpectancy ~ states_df$CollegeGradRate +
##     states_df$BingeDrinkingRate + states_df$Obesity, data = states_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9242 -0.7923 -0.2599  0.7032  2.5549
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 69.40655    1.83771  37.768  < 2e-16 ***
## states_df$CollegeGradRate    0.23578    0.03274   7.202 4.55e-09 ***
## states_df$BingeDrinkingRate  0.12891    0.05474   2.355   0.0228 *
## states_df$Obesity           -0.02295    0.04077  -0.563   0.5763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.073 on 46 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6527, Adjusted R-squared:   0.63
## F-statistic: 28.81 on 3 and 46 DF,  p-value: 1.228e-10
```

The estimated impact of College Graduation rate on Life Expectancy is .24, the impact of Binge Drinking is .13, and the impact of Obesity is -.02. This means that for every 1% increase in college graduation rate in a state, there is a predicted .24% increase in life expectancy. A 1% increase in binge drinking would predict a .13% increase in life expectancy, and a 1% increase in obesity would decrease life expectancy by -.02%.

The standard error values for the three independent variables are very small, which is good. We want our p-values to be small, ideally less than .05. A p-value under .05 means we can reject the null hypothesis, meaning our insights are valuable. College Graduation rate has the lowest p-value, and Binge Drinking has a p-value also under .05. Obesity has a high p-value of .58, making it not statistically significant. The R-squared of the model is .65, which means that 65% of the life expectancy observations can be explained by the health behaviors studied.

My research shows the strongest predictor of an increase in life expectancy at the state level is college graduation rate. Increases in binge drinking is slightly related to life expectancy increases, which seems paradoxical. I would want to study this in greater detail. Increases in obesity is negatively related to life expectancy, but the p-value in my regression was too high, making this insight not particularly valuable. My biggest takeaway from this project is the higher education is the best predictor of health outcomes.

My data set's biggest limitation is that I only used three dependent variables. In reality, there are much more than three things that contribute to health outcomes. I chose college education, binge drinking, and obesity because they were easy to gather data on and are widely understood. In a future project, I would like to gather many more dependent variables, like smoking rates and quality of hospitals.

I greatly enjoyed working on this project. I learned a lot about both health and R. It was exciting to gather my own data, as opposed to using pre-made data sets. It made me think more about how I should model my data. I liked making both data visualizations and regression models. I think good data science projects must combine compelling visuals with actual statistics. Strking this balance is difficult, but extremely valuable.