Predictors of Covid Death Rate in the United States

Econ 5321 Final Project

Ian McDonough

Introduction

Over the past 2 and a half years the Covid-19 virus has impacted nearly every aspect of our lives striking fear into the hearts of many. This fear was and still is largely driven by uncertainty and has abetted somewhat as we learn more about the virus. People wish to know who is most vulnerable, what risk factors increase their vulnerability and if possible, why this is the case. In this paper I will draw on data from a wide variety of sources to build a model which represents the impact of different risk factors on covid mortality rates by US State.

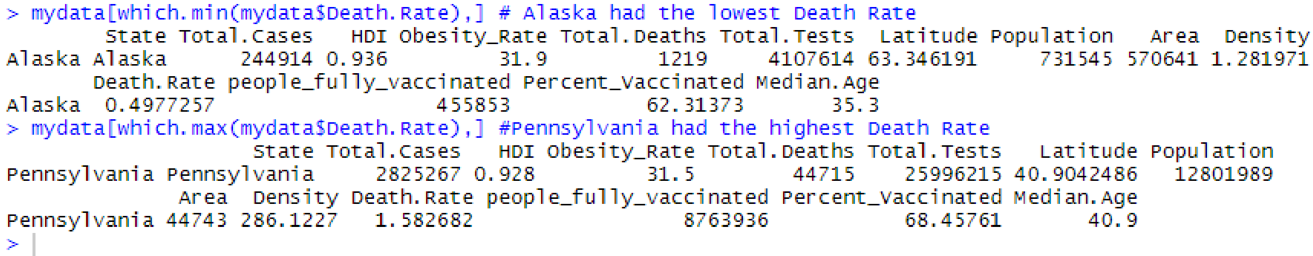
Data

The data sets draw on include USA Covid Data, USA Geography, Vaccination Data as of May 3rd, USA Obesity Statistics, Age by State, and Human Development Index calculations. By comparing the raw data to State Population, I was able to generate various population adjusted figures, i.e. Obesity Rate, Death Rate, Vaccination Rate, etc. Latitude was also used as a regressor in simple regression however it was not included in the final multiple regression model as it proved difficult to work with.

Summary Statistics

Death Rate

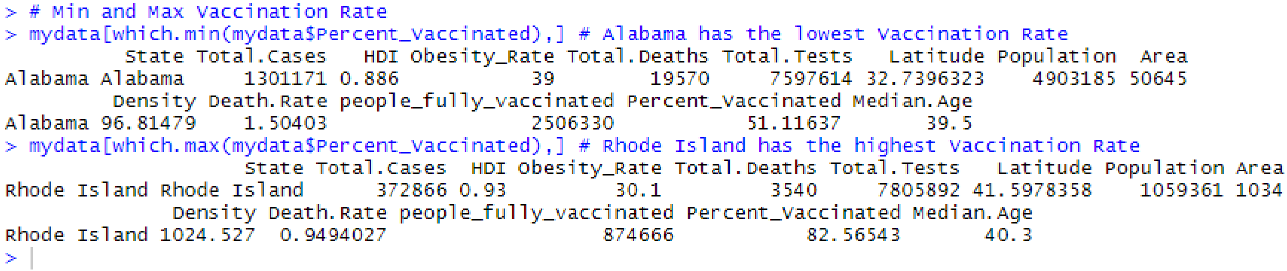
```
Min.    1st Qu. Median  Mean    3rd Qu.    Max.
0.4977  0.9507  1.2047  1.1641  1.3987     1.5827
```



```
> mydata[which.min(mydata$Death.Rate),] # Alaska had the lowest Death Rate
        State Total.Cases   HDI Obesity_Rate Total.Deaths Total.Tests  Latitude Population   Area  Density
Alaska Alaska      244914 0.936         31.9         1219     4107614 63.346191     731545 570641 1.281971
       Death.Rate people_fully_vaccinated Percent_Vaccinated Median.Age
Alaska  0.4977257                  455853           62.31373       35.3
> mydata[which.max(mydata$Death.Rate),] #Pennsylvania had the highest Death Rate
                    State Total.Cases   HDI Obesity_Rate Total.Deaths Total.Tests   Latitude Population
Pennsylvania Pennsylvania     2825267 0.928         31.5        44715    25996215 40.9042486   12801989
              Area  Density Death.Rate people_fully_vaccinated Percent_Vaccinated Median.Age
Pennsylvania 44743 286.1227   1.582682                 8763936           68.45761       40.9
> 
```

Vaccination Rate

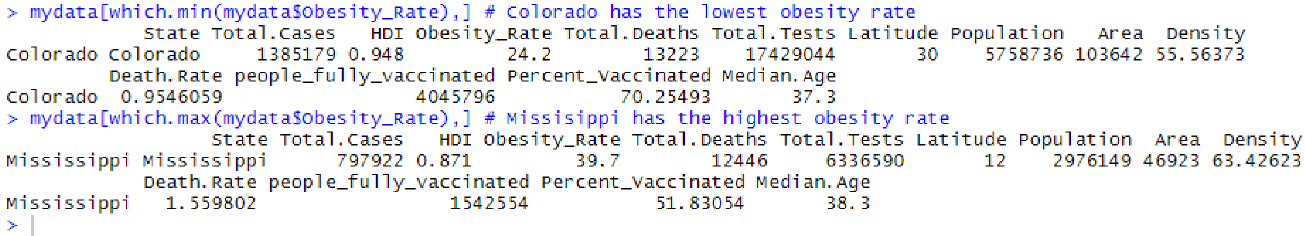
```
Min.    1st Qu.  Median    Mean    3rd Qu.    Max.
51.12   57.04    61.96     64.39   70.36      82.57       1
```



```
> # Min and Max Vaccination Rate
> mydata[which.min(mydata$Percent_Vaccinated),] # Alabama has the lowest Vaccination Rate
          State Total.Cases   HDI Obesity_Rate Total.Deaths Total.Tests   Latitude Population  Area
Alabama Alabama     1301171 0.886           39        19570     7597614 32.7396323    4903185 50645
         Density Death.Rate people_fully_vaccinated Percent_Vaccinated Median.Age
Alabama 96.81479    1.50403                 2506330           51.11637       39.5
> mydata[which.max(mydata$Percent_Vaccinated),] # Rhode Island has the highest Vaccination Rate
                    State Total.Cases  HDI Obesity_Rate Total.Deaths Total.Tests   Latitude Population Area
Rhode Island Rhode Island      372866 0.93         30.1         3540     7805892 41.5978358    1059361 1034
              Density Death.Rate people_fully_vaccinated Percent_Vaccinated Median.Age
Rhode Island 1024.527  0.9494027                  874666           82.56543       40.3
> 
```

Obesity Rate

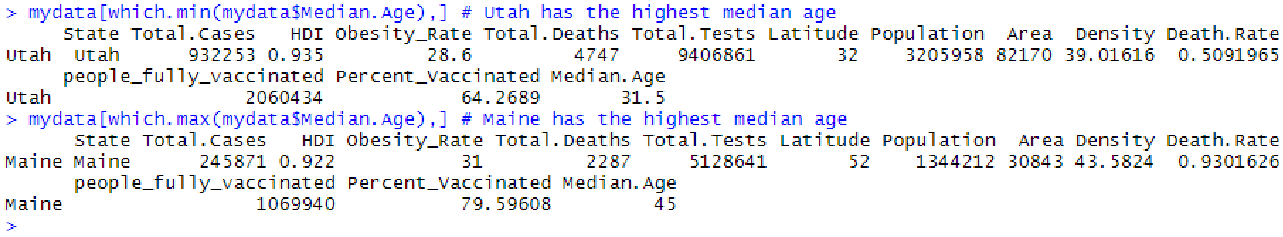
```
Min.    1st Qu. Median  Mean    3rd Qu. Max.
24.20   29.38   32.05   32.21   35.58   39.70
```



```
> mydata[which.min(mydata$Obesity_Rate),] # Colorado has the lowest obesity rate
            State Total.Cases   HDI Obesity_Rate Total.Deaths Total.Tests Latitude Population   Area  Density
Colorado Colorado     1385179 0.948         24.2        13223    17429044       30    5758736 103642 55.56373
         Death.Rate people_fully_vaccinated Percent_Vaccinated Median.Age
Colorado  0.9546059                 4045796           70.25493       37.3
> mydata[which.max(mydata$Obesity_Rate),] # Missisippi has the highest obesity rate
                  State Total.Cases   HDI Obesity_Rate Total.Deaths Total.Tests Latitude Population  Area  Density
Mississippi Mississippi      797922 0.871         39.7        12446     6336590       12    2976149 46923 63.42623
            Death.Rate people_fully_vaccinated Percent_Vaccinated Median.Age
Mississippi   1.559802                 1542554           51.83054       38.3
> 
```

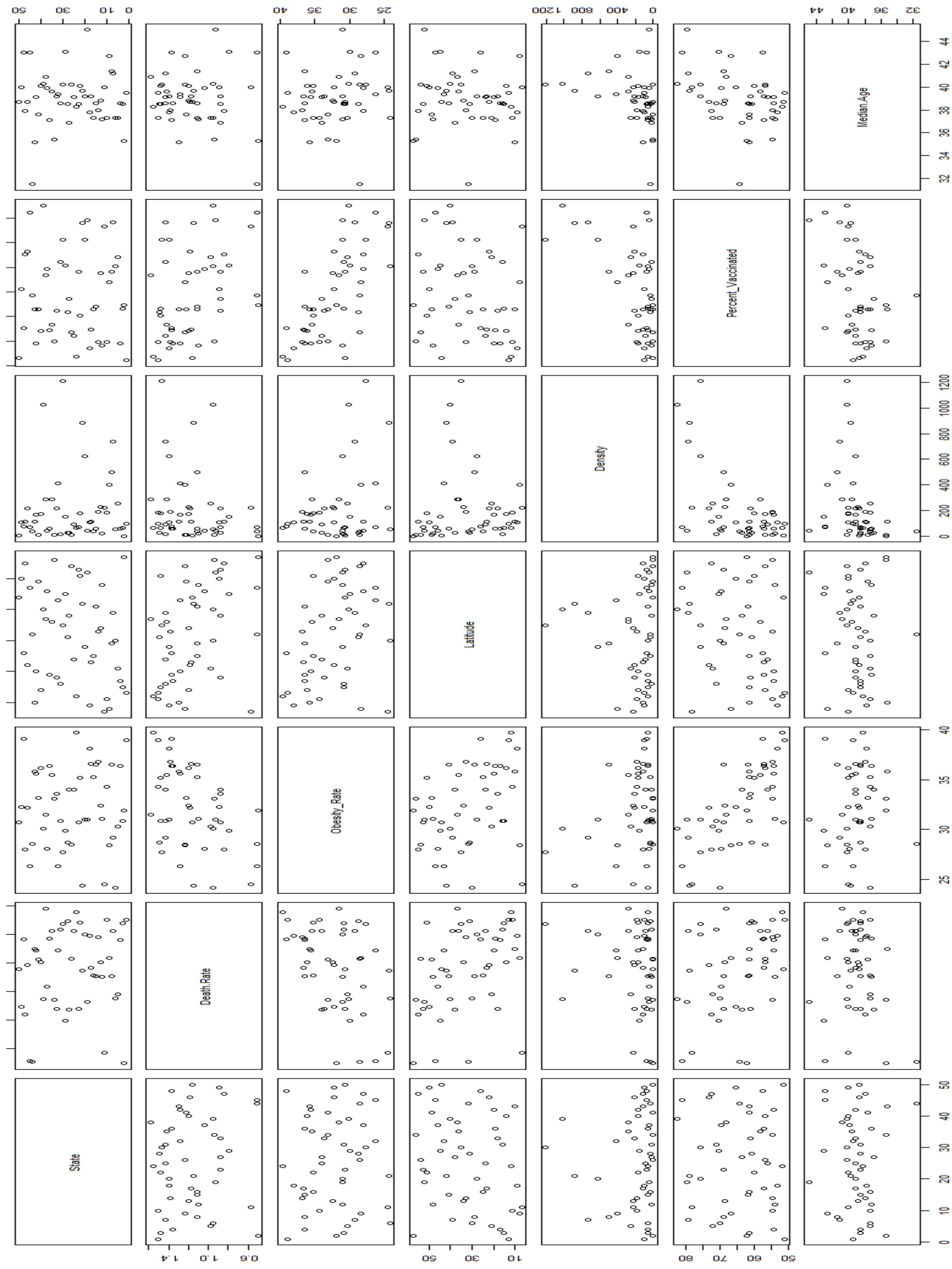
Median Age

```
Min.    1st Qu. Median  Mean    3rd Qu. Max.
31.50   37.83   38.95   39.00   40.08   45.00
```

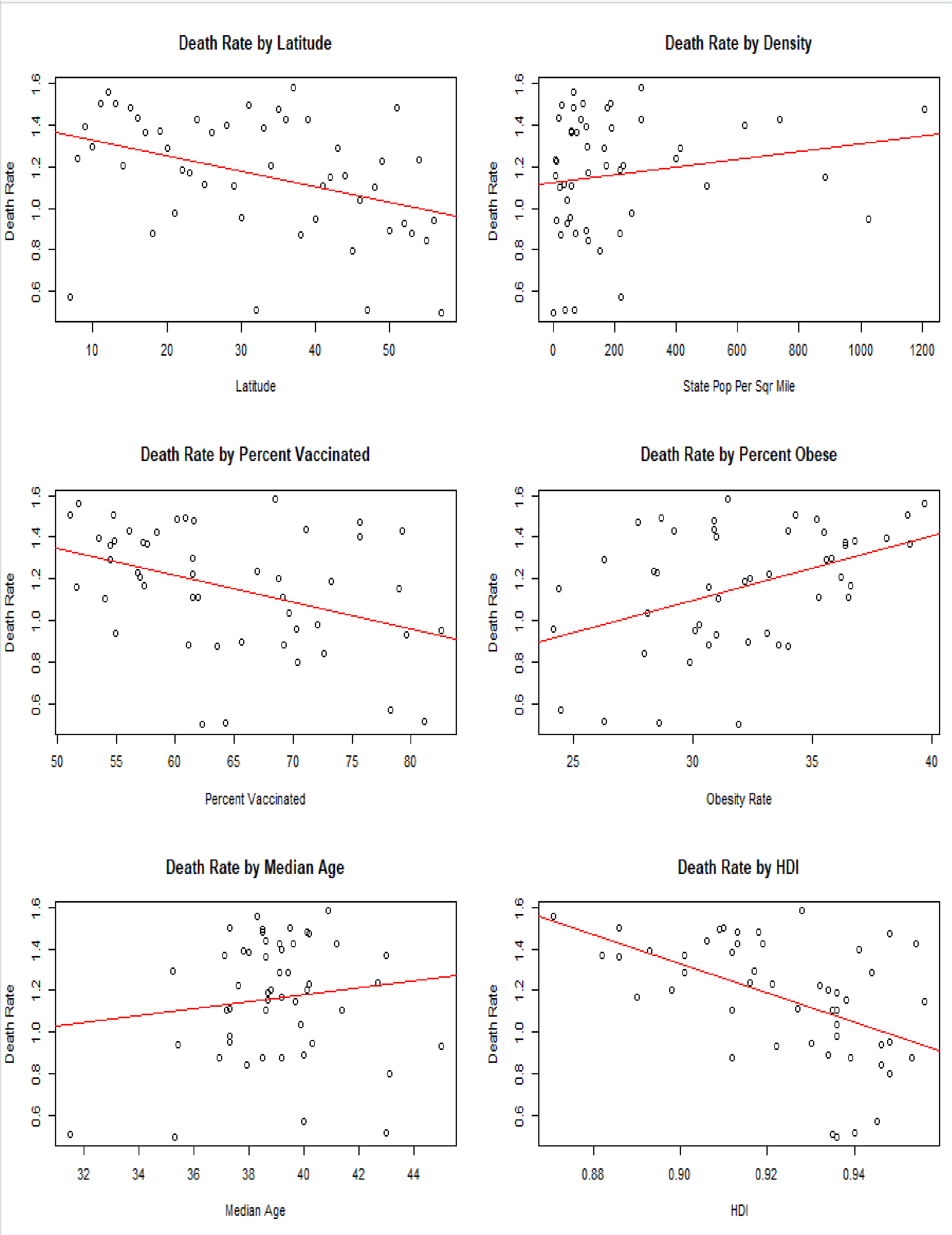


```
> mydata[which.min(mydata$Median.Age),] # Utah has the highest median age
     State Total.Cases   HDI Obesity_Rate Total.Deaths Total.Tests Latitude Population  Area  Density Death.Rate
Utah  Utah      932253 0.935         28.6         4747     9406861       32    3205958 82170 39.01616  0.5091965
     people_fully_vaccinated Percent_Vaccinated Median.Age
Utah                 2060434            64.2689       31.5
> mydata[which.max(mydata$Median.Age),] # Maine has the highest median age
      State Total.Cases   HDI Obesity_Rate Total.Deaths Total.Tests Latitude Population  Area Density Death.Rate
Maine Maine      245871 0.922           31         2287     5128641       52    1344212 30843 43.5824  0.9301626
      people_fully_vaccinated Percent_Vaccinated Median.Age
Maine                 1069940           79.59608         45
> 
```

Relation Overview



Death Rate vs. Factors



Death Rate by Latitude

Death Rate by Density

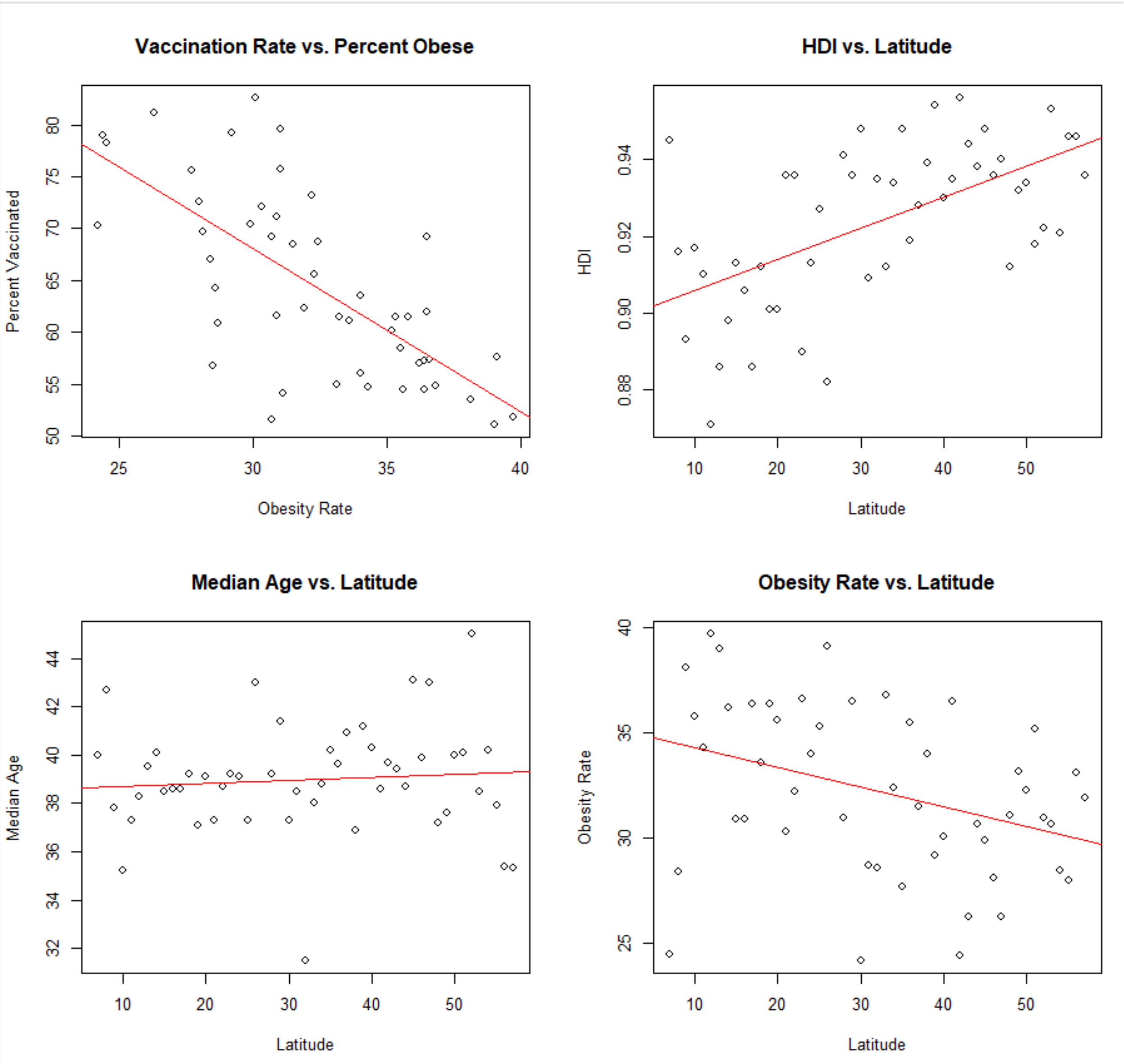
Death Rate by Percent Vaccinated

Death Rate by Percent Obese

Death Rate by Median Age

Death Rate by HDI

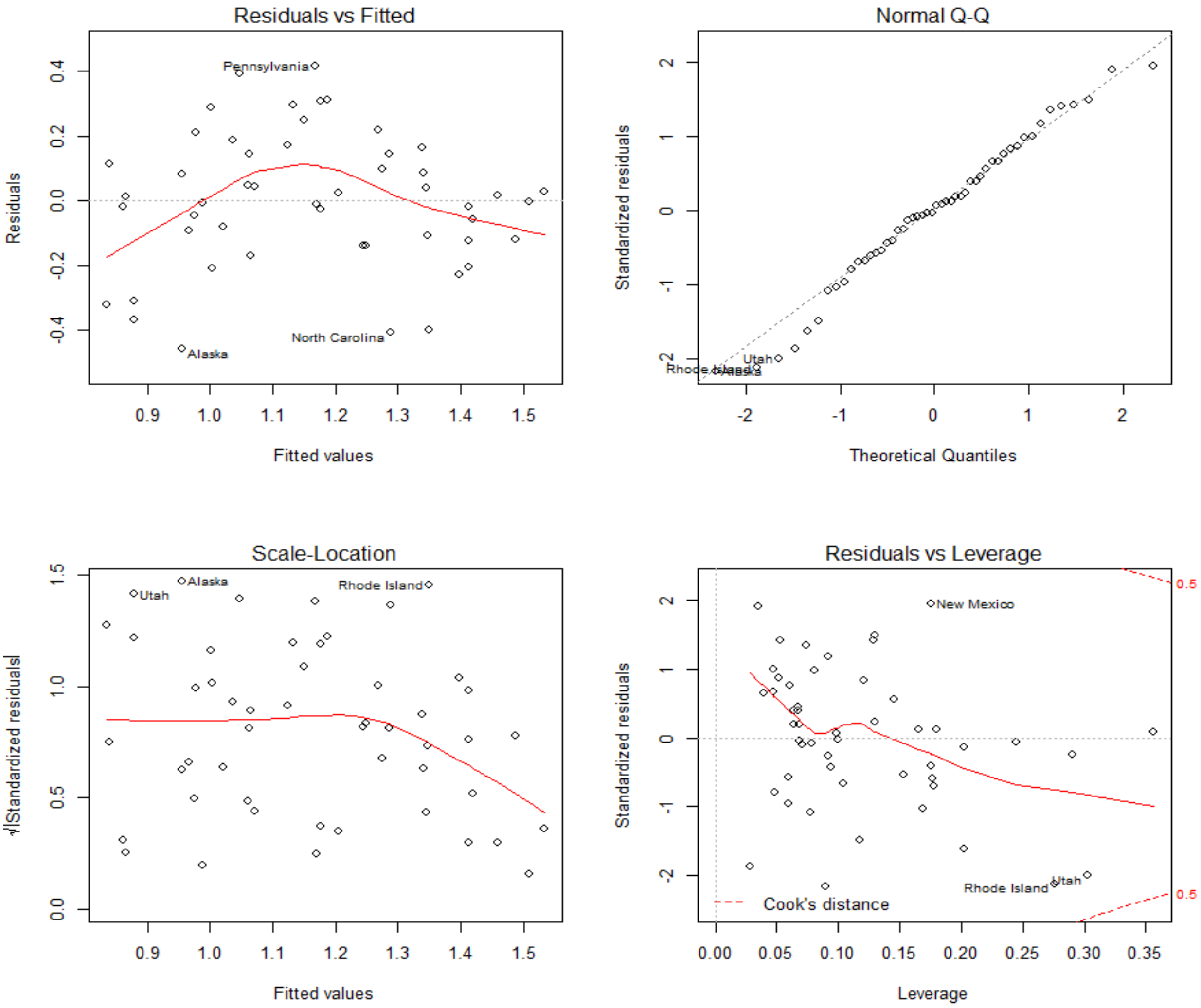
Associations Between Factors



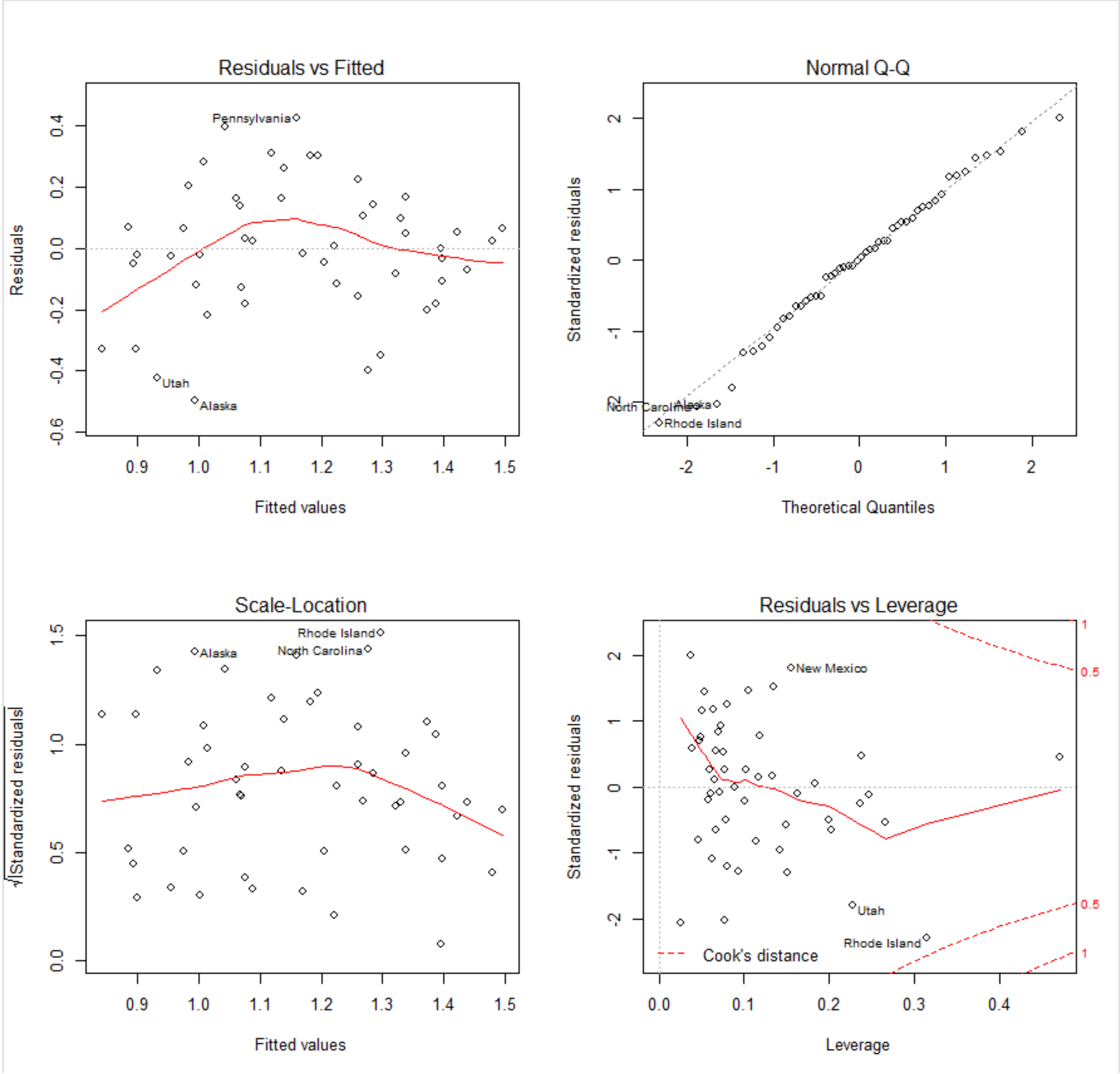
Analysis Methods

After an initial attempt to perform a multiple regression of Density, Median Age, Human Development Index, Percent Vaccinated, and Obesity Rate on Death Rate it became clear that the data was highly heteroskedastic. The relation between HDI and death rate is responsible for much of this, not only is it visually heteroskedastic but after a Breusch-Pagan test which resulted in a p-value = 0.02144, the null hypothesis of homoscedasticity had to be discarded. To deal with the heteroscedasticity I utilized a more robust method, Weighted Least Squares regression.

OLS Regression

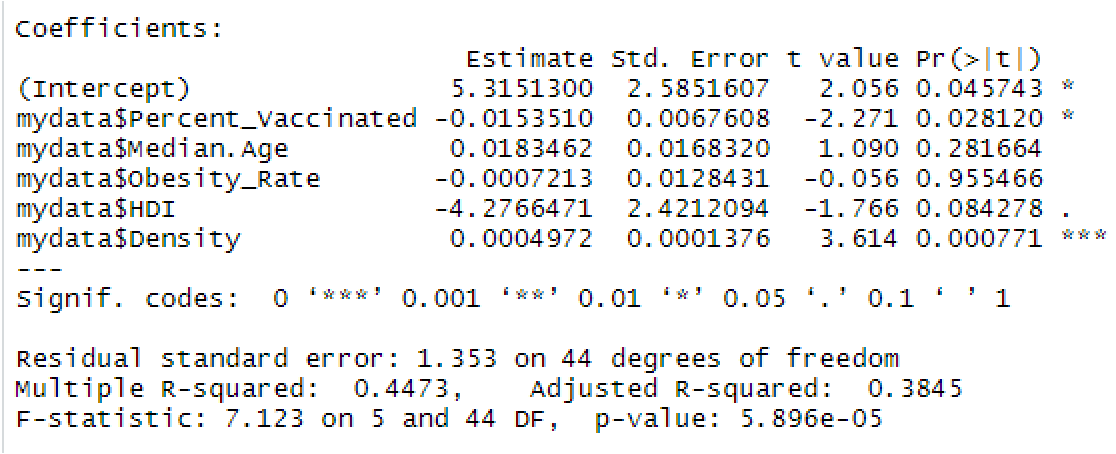


WLS Method


Residuals vs Fitted
Residuals
0.4
0.2
0.0
-0.2
-0.4
-0.6
Pennsylvania
Utah
Alaska
0.9
1.0
1.1
1.2
1.3
1.4
1.5
Fitted values
Normal Q-Q
Standardized residuals
2
1
0
-1
-2
North Carolina
Alaska
Rhode Island
-2
-1
0
1
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
Rhode Island
Alaska
North Carolina
0.9
1.0
1.1
1.2
1.3
1.4
1.5
Fitted values
Residuals vs Leverage
Standardized residuals
2
1
0
-1
-2
New Mexico
Utah
Rhode Island
Cook's distance
1
0.5
0.5
1
0.0
0.1
0.2
0.3
0.4
Leverage

Results

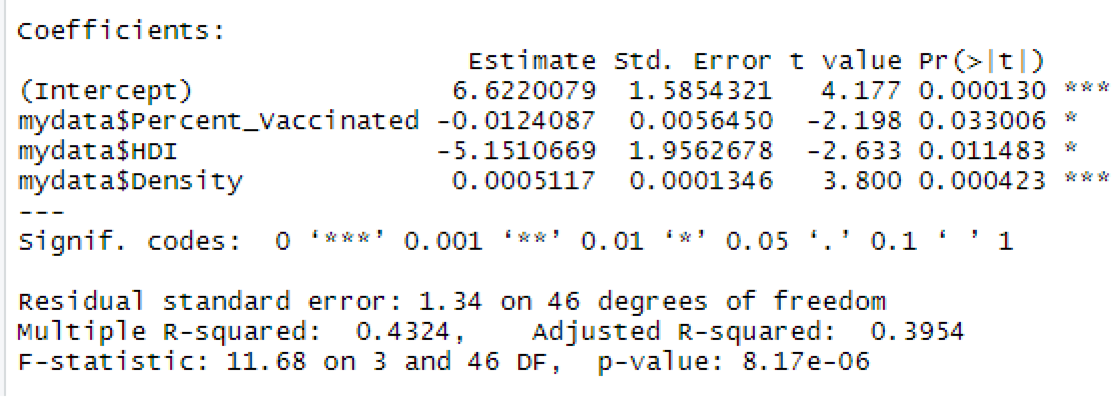
Full Model



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                5.3151300  2.5851607   2.056 0.045743 *  
mydata$Percent_Vaccinated -0.0153510  0.0067608  -2.271 0.028120 *  
mydata$Median.Age          0.0183462  0.0168320   1.090 0.281664    
mydata$Obesity_Rate       -0.0007213  0.0128431  -0.056 0.955466    
mydata$HDI                -4.2766471  2.4212094  -1.766 0.084278 .  
mydata$Density             0.0004972  0.0001376   3.614 0.000771 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.353 on 44 degrees of freedom
Multiple R-squared:  0.4473,	Adjusted R-squared:  0.3845 
F-statistic: 7.123 on 5 and 44 DF,  p-value: 5.896e-05
```

Reduced Model



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                6.6220079  1.5854321   4.177 0.000130 ***
mydata$Percent_Vaccinated -0.0124087  0.0056450  -2.198 0.033006 *  
mydata$HDI                -5.1510669  1.9562678  -2.633 0.011483 *  
mydata$Density             0.0005117  0.0001346   3.800 0.000423 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.34 on 46 degrees of freedom
Multiple R-squared:  0.4324,	Adjusted R-squared:  0.3954 
F-statistic: 11.68 on 3 and 46 DF,  p-value: 8.17e-06
```

Death Rate = 6.6220 -0.0124*(%Vaccinated) -5.15*(HDI)
+ 0.0005*(Density)

- Median age and obesity rate were removed form the full multiple regression model due to their high p-values, this resulted in:
  - Increase in Adjusted R Squared from 0.3845 to 0.3954
  - Decreased P-value from 5.896e^-5 to 8.17e^-6

Discussion

While it was initially supposed that states in higher latitudes with older populations would fare worse than states in lower latitudes, this appears too only be partially true. While States in higher latitudes do tend to be older, they also have a notably higher HDI. The negative effect of age is far outweighed by the impact of HDI on mortality.

There is a strong negative relationship between obesity and vaccination rates (Adjusted R-squared: 0.4903), while this seems counter intuitive there is an even stronger relation between HDI and obesity (Adjusted R-squared: 0.5142). This means that those in areas with low HDI have lower rates of vaccination despite being at higher risk of covid mortality due to increased obesity. This relationship between low HDI and Low vaccination rates could be a result of lower availability of vaccination or lower trust in the medial establishment.

Vaccination rates and density play a small but significant role, the Adjusted R-squared of HDI alone is only 0.2477 vs. 0.3954 for the Final (Resuced) Model, however the effect of density may be misrepresented in this model. Density was calculated by dividing the state's population by its land are. While useful this method of calculating density doesn't account for states like Alaska with a low population density over all but a comparably dense urban center with a large proportion of the state's population.

Finally, given that the dependent variable is a percentage, aka a properly stationarized series, the R squared of 39.54 % is high and indicates that the model has predictive value.

Conclusion

This analysis suggests that, with a coefficient of -5.15, HDI is far and away the best predictor of death on a state-by-state basis. Due to the strong relationship between HDI, obesity, and low vaccination rates it is possible to remove obesity and vaccination rates from the model and increase its predictive power. HDI is calculated using 3 factors: Life Expectancy at Birth, Education Index, and Income. It is not surprising that States which score low on these metrics have suffered disproportionately high mortality from the Covid-19 epidemic. This regression analysis highlights the importance of improving these factors as they have a greater predictive power (Adjusted R Squared: 0.2477) regarding covid outcomes than vaccination (Adjusted R Squared: 0.1221) or age (Adjusted R squared: -0.00392) alone.

Code Appendix

```
######################################
#Final Project Ian McDonough
######################################

# Set Working Directory
setwd("C:/Users/Ian/Desktop/Coding in R 2022 Spring/Final Project")
dir()

#install.packages('lmtest')
library(stringr)

######################################
#Get Data
######################################
```

```
#USA Covid Data
# https://www.kaggle.com/datasets/anandhuh/usa-statewise-latest-covid19-data
usacov = read.csv("USA Covid Data.csv")

# USA Geography
# https://www.census.gov/geographies/reference-files/2010/geo/state-area.html
usageo = read.csv("USA_Geo.csv")

#Vaccinantion Data
# https://ourworldindata.org/us-states-vaccinations
vac = read.csv("us_state_vaccinations.csv")

# USA Obesity
# https://www.cdc.gov/obesity/data/prevalence-maps.html
obe = read.csv("2020-overall.csv")

# Median Age
age = read.csv("Median_Age_States.csv")

# Human Development Index
# https://americatracker.com/most-developed-state-in-the-usa-by-hdi/
hdi = read.csv("HDI.csv")


####################################
# Data Processing
####################################

#HDI
```

```r
hdi = hdi[1:50,]
colnames(hdi)[3]="HDI"

# USA Geography (only States)
usageo1 = usageo[6:56, c("state.and.other.areas", "Land.Area1","Internal.Point..")]
usageo2 = subset(usageo1, state.and.other.areas != "District of Columbia")

#Rename
colnames(usageo2) <- c('State', 'Area', 'Latitude')



# Merging Data Sets
mydata = merge(usacov, usageo2, by = "State")

#Remove commas
mydata$Population = as.numeric(gsub(",","",mydata$Population))
mydata$Area = as.numeric(gsub(",","",mydata$Area))

# New Columns
mydata$Density = mydata$Population / mydata$Area

mydata$Death.Rate = (mydata$Total.Deaths / mydata$Total.Cases)*100


#Vaccine Data as of May 3rd
mayvac = vac[vac$date %in% "2022-05-03",]

mayvac1 = str_replace(mayvac$location, "New York State", "New York")
mayvac$location = mayvac1
```

```
colnames(mayvac)[2] = "State"


mydata = merge(mydata, mayvac, by='State', all.x = TRUE)


# Percent Vacinated
mydata$Percent_Vaccinated = (mydata$people_fully_vaccinated/mydata$Population)*100


# Merging Obesity Stats
mydata = merge(mydata, obe, by='State', all.x = TRUE)


#Merging Age Data
mydata = merge(mydata, age, by='State', all.x = TRUE)


# Merging HDI
mydata = merge(mydata, hdi, by='State', all.x = TRUE)



#Removing Redundent Columns
mydata = mydata[c("State", "Total.Cases","HDI","Prevalence", "Total.Deaths", "Total.Tests", "Latitude",
"Population", "Area", "Density", "Death.Rate", "people_fully_vaccinated", "Percent_Vaccinated",
"Median.Age")]


colnames(mydata)[4] = "Obesity_Rate"


row.names(mydata)= mydata$State


mydata[is.na(mydata)] = 0


class(mydata$Latitude)
mydata$Latitude = as.numeric(mydata$Latitude)
```

```
mydata$Latitude = format(round(mydata$Latiude), 1,) nsmall = 1)



###################################
# Summary Stats
##################################

# Average Death Rate Total
average_death_rate = mean(mydata$Death.Rate)
average_death_rate # Average Death Rate as of May 3rd: 1.164124%
summary (mydata$Death.Rate)

# Min and Max Death Rates
mydata[which.min(mydata$Death.Rate),] # Alaska had the lowest Death Rate

mydata[which.max(mydata$Death.Rate),] #Pennsylvania had the highest Death Rate

# Min and Max Vaccination Rate
summary(mydata$Percent_Vaccinated)
mydata[which.min(mydata$Percent_Vaccinated),] # Alabama has the lowest Vaccination Rate

mydata[which.max(mydata$Percent_Vaccinated),] # Rhode Island has the highest Vaccination Rate

# Min and Max Obesity Rate
summary(mydata$Obesity_Rate)
mydata[which.min(mydata$Obesity_Rate),] # Colorado has the lowest obesity rate

mydata[which.max(mydata$Obesity_Rate),] # Missisippi has the highest obesity rate

#Min and Max Median Age
```

```
summary(mydata$Median.Age)

mydata[which.min(mydata$Median.Age),] # Utah has the highest median age

mydata[which.max(mydata$Median.Age),] # Maine has the highest median age



###################################

#Graphing

###################################


#Focus

mydata1 = mydata[c("State","HDI", "Death.Rate", "Obesity_Rate", "Latitude", "Density",
"Percent_Vaccinated", "Median.Age")]

plot(mydata1)

# Negative correllation between Obesity_Rate and Percent_Vaccinated..




# Latitude vs. Number of Deaths

lat = as.numeric(mydata$Latitude)

dr = mydata$Death.Rate

plot (lat, mydata$Total.Deaths, xlab = "Latitude", ylab = "Total Deaths",main = "Death by Latitude" )

abline(lm(mydata$Total.Deaths~lat,data=mydata),col='red')

#Coefficients

death.reg = lm(mydata$Total.Deaths~lat,data=mydata) #B1: -504.8

summary(death.reg)


# Organize Plots


par(mfrow=c(3,2))

# Latitude vs. Death Rate
```

```
plot (lat, dr, xlab = "Latitude", ylab = "Death Rate",main = "Death Rate by Latitude" ) # Slight Negative Trend between Latitude and Death Rate, Possibley due to HDI or Density
abline(lm(dr~lat,data=mydata),col='red')
#Coefficients
lat.reg = lm(dr~lat,data=mydata) # B1: -0.007486
summary(lat.reg)



# Density vs. Death Rate
plot (mydata$Density, dr, xlab = "State Pop Per Sqr Mile", ylab = "Death Rate",main = "Death Rate by Density" )
abline(lm(dr~mydata$Density,data=mydata),col='red')
# Coefficients
den.reg = lm(dr~mydata$Density,data=mydata) #B1: 0.0001828
summary(den.reg)


# Vaccination vs. Death Rate
plot (mydata$Percent_Vaccinated, dr, xlab = "Percent Vaccinated", ylab = "Death Rate",main = "Death Rate by Percent Vaccinated" )
abline(lm(dr~mydata$Percent_Vaccinated,data=mydata),col='red')
# Coefficeints
vac.reg = (lm(dr~mydata$Percent_Vaccinated,data=mydata)) #-0.0129
summary(vac.reg)


# Obesity vs. Death Rate
plot (mydata$Obesity_Rate, dr, xlab = "Obesity Rate", ylab = "Death Rate",main = "Death Rate by Percent Obese" )
abline(lm(dr~mydata$Obesity_Rate,data=mydata),col='red')
# Coefficeints
obe.reg = lm(dr~mydata$Obesity_Rate,data=mydata) # B1: 0.03111
summary(obe.reg)
```

```
# Median Age vs. Death Rate
plot (mydata$Median.Age, dr, xlab = "Median Age", ylab = "Death Rate",main = "Death Rate by Median Age" )
abline(lm(dr~mydata$Median.Age,data=mydata),col='red')
# Coefficeints
age.reg =lm(dr~mydata$Median.Age,data=mydata) # B1: 0.01636
summary(age.reg)

# HDI vs. Death Rate
plot (mydata$HDI, dr, xlab = "HDI", ylab = "Death Rate",main = "Death Rate by HDI" )
abline(lm(dr~mydata$HDI,data=mydata),col='red')
# Coefficeints
hdi.reg =lm(dr~mydata$HDI,data=mydata) # B1: 0.01636
summary(hdi.reg)
# Appears Homoscedastic
library(lmtest)
bptest(hdi.reg) #Breusch-Pagan test - to check for Homoscedasticity
        #p-value = 0.02144



#################################
# Associations
#################################

#Focus
mydata1 = mydata[c("State", "Death.Rate", "Obesity_Rate", "Latitude", "Density", "Percent_Vaccinated", "Median.Age")]
plot(mydata1)
```

```
#Organize
par(mfrow=c(2,2))

# Obesity Rate vs. Vaccinated Rate
plot (mydata$Obesity_Rate, mydata$Percent_Vaccinated, xlab = "Obesity Rate", ylab = "Percent Vaccinated",main = "Vaccination Rate vs. Percent Obese" )
abline(lm(mydata$Percent_Vaccinated~mydata$Obesity_Rate,data=mydata),col='red')
# Coefficeints
obvac.reg = lm(mydata$Percent_Vaccinated~mydata$Obesity_Rate,data=mydata) # B1:  -1.571
summary(obvac.reg)# Negative correllation between Obesity_Rate and Percent_Vaccinated..

# Latitude vs. HDI
plot (lat, mydata$HDI, xlab = "Latitude", ylab = "HDI", main = "HDI vs. Latitude" )
abline(lm(mydata$HDI~lat,data=mydata),col='red')
# Coefficeints
hdilat.reg = lm(mydata$HDI~lat,data=mydata) #
summary(hdilat.reg)

# Median Age vs. Latitude
plot (lat, mydata$Median.Age, xlab = "Latitude", ylab = "Median Age", main = "Median Age vs. Latitude" )
abline(lm(mydata$Median.Age~lat,data=mydata),col='red')
# Coefficeints
medlat.reg = lm(mydata$Median.Age~lat,data=mydata) #
summary(medlat.reg)

# Obesity vs. Latitude
```

```
plot (lat, mydata$Obesity_Rate, xlab = "Latitude", ylab = "Obesity Rate", main = "Obesity Rate vs.
Latitude" )

abline(lm(mydata$Obesity_Rate~lat,data=mydata),col='red')

# Coefficeints

obelat.reg = lm(mydata$Obesity_Rate~lat,data=mydata) #

summary(obelat.reg)


# Obesity vs. HDI

plot (mydata$HDI, mydata$Obesity_Rate, xlab = "HDI", ylab = "Obesity Rate", main = "Obesity Rate vs.
HDI" )

abline(lm(mydata$Obesity_Rate~mydata$HDI,data=mydata),col='red')

# Coefficeints

obehdi.reg = lm(mydata$Obesity_Rate~mydata$HDI,data=mydata) #

summary(obehdi.reg)



######################################

# Multiple Regression Model

######################################


mreg = lm(mydata$Death.Rate~ mydata$Percent_Vaccinated + mydata$Median.Age +
mydata$Obesity_Rate + mydata$HDI + mydata$Density,data=mydata)

summary(mreg)


par(mfrow=c(2,2))

plot(mreg) # Very Heteroscedasic data




#define weights to use

weight <- 1 / lm(abs(mreg$residuals) ~ mreg$fitted.values)$fitted.values^2
```

```
#perform weighted least squares regression

wls_model <- lm(mydata$Death.Rate~ mydata$Percent_Vaccinated + mydata$Median.Age +
mydata$Obesity_Rate + mydata$HDI + mydata$Density,data=mydata, weights = weight)


#view summary of model

summary(wls_model)


plot(wls_model)



#reduced model

mreg1 =  lm(mydata$Death.Rate~ mydata$Percent_Vaccinated + mydata$HDI +
mydata$Density,data=mydata)


weight1 <- 1 / lm(abs(mreg1$residuals) ~ mreg1$fitted.values)$fitted.values^2


red_wls_model <- lm(mydata$Death.Rate~ mydata$Percent_Vaccinated + mydata$HDI +
mydata$Density,data=mydata, weights = weight1)


#view summary of model

summary(red_wls_model)


plot(red_wls_model)
```