STA9890 Project Proposal
Predicting Premature Death Rates in US Counties

Overview:
The 3,142 counties of the United States span a diverse range of social, economic, and health conditions. In this project, we analyze data collected from the Center for Disease Control (CDC) and the Robert Wood Johnson Foundation to examine the rate of premature deaths in American counties. We will apply different regression techniques to build a predictive model of *'Years of potential life lost rate'* based on other socioeconomic and health factors.

Data Collection:
Collected by the CDC, the SVI dataset contains metrics of "social vulnerability" to natural and man-made disasters. These metrics can be grouped into four categories of risk factors: socioeconomic, household composition and disability, minority status and language, and housing / transportation.

The County Health Rankings, a collaboration between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute, measure the health of nearly all counties in the nation and rank them within states. These measures are standardized and combined using scientifically-informed weights.

The datasets were combined for the years 2016-2018 and can be downloaded at https://www.kaggle.com/johnjdavisiv/us-counties-covid19-weather-sociohealth-data?select=us_county_sociohealth_data.csv, where some variables have been renamed to improve interpretability.

Data Description:
The response variable, '*Years of potential life lost rate' (YPLL)*, measures the number of years of potential life lost before age 75 per 100,000 population. It has been age-adjusted in order to fairly compare counties with differing age structures. YPLL is a widely used measure of the rate and distribution of premature mortality. Measuring premature mortality, rather than overall mortality, reflects the County Health Rankings' intent to focus attention on deaths that could have been prevented. YPLL emphasizes deaths of younger persons, whereas statistics that include all mortality are dominated by deaths of the elderly. For example, using YPLL-75, a death at age 55 counts twice as much as a death at age 65.

Deaths are counted in the county where the individual lived, not the county where they died.

Data Pre-processing:
To prepare the data for model building, we have imputed missing values with column means for numeric variables and column mode for one categorical variable. The numerical predictors have been standardized using equation (6.6) in the ISLR book. Consequently, all of the standardized predictors will have a standard deviation of one, so that the model fit will not depend on the scale on which the predictors are measured. The R script and output for this process is shown in the Appendix.

The final data frame for analysis has 3144 instances (n) and 176 predictors (p). The list of variable names and their indexes is shown in Table 1.

Table 1. List of variables used for analysis

Index	Variables
	Years of potential life lost rate
1	State
2	Total population
3	Area sqmi
4	Population density per sqmi
5	Num deaths
6	Percent fair or poor health
7	Average number of physically unhealthy days
8	Average number of mentally unhealthy days
9	Percent low birthweight
10	Percent smokers
11	Percent adults with obesity
12	Food environment index
13	Percent physically inactive
14	Percent with access to exercise opportunities
15	Percent excessive drinking
16	Num alcohol impaired driving deaths
17	Num driving deaths
18	Percent driving deaths with alcohol involvement
19	Num chlamydia cases
20	Chlamydia rate
21	Teen birth rate
22	Num uninsured
23	Percent uninsured
24	Num primary care physicians
25	Primary care physicians rate
26	Num dentists
27	Dentist rate
28	Num mental health providers
29	Mental health provider rate
30	Preventable hospitalization rate
31	Percent with annual mammogram
32	Percent vaccinated
33	High school graduation rate
34	Num some college
35	Population
36	Percent some college
37	Num unemployed CHR

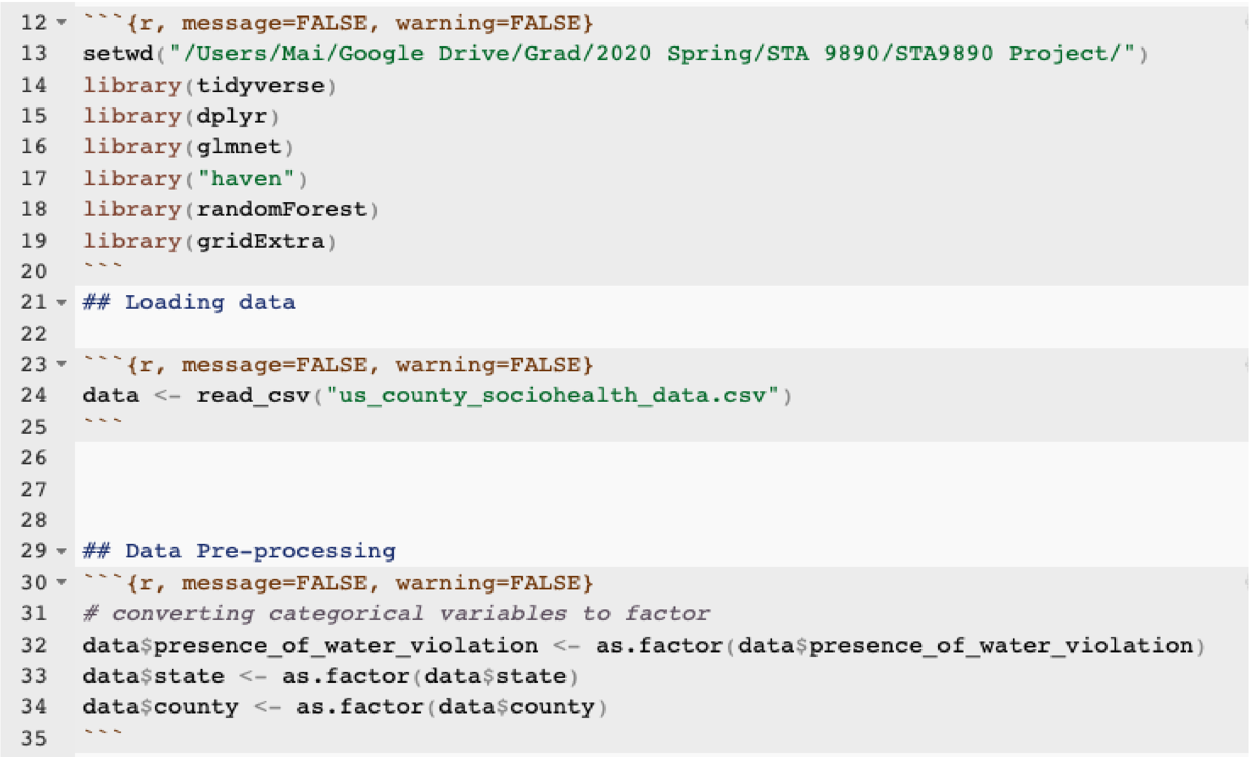
38 Labor force
39 Percent unemployed CHR
40 Percent children in poverty
41 Eightieth percentile income
42 Twentieth percentile income
43 Income ratio
44 Num single parent households CHR
45 Num households CHR
46 Percent single parent households CHR
47 Num associations
48 Social association rate
49 Annual average violent crimes
50 Violent crime rate
51 Num injury deaths
52 Injury death rate
53 Average daily pm2 5
54 Presence of water violation
55 Percent severe housing problems
56 Severe housing cost burden
57 Overcrowding
58 Inadequate facilities
59 Percent drive alone to work
60 Num workers who drive alone
61 Percent long commute drives alone
62 Life expectancy
63 Num deaths 2
64 Age adjusted death rate
65 Num deaths 3
66 Child mortality rate
67 Num deaths 4
68 Infant mortality rate
69 Percent frequent physical distress
70 Percent frequent mental distress
71 Percent adults with diabetes
72 Num HIV cases
73 HIV prevalence rate
74 Num food insecure
75 Percent food insecure
76 Num limited access
77 Percent limited access to healthy foods

78 Num drug overdose deaths
79 Drug overdose mortality rate
80 Num motor vehicle deaths
81 Motor vehicle mortality rate
82 Percent insufficient sleep
83 Num uninsured 2
84 Percent uninsured 2
85 Num uninsured 3
86 Percent uninsured 3
87 Other primary care provider rate
88 Percent disconnected youth
89 Average grade performance
90 Average grade performance 2
91 Median household income
92 Percent enrolled in free or reduced lunch
93 Segregation index
94 Segregation index 2
95 Homicide rate
96 Num deaths 5
97 Suicide rate age adjusted
98 Num firearm fatalities
99 Firearm fatalities rate
100 Juvenile arrest rate
101 Average traffic volume per meter of major roadways
102 Num homeowners
103 Percent homeowners
104 Num households with severe cost burden
105 Percent severe housing cost burden
106 Population 2
107 Percent less than 18 years of age
108 Percent 65 and over
109 Num black
110 Percent black
111 Num American Indian Alaska native
112 Percent American Indian Alaska native
113 Num Asian
114 Percent Asian
115 Num native Hawaiian other Pacific Islander
116 Percent native Hawaiian other Pacific Islander
117 Num Hispanic

118 Percent Hispanic
119 Num Non Hispanic White
120 Percent Non Hispanic White
121 Num not proficient in English
122 Percent not proficient in English
123 Percent female
124 Num rural
125 Percent rural
126 Num housing units
127 Num households CDC
128 Num below poverty
129 Num unemployed CDC
130 Per capita income
131 Num no high school diploma
132 Num age 65 and older
133 Num age 17 and younger
134 Num disabled
135 Num single parent households CDC
136 Num minorities
137 Num limited English abilities
138 Num multi unit housing
139 Num mobile homes
140 Num overcrowding
141 Num households with no vehicle
142 Num institutionalized in group quarters
143 Percent below poverty
144 Percent unemployed CDC
145 Percent no high school diploma
146 Percent age 65 and older
147 Percent age 17 and younger
148 Percent disabled
149 Percent single parent households CDC
150 Percent minorities
151 Percent limited English abilities
152 Percent multi unit housing
153 Percent mobile homes
154 Percent overcrowding
155 Percent no vehicle
156 Percent institutionalized in group quarters
157 Percentile rank below poverty

158	Percentile rank unemployed
159	Percentile rank per capita income
160	Percentile rank no high school diploma
161	Percentile rank socioeconomic theme
162	Percentile rank age 65 and older
163	Percentile rank age 17 and younger
164	Percentile rank disabled
165	Percentile rank single parent households
166	Percentile rank household comp disability theme
167	Percentile rank minorities
168	Percentile rank limited English abilities
169	Percentile rank minority status and language theme
170	Percentile rank multi unit housing
171	Percentile rank mobile homes
172	Percentile rank overcrowding
173	Percentile rank no vehicle
174	Percentile rank institutionalized in group quarters
175	Percentile rank housing and transportation
176	Percentile rank social vulnerability

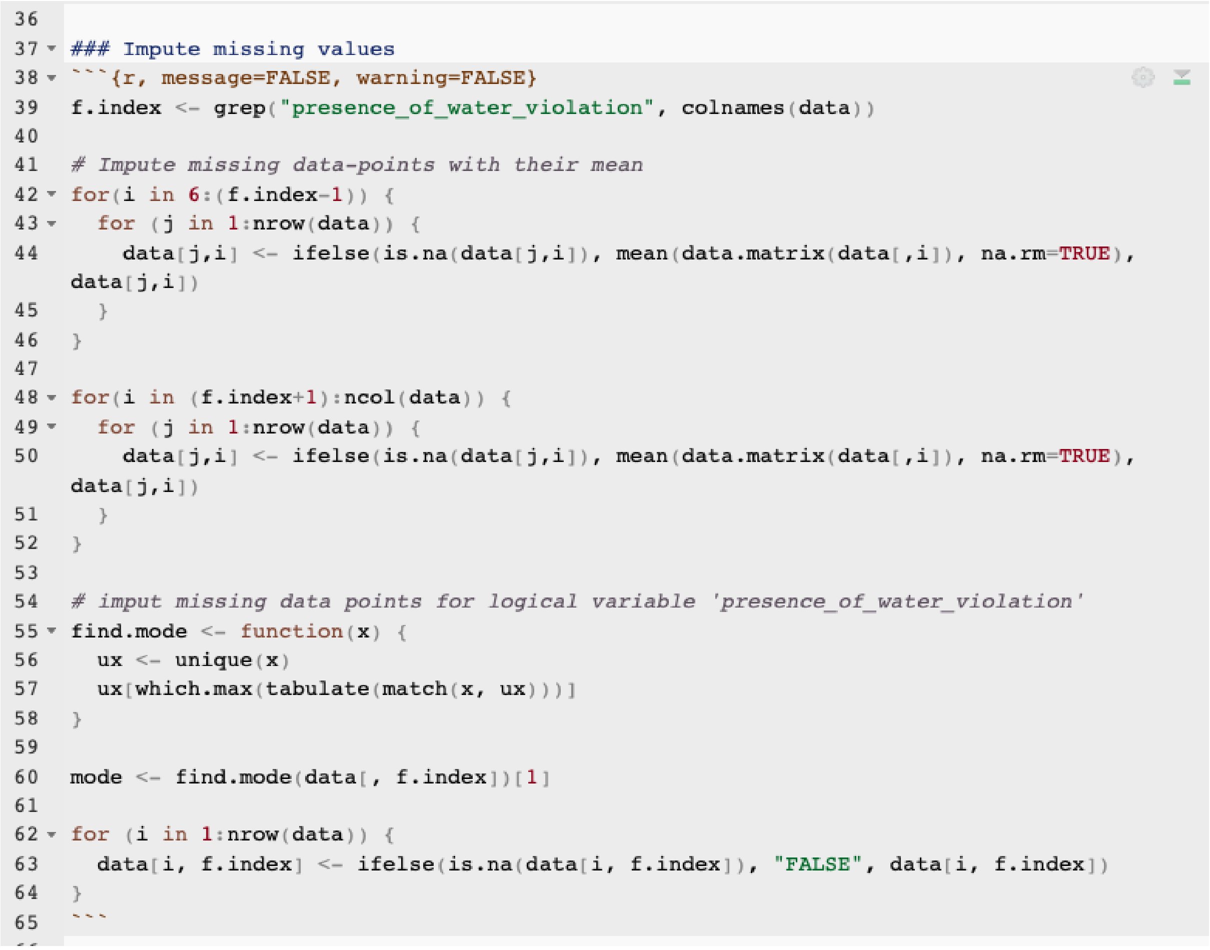
Data Processing Output



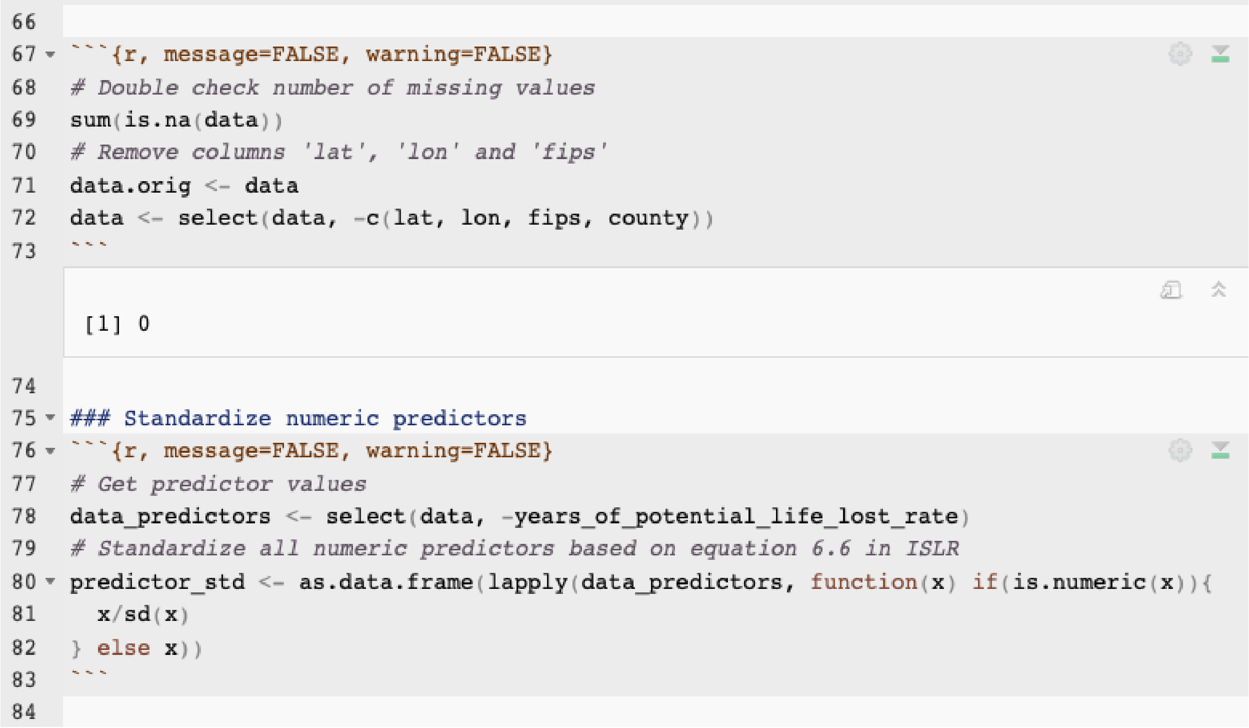
~~~
```{r, message=FALSE, warning=FALSE}
setwd("/Users/Mai/Google Drive/Grad/2020 Spring/STA 9890/STA9890 Project/")
library(tidyverse)
library(dplyr)
library(glmnet)
library("haven")
library(randomForest)
library(gridExtra)
~~~
Loading data

```{r, message=FALSE, warning=FALSE}
data <- read_csv("us_county_sociohealth_data.csv")
```

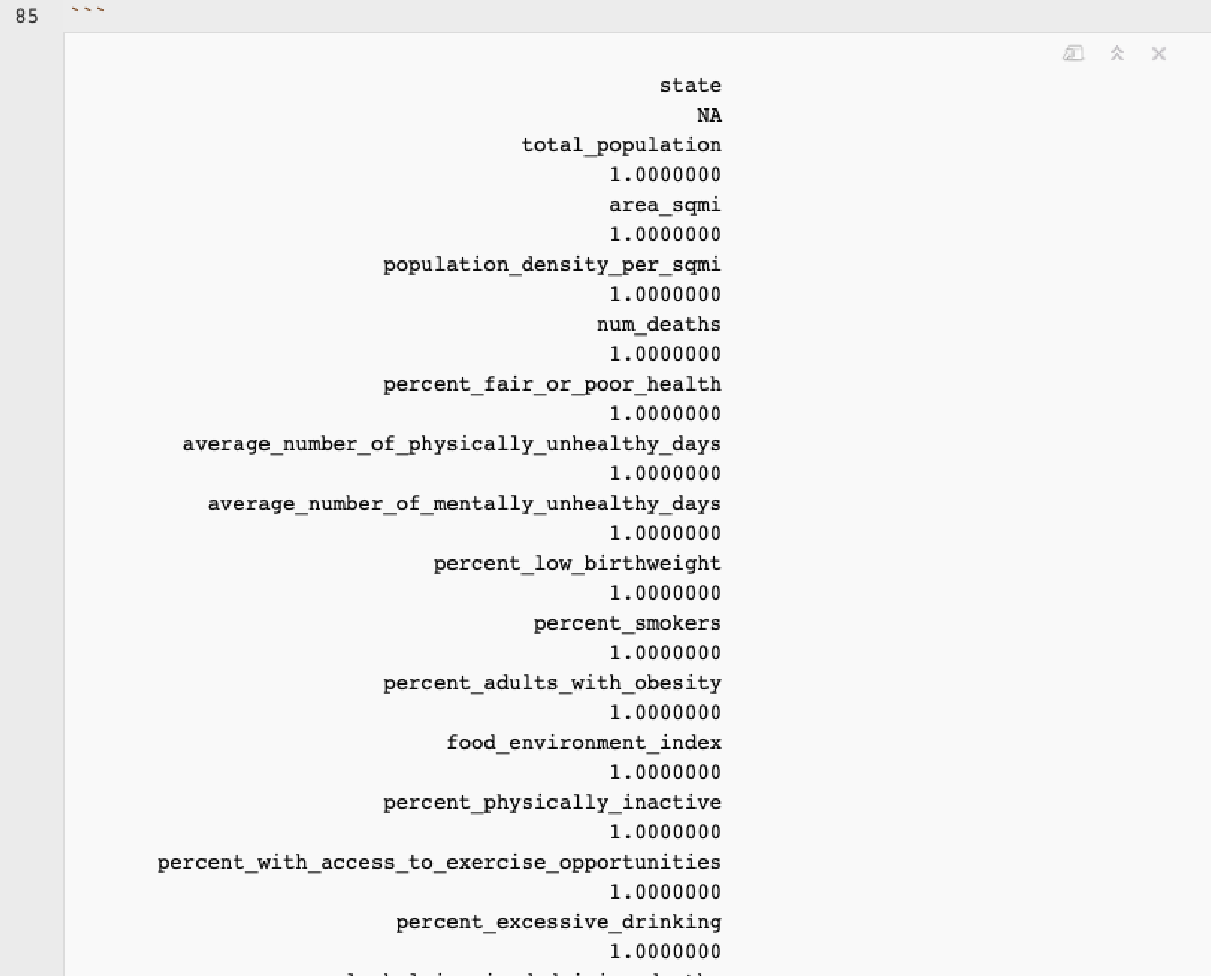
Data Pre-processing
```{r, message=FALSE, warning=FALSE}
# converting categorical variables to factor
data$presence_of_water_violation <- as.factor(data$presence_of_water_violation)
data$state <- as.factor(data$state)
data$county <- as.factor(data$county)
```
```
```



~~~
### Impute missing values
```{r, message=FALSE, warning=FALSE}
f.index <- grep("presence_of_water_violation", colnames(data))

# Impute missing data-points with their mean
for(i in 6:(f.index-1)) {
  for (j in 1:nrow(data)) {
    data[j,i] <- ifelse(is.na(data[j,i]), mean(data.matrix(data[,i]), na.rm=TRUE), data[j,i])
  }
}

for(i in (f.index+1):ncol(data)) {
  for (j in 1:nrow(data)) {
    data[j,i] <- ifelse(is.na(data[j,i]), mean(data.matrix(data[,i]), na.rm=TRUE), data[j,i])
  }
}

# imput missing data points for logical variable 'presence_of_water_violation'
find.mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

mode <- find.mode(data[, f.index])[1]

for (i in 1:nrow(data)) {
  data[i, f.index] <- ifelse(is.na(data[i, f.index]), "FALSE", data[i, f.index])
}
```
~~~



~~~
```{r, message=FALSE, warning=FALSE}
# Double check number of missing values
sum(is.na(data))
# Remove columns 'lat', 'lon' and 'fips'
data.orig <- data
data <- select(data, -c(lat, lon, fips, county))
~~~
```

```
[1] 0
```

```
### Standardize numeric predictors
```{r, message=FALSE, warning=FALSE}
# Get predictor values
data_predictors <- select(data, -years_of_potential_life_lost_rate)
# Standardize all numeric predictors based on equation 6.6 in ISLR
predictor_std <- as.data.frame(lapply(data_predictors, function(x) if(is.numeric(x)){
  x/sd(x)
} else x))
```
```
```



```
```
```

```
                                           state
                                              NA
                                total_population
                                       1.0000000
                                       area_sqmi
                                       1.0000000
                     population_density_per_sqmi
                                       1.0000000
                                      num_deaths
                                       1.0000000
                     percent_fair_or_poor_health
                                       1.0000000
    average_number_of_physically_unhealthy_days
                                       1.0000000
      average_number_of_mentally_unhealthy_days
                                       1.0000000
                         percent_low_birthweight
                                       1.0000000
                                 percent_smokers
                                       1.0000000
                     percent_adults_with_obesity
                                       1.0000000
                          food_environment_index
                                       1.0000000
                     percent_physically_inactive
                                       1.0000000
  percent_with_access_to_exercise_opportunities
                                       1.0000000
                      percent_excessive_drinking
                                       1.0000000
```
```