# DAT 200 – Final Project

Investigating the relationship between death rate, income, and other socioeconomic factors in small cities in America using Linear Regression.

Himasha Karunathilake

# Introduction

In this analysis, I will investigate the relationships between independent variables like doctor availability, hospital availability, income, and population density against the dependent variable, death rate. I decided to use this dataset to see primarily how the effect of income would impact the death rate in these small cities. Income is a well-established social determinant of health so investigating its impact on the death rate can help inform policymakers about the importance of allocating resources more effectively to areas where they are most needed (usually low-income areas) and can lead to more funding in preventative healthcare.

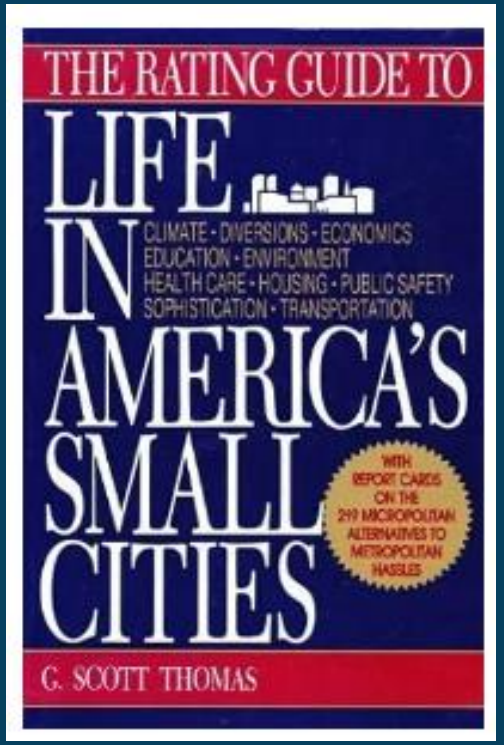
# The Dataset

- Source: The dataset I will be using is from the book "*The Rating Guide to Life in America's Small Cities by G. Scott Thomas*" (survey that rates 219 small cities on livability factors)

- This dataset compiles various health and socio-economic indicators for different cities. Each indicator is represented by a variable
- Independent Variables:
  - doc_avail: doctor availability per 100,000 residents
  - hos_avail: hospital availability per 100,000 residents
  - Income : annual per capita income in thousands of dollars
  - pop_dens: population density people per square mile

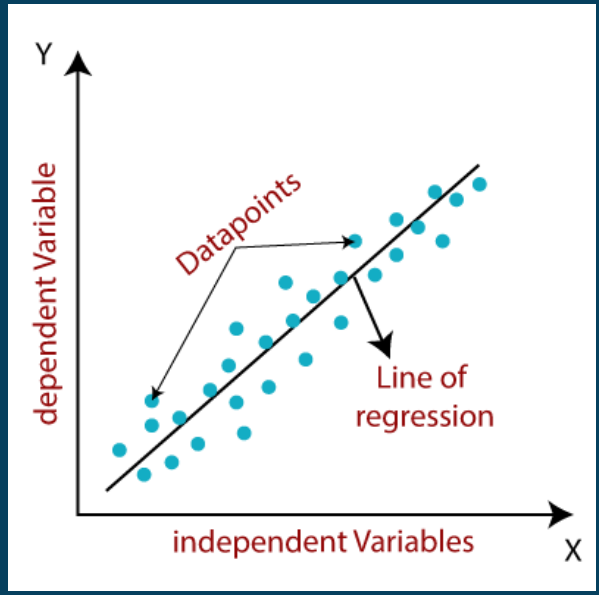
- Dependent Variable:
  - dth_rate : death rate per 100,000 residents

- Sample size: n = 53 entries



# Method of Analysis

I plan to conduct the data analysis utilizing a multiple linear regression model due to the presence of several independent variables. This statistical approach will allow me to explore the relationships and predict the influence of these variables on the dependent variable (death rate). The entire analysis, including the computation and visualization of the results, will be carried out using the R programming environment.



# Setup

- This R code installs and loads the required packages for the analysis. I have decided to rename the variables into names that better represent the data being investigates

deathrate | finalproj.R*

Source on Save | Run

```r
#packages required for analysis
install.packages("broom")
install.packages("ggpubr")
install.packages("ggplot2")
install.packages("dplyr")

#load packages
library(broom)
library(ggpubr)
library(ggplot2)
library(dplyr)

#check that imported data has been read correctly
summary(deathrate)

#change variables to names that better represent the data

#deathrate <- deathrate %>%   #make changes to var names permanent

#rename( dth_rate = V1,       #death rate per 100,000 residents                  -> dependent var
#        doc_avail = V2,      #doctor availability per 100,000 residents         -> independent var
#        hos_avail = V3,      #hospital availability per 100,000 residents       -> independent var
#        income = V4,         #annual per capita income in thousands of dollars -> independent var
#        pop_dens = V5)       #population density people per square mile        -> independent var


summary(deathrate)
```

```
> summary(deathrate)
    dth_rate        doc_avail       hos_avail         income          pop_dens
 Min.   : 3.600   Min.   : 60.0   Min.   : 190.0   Min.   : 7.200   Min.   : 35.0
 1st Qu.: 8.300   1st Qu.: 82.0   1st Qu.: 353.0   1st Qu.: 8.800   1st Qu.: 80.0
 Median : 9.400   Median :114.0   Median : 525.0   Median : 9.500   Median :103.0
 Mean   : 9.306   Mean   :116.1   Mean   : 589.8   Mean   : 9.436   Mean   :110.6
 3rd Qu.:10.300   3rd Qu.:134.0   3rd Qu.: 686.0   3rd Qu.:10.300   3rd Qu.:129.0
 Max.   :12.800   Max.   :238.0   Max.   :1792.0   Max.   :[illegible]000   Max.   :292.0
>
```

Spotify

deathrate | finalproj.R*

Filter

| | dth_rate | doc_avail | hos_avail | income | pop_dens |
|---|---|---|---|---|---|
| 1 | 8.0 | 78 | 284 | 9.1 | 109 |
| 2 | 9.3 | 68 | 433 | 8.7 | 144 |
| 3 | 7.5 | 70 | 739 | 7.2 | 113 |
| 4 | 8.9 | 96 | 1792 | 8.9 | 97 |
| 5 | 10.2 | 74 | 477 | 8.3 | 206 |
| 6 | 8.3 | 111 | 362 | 10.9 | 124 |
| 7 | 8.8 | 77 | 671 | 10.0 | 152 |
| 8 | 8.8 | 168 | 636 | 9.1 | 162 |
| 9 | 10.7 | 82 | 329 | 8.7 | 150 |
| 10 | 11.7 | 89 | 634 | 7.6 | 134 |
| 11 | 8.5 | 149 | 631 | 10.8 | 292 |
| 12 | 8.3 | 60 | 257 | 9.5 | 108 |
| 13 | 8.2 | 96 | 284 | 8.8 | 111 |
| 14 | 7.9 | 83 | 603 | 9.5 | 182 |
| 15 | 10.3 | 130 | 686 | 8.7 | 129 |
| 16 | 7.4 | 145 | 345 | 11.2 | 158 |
| 17 | 9.6 | 112 | 1357 | 9.7 | 186 |
| 18 | 9.3 | 131 | 544 | 9.6 | 177 |
| 19 | 10.6 | 80 | 205 | 9.1 | 127 |
| 20 | 9.7 | 130 | 1264 | 9.2 | 179 |
| 21 | 11.6 | 140 | 688 | 8.3 | 80 |
| 22 | 8.1 | 154 | 354 | 8.4 | 103 |
| 23 | 9.8 | 118 | 1632 | 9.4 | 101 |
| 24 | 7.4 | 94 | 348 | 9.8 | 117 |
| 25 | 9.4 | 119 | 370 | 10.4 | 88 |

- Sample of the dataset (25 entries)

# Assumptions for linear regression

- Linear regression is a statistical method that models the relationship between variables with a straight line. It identifies the optimal line through the dataset by adjusting the regression coefficient(s) to minimize the model's overall discrepancy.
- Linear regression comes in two primary forms:
  - Simple linear regression, which involves a single independent variable to predict the outcome.
  - Multiple linear regression, which incorporates two or more independent variables to explain the dependent variable. (This is the form being used for this dataset)
- Linear regression is based on four key assumptions that are necessary for the best performance and validity of the model's results:
  - 1) Independence of Observations
  - 2) Normality
  - 3) Linearity
  - 4) Homoscedasticity
- I will be checking that the data meets the four assumptions using R.

# #1) Independence of Observations

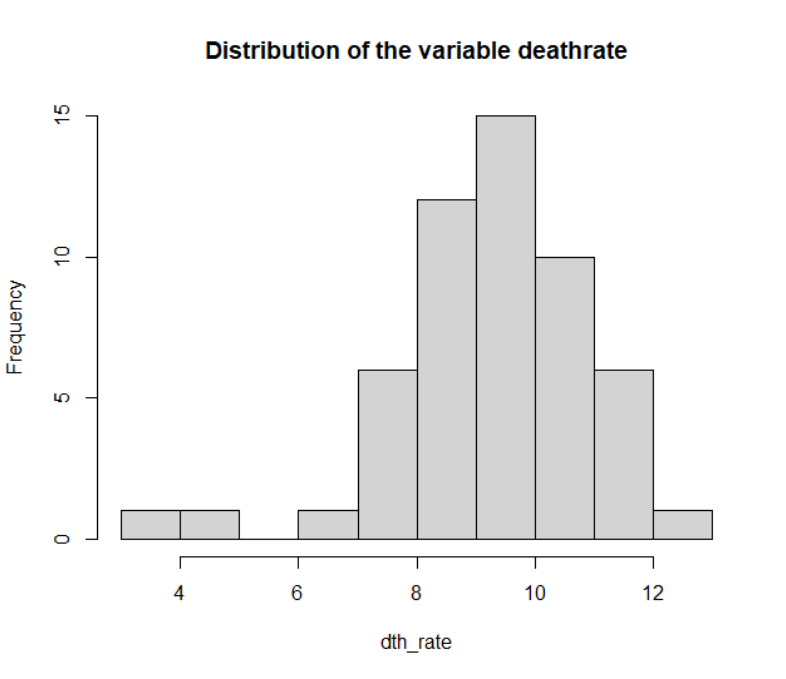
```
# I will be performing a linear regression using this dataset

# --------- Check if the four main assumptions for multiple linear regression are met  ---------

# 1) Independence of observations -> test the correlation between independent variables and check to see that they are not
#   highly correlated.

#check doc_avail against other independent variables
cor(deathrate$doc_avail, deathrate$hos_avail) # -> 0.2956284      - low correlation (between 0.3-0.5)
cor(deathrate$doc_avail, deathrate$income)    # ->  0.433288      - low correlation
cor(deathrate$doc_avail, deathrate$pop_dens)  # -> -0.01993791    - weak negative correlation (< 0.3)

#check hos_avail against other independent variables
cor(deathrate$hos_avail, deathrate$income)    # -> 0.02750354     - weak correlation (< 0.3)
cor(deathrate$hos_avail, deathrate$pop_dens)  # -> 0.1866163      - weak correlation      (< 0.3)

#check income against other independent variable
cor(deathrate$income, deathrate$pop_dens)     # -> 0.1287437      - weak correlation
```

- This code checks the correlation between each independent variable to check that they are not highly correlated
- Low correlation: (doctor availability & death rate), (doctor availability & income)
- Weak correlation: (doctor availability & population density), (hospital availability & income), (income & population density)

```
> cor(deathrate$doc_avail, deathrate$hos_avail) # -> 0.2
[1] 0.2956284
> cor(deathrate$doc_avail, deathrate$income)    # ->  0.
[1] 0.433288
> cor(deathrate$doc_avail, deathrate$pop_dens)  # -> -0.
[1] -0.01993791
> #check hos_avail against other independent variables
> cor(deathrate$hos_avail, deathrate$income)    # -> 0.0
[1] 0.02750354
> cor(deathrate$hos_avail, deathrate$pop_dens)  # -> 0.1
[1] 0.1866163
> #check income against other independent variable
> cor(deathrate$income, deathrate$pop_dens)     # -> 0.1
[1] 0.1287437
```

# #2) Normality

```
# 2) Normality -> Use a histogram to check if the dependent variable (dth_rate) follows a normal distribution

hist(deathrate$dth_rate,
     main = "Distribution of the variable deathrate",
     xlab = "dth_rate")

#distribution is bell-shaped, with only one peak and is roughly symmetric around the mean
# we can conclude that the death rate variable is normally distributed
```




- To check if the dependent variable death rate (dth_rate) follows a normal distribution, I will create a histogram of the variable in R.
- The distribution is bell-shamed, has only one peak, and is roughly symmetric around the mean, so we can conclude that the death rate is normally distributed

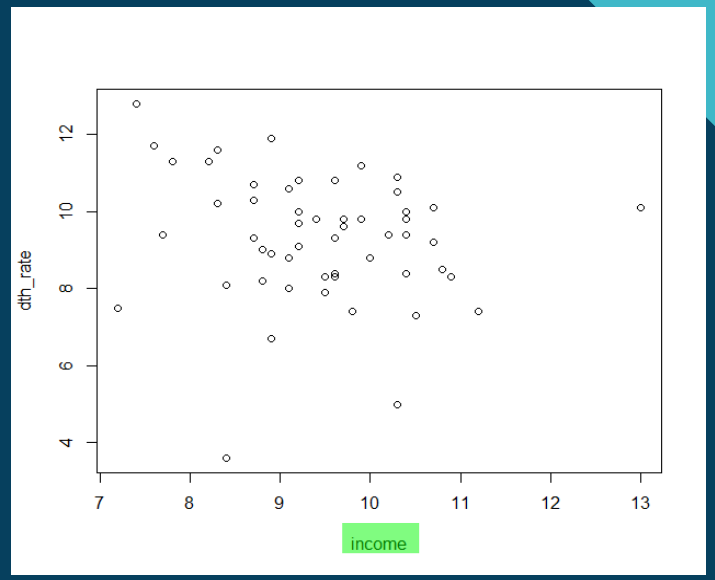
# #3) Linearity 1

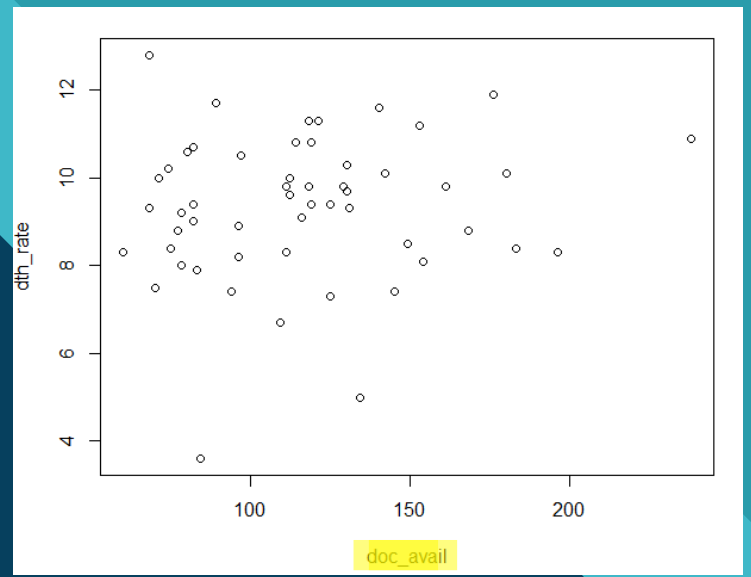
```
# 3) Linearity -> Use scatterplots to test linearity for independent variables against dependent variable

plot(dth_rate ~ income, data=deathrate)
plot(dth_rate ~ pop_dens, data=deathrate)
plot(dth_rate ~ doc_avail, data=deathrate)
plot(dth_rate ~ hos_avail, data=deathrate)

#The variable income appears to a clear linear form
```

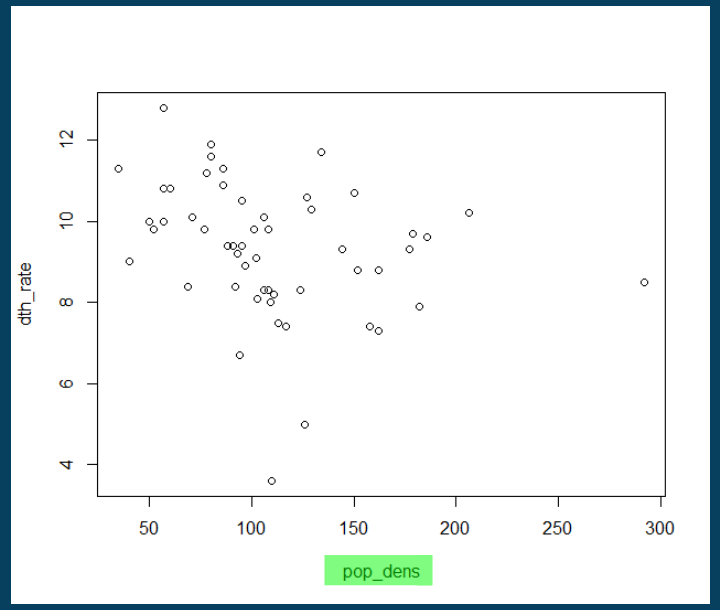
- To check for linearity, I will create scatterplots and see if they show a linear trend.

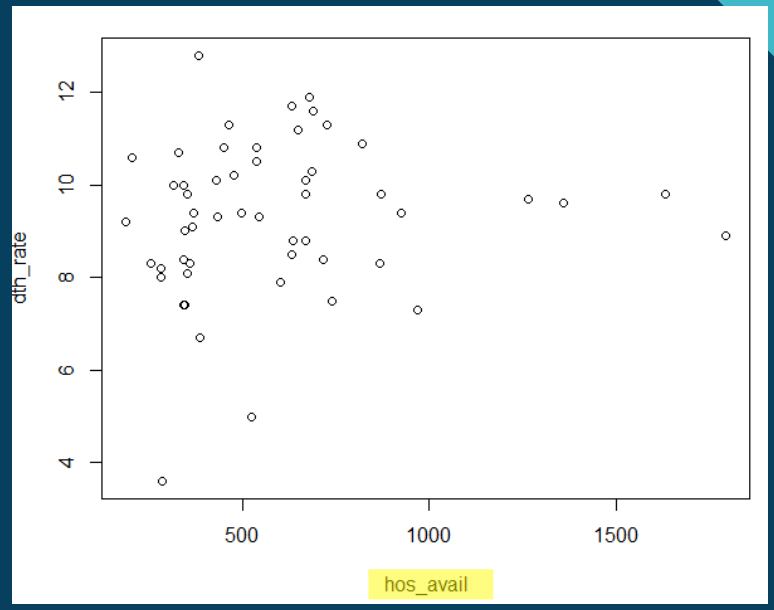
# #3) Linearity 2

- Results:
  - Looking at the scatterplots, it seems like the plots income x death rate, and population density x death rate have linear forms
  - It is more difficult to tell with the variables doctor availability and hospital availability



# Regression Analysis

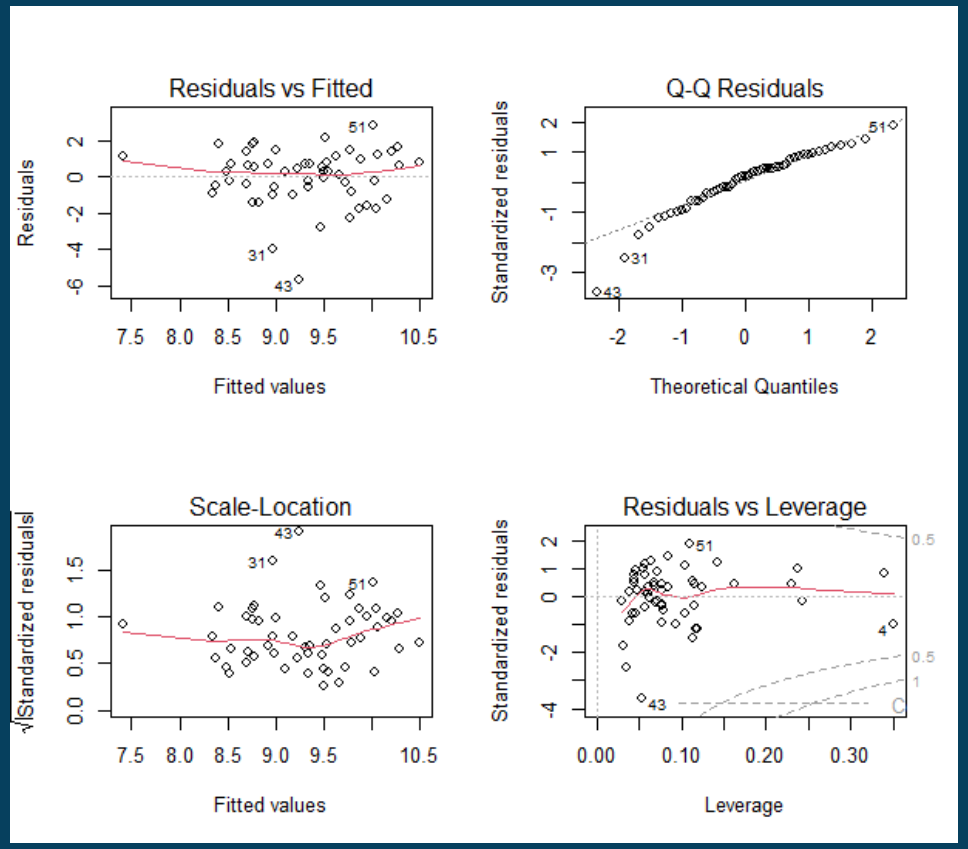
- I will perform a linear regression analysis to evaluate the relationship between the independent variables and the dependent variables
- The results suggest that the estimated effect of income on death rate is -0.33, so for every 1% increase in income (annual per capita income in thousands of dollars), there is a correlated 0.33% decrease in death rate
- However, the p-values are significantly high which means the null hypotheses cannot be rejected (I will perform another multiple regression that has a lower p-value for the visualization)

```
Call:
lm(formula = dth_rate ~ income + pop_dens + doc_avail + hos_avail,
    data = deathrate)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6404 -0.7904  0.3053  0.9164  2.7906

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.2662552  2.0201467   6.072 1.95e-07 ***
income      -0.3302302  0.2345518  -1.408    0.1656
pop_dens    -0.0094629  0.0048868  -1.936    0.0587 .
doc_avail    0.0073916  0.0069336   1.066    0.2917
hos_avail    0.0005837  0.0007219   0.809    0.4228
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.601 on 48 degrees of freedom
Multiple R-squared:  0.1437,    Adjusted R-squared:  0.07235
F-statistic: 2.014 on 4 and 48 DF,  p-value: 0.1075
```

```r
# ---------- Perform Linear Regression Analysis --------------------

#multiple regression
dth_rate.lm<-lm(dth_rate ~ income + pop_dens + doc_avail + hos_avail, data = deathrate)
summary(dth_rate.lm)

# Results of importance:
# - Estimated effect of income on death rate is -0.33, so for every 1% increase in income
# (annual per capita income in thousands of dollars), there is a correlated 0.33% decrease in death rate
#- The other independent variables have much lower percentages: pop_dens: -0.00946, doc_avail: 0.0073916,
# and hos_avail: 0.0005837
```

# #4) Homoscedasticity

```
# 4) Check for homoscedasticity -> check that there is not a large variation in the model error
par(mfrow=c(2,2))
plot(dth_rate.lm)
par(mfrow=c(1,1))
# result: mean of residuals (red line) are horizontal and centered around zero (no biases)
# so, the model fits the assumption of homoscedasticity
```




- This code helps determine that there is not a large variation in the model error
- The mean of residuals (the red lines) are horizontal and centered around zero (there are no biases)
- We can conclude that the model fits the assumption of homoscedasticity

# Visualizing the Linear Regression Model with Graphs

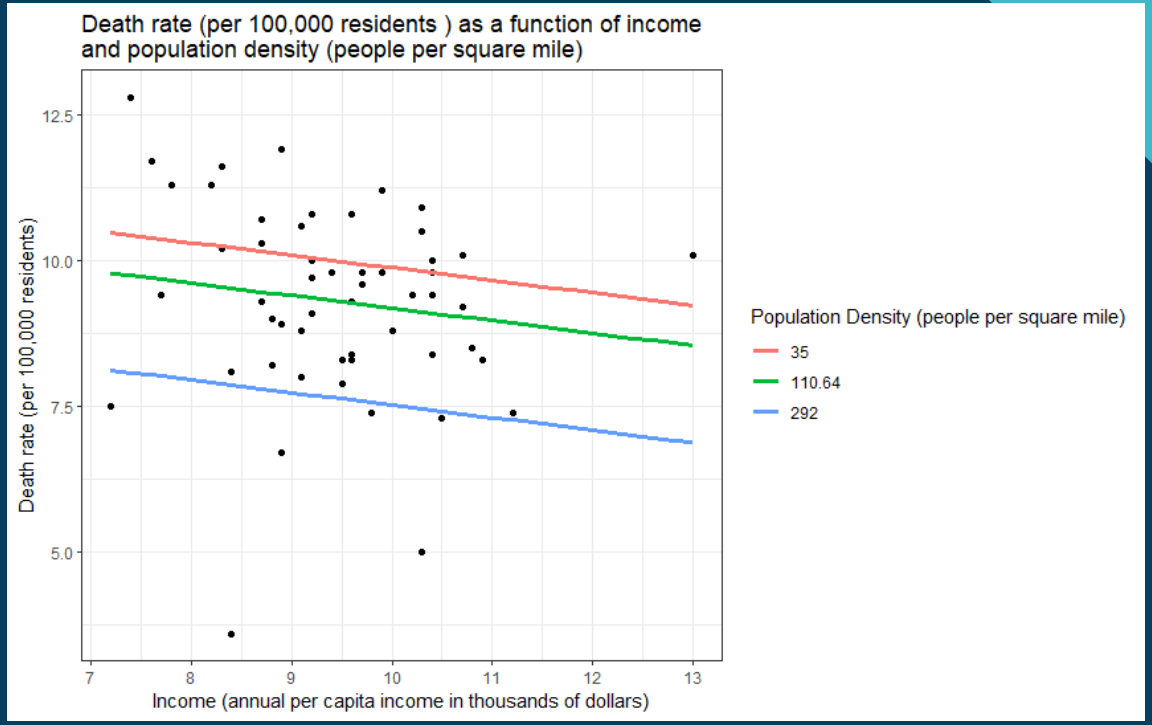
- To visualize the linear regression model, I will be creating a graph
- I will plot the relationship between income and death rate at different levels of population density as an example visualization

```
# ------------------------------------  Visualize the Linear Regression Model with Graphs ------------------------------------
# I will plot an example regression model to visualize the results

#multiple regression
dth_rate2.lm<-lm(dth_rate ~ income + pop_dens , data = deathrate)
summary(dth_rate2.lm)

# ex) Plotting the relationship between income and death rate at different levels of population density

plotting.data<-expand.grid(
  income = seq(min(deathrate$income), max(deathrate$income), length.out=30),
  pop_dens=c(min(deathrate$pop_dens), mean(deathrate$pop_dens), max(deathrate$pop_dens))) #3 levels of pop dense

plotting.data$predicted.y <- predict.lm(dth_rate2.lm, newdata=plotting.data) #predict values of deathrate based on linear model

plotting.data$pop_dens <- round(plotting.data$pop_dens, digits = 2) #round to 2 decimals

plotting.data$pop_dens <- as.factor(plotting.data$pop_dens) #make pop_denn variable into a factor (to plot at each level)

#plot data
dthrt.plot <- ggplot(deathrate, aes(x=income, y=dth_rate)) +
  geom_point()

dthrt.plot

#add regression lines
dthrt.plot <- dthrt.plot +
  geom_line(data=plotting.data, aes(x=income, y=predicted.y, color=pop_dens), linewidth=1.25)

dthrt.plot

dthrt.plot <-
  dthrt.plot +
  theme_bw() +
  labs(title = "Death rate (per 100,000 residents ) as a function of income \nand population density (people per square mile) ",
       x = "Income (annual per capita income in thousands of dollars)",
       y = "Death rate (per 100,000 residents)",
       color = "Population Density (people per square mile)")

dthrt.plot
```

# Conclusion

- In this dataset with 53 entries, there may be a relationship between the income (annual per capita income in thousands of dollars), death rate (per 100,000 residents), and population density (people per square mile)
- Specifically , the results suggest that the estimated effect of income on death rate is -0.21, so for every 1% increase in income (annual per capita income in thousands of dollars), there is a correlated 0.21% decrease in death rate
- The results also suggest that that the estimated effect of population density on death rate is -0.0092, so for every 1% increase in population density (people per square mile), there is a correlated 0.0092 % decrease in death rate

```
Call:
lm(formula = dth_rate ~ income + pop_dens, data = deathrate)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9334 -1.0161  0.0936  1.0659  2.5673

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.339452   1.992835   6.192  1.1e-07 ***
income      -0.214183   0.209616  -1.022    0.3118
pop_dens    -0.009154   0.004778  -1.916    0.0611 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.612 on 50 degrees of freedom
Multiple R-squared:  0.09594,   Adjusted R-squared:  0.05978
F-statistic: 2.653 on 2 and 50 DF,  p-value: 0.08033
```



# Limitations

- One of the constraints encountered in this data analysis was the limited sample size. An expanded dataset with a greater number of observations would likely yield more definitive results.
  - Specifically, the assessment of linearity between variables such as doctor availability, hospital availability, and death rate proved challenging. With access to a larger pool of data for evaluation, the determination of a linear relationship might have been more straightforward and clearer in the scatterplots.
  - Similarly, with a larger dataset, the regression analysis could have revealed a more defined relationship between the death rate and the independent variables.

# Appendix 1

Full source-code (Appendix 1 – 4):

```
#packages required for analysis
install.packages("broom")
install.packages("ggpubr")
install.packages("ggplot2")
install.packages("dplyr")

#load packages
library(broom)
library(ggpubr)
library(ggplot2)
library(dplyr)

#check that imported data has been read correctly
summary(deathrate)

#change variables to names that better represent the data

#deathrate <- deathrate %>%   #make changes to var names permanent

#rename( dth_rate = V1,       #death rate per 100,000 residents                    -> dependent var
#        doc_avail = V2,      #doctor availability per 100,000 residents           -> independent var
#        hos_avail = V3,      #hospital availability per 100,000 residents         -> independent var
#        income = V4,         #annual per capita income in thousands of dollars    -> independent var
#        pop_dens = V5)       #population density people per square mile           -> independent var

summary(deathrate)
```

# Appendix 2

```
# I will be performing a linear regression using this dataset

# --------- Check if the four main assumptions for multiple linear regression are met  ---------

# 1) Independence of observations -> test the correlation between independent variables and check to see that they are not
#   highly correlated.

#check doc_avail against other independent variables
cor(deathrate$doc_avail, deathrate$hos_avail) # -> 0.2956284      - low correlation (between 0.3-0.5)
cor(deathrate$doc_avail, deathrate$income)    # ->  0.433288      - low correlation
cor(deathrate$doc_avail, deathrate$pop_dens)  # -> -0.01993791    - weak negative correlation (< 0.3)

#check hos_avail against other independent variables
cor(deathrate$hos_avail, deathrate$income)    # -> 0.02750354     - weak correlation (< 0.3)
cor(deathrate$hos_avail, deathrate$pop_dens)  # -> 0.1866163      - weak correlation      (< 0.3)

#check income against other independent variable
cor(deathrate$income, deathrate$pop_dens)     # -> 0.1287437      - weak correlation


# 2) Normality -> Use a histogram to check if the dependent variable (dth_rate) follows a normal distribution

hist(deathrate$dth_rate,
     main = "Distribution of the variable deathrate",
     xlab = "dth_rate")

#distribution is bell-shaped, with only one peak and is roughly symmetric around the mean
# we can conclude that the death rate variable is normally distributed

# 3) Linearity -> Use scatterplots to test linearity for independent variables against dependent variable

plot(dth_rate ~ income, data=deathrate)
plot(dth_rate ~ pop_dens, data=deathrate)
plot(dth_rate ~ doc_avail, data=deathrate)
plot(dth_rate ~ hos_avail, data=deathrate)

#The variable income appears to a clear linear form
```

# Appendix 3

```
# ---------- Perform Linear Regression Analysis --------------------

#multiple regression
dth_rate.lm<-lm(dth_rate ~ income + pop_dens + doc_avail + hos_avail, data = deathrate)
summary(dth_rate.lm)

# Results of importance:
# - Estimated effect of income on death rate is -0.33, so for every 1% increase in income
# (annual per capita income in thousands of dollars), there is a correlated 0.33% decrease in death rate
#- The other independent variables have much lower percentages: pop_dens: -0.00946, doc_avail: 0.0073916,
# and hos_avail: 0.0005837


# 4) Check for homoscedasticity -> check that there is not a large variation in the model error
par(mfrow=c(2,2))
plot(dth_rate.lm)
par(mfrow=c(1,1))
# result: mean of residuals (red line) are horizontal and centered around zero (no biases)
# so, the model fits the assumption of homoscedasticity
```

# Appendix 4

```
# ----------------------------------  Visualize the Linear Regression Model with Graphs ----------------------------------
# I will plot an example regression model to visualize the results

#multiple regression
dth_rate2.lm<-lm(dth_rate ~ income + pop_dens , data = deathrate)
summary(dth_rate2.lm)

# ex) Plotting the relationship between income and death rate at different levels of population density

plotting.data<-expand.grid(
  income = seq(min(deathrate$income), max(deathrate$income), length.out=30),
  pop_dens=c(min(deathrate$pop_dens), mean(deathrate$pop_dens), max(deathrate$pop_dens))) #3 levels of pop dense

plotting.data$predicted.y <- predict.lm(dth_rate2.lm, newdata=plotting.data) #predict values of deathrate based on linear model

plotting.data$pop_dens <- round(plotting.data$pop_dens, digits = 2) #round to 2 decimals

plotting.data$pop_dens <- as.factor(plotting.data$pop_dens) #make pop_denn variable into a factor (to plot at each level)

#plot data
dthrt.plot <- ggplot(deathrate, aes(x=income, y=dth_rate)) +
  geom_point()

dthrt.plot

#add regression lines
dthrt.plot <- dthrt.plot +
  geom_line(data=plotting.data, aes(x=income, y=predicted.y, color=pop_dens), linewidth=1.25)

dthrt.plot

dthrt.plot <-
  dthrt.plot +
  theme_bw() +
  labs(title = "Death rate (per 100,000 residents ) as a function of income \nand population density (people per square mile) ",
       x = "Income (annual per capita income in thousands of dollars)",
       y = "Death rate (per 100,000 residents)",
       color = "Population Density (people per square mile)")

dthrt.plot
```