Code

Analyzing the Impact of Pollution on Mortality Rates Across U.S. Cities: An Examination of Sulfur Monoxide, Nitrogen Dioxide, and Confounding Factors

Hide

```
pollution <- read.csv(file.choose(), stringsAsFactors = T)
```

This data set is concerned about the relationship between pollution in the form of sulfur monoxide and nitrogen dioxide on mortality in various United States cities, as well as confounding variables for precipitation, education, and percentage of the population that is non-white.

Hide

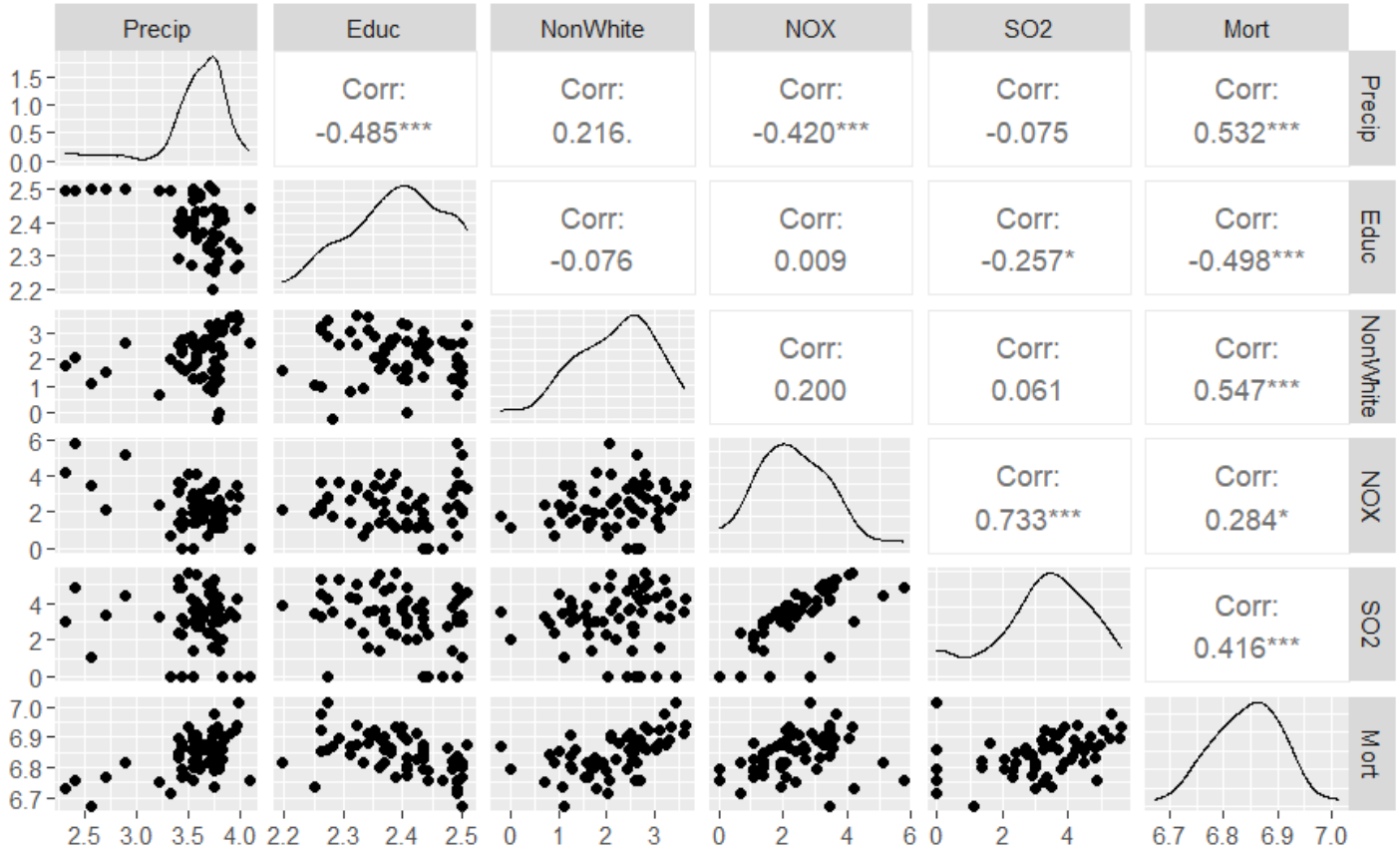
```
summary(pollution)
```

```
            City         Mort            Precip           Educ          NonWhite          NOX
SO2
 Akron, OH     : 1   Min.   : 790.7   Min.   :10.00   Min.   : 9.00   Min.   : 0.80   Min.   :
1.00   Min.   :  1.00
 Albany, NY    : 1   1st Qu.: 898.4   1st Qu.:32.75   1st Qu.:10.40   1st Qu.: 4.95   1st Qu.:
4.00   1st Qu.: 11.00
 Allentown, PA : 1   Median : 943.7   Median :38.00   Median :11.05   Median :10.40   Median :
9.00   Median : 30.00
 Atlanta, GA   : 1   Mean   : 940.4   Mean   :37.37   Mean   :10.97   Mean   :11.87   Mean   : 2
2.65   Mean   : 53.77
 Baltimore, MD : 1   3rd Qu.: 983.2   3rd Qu.:43.25   3rd Qu.:11.50   3rd Qu.:15.65   3rd Qu.: 2
3.75   3rd Qu.: 69.00
 Birmingham, AL: 1   Max.   :1113.1   Max.   :60.00   Max.   :12.30   Max.   :38.50   Max.   :31
9.00   Max.   :278.00
 (Other)       :54
```

In this dataset, we investigate the relationship between pollution, mortality, and various demographic factors across 60 cities in the United States. The mortality rate ranges from 790.7 to 1113.1 people, with an average rate of 940.4. Annual precipitation varies considerably among the cities, with values ranging from 10 to 60 inches and an average of 37.37 inches. The education level, measured by the average years of education completed by residents, also shows some variation across the cities. The minimum average education level is 9 years, while the maximum is 12.3 years, and the mean education level is 10.97 years. The non-white population percentage ranges from 0.8% to 38.5% among the cities, with an average of 11.87%. Regarding pollution, the dataset contains information on nitrogen dioxide (NOX) and sulfur dioxide (SO2) concentrations. Nitrogen dioxide concentrations range from 1 to 319 micrograms per cubic meter, with an average concentration of 22.65 micrograms per cubic meter. Sulfur dioxide concentrations show a range from 1 to 278 micrograms per cubic meter, and an average concentration of 53.77 micrograms per cubic meter.

Hide

```
library(GGally)
ggpairs(log(pollution[,c(3,4,5,6,7,2)]))
```



Since we have two pollution variables, we can create two scatterplots, one for NOX (nitrogen dioxide) and one for SO2 (sulfur monoxide), each with Mortality on the y-axis.

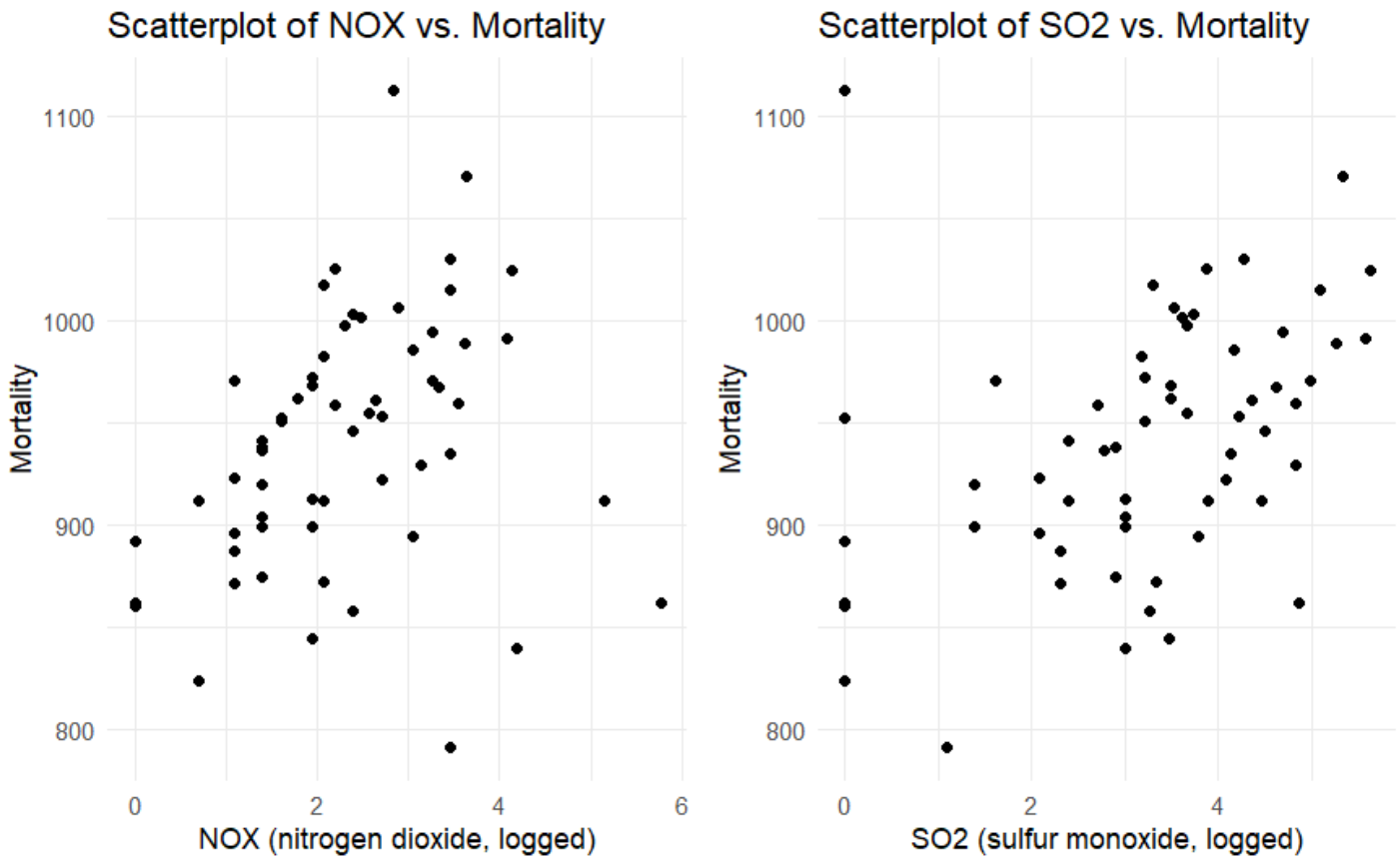
Hide

```
library(ggplot2)
library(gridExtra)

# Scatterplot for NOX vs. Mortality
nox_plot <- ggplot(pollution, aes(x = log(NOX), y = Mort)) +
  geom_point() +
  labs(title = "Scatterplot of NOX vs. Mortality",
       x = "NOX (nitrogen dioxide, logged)",
       y = "Mortality") +
  theme_minimal()

# Scatterplot for SO2 vs. Mortality
so2_plot <- ggplot(pollution, aes(x = log(SO2), y = Mort)) +
  geom_point() +
  labs(title = "Scatterplot of SO2 vs. Mortality",
       x = "SO2 (sulfur monoxide, logged)",
       y = "Mortality") +
  theme_minimal()

# Display the scatterplots side by side
grid.arrange(nox_plot, so2_plot, ncol = 2)
```



After looking at different variations of the variables, whether that was inverting them, square rooting them, and logging them, it seems that logging the data makes for the most linear scatter plot. Therefore, we will use log(S02) and log(NOX) for our analysis.

Hide

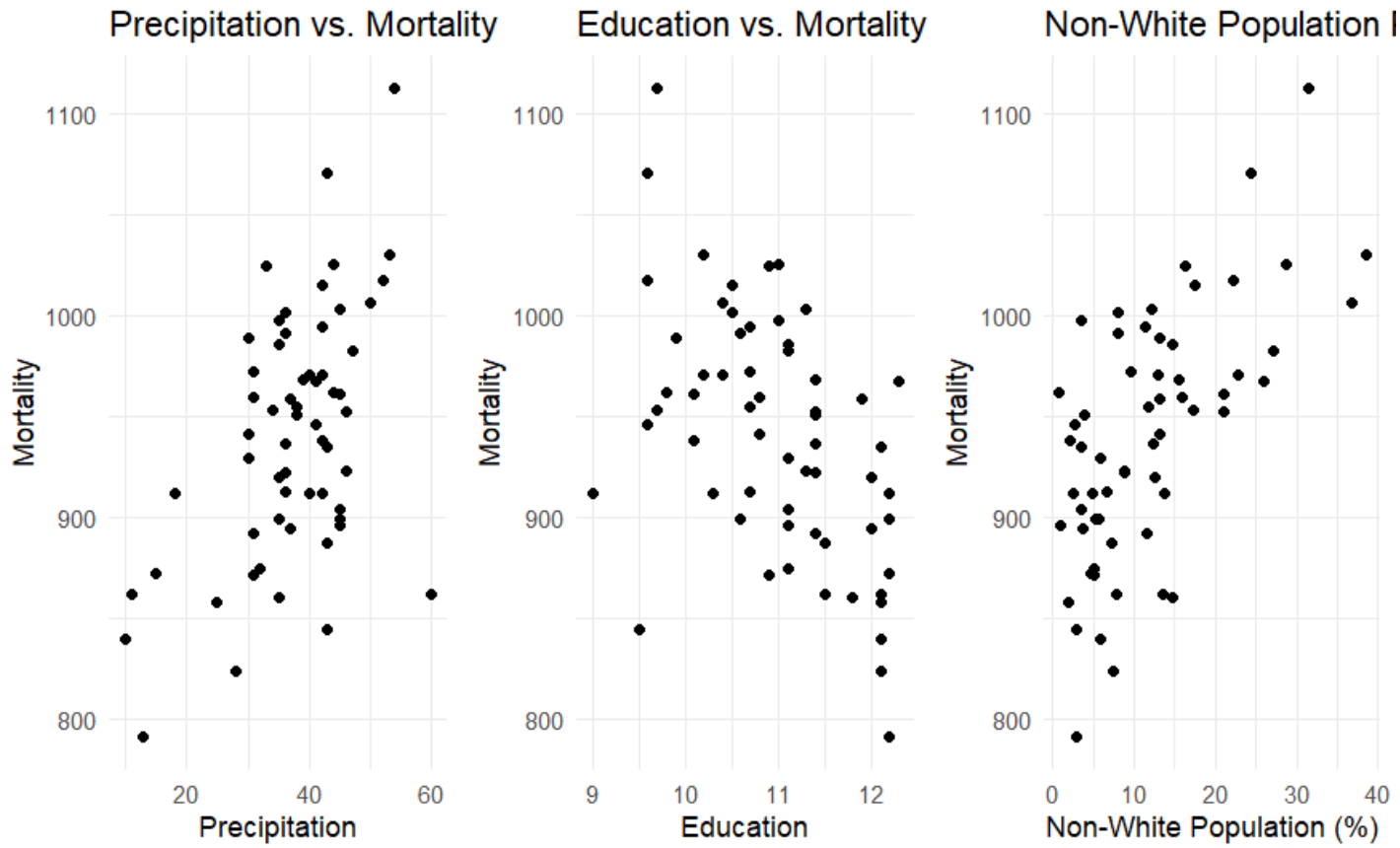
```
library(ggplot2)
library(gridExtra)

# Scatter plot for Precip vs. Mortality
precip_plot <- ggplot(pollution, aes(x = Precip, y = Mort)) +
  geom_point() +
  labs(title = "Precipitation vs. Mortality",
       x = "Precipitation",
       y = "Mortality") +
  theme_minimal()

# Scatter plot for Educ vs. Mortality
educ_plot <- ggplot(pollution, aes(x = Educ, y = Mort)) +
  geom_point() +
  labs(title = "Education vs. Mortality",
       x = "Education",
       y = "Mortality") +
  theme_minimal()

# Scatter plot for NonWhite vs. Mortality
nonwhite_plot <- ggplot(pollution, aes(x = NonWhite, y = Mort)) +
  geom_point() +
  labs(title = "Non-White Population Percentage vs. Mortality",
       x = "Non-White Population (%)",
       y = "Mortality") +
  theme_minimal()

# Display the scatter plots in a grid
grid.arrange(precip_plot, educ_plot, nonwhite_plot, ncol = 3)
```



We did the same thing with the confounding variables Precip, Educ, and NonWhite. We found that it is best to use the normal versions of Precip, Educ, and NonWhite since they are all relatively linear.

Hide

```
pollutionmodel.lm <- lm(Mort ~ Precip + Educ + NonWhite * log(SO2) + log(NOX), pollution)
summary(pollutionmodel.lm)
```

```
Call:
lm(formula = Mort ~ Precip + Educ + NonWhite * log(SO2) + log(NOX),
    data = pollution)

Residuals:
    Min      1Q  Median      3Q     Max
-98.720 -20.097   0.758  18.904  75.157

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       844.4558    91.5839   9.221 1.33e-12 ***
Precip              1.5043     0.6599   2.280 0.026687 *
Educ              -10.1222     6.5444  -1.547 0.127888
NonWhite            7.7807     1.5853   4.908 9.17e-06 ***
log(SO2)           36.3880     9.1113   3.994 0.000202 ***
log(NOX)           -2.8442     7.4251  -0.383 0.703209
NonWhite:log(SO2)  -1.3283     0.4083  -3.253 0.001989 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.45 on 53 degrees of freedom
Multiple R-squared:  0.7402,    Adjusted R-squared:  0.7107
F-statistic: 25.16 on 6 and 53 DF,  p-value: 6.802e-14
```

Hide

```
pollutionmodel2.lm <- lm(Mort ~ Precip + Educ + NonWhite * log(SO2) + log(NOX), pollution[(-6
0),])
summary(pollutionmodel2.lm)
```

```
Call:
lm(formula = Mort ~ Precip + Educ + NonWhite * log(SO2) + log(NOX),
    data = pollution[(-60), ])

Residuals:
    Min      1Q  Median      3Q     Max
-93.370 -21.661   1.464  18.919  74.637

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       826.8268    87.4165   9.458 6.86e-13 ***
Precip              1.2867     0.6337   2.031 0.047423 *
Educ               -5.5740     6.4762  -0.861 0.393363
NonWhite            5.3045     1.7926   2.959 0.004636 **
log(SO2)           34.5926     8.6980   3.977 0.000217 ***
log(NOX)          -10.6300     7.6934  -1.382 0.172967
NonWhite:log(SO2)  -0.6338     0.4740  -1.337 0.187054
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.83 on 52 degrees of freedom
Multiple R-squared:  0.7338,    Adjusted R-squared:  0.7031
F-statistic: 23.89 on 6 and 52 DF,  p-value: 2.373e-13
```

In this linear model, we are analyzing the relationship between pollution in the form of sulfur monoxide (SO2) and nitrogen dioxide (NOX) on mortality in various United States cities, considering confounding variables such as precipitation, education, and the percentage of the non-white population. We decided to remove outlier 60 (New Orleans) because of its abnormally high cook’s distance, residuals, and leverage. Additionally, New Orleans had an abnormally high percentage of nonwhite citizens according to the mean, a considerably lower S02 level than the mean, and the highest mortality rate in the data set. We also see that all the confounding variables were added in the model. We decided to control for these variables because we know each one has an effect on mortality, so including them helps us solely focus on the effect pollution has on mortality rates in the U.S. We also decided to make an interaction with NonWhite and log(SO2). The primary question we seek to answer is whether mortality is related to these two forms of pollution, using a statistical significance threshold of $p < 0.01$. The Multiple R-squared value is 0.7338, and the Adjusted R-squared value is 0.7031. These values indicate that the model explains approximately 73.38% of the variability in mortality, with 70.31% being adjusted for the number of predictors. The F-statistic is 23.89 with a p-value of 2.373e-13, suggesting that the model is statistically significant as a whole. The coefficient of Precip has an estimated value of 1.2867 and a p-value of .047423, which demonstrates that the relationship between precipitation and mortality rates is not statistically significant. The Educ coefficient has an estimated value of -5.5740 and a p-value of .393363, which shows that the relationship between years of education and mortality rate is not significant. The NonWhite coefficient has a estimated value of 5.3045 and a p-value of .004636, which demonstrates the relationship between the percentage of non-white people in a particular city is related to mortality rates. The estimated value for the log(SO2) variable is 34.5926 and has a p-value of 0.000217, which tell us that the relationship between SO2 and mortality is statistically significant. In contrast, the coefficient of log(NOX) has an estimate of -10.6300 and a p-value of 0.172967, which shows that the relationship between NOX and mortality is not statistically significant. Based on this analysis, we can conclude that mortality is related to sulfur monoxide (SO2) pollution, as its relationship with mortality is statistically significant. However, we cannot establish a statistically significant relationship between mortality and nitrogen dioxide (NOX) pollution at the given threshold ($p < .01$).

Hide

```
par(mfrow = c(2,2))
plot(pollutionmodel2.lm, c(1,2,4,5))
```