PSTAT 126 Project

Joseph Chang

December 8th, 2021

Background

By now it is widely recognized that air quality impacts health, but this was not always the case. The file `pollution.csv` contains data from an early observational study investigating the relationship between specific pollutants and mortality in U.S. cities. Variable descriptions and units are recorded in the metadata file `pollution-metadata.csv`. All measurements were taken for the period 1959 - 1961.

> McDonald, G.C. and Schwing, R.C. (1973). Instabilities of Regression Estimates Relating Air Pollution to Mortality. Technometrics, 15: 463-481.

Data Importation

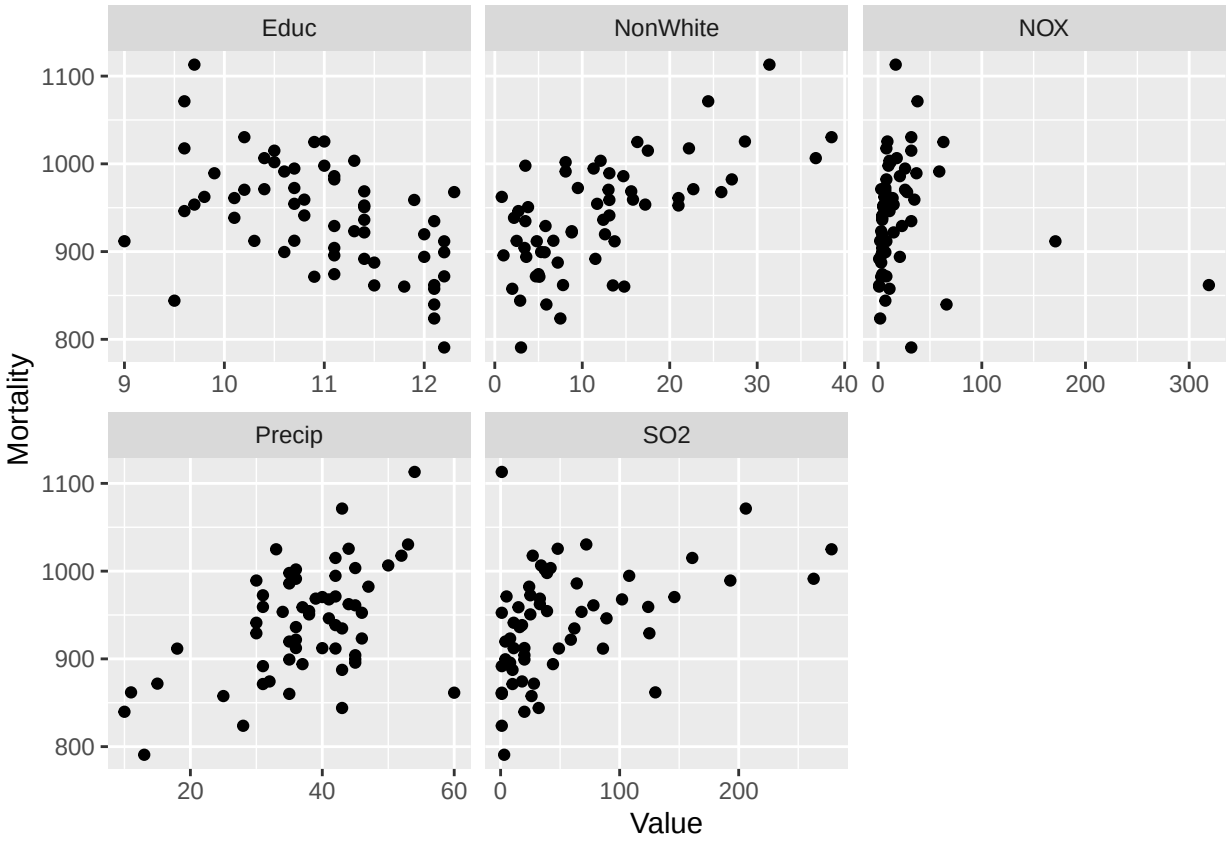
This shows the data's first 5 rows (out of 60).

```
## # A tibble: 5 x 7
##   City             Mort Precip  Educ NonWhite   NOX   SO2
##   <chr>           <dbl>  <dbl> <dbl>    <dbl> <dbl> <dbl>
## 1 San Jose, CA     791.     13  12.2      3      32     3
## 2 Wichita, KS      824.     28  12.1      7.5     2     1
## 3 San Diego, CA    840.     10  12.1      5.9    66    20
## 4 Lancaster, PA    844.     43   9.5      2.9     7    32
## 5 Minneapolis, MN  858.     25  12.1      2      11    26
```

In this data the presence of pollutants is reported as *relative pollution potential*, which is calculated by scaling emissions (tons per day per square kilometer) by a dispersion factor based on local conditions (mixing, wind, area, and the like).

Plot of relationships

I will construct a plot of the marginal relationships among the raw data



Based from the graph, it appears that Precip and NonWhite have a linear trend. Educ appears to be in a negative linear trend, while NOX has no noticeable or immediate trend. SO2 appears to be in a log pattern, so a log transformation may be needed.

Association between mortality and other variables

Here is where I estimate the association between mortality and each of the two pollutants.

```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + NOX + log(SO2),
##     data = pollution)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.988 -19.940   1.057  17.499 115.431
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 944.17471   94.17758  10.025 6.25e-14 ***
## Precip        1.42242    0.68221   2.085   0.0418 *
## Educ        -13.35817    6.95943  -1.919   0.0602 .
## NonWhite      3.43512    0.60693   5.660 5.95e-07 ***
## NOX          -0.09677    0.12989  -0.745   0.4595
## log(SO2)     15.96124    3.67450   4.344 6.21e-05 ***
```

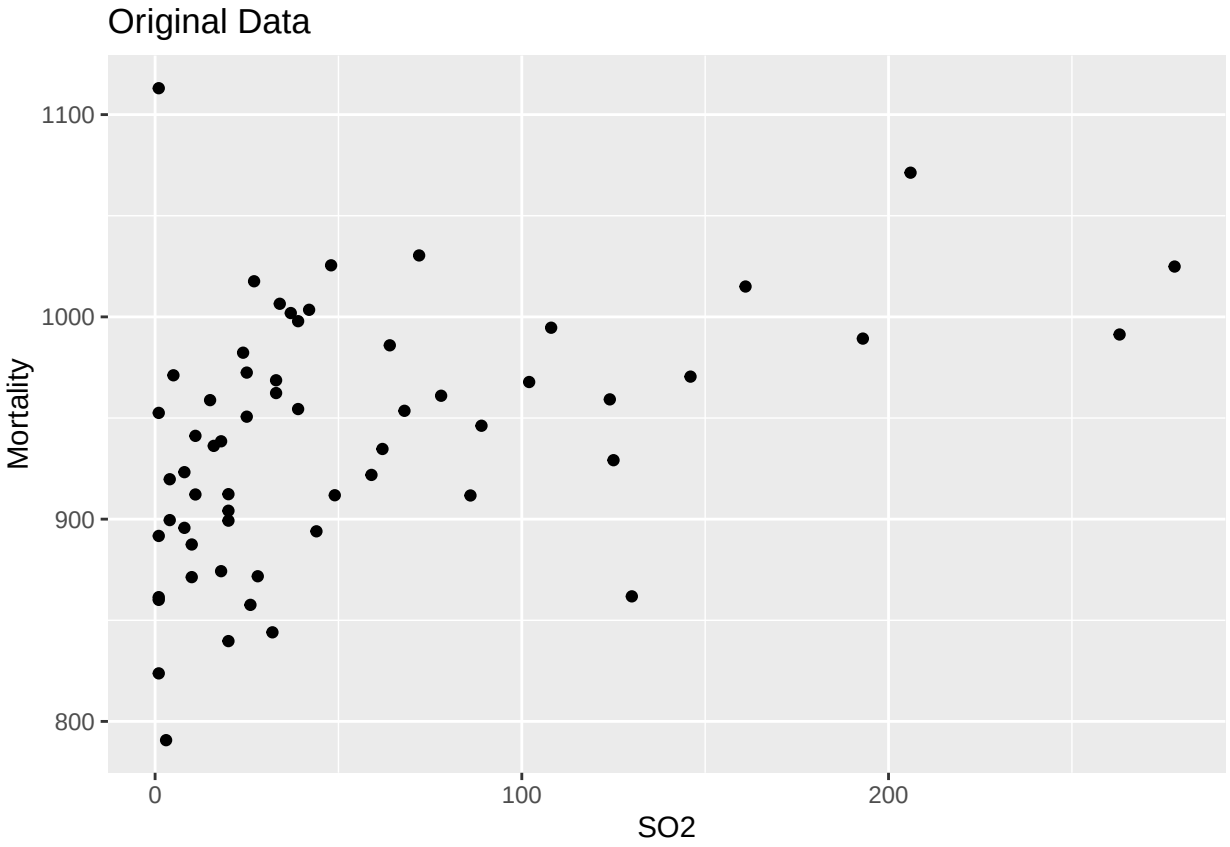
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.39 on 54 degrees of freedom
## Multiple R-squared:  0.6867, Adjusted R-squared:  0.6577
## F-statistic: 23.68 on 5 and 54 DF,  p-value: 1.615e-12
```

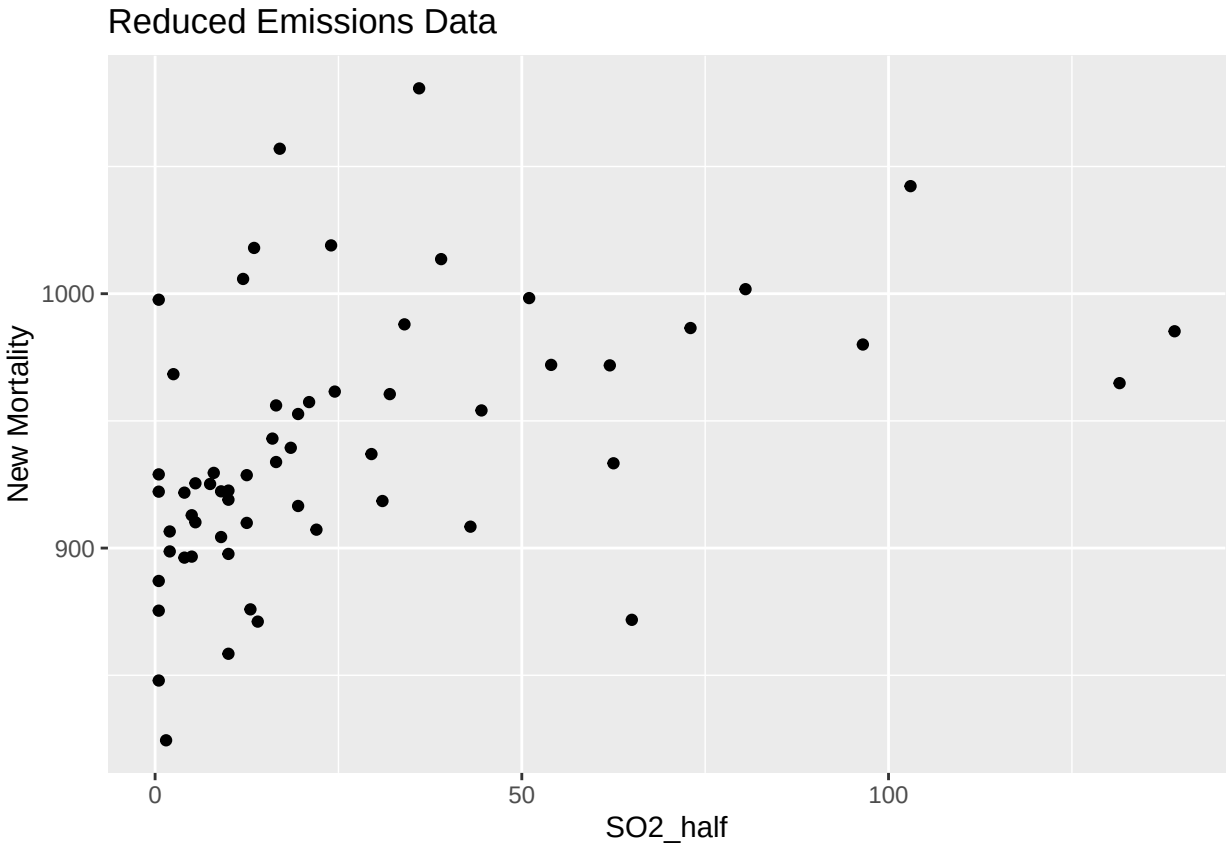
Using a fit model, I can find the association between mortality and the two pollutants by using the summary of fit and locating the estimate portion of the summary. For every increase in 1 ton per day per km of NOX, mortality decreases by 0.09677. For every increase of 1 ton per day per km of log(SO2), mortality increases by 15.96124.

Estimation of potential lives saved

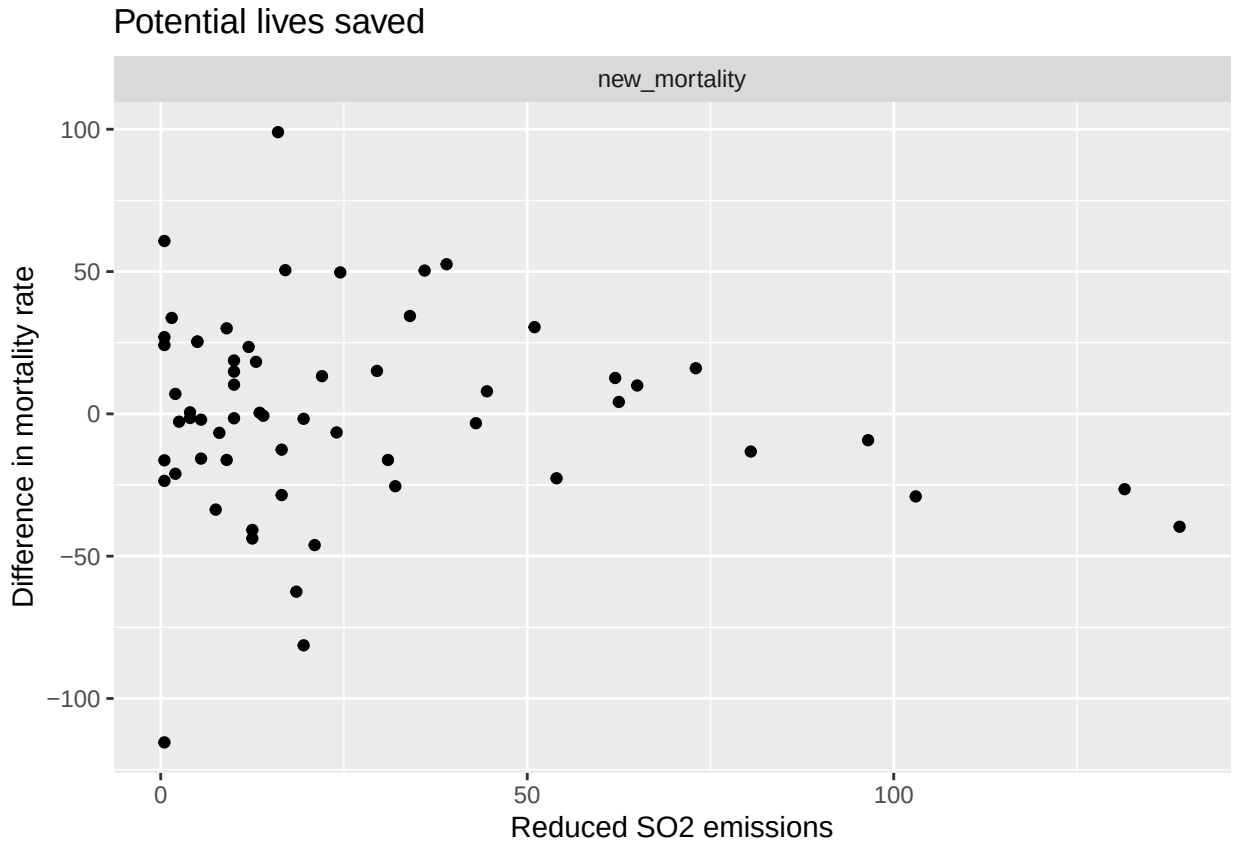
If SO2 emissions were reduced by half, I can estimate the potential lives saved and visualize it on a plot

Here is a comparison of Mortality with SO2 reduced and without SO2 reduced





Here I will calculate the difference in mortality and see the the two graphs generated above



Note: negative difference means death rate decreased. Less people died and more people lived.

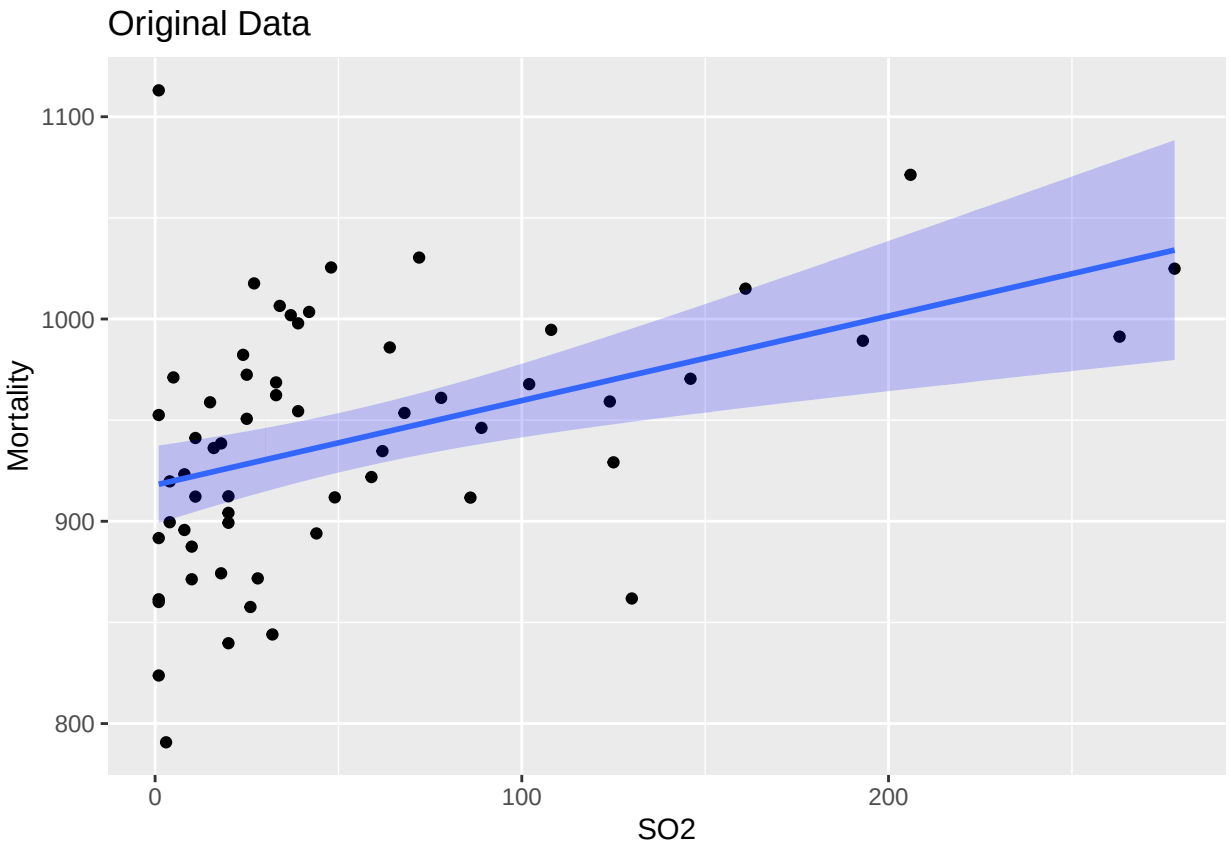
Based from the graph, it appears some city's new mortality rate did not decrease, but rather increased. Overall, it seems like there were more city that had a decreased mortality rate than that of increased. Therefore, one could assume that reduced SO2 emissions does save lives. There does not seem to be any noticeable patterns.

Hypothetical Set-up:

The EPA reports a 94% decrease in the national average sulfur dioxide concentration between 1980 and 2020.

```
##                       2.5 %       97.5 %
## (Intercept)     808.7196674 1169.4408142
## Precip            0.0546692    2.7901617
## Educ            -27.3109846    0.5946529
## NonWhite          2.2183023    4.6519378
## NOX              -0.3571941    0.1636493
## log(0.06 * SO2)   8.5943055   23.3281827
```

```
##        fit      lwr      upr
## 1 939.7604 929.2136 950.3072
```



original model mort_0 = 944.17471 + 1.432242(percip) - 13.35817(Educ) + 3.43512(NonWhite) - 0.09677(NOX) + 15.96124(SO2) + ϵ_i.

new model with 94% decrease in SO2 emissions mort_1 = 944.17471 + 1.432242(percip) - 13.35817(Educ) + 3.43512(NonWhite) - 0.09677(NOX) + 15.96124 * log(SO2 * (1-0.94)) + ϵ_i.

difference between mort_0 and mort_1 mort_0 - mort_1 = 15.96log(SO2) - 15.96log(0.06*SO2) = 15.96log(SO2) - (15.96log(0.06) + 15.96log(SO2)) = 15.96log(0.06) = 19.5 .

```
There is about 19.5 lives saved.
```

Based from the confidence interval, a one percentage increase in mortality is significantly associated with a change in 8.5943055 to 23.3281827 increase in SO2. Therefore, an estimate of 19.5 is within the confidence interval calculated above.

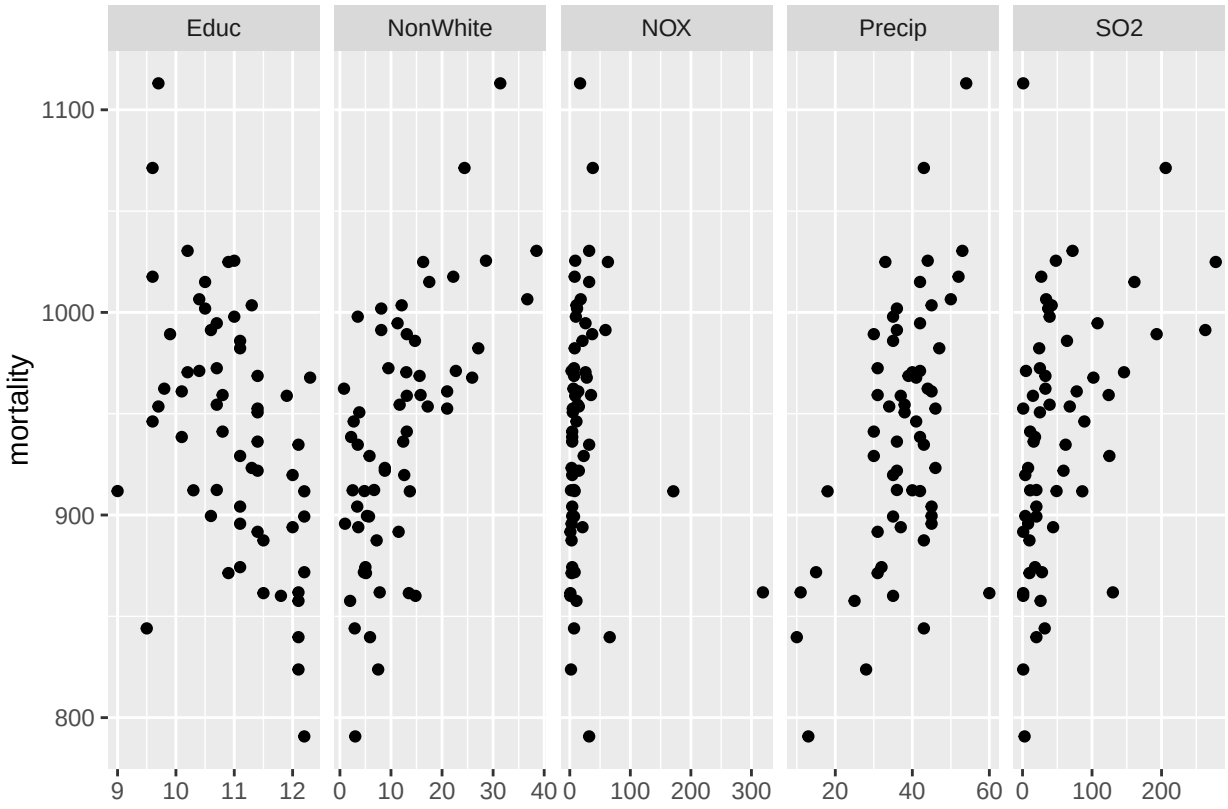
Note: Assumptions needed about this data One assumption is that the trend of Precip, Educ, NonWhite, NOX, and SO2 should all display similar trends in the years 1980-2020 as it was from 1959 to 1961. In other words, the pattern should hold across time.

Another assumption is same trend for different location. Since each city represents its own data, and that the EPA's data is the national average, the trend of Precip, Educ, NonWhite, NOX, and SO2 should hold same for each of the two datas.

Another assumption is same measurement condition and the same collection of data across time and space. The accuracy should be similar. From 1980-2020, there may be better collection of data due to technological advances. Thus, accuracy of measurement should be similar.

These assumptions are reasonable because it is difficult to collect data from every metropolitan area. Additionally, collecting these data may be different because of technological advances as time goes on. Therefore, we must rely on assumptions in order for the data to be as close to error-free as possible.

Association with mortality




```
##
## Call:
## lm(formula = Mort ~ Precip + Educ + NonWhite + NOX + log(SO2),
##     data = pollution)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.988 -19.940   1.057  17.499 115.431
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 944.17471   94.17758  10.025 6.25e-14 ***
## Precip        1.42242    0.68221   2.085   0.0418 *
## Educ        -13.35817    6.95943  -1.919   0.0602 .
## NonWhite      3.43512    0.60693   5.660 5.95e-07 ***
```

```
## NOX          -0.09677    0.12989  -0.745   0.4595
## log(SO2)     15.96124    3.67450   4.344 6.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.39 on 54 degrees of freedom
## Multiple R-squared:  0.6867, Adjusted R-squared:  0.6577
## F-statistic: 23.68 on 5 and 54 DF,  p-value: 1.615e-12
```

Using hypothesis testing, we can check if there are variables associated with mortality. We set the following:

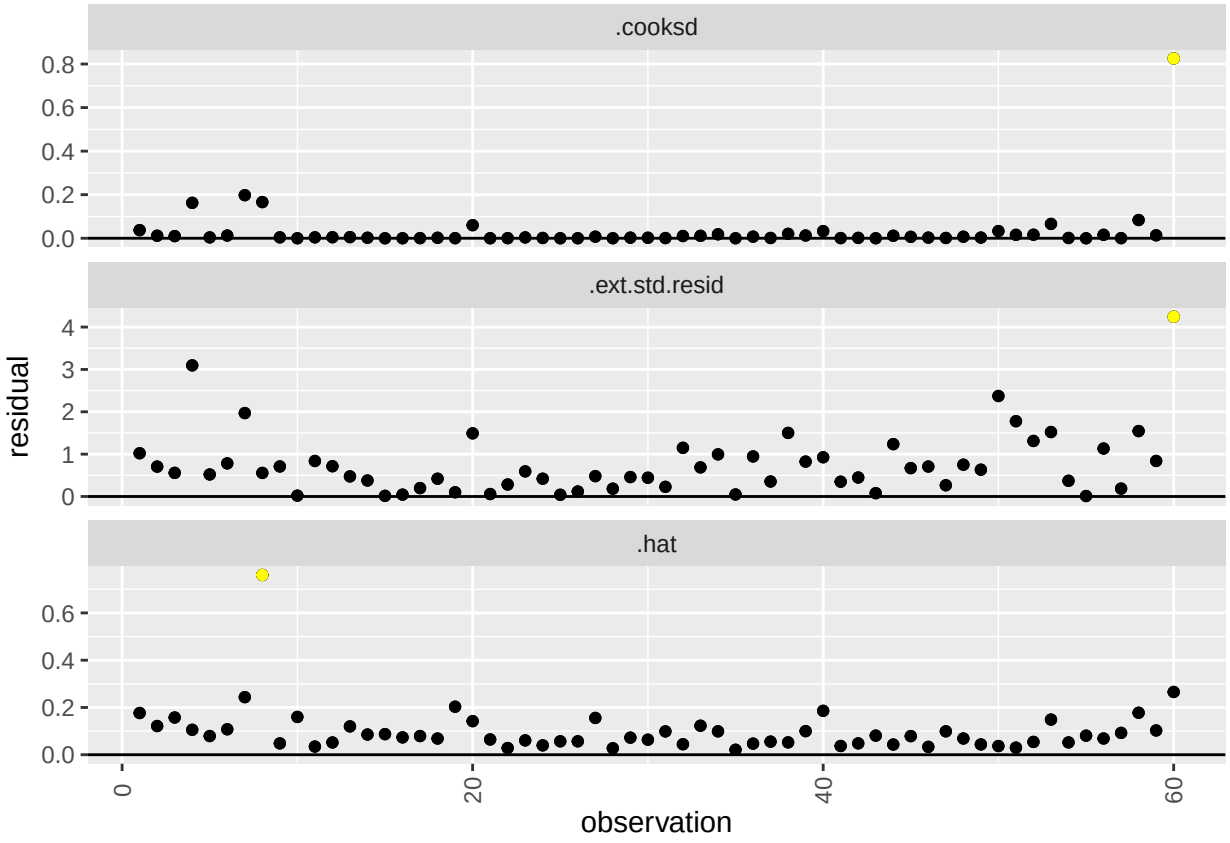
β_1 = Precip. β_2 = Educ. β_3 = NonWhite. $\beta 4$ = NOX. $\beta 5$ = SO2.

$$Ho : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$Ha : \text{at least one is not } 0$$

Based on the p-values from the hypothesis testing, we see that precip has 0.0418, Educ has 0.0602, nonWhite has 5.95e-7, NOX has 0.4595, and SO2 has 6.21e-5. With alpha set as 0.05, we can see that only the p-values for NOX and Educ are greater than alpha. As a result, we will fail to reject NOX and Educ, and reject all others, concluding that only Precip, NonWhite, and SO2 are statistically significant and have an association with mortality. NonWhite and Precip seem to have a linear association, while SO2 has a log pattern.

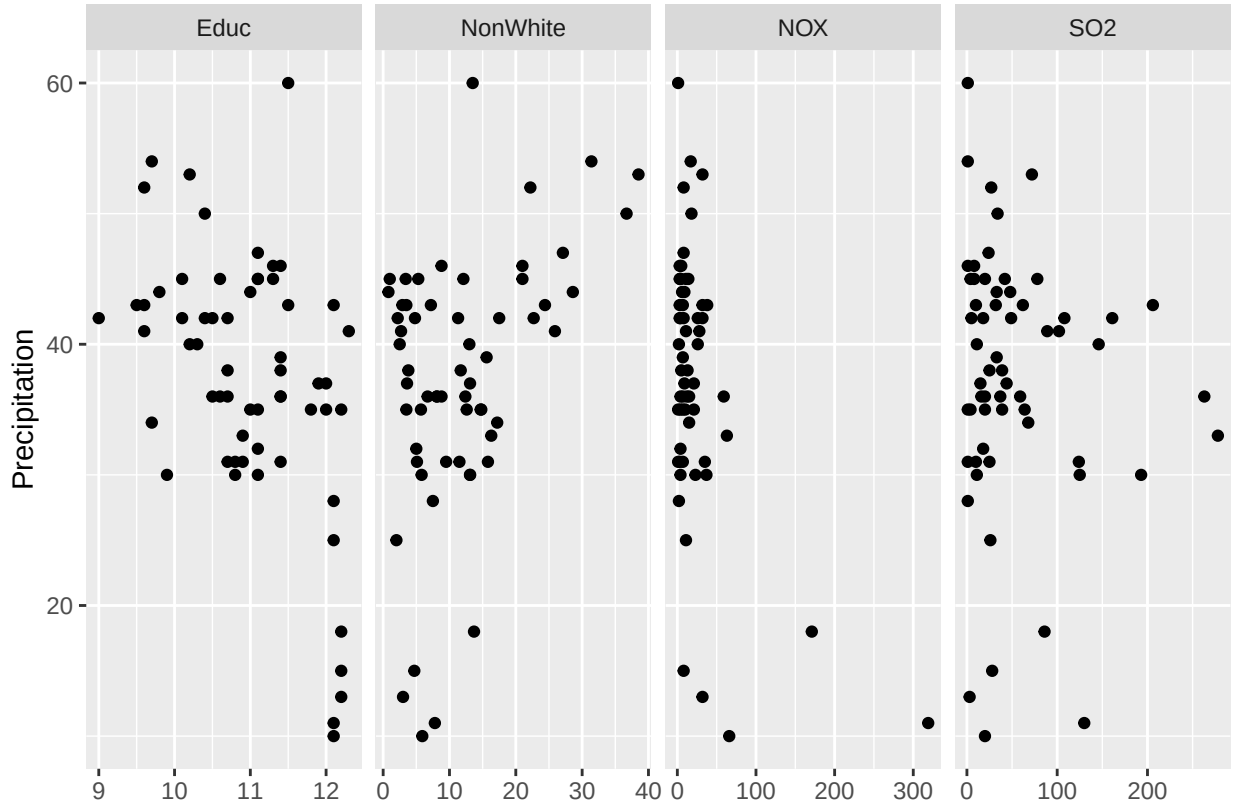
Check for Unusual points

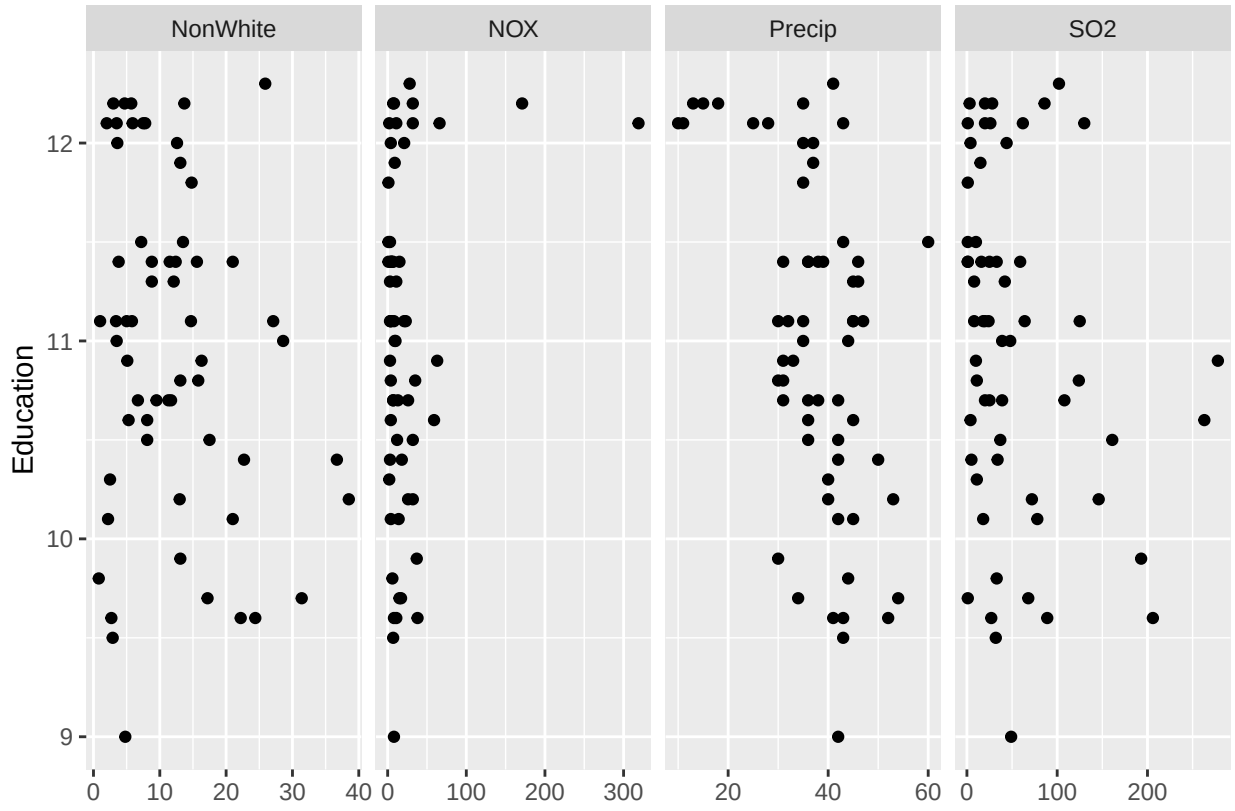


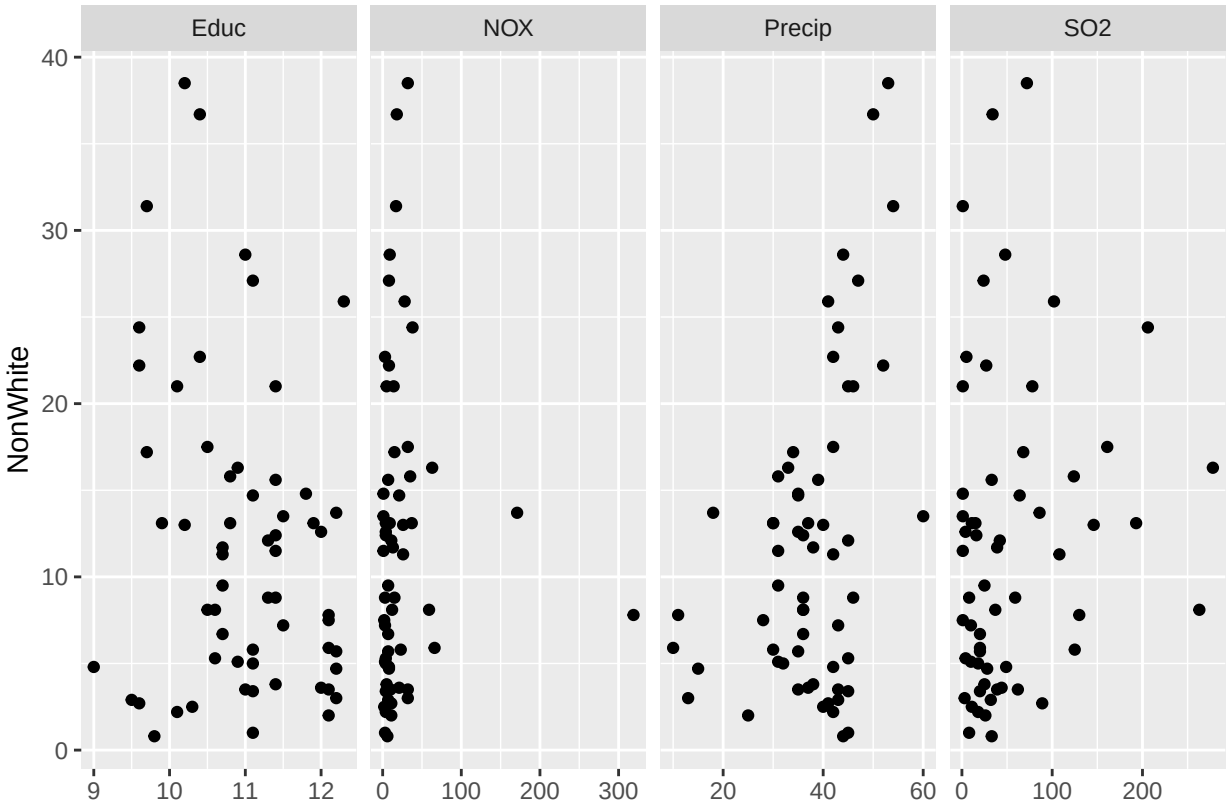
```
## # A tibble: 3 x 1
##   City
##   <chr>
## 1 New Orleans, LA
## 2 New Orleans, LA
## 3 Los Angeles, CA
```

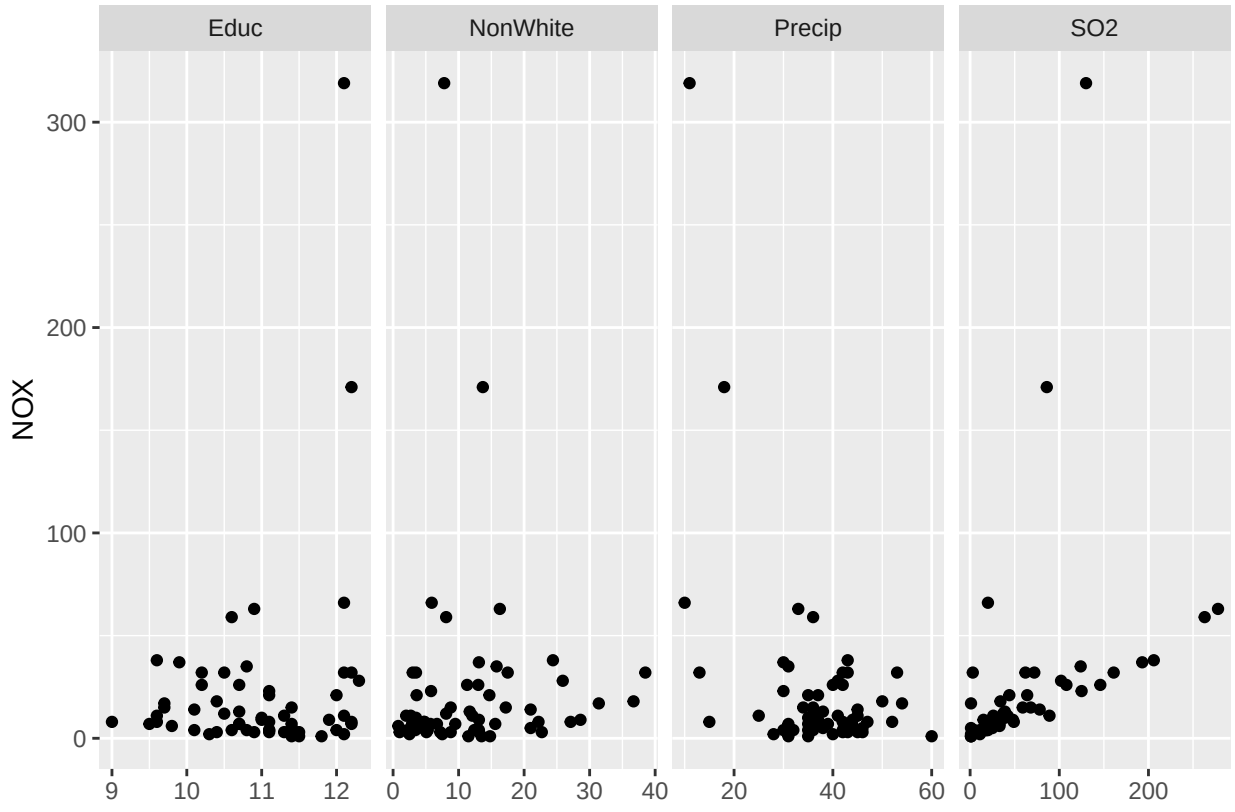
First, I will check The cities that are unusual. After highlighting those points in yellow, I printed the unusual_obs cities. Finally, I can conclude that the unusual cities are New Orleans and Los Angeles. Although these cities are the outliers, they don't seem to have too big of an influence on the data and trend. Since these are only 2 cities in a sample size of 50, there is not a big effect on the data.

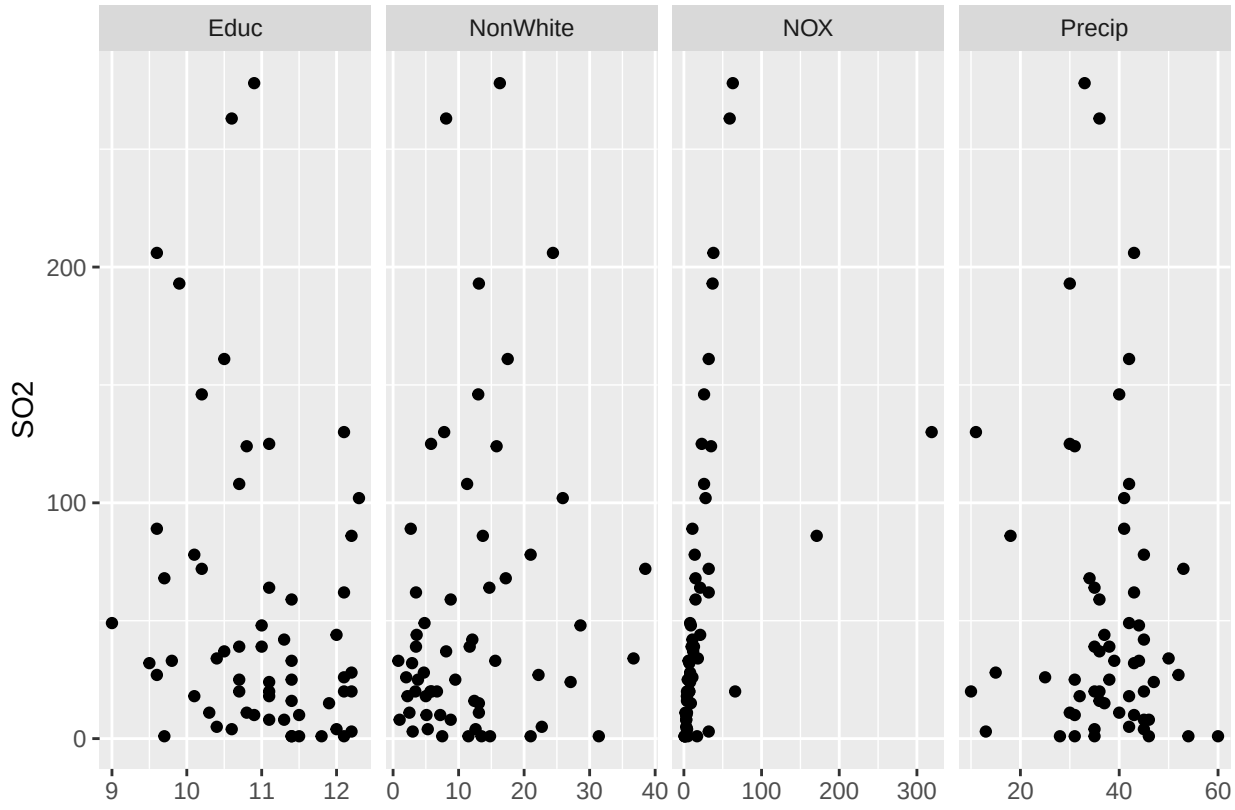
Variables closely related to one another






Educ
NOX
Precip
SO2
NonWhite
40
30
20
10
0
9
10
11
12
0
100
200
300
20
40
60
0
100
200


Educ
NonWhite
Precip
SO2
NOX
300
200
100
0
9
10
11
12
0
10
20
30
40
20
40
60
0
100
200




```
##             Precip     Educ NonWhite      NOX      SO2
## Precip    1.00000 -0.49043  0.41320 -0.48732 -0.10692
## Educ     -0.49043  1.00000 -0.20877  0.22440 -0.23435
## NonWhite  0.41320 -0.20877  1.00000  0.01839  0.15929
## NOX      -0.48732  0.22440  0.01839  1.00000  0.40939
## SO2      -0.10692 -0.23435  0.15929  0.40939  1.00000
```

To check if variables closely related to one another besides mortality, I used the plot based from data. For each individual variable, I set one as the response and rest as the independent. Based from the plots visually, I determined that Precipitation and NonWhite, Precip and Educ, Precip and NOX, and SO2 and NOX all seem to have correlation. So, I used correlation to see if there were linear. It turns out that Precip and Educ have the highest correlation, even though it is a negative association, followed by Percip and NOX. Percipitation also has the most frequent association with other variables. Having closely related variables could affect the analysis by having or incomprehensive data. There will be interactive effects caused by both predictor variables and lead to the conclusion being invalid.

Code appendix

```
# knit options
knitr::opts_chunk$set(echo = F, results = 'markup', message = F, warning = F)

# packages
library(tidyverse)
library(broom)
library(faraway)
library(modelr)
library(tinytex)
pollution <- read_csv('pollution.csv')
head(pollution, 5)
pollution %>% pivot_longer(cols = c(Precip, Educ, NonWhite, NOX, SO2)) %>%
ggplot(aes(x = value, y = Mort)) +
  facet_wrap(~ name, scales = 'free_x') +
  geom_point() +
  labs(x = 'Value', y = 'Mortality')
fit1 <- lm(Mort ~ Precip + Educ + NonWhite + NOX + log(SO2), data = pollution)
summary(fit1)
# create a new column and establish data for reduced SO2 emissions called SO2_half
pollution <- pollution %>% mutate(SO2_half = SO2/2)
new_fit <- lm(Mort ~ Precip + Educ + NonWhite + NOX + log(SO2_half), data = pollution)

# predict the new_mortality rate
without <- pollution %>% select(-SO2)
pollution$new_mortality <- predict(new_fit, newdata= without)
# graph for SO2 vs Mort (original data)
original <- pollution %>% ggplot(aes(x = SO2, y = Mort)) +
  geom_point() +  labs(x = 'SO2', y = 'Mortality', title = "Original Data")
original

# graph for reduced SO2 emissions and the new mortality
updated <- pollution %>% ggplot(aes(x = SO2_half, y = new_mortality)) +
  geom_point() +  labs(x = 'SO2_half', y = 'New Mortality', title = "Reduced Emissions Data")
updated
 pollution <- pollution %>% mutate(diff = new_mortality - Mort)
 pollution %>% pivot_longer(cols = new_mortality) %>%
  ggplot(aes(x = SO2_half, y = diff)) +
  facet_wrap(~ name, scales = 'free_x') +
  geom_point() +
  labs(x = 'Reduced SO2 emissions', y = 'Difference in mortality rate', title = "Potential lives saved"
# Using confidence intervals, I will estimate the number of lives saved each year
fit2 <- lm(Mort ~ Precip + Educ + NonWhite + NOX + log(0.06*SO2), data = pollution)
confint(fit2)

# graph to show
pred_df_pollution <- pollution %>%
  data_grid(.model = fit2) %>%
  add_predictions(model = fit2)

predict(fit1, newdata = pred_df_pollution, interval = 'confidence', level = 0.95)
```

```r
pred_df_pollution_ci <- pred_df_pollution %>%
  cbind(ci = predict(fit2, newdata = pred_df_pollution,
        interval = 'confidence', level = 0.95))

original + geom_path(aes(y = pred), data = pred_df_pollution,
              color = 'red') +
  geom_ribbon(aes(ymin = ci.lwr, ymax = ci.upr,y = ci.fit),
              data = pred_df_pollution_ci, fill = 'red',
              alpha = 0.2) +
  geom_smooth(method = 'lm', se = T, fill = 'blue', alpha =
                0.2, linetype = 1)
# I will see what varaibles seem to be associated with mortality
pollution %>% pivot_longer(cols = c(Precip, Educ, NonWhite, NOX, SO2)) %>%
ggplot(aes(x = value, y = Mort)) +
  facet_wrap(~ name, scales = 'free_x', nrow=1) +
  geom_point() +
  labs(x = '', y = 'mortality')

fit1 <- lm(Mort ~ Precip + Educ + NonWhite + NOX + log(SO2), data = pollution)
summary(fit1)

# I will see if any of the cities in the dataset unusual relative to the others
# check for unusual outliers
studentize <- function(resid, n, p){ resid*sqrt((n - p - 1)/(n - p - resid^2))}
  n <- nrow(model.matrix(fit1))
  p <- ncol(model.matrix(fit1)-1)

fit_df <- augment(fit1, pollution) %>%
  mutate(obs_ix = row_number(), .ext.std.resid = studentize(.std.resid, n, p))

plot <- fit_df %>%
  pivot_longer(cols = c(.ext.std.resid, .hat, .cooksd)) %>%
  ggplot(aes(x = obs_ix, y = abs(value))) +
  facet_wrap(~ name, scales = 'free_y', nrow=3) +
  geom_point() +
  geom_hline(aes(yintercept = 0)) +
  labs(x = 'observation', y = 'residual') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) +
  geom_point()

unusual_obs <- fit_df %>%
  pivot_longer(cols = c(.hat,.ext.std.resid, .cooksd)) %>%
  group_by(name) %>%
  slice_max(order_by = abs(value), n=1) %>%
  ungroup()

# highlight the unusual_obs in yellow
plot+geom_point(data=unusual_obs, color='yellow')

# print the unusual_obs cities
unusual_obs['City']

# I will see if there are any of the variables besides mortality closely related with one another
```

```r
# check to see if Precip and the rest are related
pollution %>% pivot_longer(cols = c(Educ, NonWhite, NOX, SO2)) %>%
ggplot(aes(x = value, y = Precip)) +
  facet_wrap(~ name, scales = 'free_x', nrow=1) +
  geom_point() +
  labs(x = '', y = 'Precipitation')
# It appears NonWhite has somewhat of a linear pattern, and given a few outliers, NOX has a linear patt


# check to see if Education and the rest are related
pollution %>% pivot_longer(cols = c(Precip, NonWhite, NOX, SO2)) %>%
ggplot(aes(x = value, y = Educ)) +
  facet_wrap(~ name, scales = 'free_x', nrow=1) +
  geom_point() +
  labs(x = '', y = 'Education')
# It appears Precipitation has a negative and log trend with Educ


# check to see if NonWhite and the rest are related
pollution %>% pivot_longer(cols = c(Precip, Educ, NOX, SO2)) %>%
ggplot(aes(x = value, y = NonWhite)) +
  facet_wrap(~ name, scales = 'free_x', nrow=1) +
  geom_point() +
  labs(x = '', y = 'NonWhite')
# It appears Precipitation has a linear trend with NonWhite


# check to see if NOX and the rest are related
pollution %>% pivot_longer(cols = c(Precip, Educ, NonWhite, SO2)) %>%
ggplot(aes(x = value, y = NOX)) +
  facet_wrap(~ name, scales = 'free_x', nrow=1) +
  geom_point() +
  labs(x = '', y = 'NOX')
# If excluded some outliers, it appears SO2 has very small trace of a linear trend with NOX

# check to see if SO2 and the rest are related
pollution %>% pivot_longer(cols = c(Precip, Educ, NonWhite, NOX)) %>%
ggplot(aes(x = value, y = SO2)) +
  facet_wrap(~ name, scales = 'free_x', nrow=1) +
  geom_point() +
  labs(x = '', y = 'SO2')
# If excluded some outliers, it appears NOX has very small trace of a linear trend with SO2

# check correlations between variables
pollution %>% select(-c(Mort, City,SO2_half,new_mortality,diff)) %>% cor() %>% round(5)
```