Final Project

Shaan Patel

1. Learning Objective 1.

1.1. [4 points] What is your resarch question?

Air pollution in urban areas of India contribute to increased infant mortality rates more than water pollution while considering what cities are most affected by that air and water pollution as well as an increasing infant mortality rate?

1.2. [8 points] How is this research question connected to environmental economics? Complete the following bullet points with complete sentences.

- This research is based on analyzing air pollution in given areas
- This research is based on analyzing water pollution in given areas
- This research is related to the environmental justice and who is most affected by environmental issues
- This research is founded on providing proof of causaulity as well as measures of dissolving the problem

2. Describe your data

2.1 [8 points] Source:

- Where did you find your data? Open ICR Dataset Research Paper
- Who collected the data? Hanna-Greenston
- When was the data collected? From 1986-2004
- Who is studied in this data? Households in cities
- Where was the data collected? Indian cities
- Why was the data collected? To provide clarity around the environmental soundness of India's municipalities

2.2 [4 points] Data Structure:

- Type of file: csv
- Number of observations (N) = 1997
- Number of Variables = 58
- Unit of observation (e.g. students, people, states, countries?): Year, City, State, District

3. Learning Objective 4.1: Import the data

3.1. [2 points] Set working directory:

3.2 [3 points] Call the libraries that will be used in the analysis.

If the data is in a different format, is there are packaged that imports it into R? If so, install package in console and call the library:

Packages:

```
library(ggplot2)
library(MASS)
library(stargazer)
library(AER)
library(plm)
library(tidyverse)
library(lmtest)
library(sandwich)
library(dplyr)
library(ggdark)
library(readr)
library(cowplot)
library(plotly)
library(ggpubr)
```

3.3. Import data:

Import the data manually and don't forget to copy and paste the code within the Rchunk below:

```
library(haven)
im_air <- read_dta("/Volumes/GoogleDrive/My Drive/Econ 432/Final Project1/Sustainablity Data/Infant-Mor
im_water <- read_dta("/Volumes/GoogleDrive/My Drive/Econ 432/Final Project1/Sustainablity Data/Infant-M
```

Learning Objective 3

4.1 [10 points] Variable Descriptions:

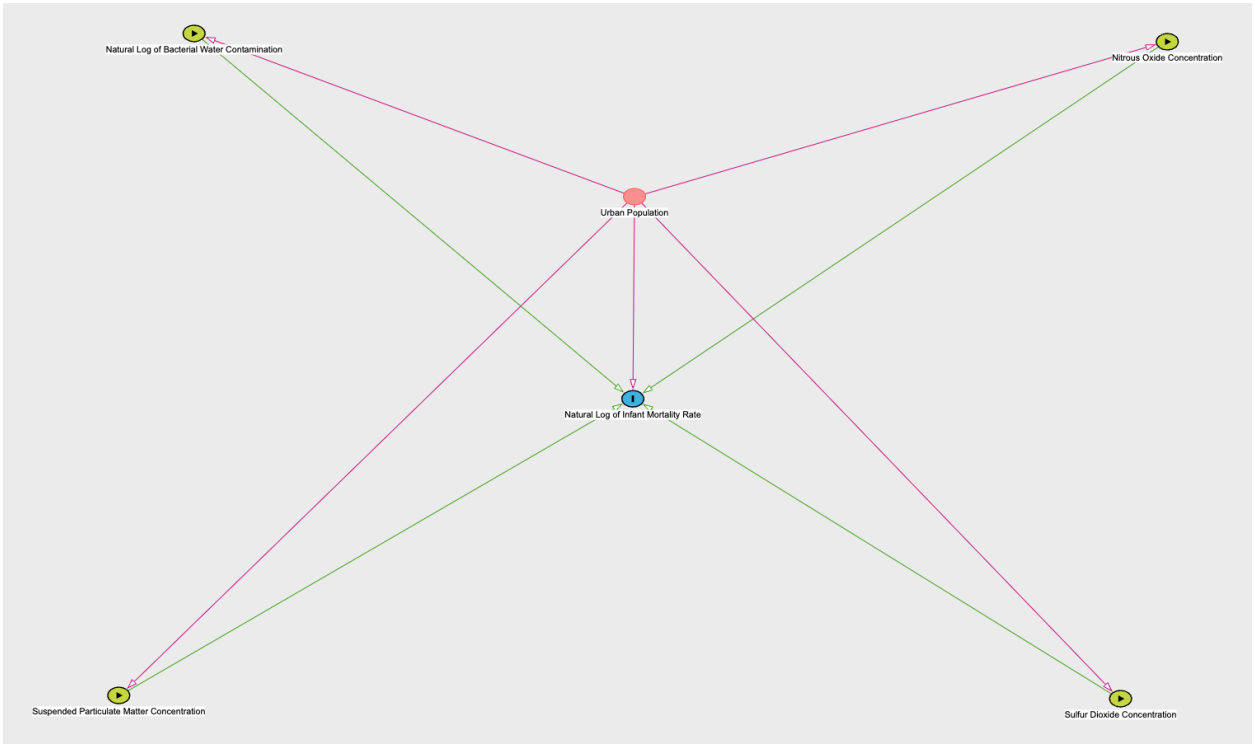
Describe variables you plan on using for your analysis including type of variable (e.g. numeric, factor, dummy) and its units if numeric

- Y (outcome) = c_IM

- $X1$ = lnfcoli
- $X2$ = e_spm_mean
- $X3$ = e_so2_mean
- $X4$ = e_no2_mean
- $X5$ = do
- $X6$ = bod
- $X7$ = state
- $X8$ = year
- $X9$ = pop_urban

5. [10 points] Draw a DAG to illustrate your research question and your identification strategy.

You may use:

DAGitty 

6. Summaries

summarize your key variables

```
summary(im_air$e_spm_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     9.0   140.2   216.3   228.0   302.1   838.4    1183
```

```
summary(im_air$e_so2_mean)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##   0.4392   8.3235  13.9850  17.4624  21.9975 125.9810     1200
```

```
summary(im_air$e_no2_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.011  15.856  24.200  27.959  34.692 209.607    1171
```

```
summary(im_air$c_IM)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##   0.0594   9.0238  19.9659  23.4621  33.0903 499.8054     1084
```

```
summary(im_water$lnfcoli)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.963   5.886   5.668   7.643  14.557     677
```

```
summary(im_water$bod)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##   0.2667   1.7239   2.6818   5.2090   5.0909 100.0000      358
```

```
summary(im_water$do)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.400   6.250   7.038   6.833   7.721  11.000     355
```

```
summary(im_water$c_IM)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##   0.0594   5.9762  17.7755  21.2724  32.3459 168.2521      954
```

7. Possible Regression Model:

You are likely going to revise this as you continue your data analysis. However, it is important to start thinking about the model:

7.1. [2 points] Model Specification:

Choose a specification (level-level, log-log, log-level, level-log). In other words, is your outcome variable in log or level form? What about your control variables? (you may have level and log Xs in a model)

log-level because the y variable is needed to be log in order for the result to be statistically significant same for x variables.

7.2. [4 points] Regression Type

Choose type of regression/model (multiple regression, panel regression, binary regression, multinomial logit regression). Explain your choice.

Multiple regression because we will be using multiple regression to evaluate statistically significance of the variables

7.3. [15 points] Regression Model

Copy and Modify the following code to propose 3 different regression models. In other words, replace Y and the X's for the names of your variables (e.g. instead of X_{1i}, write age_i). Start from the simplest model and then add variables to make it more complex. You will be assessed based on your choice of variables (You should not include variables that are endogenous)

model.1: $cIM = \beta_0 + \beta_1 e.spm.mean_{1i} + YearFixedEffects + u_{it} + StateFixedEffects + u_{it} + Pop.UrbanFixedEffects + u_{it}$

model.2: $cIM = \beta_0 + \beta_1 e.spm.mean_{1i} + \beta_2 e.so2.mean_{2i} + \beta_3 e.no2.mean_{3i} + YearFixedEffects + u_{it} + StateFixedEffects + u_{it} + Pop.UrbanFixedEffects + u_{it}$

model.3: $cIM = \beta_0 + \beta_1 nfcoli_{1i} + YearFixedEffects + u_{it} + StateFixedEffects + u_{it} + Pop.UrbanFixedEffects + u_{it}$

model.4: $cIM = \beta_0 + \beta_1 nfcoli_{1i} + \beta_5 bod_{2i} + \beta_6 do_{3i} + YearFixedEffects + u_{it} + StateFixedEffects + u_{it} + Pop.UrbanFixedEffects + u_{it}$

7. Learning Objective 4.1: Clean the data

7.1 [3 points] Subsetting your data

Most likely, the chosen data comes with multiple variables. Given your answer to 4, you want to use `tidyverse` to subset your data and select the columns (both y and x's that you will use). Name this subset: `data`

```
water = im_water %>% select(c_IM, bod, do, lnfcoli, pop_urban, state, year)
water = na.omit(water)
dim(water)
```

```
## [1] 578   7
```

```
air = im_air %>% select(e_spm_mean, e_so2_mean, e_no2_mean, state, year, pop_urban, c_IM)
air = na.omit(air)
water = water[-c(530:578),]
dim(air)
```

```
## [1] 529   7
```

```
a <- data(1:10, start=c(1987), frequency=12)
b <- data(1:12, start=c(2015,1), frequency=12)

library(zoo)
m <- merge(a = as.zoo(a), b = as.zoo(b))
```

7.3 [4 points] Renaming variables

Rename variables in an optimal way. Think about the name of the variables in your data and consider whether there is any improvement. Use `tidyverse` and `mutate()` to change the name of a variable.

For example, if the name of the variable is long or includes spaces, you should modify the name to a single word (e.g. if the name of column is `what's your age`, modify the name to `age`)

#“‘{r} data = data %>% mutate(lnc_IM = lnc_IM...43)

data = data %>% mutate(pop_urban = pop_urban...25)

data = data %>% mutate(year = year...3)

data = data %>% mutate(state = state...1)

data = data %>% mutate(city = city...2)

data <- data %>% mutate(lne_spm_mean = log(e_spm_mean))

data <- data %>% mutate(lne_so2_mean = log(e_so2_mean))

data <- data %>% mutate(lne_no2_mean = log(e_no2_mean))

data = select(data, -1, -2, -4, -5, -6, -7,-8,-9) #“‘

8. Learning Objective 4.2. Summarize key variables

Write code to summarize the variables in your model from question 4 and write a sentence describing each. Points will be taken off if you use the wrong function.

Remember, use `summary()` or `table()` depending on the type of variable.

8.1. [2 points] Outcome variable y

```
summary(air$c_IM)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.05936 11.15546 20.57571 22.59976 32.45787 99.03008
```

```
summary(water$c_IM)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.05936  5.57547 17.22860 20.25643 29.87463 89.40035
```

Description:

- The max average concentration of SPM in the air is 548.63
- The max average concentration of F-Coli in the water is 14.5574
- The natural log of infant mortality can reach a max of 4.354 measure in thousands
- The max urban population is 15122 measured in thousand

8.2. [10 points] Control Variables x's

Water and Air Datasets control variables are the same

For each x, complete the following:

```
summary(air$pop_urban)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.2   715.6  1310.9  2170.4  2396.3 12905.8
```

Description:

```
summary(air$state)
```

```
##    Length     Class      Mode
##       529 character character
```

Description:

```
summary(air$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1987    1992    1996    1996    2000    2004
```

Description:

Do these steps for all the $x's$ in your model

10. [25 points] Learning Objective 4.4: Visualize key variables using `tidyverse` and `ggplot2`

Generate **5 different** plots to summarize key variables from your data. Focus on the outcome variable and how it may be related to key control variables. Think about your DAG and equations from 6. Each graph should look different. I expect at least one plot that illustrates a single variable, two variables, and two variables + facet.

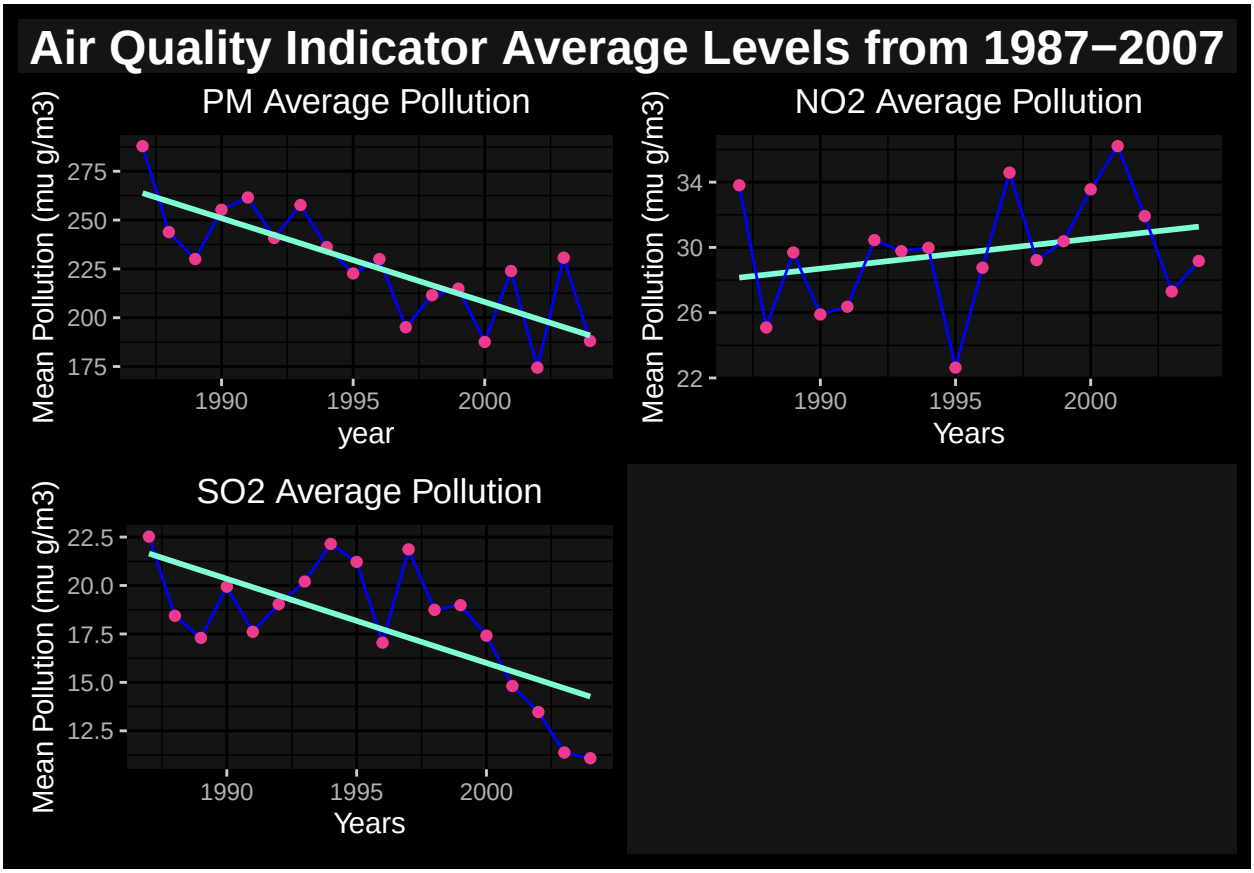
Each plot will be graded by checking:

1. [**0.5 point**] Did you combine `tidyverse` and `ggplot2` to generate the plot? Specifically, did follow any filtering instructions and/or remove `NA`s from the data using `tidyverse`?
2. [**1 points**] The choice of plot - did you choose the appropriate plot to visualize the type(s) of variable(s)?
3. [**1 point**] the correct mapping of the variable(s)
4. [**1.5 points**] The design of the plot: make sure you check the following items for your plots
    a. Plot title
    b. vertical axis title
    c. horizontal axis title
    d. appropriate y scale
    e. appropriate x scale
5. [**1 points**] After every visualization plot, I will ask you to describe the plot or answer a question about it.

10.1 Plot 1

```
year_mean1 = air %>%
  group_by(year) %>%
  summarise_at(vars(e_spm_mean), list(e_spm_mean = mean))

p3.0 <- year_mean1 %>%
  ggplot(aes(x = year, y = e_spm_mean)) +
  geom_line(color="blue")+
  geom_point(color="violetred2")+
  geom_smooth(se = FALSE, method = lm, color = "aquamarine") +
  ggtitle(paste("PM Average Pollution")) +
  theme_bw() +
  scale_x_continuous() +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=12)) +
  theme(axis.text.y=element_text(size=16)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=14)) +
  theme(legend.text=element_text(size=12)) +
  ylab("Mean Pollution (mu g/m3)") + dark_mode()+
  theme(plot.title = element_text(hjust = 0.5))


year_mean3.0 = air %>%
  group_by(year) %>%
  summarise_at(vars(e_so2_mean), list(e_so2_mean = mean))

p5.0 <- year_mean3.0 %>%
  ggplot(aes(x = year, y = e_so2_mean)) +
  geom_line(color="blue")+
  geom_point(color="violetred2")+
  geom_smooth(se = FALSE, method = lm, color = "aquamarine") +
  ggtitle(paste("SO2 Average Pollution")) +
  theme_bw() +
  scale_x_continuous() +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=12)) +
  theme(axis.text.y=element_text(size=16)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=14)) +
  theme(legend.text=element_text(size=12)) +
  ylab("Mean Pollution (mu g/m3)") +
  xlab("Years") + dark_mode()+
  theme(plot.title = element_text(hjust = 0.5))


year_mean4.0 = air %>%
  group_by(year) %>%
```

```
  summarise_at(vars(e_no2_mean), list(e_no2_mean = mean))

p4.0 <- year_mean4.0 %>%
  ggplot(aes(x = year, y = e_no2_mean)) +
  geom_smooth(se = FALSE, method = lm, color = "aquamarine") +
  geom_line(color="blue")+
  geom_point(color="violetred2")+
  ggtitle(paste("NO2 Average Pollution")) +
  theme_bw() +
  scale_x_continuous() +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=12)) +
  theme(axis.text.y=element_text(size=16)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=14)) +
  theme(legend.text=element_text(size=12)) +
  ylab("Mean Pollution (mu g/m3)") +
  xlab("Years") + dark_mode() +
  theme(plot.title = element_text(hjust = 0.5))



figure3 <- plot_grid(
  p3.0, p4.0, p5.0,
  align="center"
)

annotate_figure(figure3,
                top = text_grob("Air Quality Indicator Average Levels from 1987-2007", color = "white",
                fig.lab.face = "bold"
                ) + dark_mode()
```



Description:

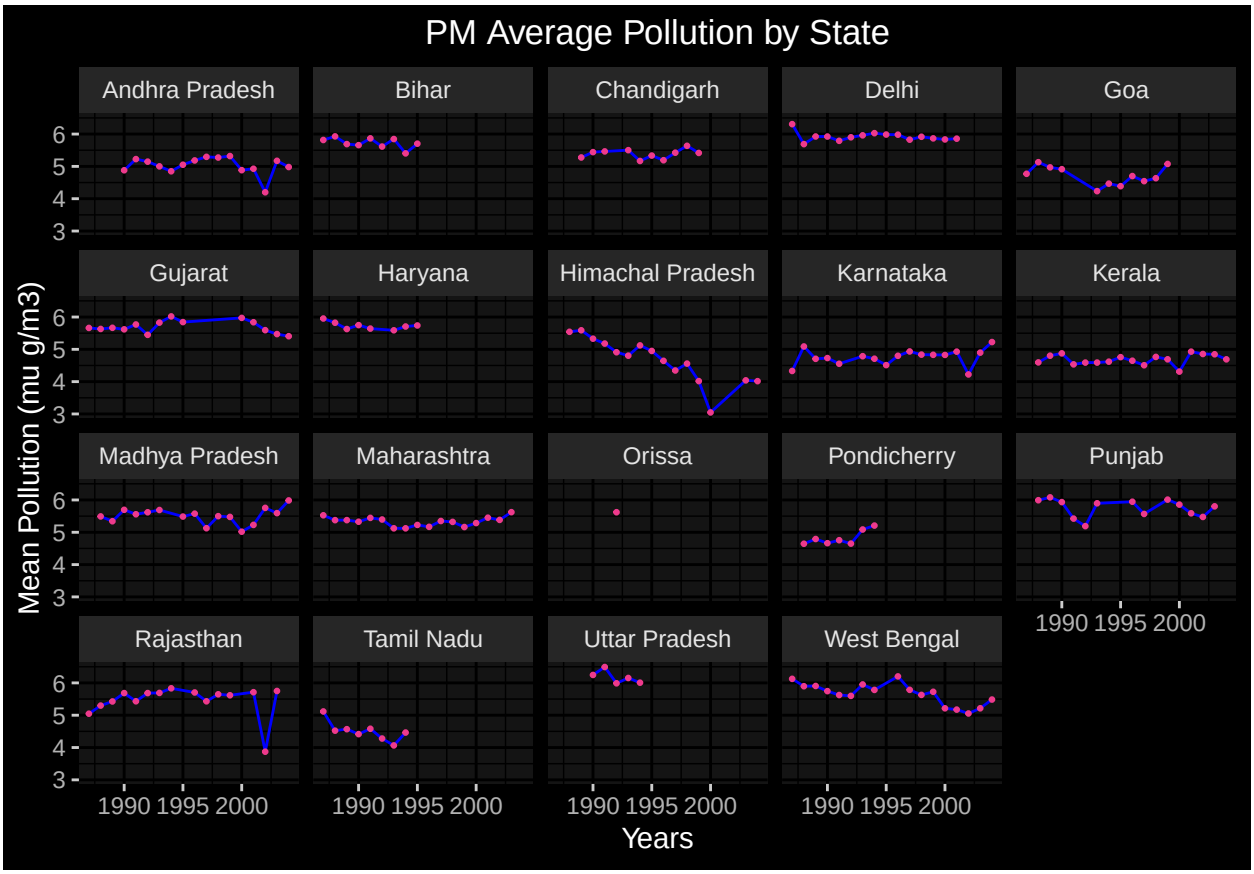
In this graph we are looking at the overall trends in average air quality in different states within India between 1987 and 2007. From the 1987 to 2007 SO2 dropped dramatically from ~22.5 micrograms to ~10 micrograms. Similarly, the particulate matter levels have decreased from ~287.5 micrograms to ~187.5 micrograms. In contrast, NO2 average pollution have been volatile where the mean pollution levels went from 34 micrograms in 1987 to 22 micrograms in 1995 then back to around 34 micrograms in the early 2000s (overall NO2 Pollution have increased.

```
air1 = air %>%
  group_by(state, year) %>%
  summarise_at(vars(e_spm_mean), list(e_spm_mean = mean))

p3.1 <- air1 %>%
  ggplot(aes(x = year, y = log(e_spm_mean))) +
  geom_line(color="blue") +
  geom_point(color="violetred2", size = .5 )+
  ggtitle(paste("PM Average Pollution by State")) +
  theme_bw() +
  scale_x_continuous() +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=12)) +
  theme(axis.text.y=element_text(size=16)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=14)) +
```

```
  theme(legend.text=element_text(size=12)) +
  ylab("Mean Pollution (mu g/m3)")+
  xlab("Years") +
  facet_wrap(~state) +
  dark_mode() +
  theme(plot.title = element_text(hjust = 0.5))


air2 = air %>%
  group_by(state, year) %>%
  summarise_at(vars(e_so2_mean), list(e_so2_mean = mean))


p5.1 <- air2 %>%
  ggplot(aes(x = year, y = log(e_so2_mean))) +
  geom_line(color="blue") +
  geom_point(color="violetred2", size = .5 )+
  ggtitle(paste("SO2 Average Pollution by State")) +
  theme_bw() +
  scale_x_continuous() +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=12)) +
  theme(axis.text.y=element_text(size=16)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=14)) +
  theme(legend.text=element_text(size=12)) +
  ylab("Mean Pollution (mu g/m3)") +
  xlab("Years") +
  facet_wrap(~state) +
  dark_mode()+
  theme(plot.title = element_text(hjust = 0.5))


air3 = air %>%
  group_by(state, year) %>%
  summarise_at(vars(e_no2_mean), list(e_no2_mean = mean))


p4.1 <- air3 %>%
  ggplot(aes(x = year, y = log(e_no2_mean))) +
  geom_line(color="blue")+
  geom_point(color="violetred2", size = .5 )+
  ggtitle(paste("NO2 Average Pollution by State")) +
  theme_bw() +
  scale_x_continuous() +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=12)) +
  theme(axis.text.y=element_text(size=16)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
```

```
  theme(plot.title=element_text(size=14)) +
  theme(legend.text=element_text(size=12)) +
  ylab("Mean Pollution (mu g/m3)") +
  xlab("Years") +
  facet_wrap(~state, scales="fixed" ) +
  dark_mode()+
  theme(plot.title = element_text(hjust = 0.5))
```

```
p3.1
```

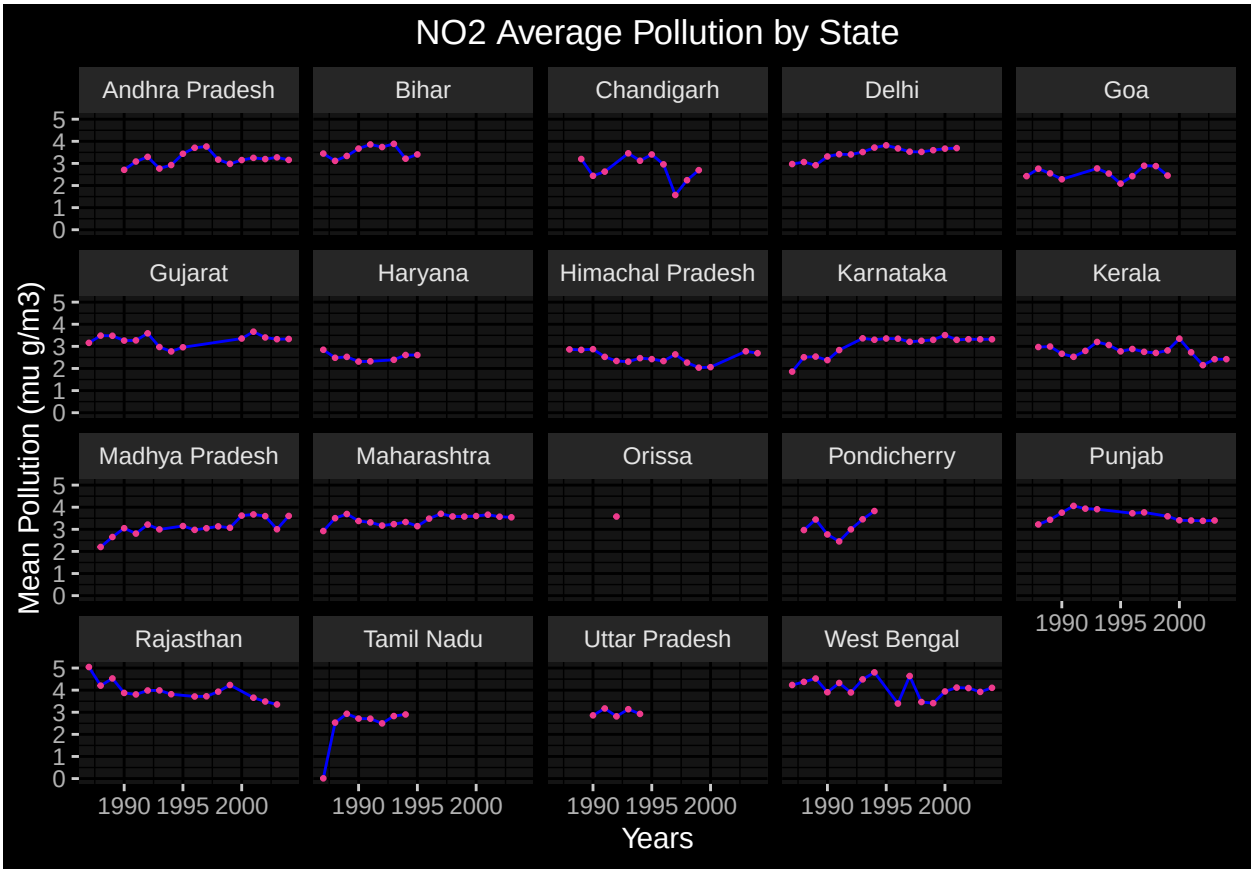


Description:

This graph is showing us the average particulate matter levels in India's largest states from around 1887 to 2007. Look closely we can see a similar decreasing or stable trend for each state. The most drastic fall in average particulate matter was in Himachal Pradesh, decreasing at ~5.5 micrograms in 1987 to ~3 micrograms in the early 2000s. The two states that are suffering from relatively increasing particulate matter pollution is Karnataka and Rajasthan . From 1987 to 2007, Karnataka and Rajasthan's particulate matter polution had increased almost 20%.

Plot 2.1

```
p4.1
```

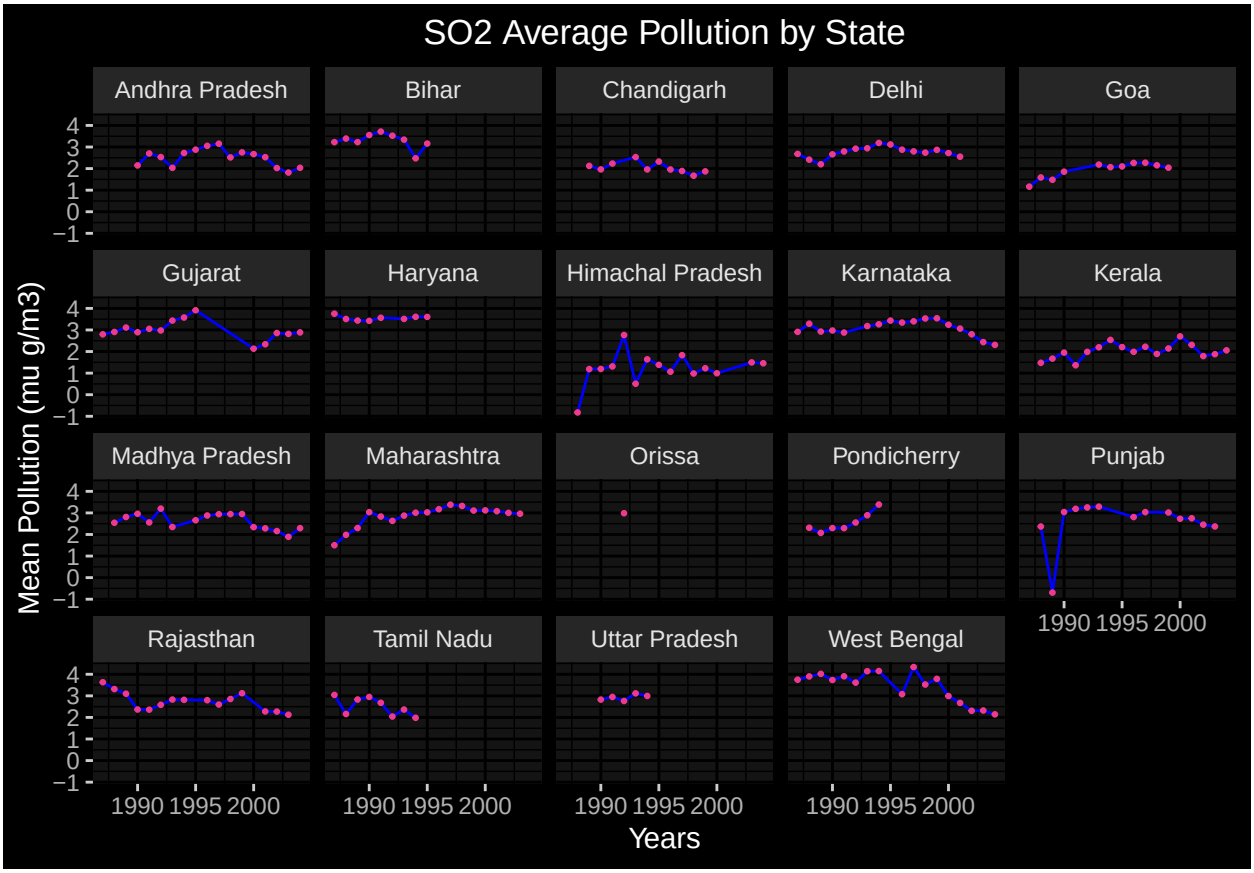


Description:

This graph is showing us the average nitrous oxide levels in India's largest states from around 1887 to 2007. Look closely we can see a similar steadily increasing trend for each state. The most drastic increase in average NO2 was in Madhya Pradesh and Karnataka, increasing around 2 micrograms from 1987 to 2007 (almost 100% increase). The two states that are suffering from relatively constant NO2 pollution are West Bengal and Delhi. Rajasthan has managed to reduce NO2 pollution from 1987 to 2007, where NO2 mean pollution dropped from ~5 micrograms to ~3 micrograms.

Plot 2.2

```
p5.1
```



Description:

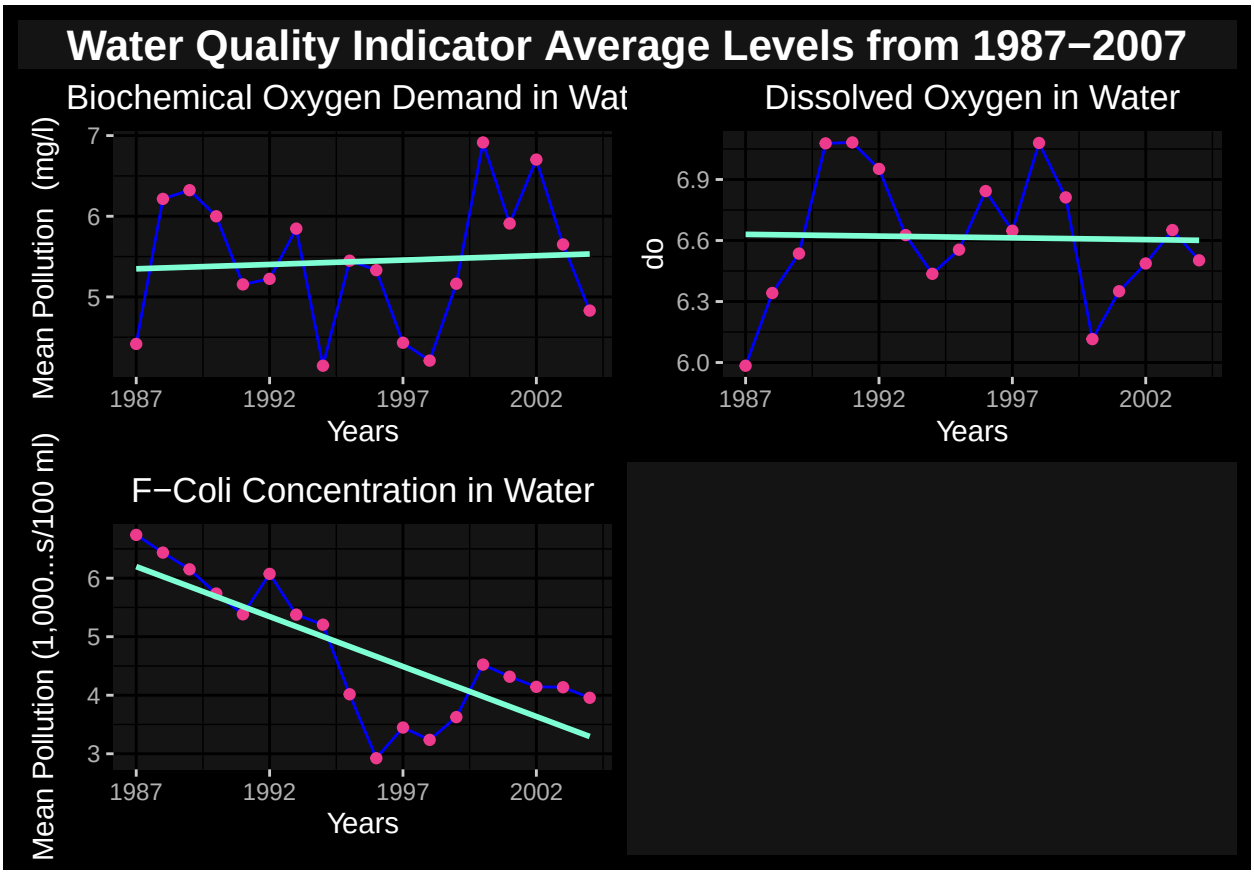
This graph is showing us the average sulfur dioxide levels in India's largest states from around 1887 to 2007. Look closely we can see a similar decreasing or stable trend for each state. The largest decrease in average SO2 levels was in West Bengal and Rajasthan, decreasing at ~4 micrograms in 1987 to ~2 micrograms in the late 2000s. The two states that are suffering from relatively increasing particulate matter pollution are Maharashtra and Kerala. From 1987 to 2007, Maharashtra and Kerala's SO2 mean polution had increased almost 50%.

10.3 Plot 3

```
year_mean2 = water %>%
  group_by(year) %>%
  summarise_at(vars(bod), list(bod = mean))

p3 <- year_mean2 %>%
  ggplot(aes(x = year, y = bod)) +
  geom_line(color = "blue")+
  geom_point(color="violetred2", size = 1.5 )+
  geom_smooth(se = FALSE, method = lm, color = "aquamarine") +
  ggtitle(paste("Biochemical Oxygen Demand in Water")) +
  theme_bw() +
  scale_x_continuous(breaks = seq(1987,2007, 5)) +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
```

```
  theme(axis.text.x=element_text(size=5)) +
  theme(axis.text.y=element_text(size=6)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=12)) +
  theme(legend.text=element_text(size=18)) +
  ylab("Mean Pollution  (mg/l) ") +
  xlab("Years")  + dark_mode() +
  theme(plot.title = element_text(hjust = 0.5))


year_mean3 = water %>%
  group_by(year) %>%
  summarise_at(vars(lnfcoli), list(lnfcoli = mean))

p5 <- year_mean3 %>%
  ggplot(aes(x = year, y = lnfcoli)) +
  geom_line(color = "blue")+
  geom_point(color="violetred2", size = 1.5 )+
  geom_smooth(se = FALSE, method = lm, color = "aquamarine") +
  ggtitle(paste("F-Coli Concentration in Water")) +
  theme_bw() +
  scale_x_continuous(breaks = seq(1987,2007, 5)) +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=5)) +
  theme(axis.text.y=element_text(size=6)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=12)) +
  theme(legend.text=element_text(size=18)) +
  ylab("Mean Pollution (1,000's/100 ml)") +
  xlab("Years")  + dark_mode() +
  theme(plot.title = element_text(hjust = 0.5))


year_mean4 = water %>%
  group_by(year) %>%
  summarise_at(vars(do), list(do = mean))

p4 <- year_mean4 %>%
  ggplot(aes(x = year, y = do)) +
  geom_line(color = "blue")+
  geom_point(color="violetred2", size = 1.5 )+
  geom_smooth(se = FALSE, method = lm, color = "aquamarine") +
  ggtitle(paste("Dissolved Oxygen in Water")) +
  theme_bw() +
  scale_x_continuous(breaks = seq(1987,2007, 5)) +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=5)) +
  theme(axis.text.y=element_text(size=6)) +
  theme(axis.title.x=element_text(size=16)) +
```

```
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=12)) +
  theme(legend.text=element_text(size=18)) +
  xlab("Years")  + dark_mode() +
  theme(plot.title = element_text(hjust = 0.5))



figure2 <- plot_grid(
  p3, p4, p5,
  align="hv"
)

annotate_figure(figure2,
                top = text_grob("Water Quality Indicator Average Levels from 1987-2007", color = "white
                fig.lab.face = "bold"
                ) + dark_mode()
```



Description

In this graph we are looking at the overall trends in average water quality in different states within India between 1987 and 2007. The overall trends are pretty mixed. Plot 3 demonstrates that BOD steadily got worse throughout the late 1980s to mid 1990s but began to improve after 1997, but the average shows us that BOD has only really gotten around 20% better from 1987 to 2007. DO has been on average decreasing over time suggesting worsening water quality. F-coli drops dramatically in the 1990s, but when going into the 2000s it begins to increase slightly. F-Coli's decrease is interesting suggesting that despite the alarmingly fast growth in sewage generation, domestic water pollution is still decreasing.

```r
water1 = water %>%
  group_by(state, year) %>%
  summarise_at(vars(bod), list(bod = mean))

p3.1.0 <- water1 %>%
  ggplot(aes(x = year, y = bod)) +
  geom_line(color = "blue")+
  geom_point(color="violetred2", size = .5 )+
  ggtitle(paste("Biochemical Oxygen Demand in Water by State")) +
  theme_bw() +
  scale_x_continuous(breaks = seq(1987,2007, 5)) +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=5)) +
  theme(axis.text.y=element_text(size=6)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=12)) +
  theme(legend.text=element_text(size=18)) +
  ylab("Mean Pollution  (mg/l) ") +
  xlab("Years")+
  facet_wrap(~state, scales="fixed" ) +
  dark_mode()+
  theme(plot.title = element_text(hjust = 0.5))



water2 = water %>%
  group_by(state, year) %>%
  summarise_at(vars(lnfcoli), list(lnfcoli = mean))


p5.1.0 <- water2 %>%
  ggplot(aes(x = year, y = lnfcoli)) +
  geom_line(color = "blue")+
  geom_point(color="violetred2", size = .5 )+
  ggtitle(paste("F-Coli Concentration in Water by State")) +
  theme_bw() +
  scale_x_continuous(breaks = seq(1987,2007, 5)) +
  scale_y_continuous(breaks = seq(1.9, 12, 2.5))+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=5)) +
  theme(axis.text.y=element_text(size=6)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=12)) +
  theme(legend.text=element_text(size=18)) +
  ylab("Mean Pollution (1,000's/100 ml)") +
  xlab("Years") +
  facet_wrap(~state, scales="fixed" ) +
  dark_mode()+
  theme(plot.title = element_text(hjust = 0.5))
```

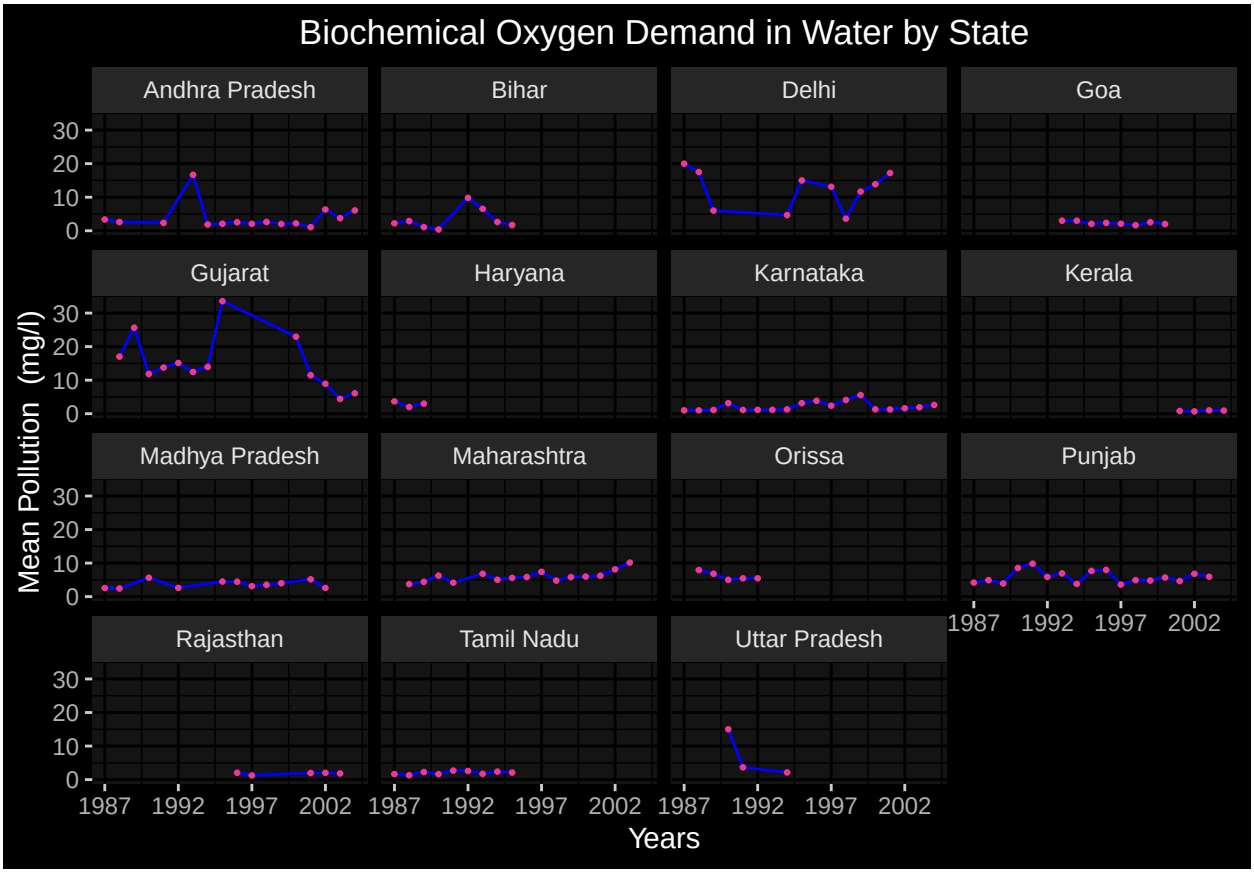
```r
water3 = water %>%
  group_by(state, year) %>%
  summarise_at(vars(do), list(do = mean))

summary(water3$do)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.950   6.021   6.724   6.472   7.242   9.058
```

```r
p4.1.0 <- water3 %>%
  ggplot(aes(x = year, y = do)) +
  geom_line(color = "blue")+
  geom_point(color="violetred2", size = .5 )+
  ggtitle(paste("Dissolved Oxygen in Water by State")) +
  theme_bw() +
  scale_x_continuous(breaks = seq(1987,2007, 5)) +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=5)) +
  theme(axis.text.y=element_text(size=6)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=12)) +
  theme(legend.text=element_text(size=18)) +
  xlab("Years") +
  ylab("Mean Pollution (mg/l)") +
  facet_wrap(~state, scales="fixed" ) +
  dark_mode() +
  theme(plot.title = element_text(hjust = 0.5))
```

Plot 4.0

```r
p3.1.0
```

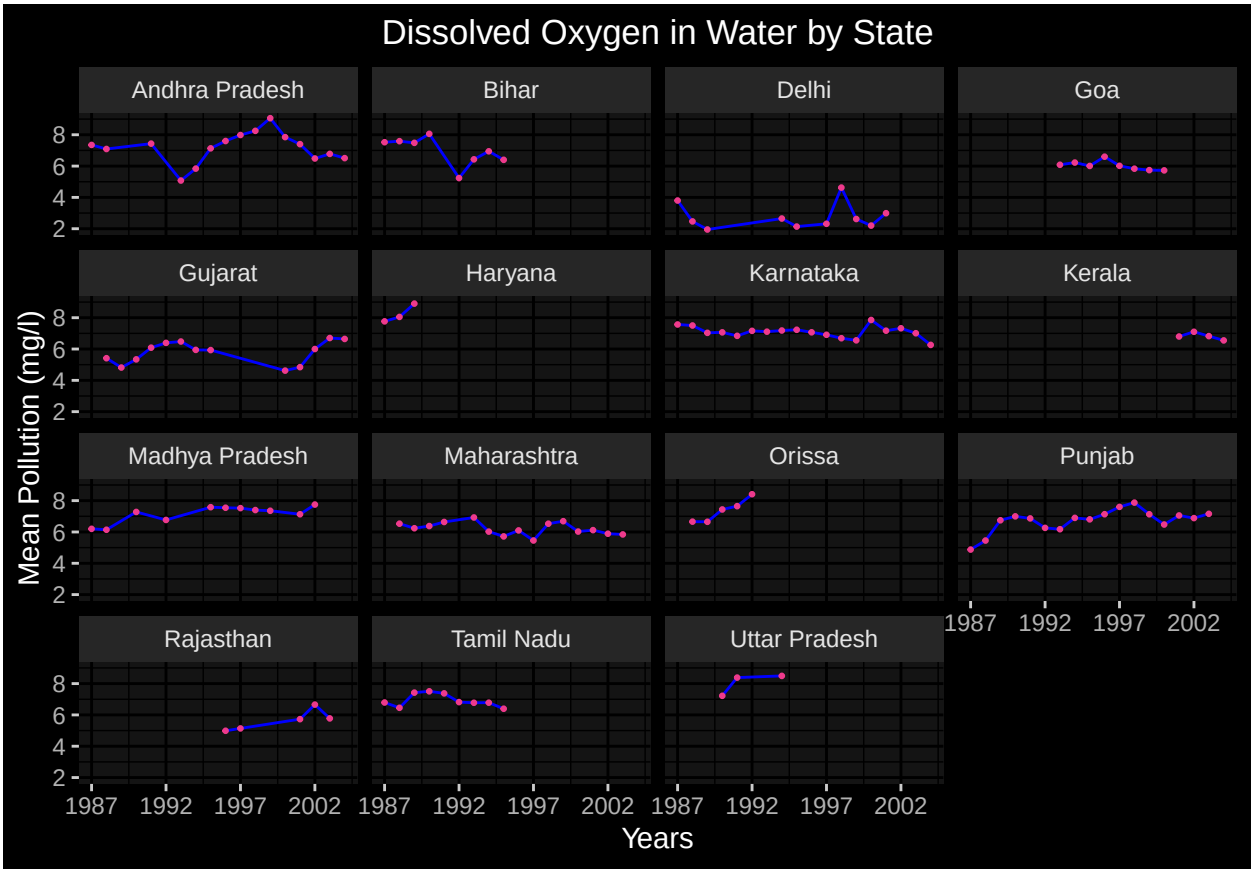


Description:

This graph is showing us the average biochemical oxygen demand in India's largest states from around 1887 to 2007. Look closely we can see a similar slightly increasing or low decreased trend for each state. The largest decrease in average BOD was in Gujurat, decreasing at ~35mg/l in 1995 to ~5 mg/l in the late 2000s. The two states that are suffering from relatively increasing BOD are Maharashtra and Delhi. From 1987 to 2007, Maharashtra and Delhi's BOD has increased more than 5% on average.

Plot 4.1

```
p4.1.0
```

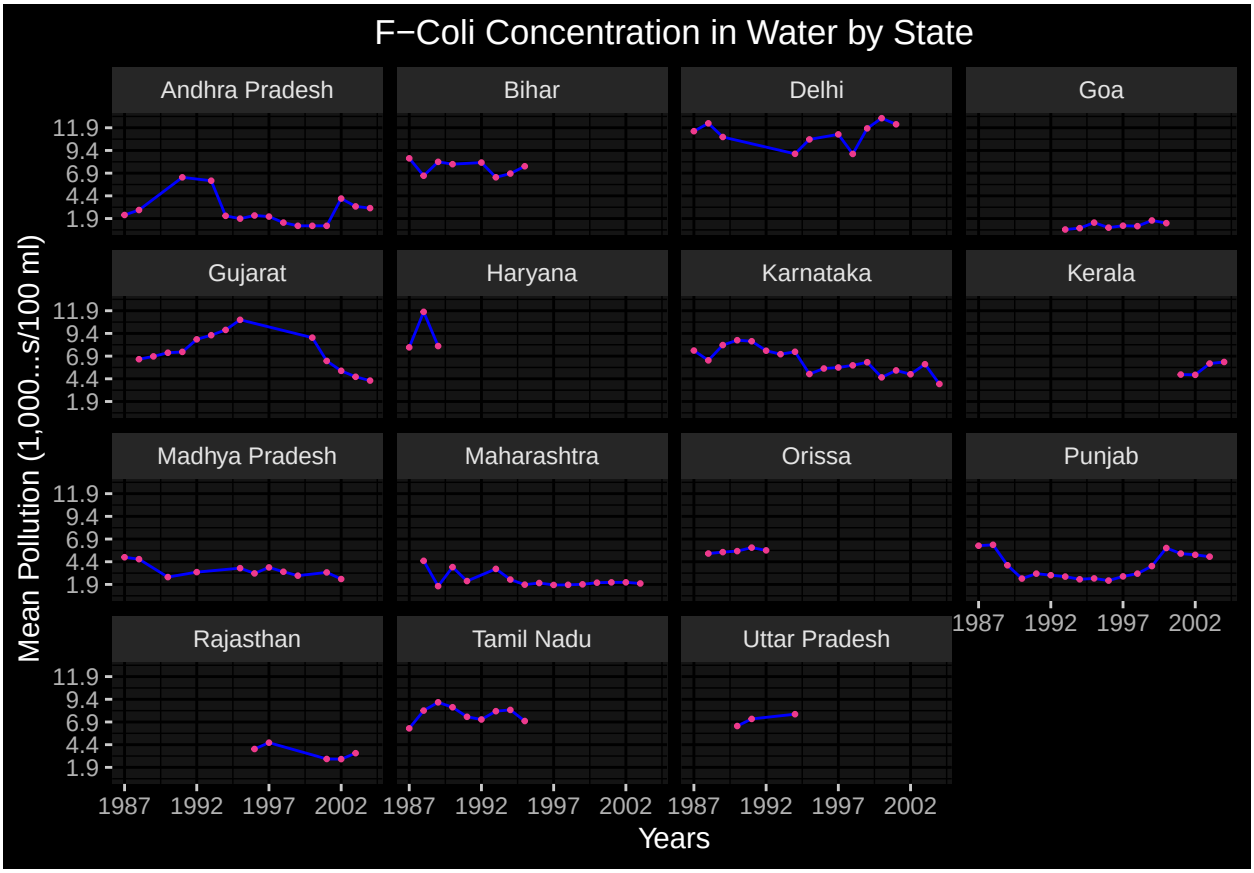


Description:

This graph is showing us the average dissolved oxygen in India's largest states from around 1887 to 2007. Look closely we can see a similar slightly decreasing or relatively constants trends for each state. Delhi has the worst water quality with their average DO levels below ~4 mg/l from 1987 to 2007 meaning water quality is so bad fish can barely survive to decompose particulate matter. The two states that are suffering from relatively increasing DO are Punjab and Gujurat. From 1987 to 2007, Punjab's DO has inxrease more than 25% and Gujurat's DO has increased more than 12.5% on average. The largest decrease in average DO was in Bihar, decreasing at ~8mg/l in 1990 to ~5 mg/l in 1992 .

Plot 4.2

```
p5.1.0
```



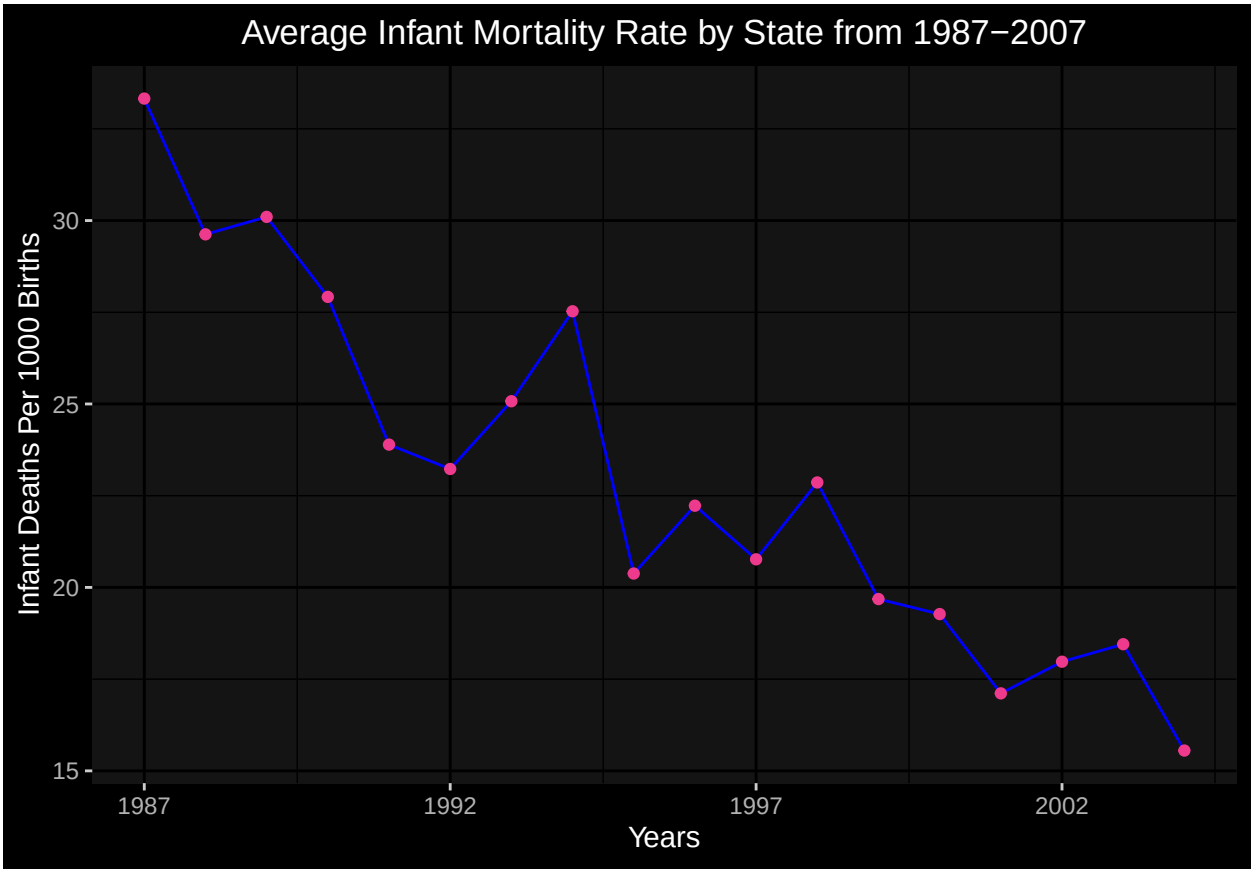
Description:

This graph is showing us the average F-Coli levels in India's largest states from around 1887 to 2007. We can see a similar decreasing or relatively decreasing trend for each state. The two states that have decreased F-Coli levels the most are Karnataka and Gujurat. In 1987, Karnataka had ~6.9 micrograms where in 2007 it had decreases to ~4.4 micrograms.

10.4 Plot 5

```
year_mean = air %>%
  group_by(year) %>%
  summarise_at(vars(c_IM), list(c_im = mean))
year_mean %>%
  ggplot(aes(x = year, y = c_im)) +
  geom_line(color = "blue")+
  geom_point(color="violetred2", size = 1.5 )+
  ggtitle(paste("Average Infant Mortality Rate by State from 1987-2007")) +
  theme_bw() +
  scale_x_continuous(breaks = seq(1987,2007, 5)) +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=12)) +
  theme(axis.text.y=element_text(size=9)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
```

```
  theme(plot.title=element_text(size=14)) +
  theme(legend.text=element_text(size=18)) +
  ylab("Infant Deaths Per 1000 Births") +
  xlab("Years") + dark_mode()+
  theme(plot.title = element_text(hjust = 0.5))
```

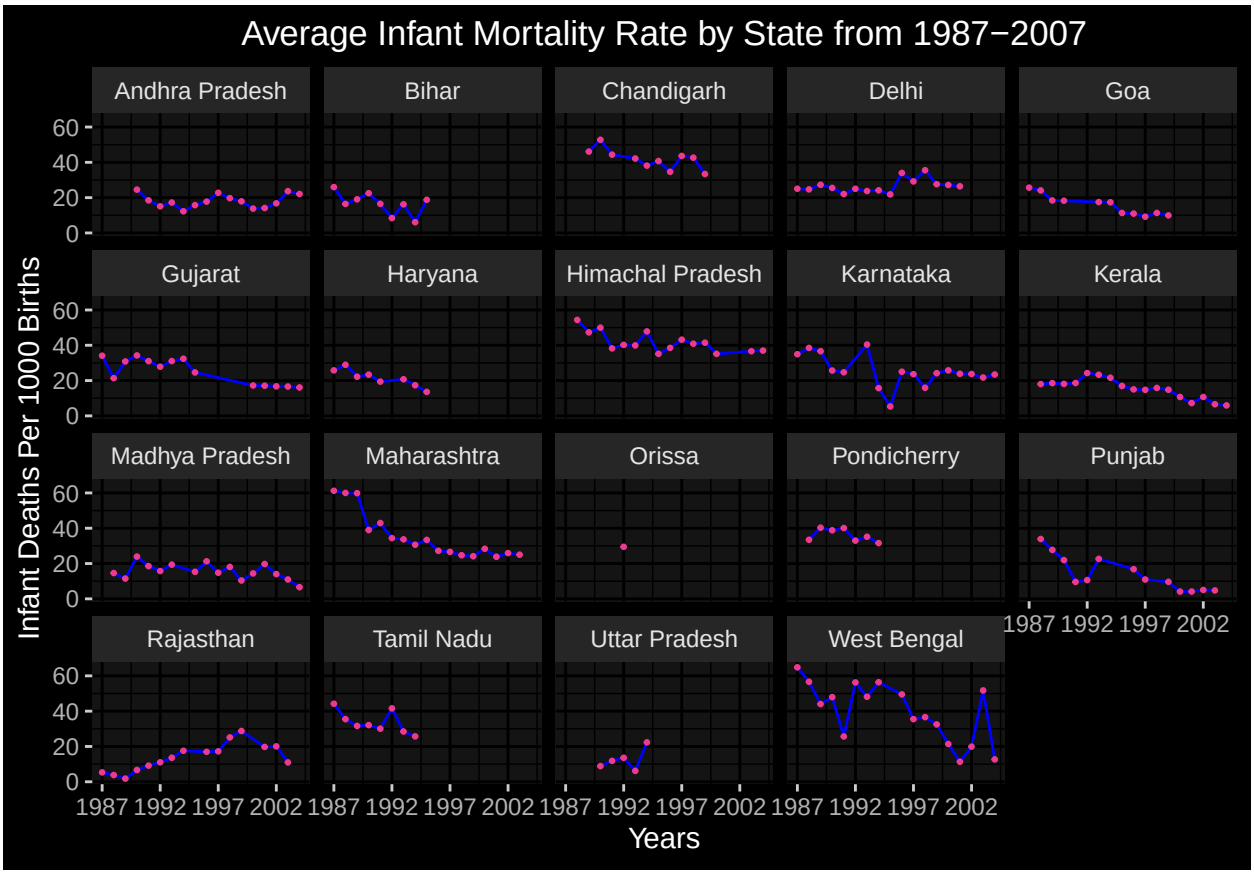


Description: This graph is showing us the average infant deaths per 1000 births(c_IM) in India from around 1887 to 2007. Infant mortality rates are a good way to measure how effective environmental regulations are for several reasons. One is that, compared to measures of adult health, infant health is probably more affected by short- and medium-term changes in pollution. Another reason is that the first year is a vulnerable time, so a loss of life expectancy can be large. Infant mortality in urban India fell a lot between 1987 and 2004, from ~34.5 deaths per 1,000 live births to ~15.7. ## 10.5 Plot 6

```
state_mean = air %>%
  group_by(state, year) %>%
  summarise_at(vars(c_IM), list(c_im = mean))
state_mean %>%
  ggplot(aes(x = year ,y = c_im)) +
  geom_line(color = "blue")+
  geom_point(color="violetred2", size = .5 )+
  ggtitle(paste("Average Infant Mortality Rate by State from 1987-2007")) +
  theme_bw() +
  scale_x_continuous(breaks = seq(1987,2007, 5)) +
  scale_y_continuous()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x=element_text(size=5)) +
```

```
  theme(axis.text.y=element_text(size=6)) +
  theme(axis.title.x=element_text(size=16)) +
  theme(axis.title.y=element_text(size=16, angle=90)) +
  theme(plot.title=element_text(size=14)) +
  theme(legend.text=element_text(size=18)) +
  ylab("Infant Deaths Per 1000 Births") +
  xlab("Years") +
  facet_wrap(~state) + dark_mode()+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```



Description:

This graph is showing us the average infant deaths per 1000 births(c_IM) in India's largest states from around 1887 to 2007. Look closely we can see a similar decreasing trend for each state. The most drastic fall in average c_IM was West Bengal, decreasing from a ~6% to ~1% infant mortality rate. The two states that are suffering from a relatively constant infant mortality rate are Delhi and Andhra Pradesh. FroM 1987 to 2007, Andhra Pradesh and Delhi had around a 2% infant mortality rate.

11. Learning Objectives 3 and 4.5. Write econometric models as equations

Revise your answer for question 6:

11.1. [2 points] Model Specification

Choose a specification (e.g level-level, log-log, log-level, level-log). In other words, is your outcome variable in log or level form? What about your control variables? (you may have level and log Xs in a model)

level-level because the most important x factors are in level form. The control variables are in in level form in order to group accordingly to particular regions in India.

11.2. [4 points] Regression Type

Choose type of regression/model (multiple regression, panel regression, binary regression, multinomial logit regression). Explain your choice

Multiple regression because we are looking at how much the levels of SPM in the air and F-Coli in the water are good estimators of infant mortality.

11.3 [15 points] Regression Model

Copy and Modify the following code to propose 3 different regression models. In other words, replace Y and the X's for the names of your variables (e.g. instead of X_{1i}, write age_i). Start from the simplest model and then add variables to make it more complex. You will be assessed based on your choice of variables (You should not include variables that are endogenous)

model.1: $cIM = \beta_0 + \beta_1 e.spm.mean_{1i} + YearFixedEffects + u_{it} + StateFixedEffects + u_{it} + Pop.UrbanFixedEffects + u_{it}$

model.2: $cIM = \beta_0 + \beta_1 e.spm.mean_{1i} + \beta_2 e.so2.mean_{2i} + \beta_3 e.no2.mean_{3i} + YearFixedEffects + u_{it} + StateFixedEffects + u_{it} + Pop.UrbanFixedEffects + u_{it}$

model.3: $cIM = \beta_0 + \beta_1 nfcoli_{1i} + YearFixedEffects + u_{it} + StateFixedEffects + u_{it} + Pop.UrbanFixedEffects + u_{it}$

model.4: $cIM = \beta_0 + \beta_1 nfcoli_{1i} + \beta_5 bod_{2i} + \beta_6 do_{3i} + YearFixedEffects + u_{it} + StateFixedEffects + u_{it} + Pop.UrbanFixedEffects + u_{it}$

12. Learning Objective 4.6 Estimate Regression Models

12.1 [9 points] Write code to estimate the models above.

```
mult.mod.1.0 = lm(c_IM ~ e_spm_mean  +  pop_urban -1 + year -1
                  , (as.factor(state))
                  , data = air)

mult.mod.1.1 = lm(c_IM ~ e_spm_mean + e_so2_mean + e_no2_mean + pop_urban -1 + year -1
                  , (as.factor(state))
                  , data = air)
```

```
mult.mod.1.2 = lm(c_IM ~ lnfcoli +  pop_urban -1 + year-1
                  , (as.factor(state))
                  , data = water )

mult.mod.1.3 = lm(c_IM ~ lnfcoli + bod + do +  pop_urban -1 + year -1
                  , (as.factor(state))
                  , data = water )
```

12.2. [6 points] Write code to gather the robust standard errors in a list.

```
rob_se2.1.0 <- list(sqrt(diag(vcovHC(mult.mod.1.0, type = "HC1"))),
                    sqrt(diag(vcovHC(mult.mod.1.1, type = "HC1"))),
                    sqrt(diag(vcovHC(mult.mod.1.2, type = "HC1"))),
                    sqrt(diag(vcovHC(mult.mod.1.3, type = "HC1"))))
```

13. [10 points] Learning Objective 4.7: Summarize regression results in a table using stargazer (including robust standard errors)

Use stargazer to generate a table with the results (Hint: make sure you include type = "latex' (if you are knitting a pdf) or"html" (if you are knitting an html) within stargazer()):

Warning: Include 3 regressions per Stargazer table. Thus, if you are estimating more models, you need to separate the rob_se and the stargazer codes.

```
stargazer(mult.mod.1.0, mult.mod.1.1, mult.mod.1.2, mult.mod.1.3,
digits = 3,
header = FALSE,
type = "latex",
se = rob_se2.1.0,
title = "OLS of Determinants of Infant Mortality Rate in India",
model.numbers = FALSE,
column.labels = c("(1)", "(2)", "(3)",  "(4)"),
single.row = TRUE, # to put coefficients and standard errors on same line
no.space = TRUE, # to remove the spaces after each line of coefficients
olumn.sep.width = "5pt", # to reduce column width
font.size = "small",
column.sep.width = "-15pt")# to make font size smaller
```

14. [7 points] Learning Objective 4.8 Assess the goodness of fit of the regression

Complete the following bullet points describing the goodness of fit of the regression and state which regression is better. If you estimated more than 3 regressions, please specify model name for each bullet point. Add bullet points as needed. Only compare models with the same Y.

- (I): Model 2

Table 1: OLS of Determinants of Infant Mortality Rate in India

	Dependent variable:			
	c_IM			
	(1)	(2)	(3)	(4)
e_spm_mean	0.009 (0.006)	0.120^{***} (0.005)		
e_so2_mean		-1.219^{***} (0.058)		
e_no2_mean		-0.013 (0.019)		
lnfcoli			-1.085^{***} (0.188)	-1.849^{***} (0.303)
bod				-0.301 (0.193)
do				7.064^{***} (0.495)
pop_urban	-0.001^{***} (0.0001)	-0.003^{***} (0.0001)	0.001^{***} (0.0002)	0.005^{***} (0.001)
year	0.012^{***} (0.001)	0.016^{***} (0.0002)	0.010^{***} (0.0005)	-0.015^{***} (0.001)
Observations	529	529	529	529
R^2	0.925	0.971	0.764	0.812
Adjusted R^2	0.925	0.971	0.763	0.810
Residual Std. Error	6.194 (df = 526)	3.839 (df = 524)	10.260 (df = 526)	9.181 (df = 524)
F Statistic	$2{,}168.569^{***}$ (df = 3; 526)	$3{,}555.695^{***}$ (df = 5; 524)	568.032^{***} (df = 3; 526)	452.136^{***} (df = 5; 524)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2: OLS of Determinants of Infant Mortality Rate in India

5pt

- (II): Model 4
- (III): Model 1
- (IIII): Model 3
- The best regression is: The best regression is model 2 with an adjusted R2 of 0.971

15. Learning Objective 4.9: Perform Hypothesis Testing:

15.1 [5 points]

Use `linearHypothesis()` to test simultaneously whether the two control variables should be in one of your models from question 11

```
linearHypothesis(mult.mod.1.1, c("pop_urban=0", "year=0") , white.adjust = "hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## pop_urban = 0
## year = 0
##
## Model 1: restricted model
## Model 2: c_IM ~ e_spm_mean + e_so2_mean + e_no2_mean + pop_urban - 1 +
##     year - 1
##
## Note: Coefficient covariance matrix supplied.
##
```

```
##   Res.Df Df      F    Pr(>F)
## 1    526
## 2    524  2 5382.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

15.2 [2 points] Interpret the results from the test above. Should you keep both variables?

This is the overall regression F-statistic and the null hypothesis is obviously different from testing when the control variables are zero. We should keep both variables as they both add a high level of significance to the model.

16. Learning Objective 4.10 Interpret the Coefficients (or Average Marginal Effects) from the best model

16.1 [2 points] Can you interpret the estimated coefficients of your best model? Why?

Yes, because each variable has significant coeffecient values especially e_spm_mean and e_so2_mean because they have allowed me to reject the null hypothesis.

16.2 [25 points] Interpretation

Interpret at least 5 estimated coefficients or AMEs from your best model (Think about your research question).

Write a sentence explaining the intuition behind the result. Complete the following bullet points with your answers. Replace `x` with the name of the variable you are interpreting. For example, `age`. Consider the model specification in your interpretation (e.g. level-level, level-log, log-level, and log-log)

```
mult.mod.1.1
```

```
##
## Call:
## lm(formula = c_IM ~ e_spm_mean + e_so2_mean + e_no2_mean + pop_urban -
##     1 + year - 1, data = air, subset = (as.factor(state)))
##
## Coefficients:
## e_spm_mean  e_so2_mean  e_no2_mean   pop_urban        year
##   0.119990   -1.219331   -0.012742   -0.003133    0.015993
```

```
cor(x = air$e_spm_mean, y=air$c_IM)
```

```
## [1] 0.05386448
```

(Infant mortality rate .1% change means 1 infant per 1000) - `x`: e_spm_mean

```
-   interpretation: For every 1 unit decrease in e_spm_mean, the infant mortality rate decreases by .1%
-   intuition: Particulate matter levels have had a positive correlation with infant mortality rate.
```

- `x`: e_so2_mean
  - interpretation: For every 1 unit decrease in e_so2_mean, the infant mortality rate increases by .1%.
  - intuition: Dulfur dioxide levels have had a negative correlation with infant mortality rate.
- `x`: e_no2_mean
  - interpretation: For every 1 unit decrease in e_no2_mean, the infant mortality rate increases by .1%.
  - intuition: Nitrous oxide levels have had a negative correlation with infant mortality rate.
- `x`: pop_urban
  - interpretation: For every 1 unit decrease in urban population, the infant mortality rate increases by .1%.
  - intuition: Urban population levels have had a negative correlation with infant mortality rate.
- `x`: year
  - interpretation: For every 1 unit change in years the infant mortality rate decreases by .1%.
  - intuition: The dates have had a positive correlation with infant mortality rate.

17. [9 points] Learning Objective 4.11. Discuss potential sources of biases in your analysis:

The potential biases that occured during this research

- Potential Bias 1: Selection Bias

The potential for a selection bias is evident because the infant mortality rates vary greatly from the state level surveys to the vital statistics data while being highly correlated.

- Potential Bias 2: Downward Bias

The potential for a downward bias is further evident because most infant deaths are not reported in India, with research also suggesting the levels are much higher. Which creates a equality gap that is represented in a bias distribution of partial observations that ultimately affect the individual outcomes.

- Potential Bias 3: Confounding Bias

The reason a confounding bias might exist is because there is almost always a systematic discrepancy in the measures of exposure effects and public health outcomes caused by the interchanging effect of the exposure to external risk factors. This can be assumed to be happening through the effect at which infant mortality rate has gone down because of better air quality but is volatily getting worst with poot water quality.

18. [5 points] Learning Objective 4.12: Summarize the main takeaways from the analysis.

Research question answer: Air pollution does contribute more to increased infant mortality rates than water pollution.

The interesting thing I noticed first was the drastic decreasing trend in Indias infant mortality rate as well as air pollution. This lead me to find out that worse air quality ontop of India's rising heat levels exacerbates infant mortality rates the most, so the correlation between air pollution and infant deaths decreasing is no coincidence. I went on to estimate whether the water quality levels were being realized as much as the air quality was and from creating to models to estimate infant mortality, the air pollution model ended up have a goodness of fit of 0.971 as opposed to a 0.81 for the water pollution model. This pushed me further to see where is being the most affected and lead me to seeing how much it varys between large and small states in India. Specifically, seeing the fact that Delhi has the worst water quality yet a rather steadily low infant mortality rate. Furthermore, looking into places like Himachal Pradesh where air quality has gotten worse on average, yet infant mortality rate is decreasing. This has encouraged me to research further the low-tech solutions we can implement to close the gap of unreported and unobserved public health deaths in relation to increasing levels of pollution and climate hazard.