Final Report

due November 16, 2021 by 11:59 PM

Kaitlyn Lewars and Katie Meehl: The Exposure Experience

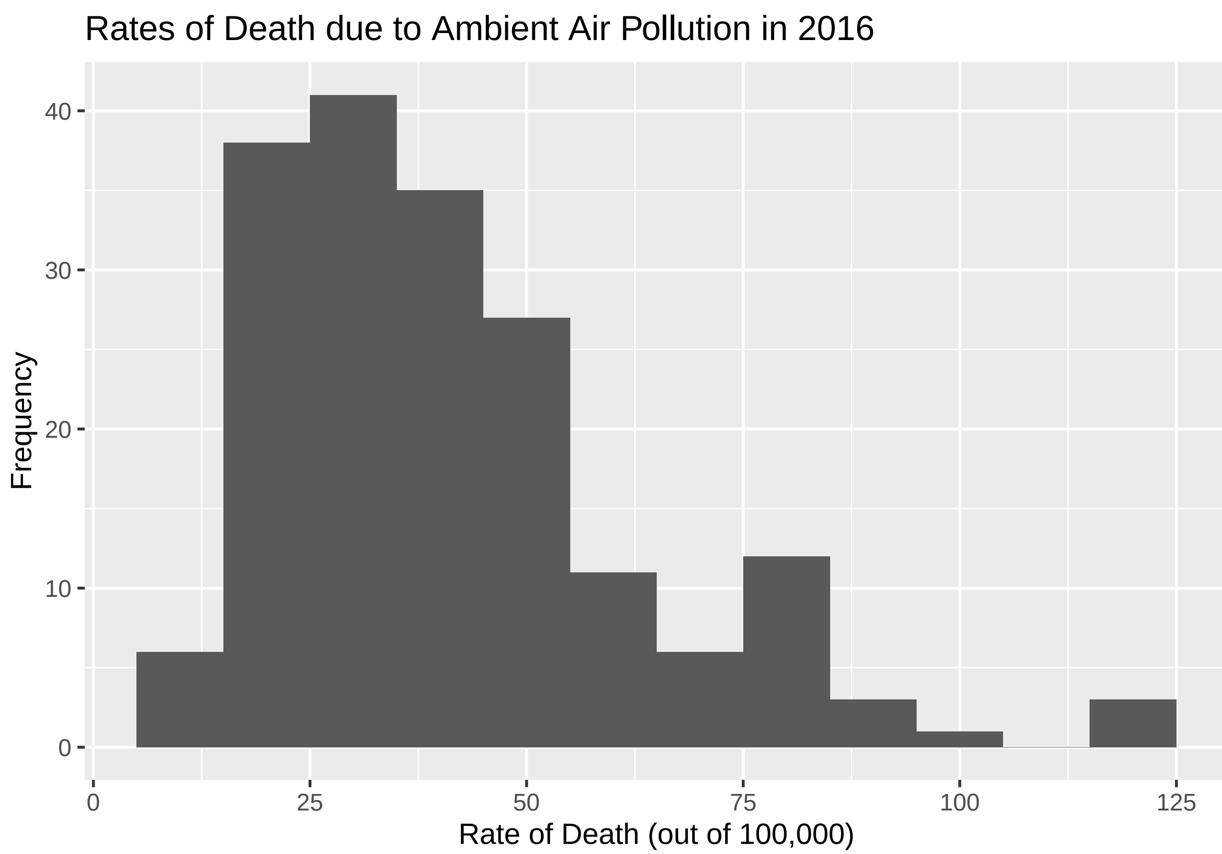
November 16, 2021

Climate change has been a recurring topic in the news in recent years as it becomes a more pressing problem. One of the important factors of climate change is air pollution. In 2017, air pollution was the 4th leading cause of mortality and the 5th leading cause of morbidity worldwide. As air pollution is a leading cause of morbidity and mortality, we thought it would be important to explore a data set investigating this problem. [Edit, add sources]

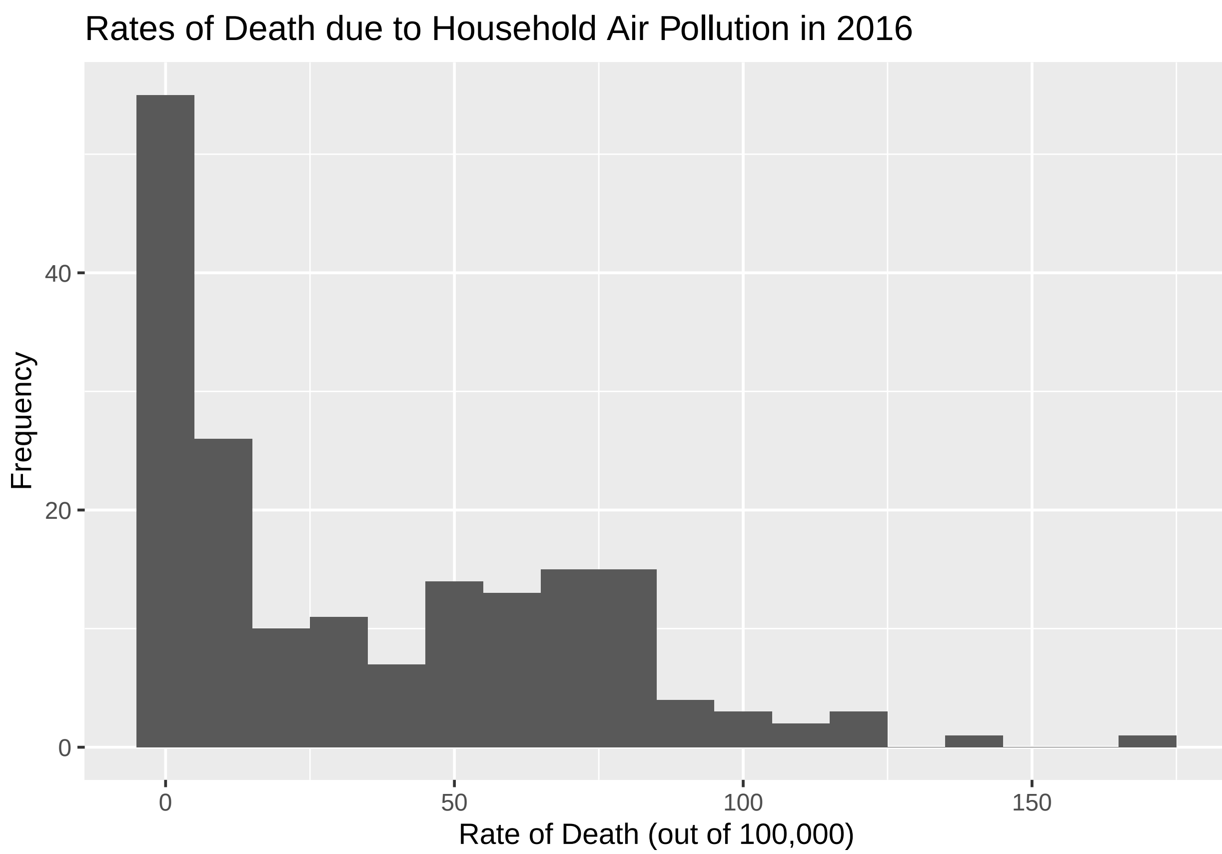
In general we would like to investigate air pollution as a cause of mortality globally. There are several different types of air pollution, but we will look at household pollution and ambient matter pollution. With these two variables we will compare them to see which air pollution is the most fatal. We would also like to look into the trend of air pollution over the last 27 years. Lastly we would like to compare air pollution as a risk factor to other common risk factors. We downloaded this data from the World Health Organization Data Collections. There are several variables in this data including year, country, deaths by each type of air pollution, and deaths by other risk factors.

The data collection is a bit complicated. In order to estimate deaths caused by pollution they use "mathematical functions, derived from epidemiological studies from countries around the world, that relate different levels of exposure to the increased risk of death or disability from each cause, by age and sex, where applicable, estimates of population exposure to PM2.5, ozone, and household air pollution, country-specific data on underlying rates of disease and death for each pollution-linked disease, and a comprehensive set of population data, adjusted to match the UN2015 Population Prospectus and obtained from the Gridded Population of the World (GPW) database for each country" (https://www.stateofglobalair.org/data/estimate-burden).

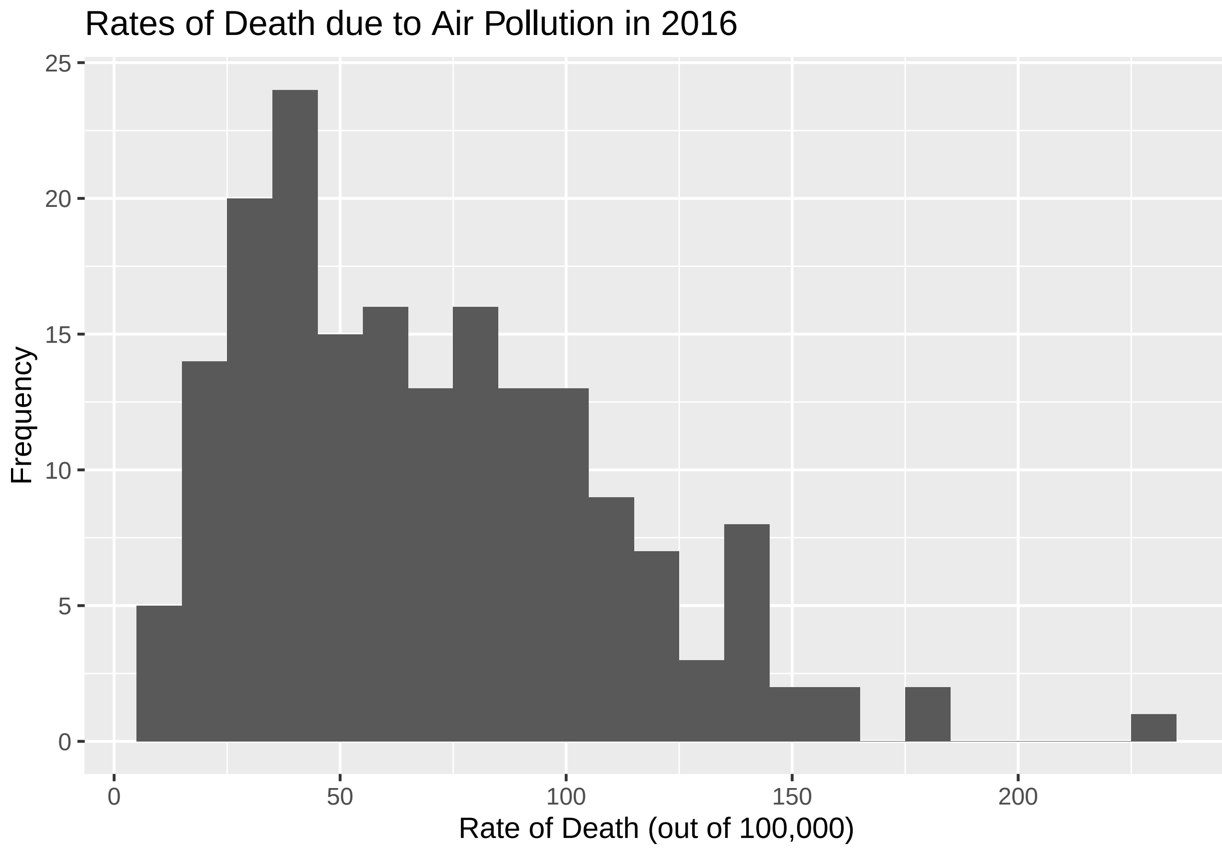
```
Ambient_Air_Pollution_Rate %>%
  filter(Dim2 == "Total", Dim1 == "Both sexes") %>%
  ggplot(aes(x = FactValueNumeric)) +
  geom_histogram(binwidth = 10) +
  scale_x_continuous(labels = label_comma()) +
  labs(x = "Rate of Death (out of 100,000)",
       y = "Frequency",
       title = "Rates of Death due to Ambient Air Pollution in 2016")
```




```
Household_Air_Pollution_Rate %>%
  filter(Dim2 == "Total", Dim1 == "Both sexes") %>%
  ggplot(aes(x = FactValueNumeric)) +
  geom_histogram(binwidth = 10) +
  scale_x_continuous(labels = label_comma()) +
  labs(x = "Rate of Death (out of 100,000)",
       y = "Frequency",
       title = "Rates of Death due to Household Air Pollution in 2016")
```




```
Total_Air_Pollution_Death_Rate %>%
  filter(Dim2 == "Total", Dim1 == "Both sexes") %>%
  ggplot(aes(x = FactValueNumeric)) +
  geom_histogram(binwidth = 10) +
  scale_x_continuous(labels = label_comma()) +
  labs(x = "Rate of Death (out of 100,000)",
       y = "Frequency",
       title = "Rates of Death due to Air Pollution in 2016")
```



The first thing we wanted to look at was the frequency of higher death rates due to the ambient and household air pollution. As seen in the first visualization showing the rates of death due to Ambient Air Pollution, there tends to be a great amount of countries that have death rates of around 25-30 out of every 100,000 in their population, with very few countries have less than 15 or greater than 75 deaths out of every 100,000.

This is much more alarming than the household air pollution death rates, which tend to center around 0 for most countries. However, this visualization also shows quite a few countries that have death rates between 50 and 100 deaths out of every 100,000. Our third visualization shows the total deaths due to air pollution in selected countries around the globe in 2016. Here, we are able to see the sheer amount of countries that have had rates of roughly 50 up to roughly 150 deaths out of 100,000, indicating a serious global issue.

Out of these three plots, there are four countries that should be noted: Democratic People's Republic of Korea, Georgia, Chad, and Bosnia and Herzegovina. These countries are outliers and show much higher death rates due to air pollution than other countries.

```
joinedambient1 <- Ambient_Air_Pollution_Rate %>%
  rename(AmbientDeathRate = FactValueNumeric) %>%
  rename(Sex = Dim1) %>%
  rename(CauseofDeath = Dim2) %>%
  select(Location, Sex, CauseofDeath, AmbientDeathRate) %>%
  left_join(Ambient_Air_Pollution_Total_Deaths) %>%
  select(Location, Sex, CauseofDeath, AmbientDeathRate, FactValueNumeric) %>%
filter(CauseofDeath == "Total") %>%
  mutate(totalpopulation = (100000*FactValueNumeric)/AmbientDeathRate) %>%
  rename(Totaldeathsambient = FactValueNumeric)
```

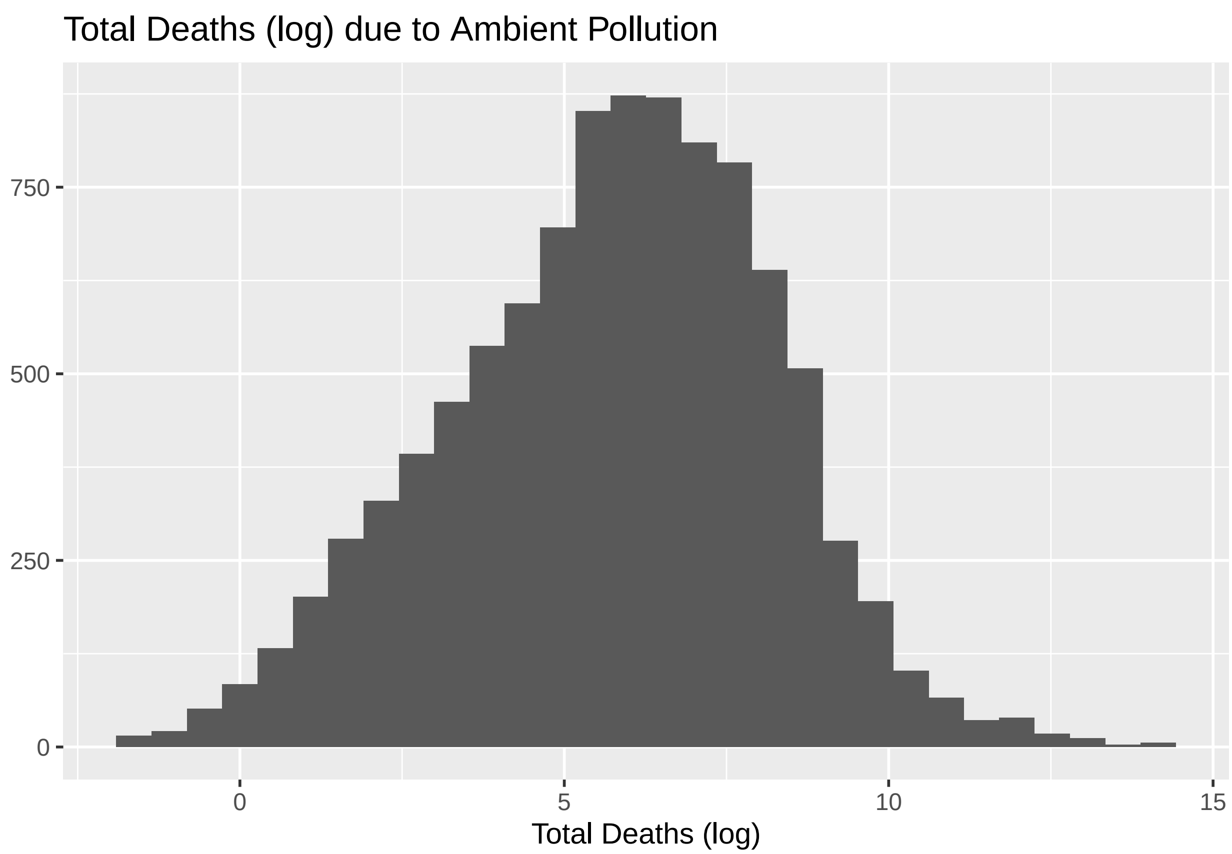
```
ambient <- Particulate_Ambient_Concentration %>%
  rename(AmbientAirConcentration = FactValueNumeric) %>%
```

```
  select(Location, AmbientAirConcentration) %>%
  left_join(joinedambient1, by = "Location") %>%
  select(Location, Sex, CauseofDeath, AmbientDeathRate, Totaldeathsambient, AmbientAirConcentration, to
filter(CauseofDeath == "Total", Sex == "Both sexes")
```

Here, we have created a new datatset to explore our discovery that ambient air pollution had a higher average death rate globally than household air pollution. This includes the total deaths due to to ambient pollution, the rates at which countries have had deaths due to ambient air pollution, ambient air concentration, and other factors which will help understand how ambient air pollution affect countries around the world.

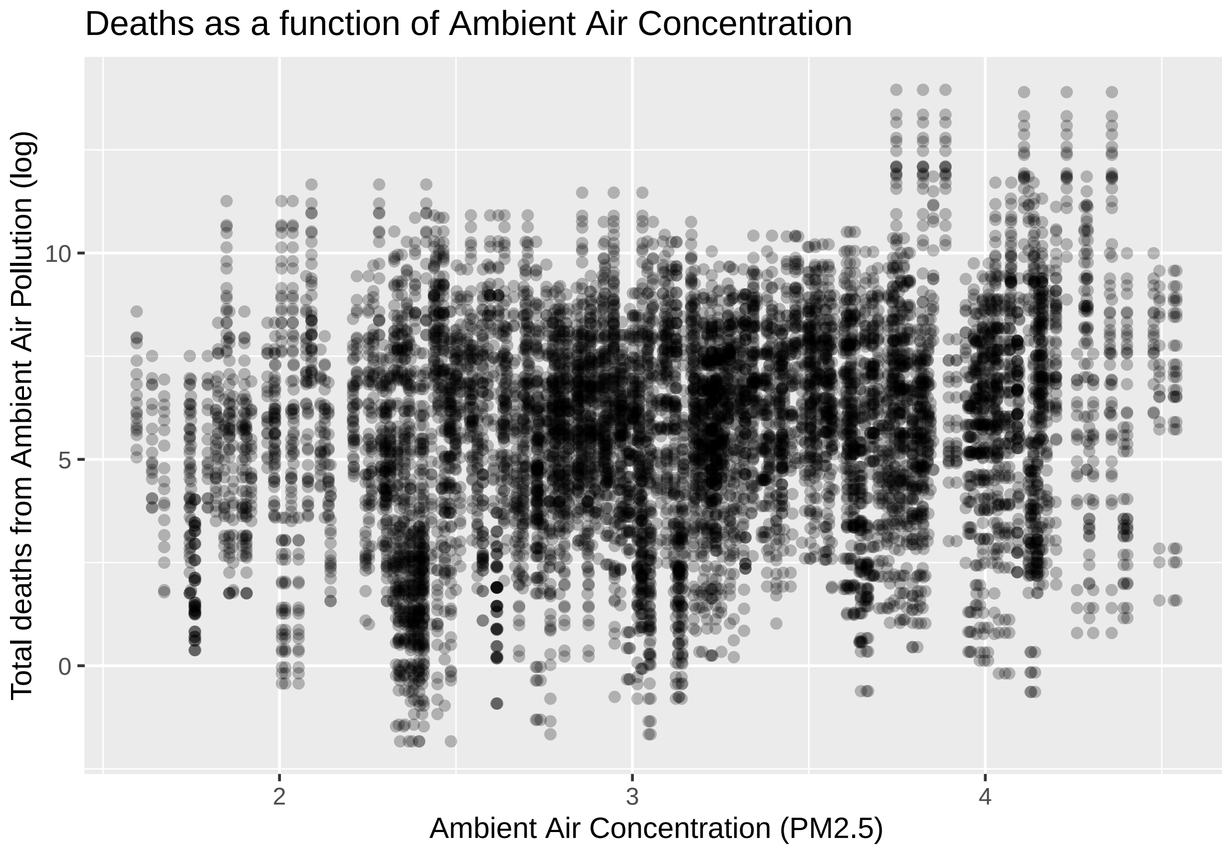
```
ggplot(data = ambient, aes(x = log(Totaldeathsambient))) +
  geom_histogram() +
  labs(x = "Total Deaths (log)", y = NULL, title="Total Deaths (log) due to Ambient Pollution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



[Insert explanation]

```
ggplot(data = ambient, aes(x = log(AmbientAirConcentration), y = log(Totaldeathsambient))) +
geom_point(alpha = 0.25) +
labs(title = "Deaths as a function of Ambient Air Concentration",
x = "Ambient Air Concentration (PM2.5)",
y = "Total deaths from Ambient Air Pollution (log)")
```

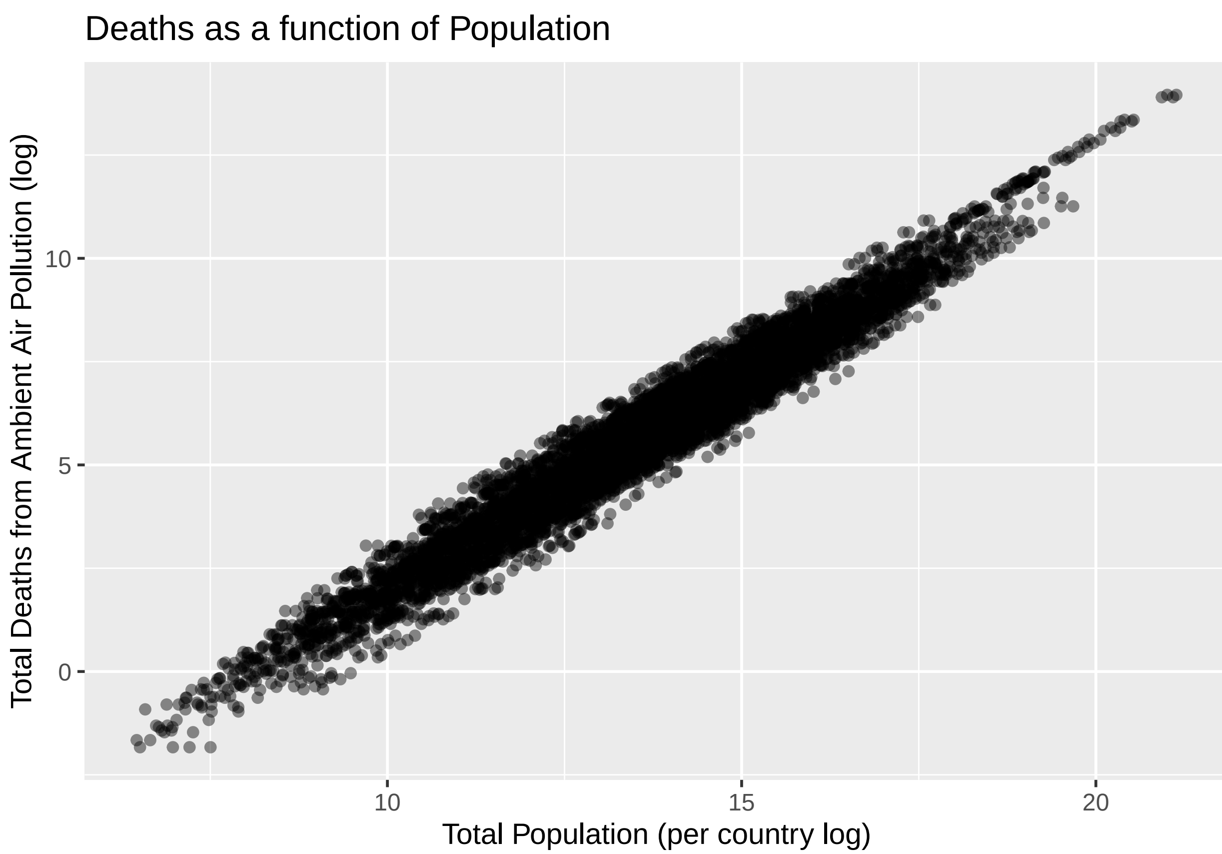


We want to explore the different factors affecting deaths due to ambient air pollution. A potential factor in the rise in environmental air pollution could be the concentration of air pollution. In the plot above, we examine the connection between total deaths from ambient air pollution vs. ambient air concentration in fine particulate matter, both of which are on a log scale for normalization of the data. Although slightly centered around the middle, the plot clearly shows a lack of strong relationship between these two variables, indicating that there are likely other factors influencing the data.

```
ambient2 <- Particulate_Ambient_Concentration %>%
  rename(AmbientAirConcentration = FactValueNumeric) %>%
  rename(Residencetype = Dim1) %>%
  select(Location, AmbientAirConcentration, Residencetype) %>%
  left_join(joinedambient1, by = "Location") %>%
  select(Location, Sex, CauseofDeath, AmbientDeathRate, Totaldeathsambient, AmbientAirConcentration, tot
filter(CauseofDeath == "Total", !Sex %in% c("Both sexes"), !Residencetype %in% c("Total"))
```

[Explain]

```
ggplot(data = ambient2, aes(x = log(totalpopulation), y = log(Totaldeathsambient))) +
geom_point(alpha = 0.25) +
  labs(title = "Deaths as a function of Population",,
        x = "Total Population (per country log)",
        y = "Total Deaths from Ambient Air Pollution (log)")
```



For the next 4 sections we created two different models for a linear regression. The first one we did just based off the outdoor air concentration. The equation we got was deaths = -2601.0 + 274.1(AirConcentration). The air concentration is in PM2.5.

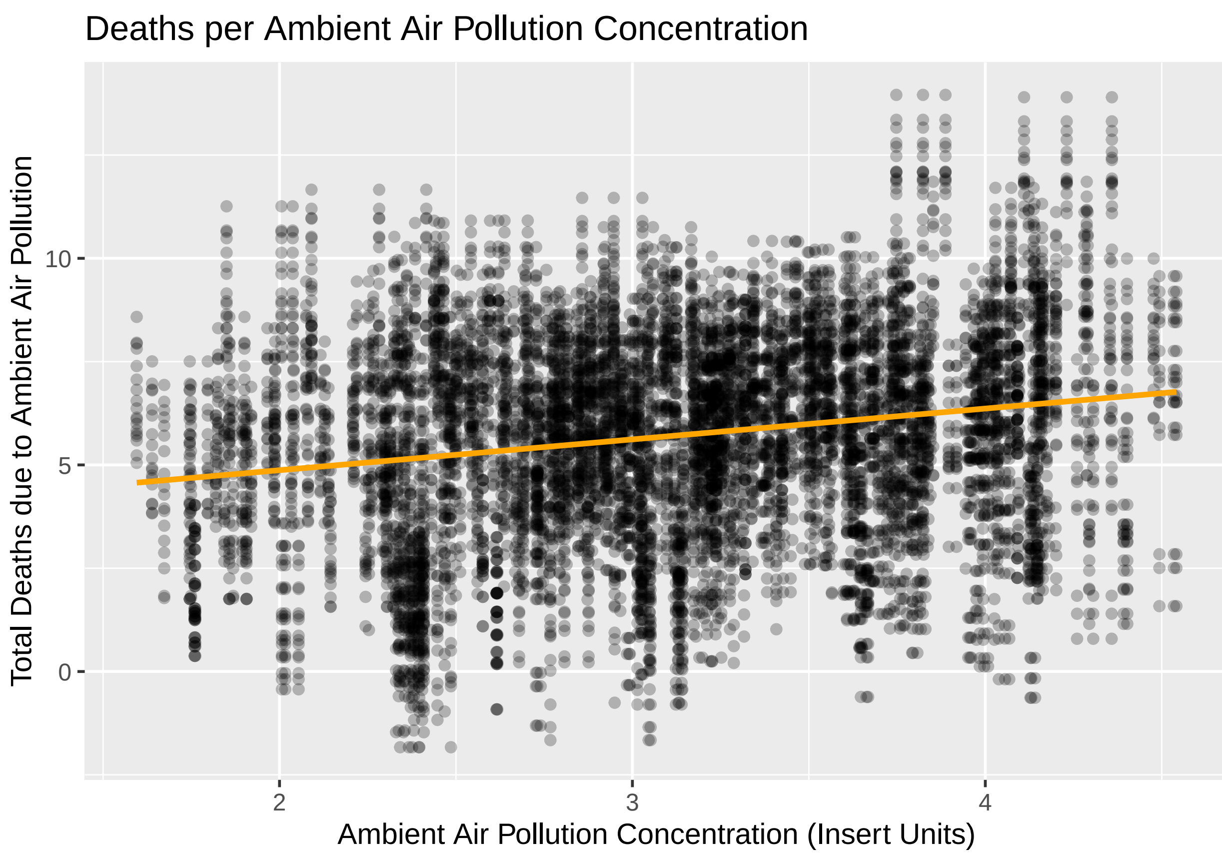
```
death_ambientairpol <- linear_reg() %>%
  set_engine("lm") %>%
  fit(log(Totaldeathsambient) ~ log(AmbientAirConcentration), data = ambient)
  tidy(death_ambientairpol, conf.int=TRUE, exponentiate=TRUE)
```

```
## # A tibble: 2 x 7
##   term                      estimate std.error statistic   p.value conf.low conf.high
##   <chr>                        <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)                  3.38      0.120      28.2 7.31e-169    3.15      3.62
## 2 log(AmbientAirConce~         0.746     0.0374     20.0 6.29e- 87    0.673     0.819
```

deaths = 4.979.0 + 0.0261(AirConcentration)

[evaluate what this means]

```
ggplot(data = ambient, aes(x = log(AmbientAirConcentration), y = log(Totaldeathsambient))) +
  geom_point(alpha = 0.25) +
  geom_smooth(method = "lm", se = FALSE, color = "orange") +
  labs(title = "Deaths per Ambient Air Pollution Concentration",
       x = "Ambient Air Pollution Concentration (Insert Units)",
       y = "Total Deaths due to Ambient Air Pollution")
```



[Add narrative about the graph - most deaths are concentrated at mid concentrations - lethal dose, what are the outliers?]

```
glance(death_ambientairpol)$r.squared
```

```
## [1] 0.03876003
```

```
death_ambientairpol_totalpopulation <- linear_reg() %>%
    set_engine("lm") %>%
    fit(log(Totaldeathsambient) ~ log(AmbientAirConcentration) + Sex + Residencetype, data = ambient2)

tidy(death_ambientairpol_totalpopulation, conf.int=TRUE, exponentiate=TRUE)
```

```
## # A tibble: 4 x 7
##   term                 estimate std.error statistic   p.value conf.low conf.high
##   <chr>                   <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)          3.45e+ 0    0.106   3.26e+ 1 7.91e-224   3.24      3.66
## 2 log(AmbientAirConc~  7.35e- 1    0.0322  2.28e+ 1 3.09e-113   0.672     0.798
## 3 SexMale             -1.30e-17    0.0426 -3.05e-16 1    e+  0  -0.0835    0.0835
## 4 ResidencetypeUrban  -6.70e- 2    0.0427 -1.57e+ 0 1.17e-   1  -0.151     0.0167
```

```
glance(death_ambientairpol_totalpopulation)$r.squared
```

```
## [1] 0.03808491
```

Here we created a binomial regression model to predict the likelihood of a person dying based on the level of ambient air concentration. (will add explanation later)

```
Alive <- ambient2$totalpopulation - ambient2$Totaldeathsambient
```

```r
modelagg1<-glm(cbind(round(Totaldeathsambient), round(Alive)) ~ AmbientAirConcentration, data=ambient2,

summary(modelagg1)
```

```
##
## Call:
## glm(formula = cbind(round(Totaldeathsambient), round(Alive)) ~
##     AmbientAirConcentration, family = binomial, data = ambient2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -161.24   -12.68    -3.56     0.39   430.08
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -8.005e+00  2.763e-04  -28975   <2e-16 ***
## AmbientAirConcentration  1.237e-02  5.419e-06    2282   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14254994  on 13175  degrees of freedom
## Residual deviance:  9023120  on 13174  degrees of freedom
## AIC: 9122756
##
## Number of Fisher Scoring iterations: 4
```

Here we created a binomial regression model to predict the likelihood of a person dying based on the level of ambient air concentration, gender, and where they live. (will add explanation later)

```r
Alive <- ambient2$totalpopulation - ambient2$Totaldeathsambient

modelagg2<-glm(cbind(round(Totaldeathsambient), round(Alive)) ~ AmbientAirConcentration + Sex + Residen

summary(modelagg2)
```

```
##
## Call:
## glm(formula = cbind(round(Totaldeathsambient), round(Alive)) ~
##     AmbientAirConcentration + Sex + Residencetype, family = binomial,
##     data = ambient2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -143.52   -12.84    -3.78     0.17   347.88
##
## Coefficients:
##                           Estimate Std. Error  z value Pr(>|z|)
## (Intercept)             -8.058e+00  3.130e-04 -25742.6   <2e-16 ***
## AmbientAirConcentration  1.261e-02  5.509e-06   2288.9   <2e-16 ***
## SexMale                  1.674e-01  2.427e-04    689.6   <2e-16 ***
## ResidencetypeUrban      -7.582e-02  2.458e-04   -308.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14254994  on 13175  degrees of freedom
## Residual deviance:  8452356  on 13172  degrees of freedom
## AIC: 8551996
##
## Number of Fisher Scoring iterations: 4
```

Finally we conducted a t-test to see if the impact on mortality is different depending on the type of air pollution (household vs. ambient)

```
t.test(Household_Air_Pollution_Rate$FactValueNumeric,Ambient_Air_Pollution_Rate$FactValueNumeric, var.e
```

```
##
##  Welch Two Sample t-test
##
## data:  Household_Air_Pollution_Rate$FactValueNumeric and Ambient_Air_Pollution_Rate$FactValueNumeric
## t = -3.9975, df = 6328.8, p-value = 6.474e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.8883144 -0.9876038
## sample estimates:
## mean of x mean of y
##  11.87733  13.81529
```