Final Report

due November 16, 2021 by 11:59 PM

Kaitlyn Lewars and Katie Meehl: The Exposure Experience

November 16, 2021

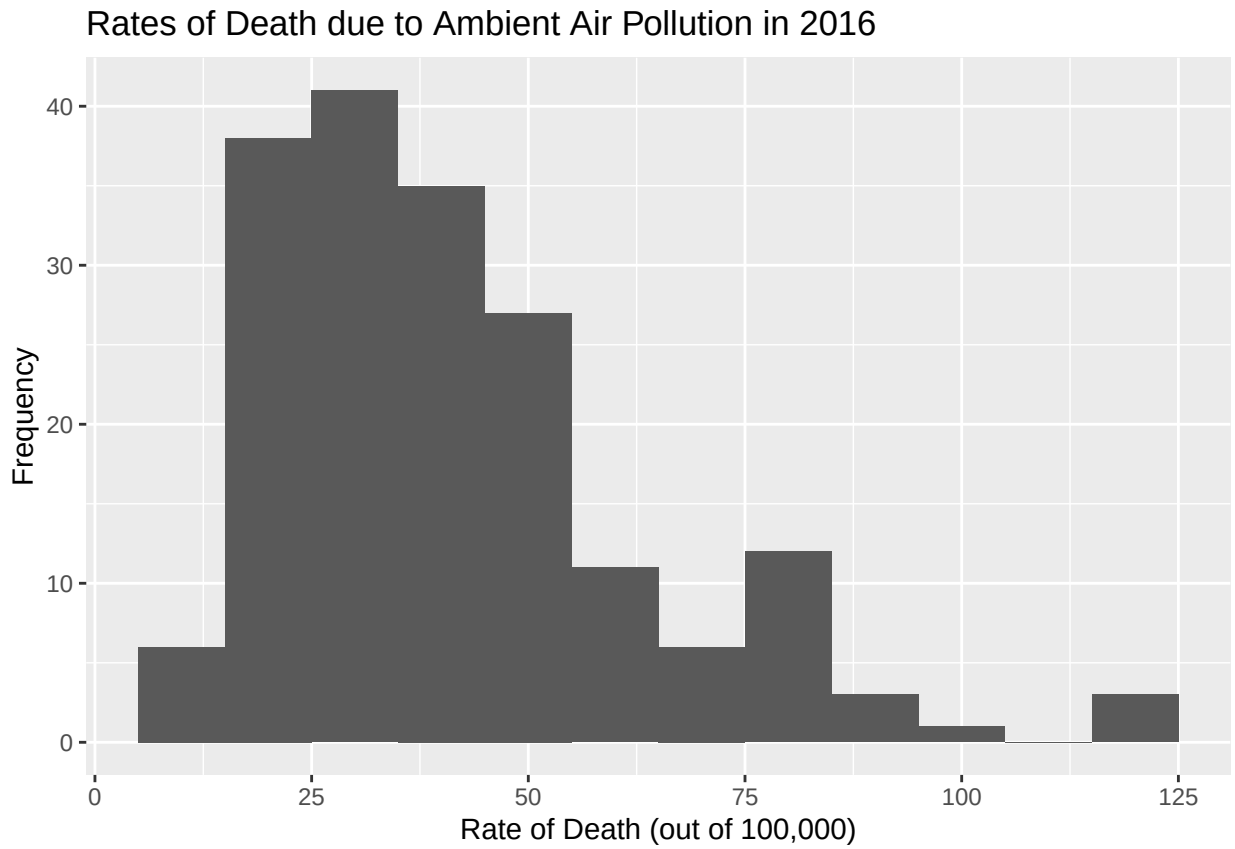
```
library(tidyverse)
library(readr)
library(scales)
library(tidymodels)
library(knitr)

Total_Air_Pollution_Death_Rate <- read_csv("~/R/Project Proposal/data/Total Air Pollution Death Rate.csv
Household_Air_Pollution_Total_Deaths <- read_csv("~/R/Project Proposal/data/Household Air Pollution Tota
Household_Air_Pollution_Rate <- read_csv("~/R/Project Proposal/data/Household Air Pollution Rate.csv")
Ambient_Air_Pollution_Rate <- read_csv("~/R/Project Proposal/data/WHO Ambient Air Pollution Rate.csv")
Ambient_Air_Pollution_Total_Deaths <- read_csv("~/R/Project Proposal/data/Ambient Air Pollution Total De
Particulate_Ambient_Concentration <- read_csv("~/R/Project Proposal/data/Ambient Particulate Concentrati
```

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.

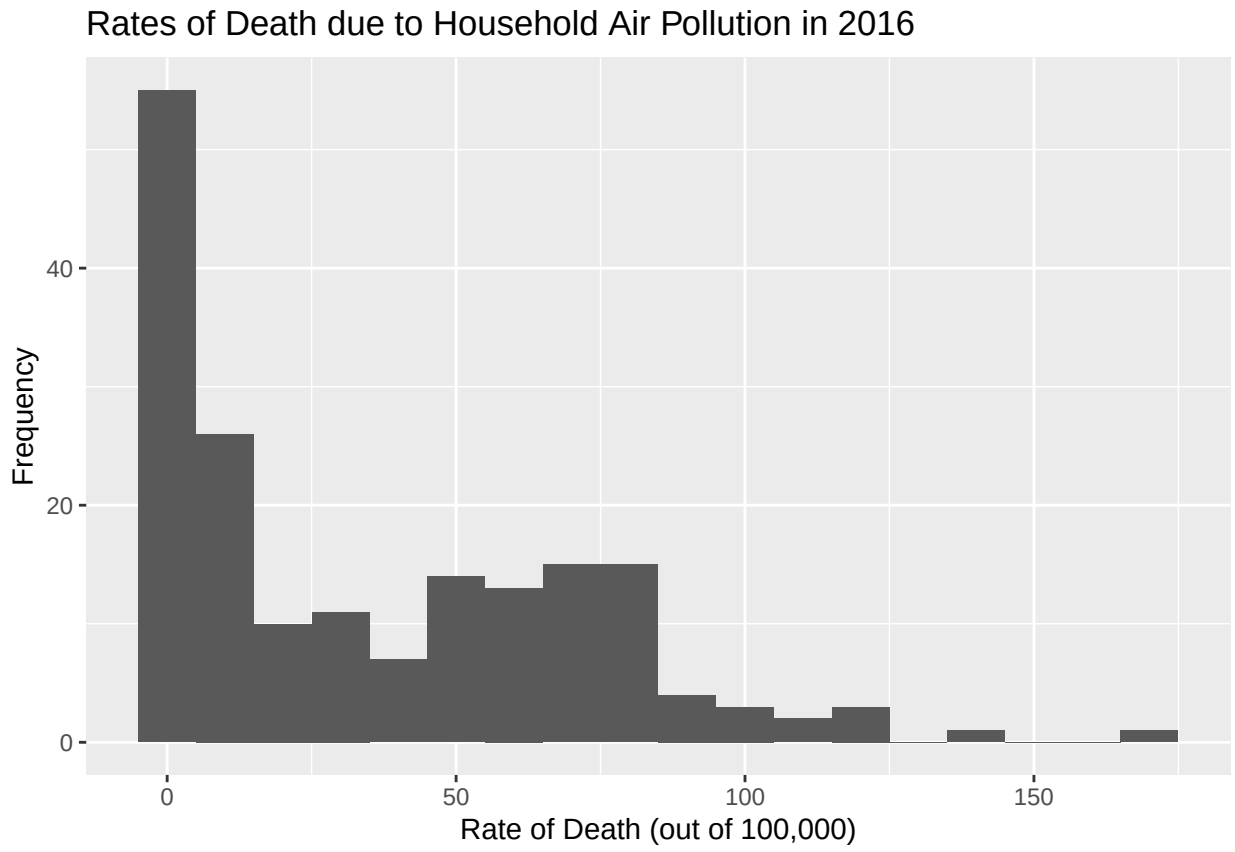
In general we would like to investigate air pollution as a cause of mortality globally. There are several different types of air pollution, but we will look at household pollution, ambient matter pollution, and ambient ozone pollution. With these variables we will compare them to see which air pollution is the most dangerous. 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 kaggle. 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).

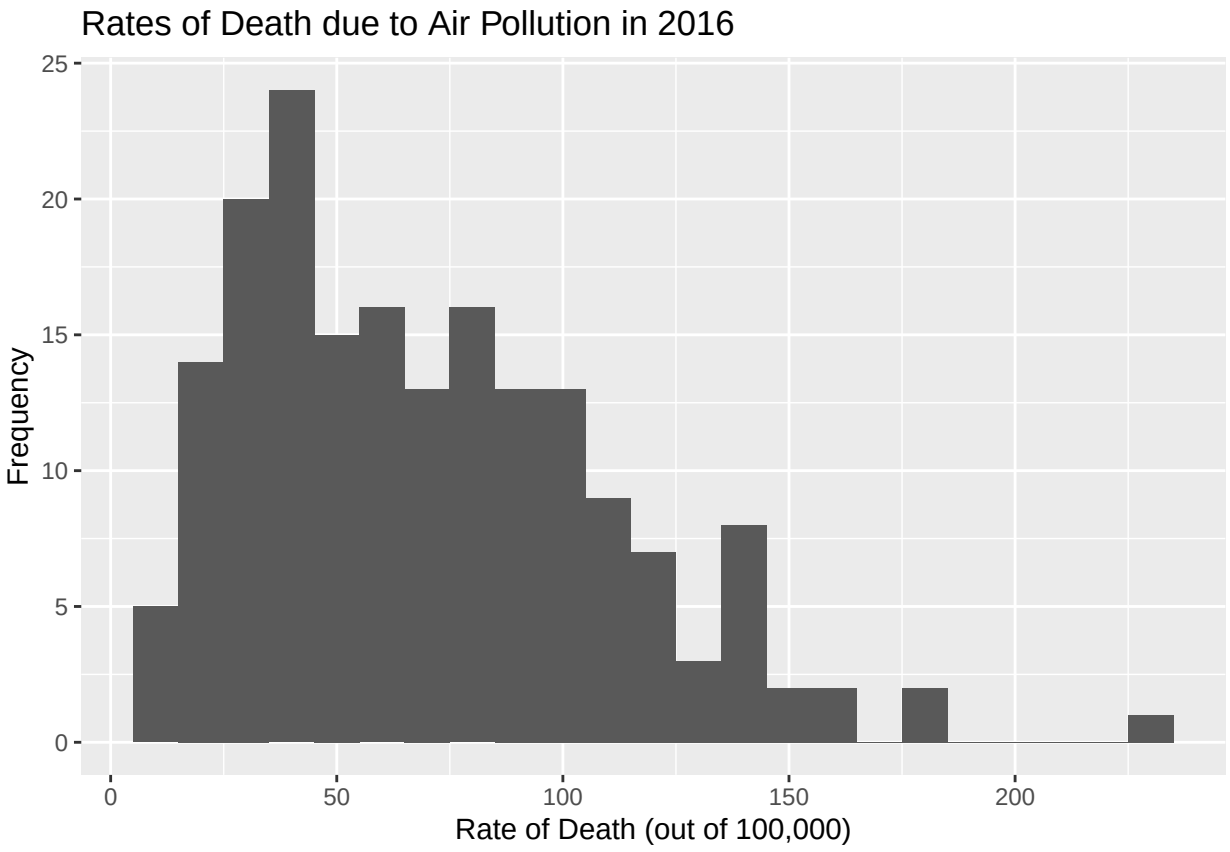
```
Ambient_Air_Pollution_Rate %>%
  filter(Dim2 == "Total", Dim1 == "Both sexes") %>%
  ggplot(aes(x = FactValueNumeric)) +
  geom_histogram(binwidth = 10) +
  scale_x_continuous(labels = label_comma()) +
  labs(x = "Rate of Death (out of 100,000)",
       y = "Frequency",
       title = "Rates of Death 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 four countries of interest are Democratic People's Republic of Korea, Georgia, Chad, and Bosnia and Herzegovina.

[Explore these?]

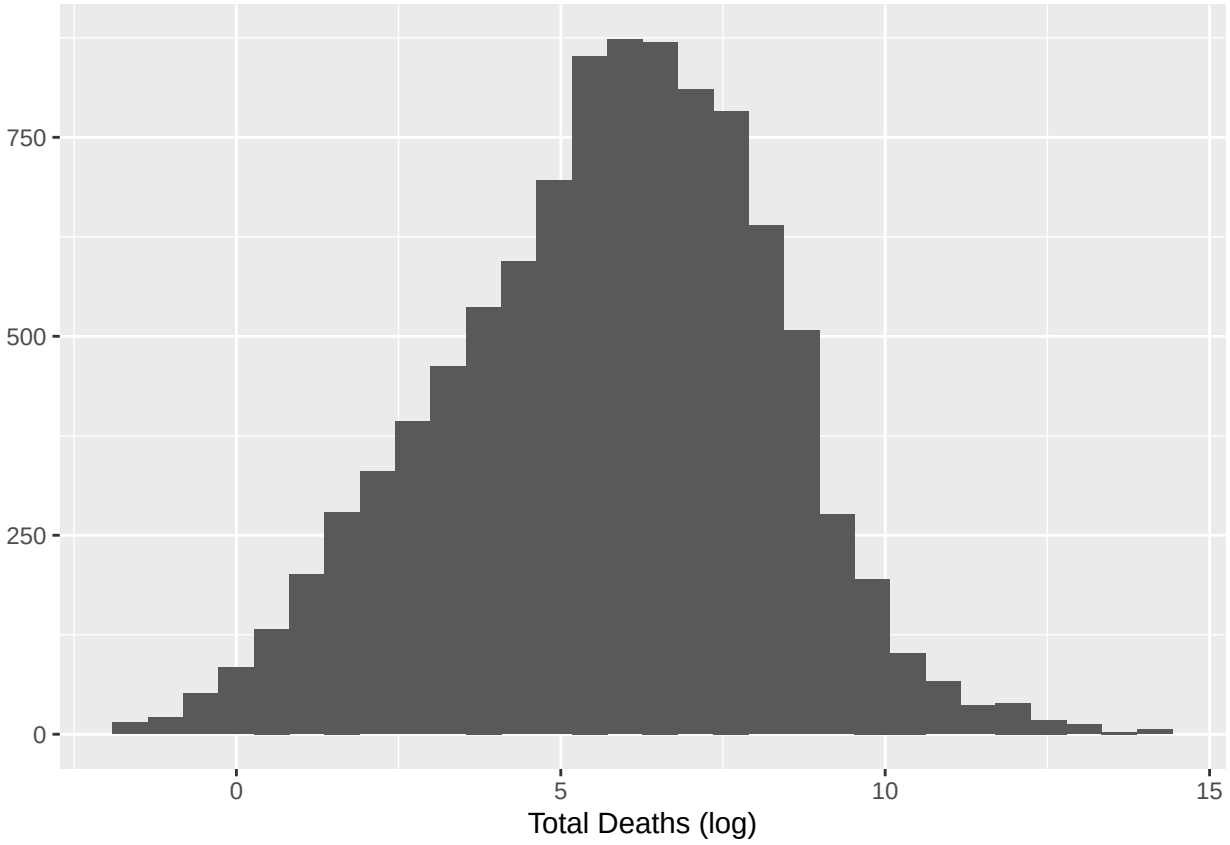
```
joinedambient1 <- Ambient_Air_Pollution_Rate %>%
  rename(AmbientDeathRate = FactValueNumeric) %>%
  rename(Sex = Dim1) %>%
  rename(CauseofDeath = Dim2) %>%
  select(Location, Sex, CauseofDeath, AmbientDeathRate) %>%
  left_join(Ambient_Air_Pollution_Total_Deaths) %>%
  select(Location, Sex, CauseofDeath, 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")
```

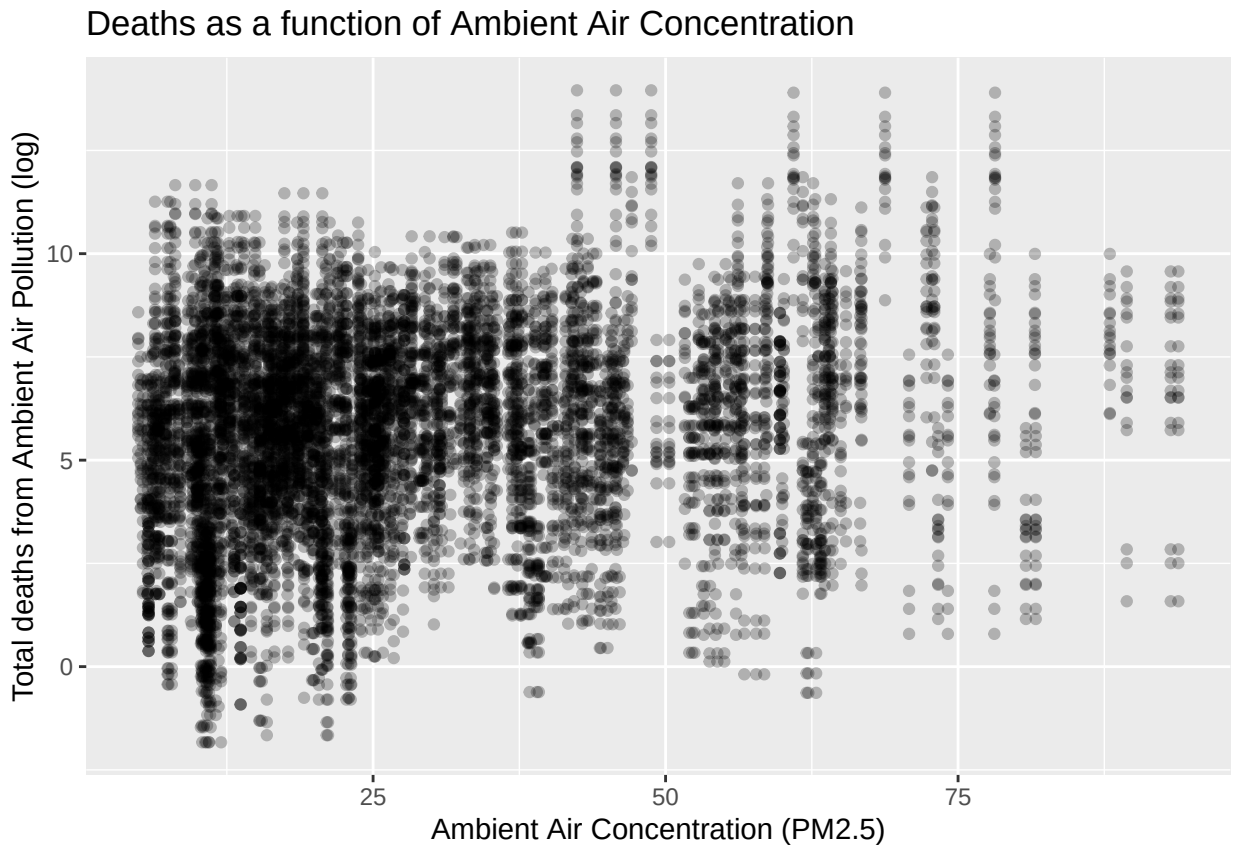
[Evaluate, explain, need to edit scale for both axes]

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

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

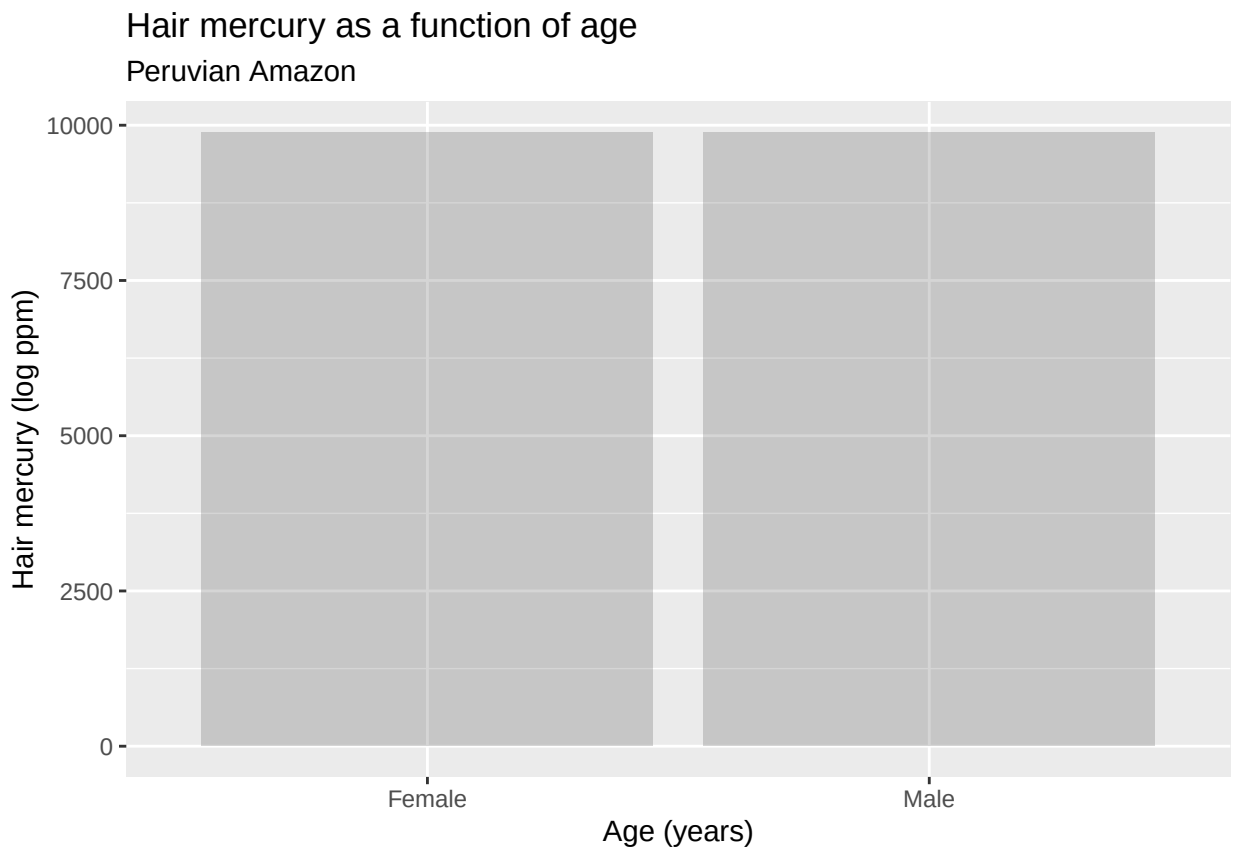



```
ggplot(data = ambient, aes(x = AmbientAirConcentration, y = log(Totaldeathsambient))) +
geom_point(alpha = 0.25) +
labs(title = "Deaths as a function of Ambient Air Concentration",,
x = "Ambient Air Concentration (PM2.5)",
y = "Total deaths from Ambient Air Pollution (log)")
```

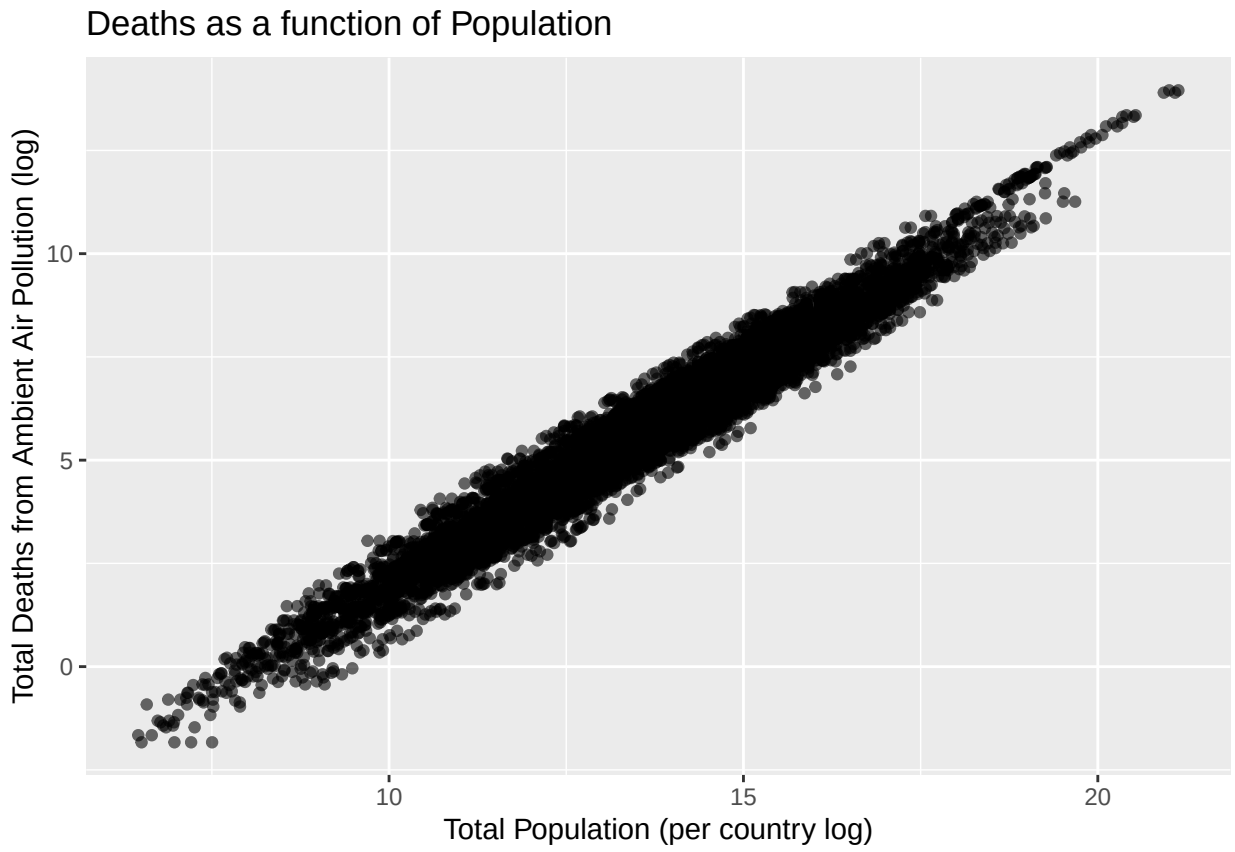



```
ambient2 <- Particulate_Ambient_Concentration %>%
  rename(AmbientAirConcentration = FactValueNumeric) %>%
  rename(Residencetype = Dim1) %>%
  select(Location, AmbientAirConcentration, Residencetype) %>%
  left_join(joinedambient1, by = "Location") %>%
  select(Location, Sex, CauseofDeath, AmbientDeathRate,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

```
ggplot(data = ambient2, aes(x = Sex)) +
geom_bar(alpha = 0.25) +
labs(title = "Hair mercury as a function of age",
subtitle = "Peruvian Amazon",
x = "Age (years)",
y = "Hair mercury (log ppm)")
```




```
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.

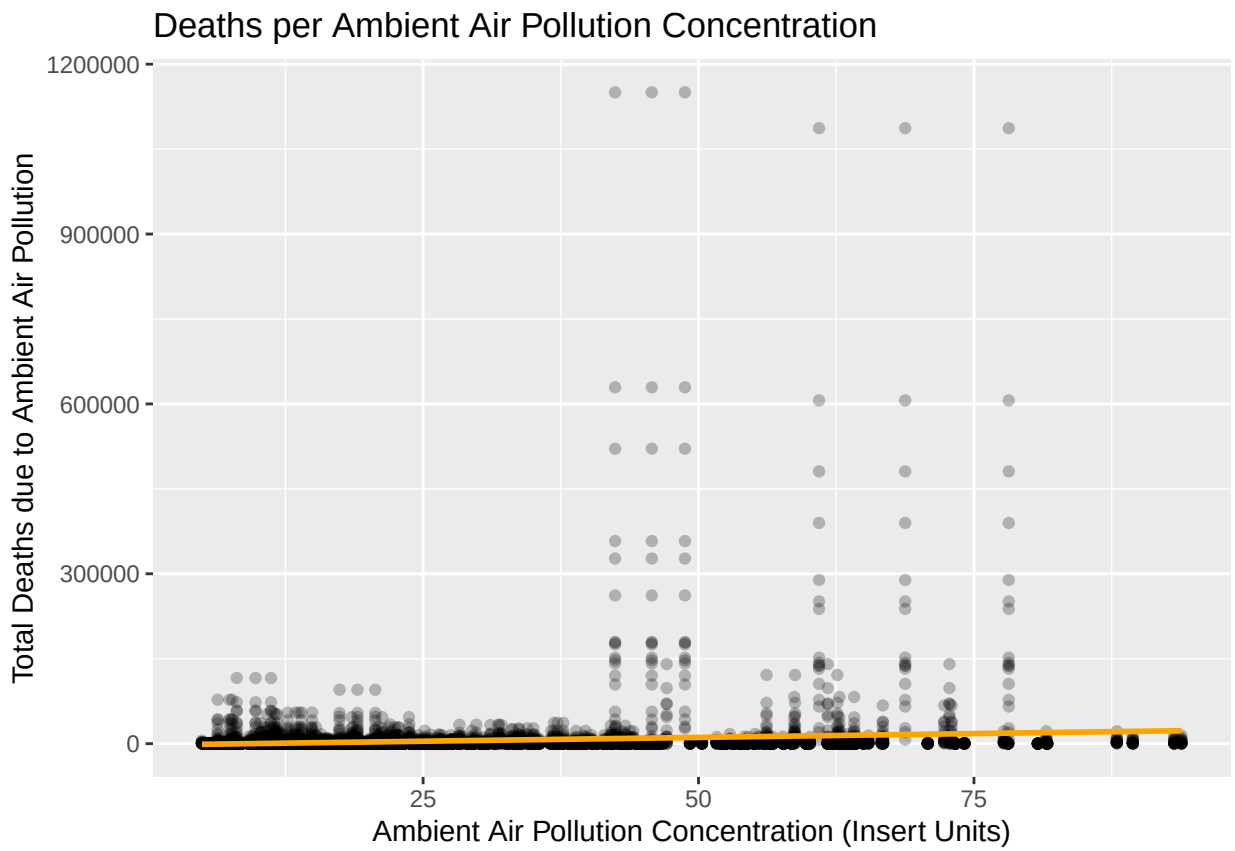
```
death_ambientairpol <- linear_reg() %>%
  set_engine("lm") %>%
  fit(log(Totaldeathsambient) ~ AmbientAirConcentration, data = ambient)
  tidy(death_ambientairpol, conf.int=TRUE, exponentiate=TRUE)
```

```
## # 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)               4.99     0.0454      110.  0          4.90      5.08
## 2 AmbientAirConcentration   0.0259   0.00135      19.2 8.51e-81   0.0232    0.0285
```

deaths = 4.979.0 + 0.0261(AirConcentration)

[evalue what this means]

```
ggplot(data = ambient, aes(x = AmbientAirConcentration, y = 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 lower concentrations - lethal dose, what are the outliers?]

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

```
## [1] 0.03601614
```

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

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

```
## # A tibble: 5 x 7
##   term                      estimate std.error statistic   p.value conf.low conf.high
##   <chr>                        <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)              5.01e+ 0  0.0435    1.15e+ 2 0           4.92      5.10
## 2 AmbientAirConcentration  2.59e- 2  0.000953  2.72e+ 1 6.41e-160   0.0241    0.0278
## 3 SexMale                  2.86e-17  0.0348    8.21e-16 1.00e+  0  -0.0683    0.0683
## 4 ResidencetypeTotal      -2.55e- 2  0.0427   -5.98e- 1 5.50e-  1  -0.109     0.0581
## 5 ResidencetypeUrban      -4.89e- 2  0.0427   -1.15e+ 0 2.52e-  1  -0.133     0.0347
```

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

```
## [1] 0.03608026
```

```
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
```

[Interpret, write formula, 96.5% variance explained]