Final Report

due November 16, 2021 by 11:59 PM

Kaitlyn Lewars and Katie Meehl: The Exposure Experience

November 16, 2021

Background and Significance

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. Air pollution refers to the release of pollutants into the air, To be more specific pollutants are gases and chemicals that are released in the air usually due to energy use and production. Air pollutants like carbon dioxide and methane raise the earth's temperature which contributes to climate change. Other air pollutants like smog and ozone are worsened by increased heat. Air pollution has varying effects on the human body depending on the type of air pollutant and length of exposure. However we do know that overall air pollution is detrimental to human health and the planet as a whole. In 2017, air pollution was the 4th leading cause of mortality and the 5th leading cause of morbidity worldwide. Currently, WHO estimates that 9 out of 10 humans breathe air that exceeds the WHO's guideline limits for pollutants, with those living in low- and middle-income countries suffering the most. 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.

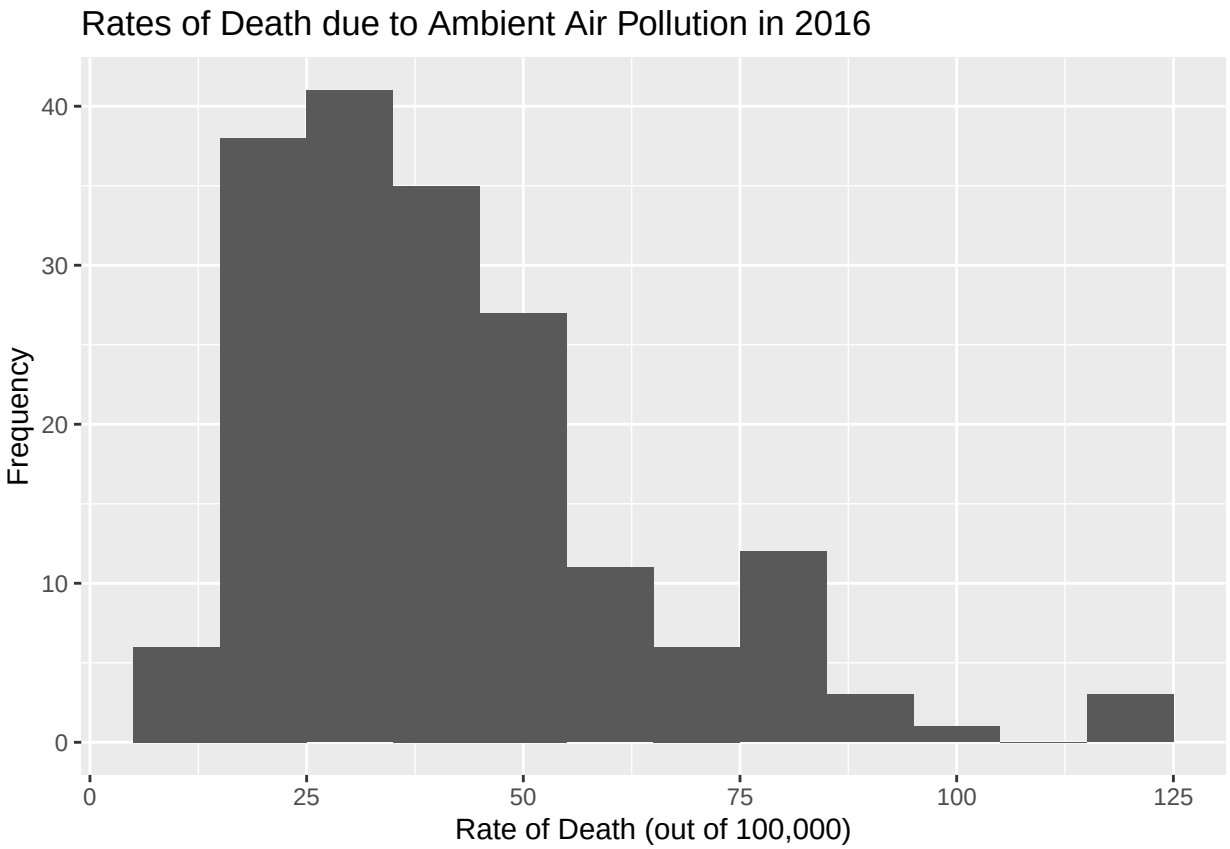
In general we would like to investigate air pollution as a cause of mortality globally to see 1) if the level of ambient air pollution concentration is correlated with number of deaths in a population or likelihood of an individual dying 2) what type of air pollution has the most effect on mortality. We hypothesize that ambient air pollution concentration and the likelihood of dying or number of deaths in a population are associated, with the higher the concentration of ambient air pollution the more likely an individual will die or more people from a population will die. We also hypothesize that ambient air pollution will have a greater effect on mortality because household air pollution is mainly a problem in low-income countries.

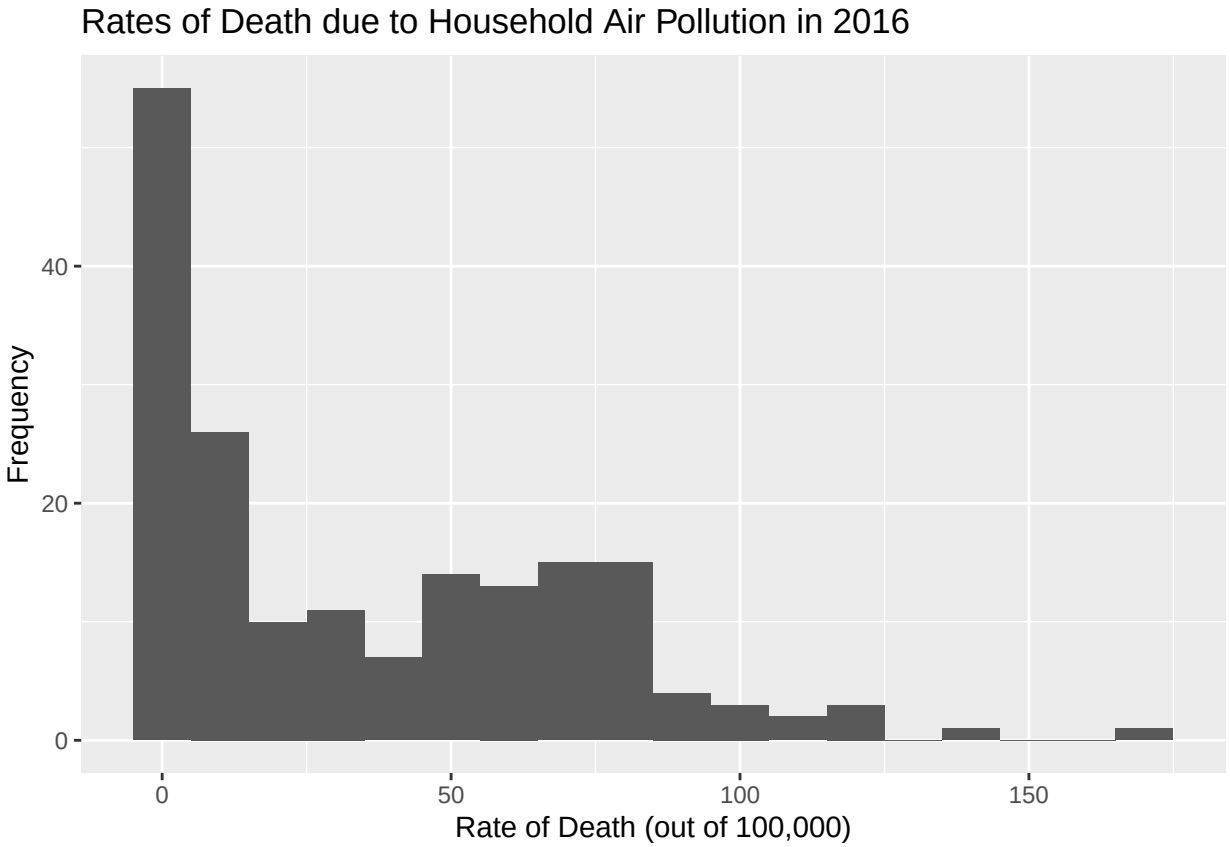
Data Collection and Variables

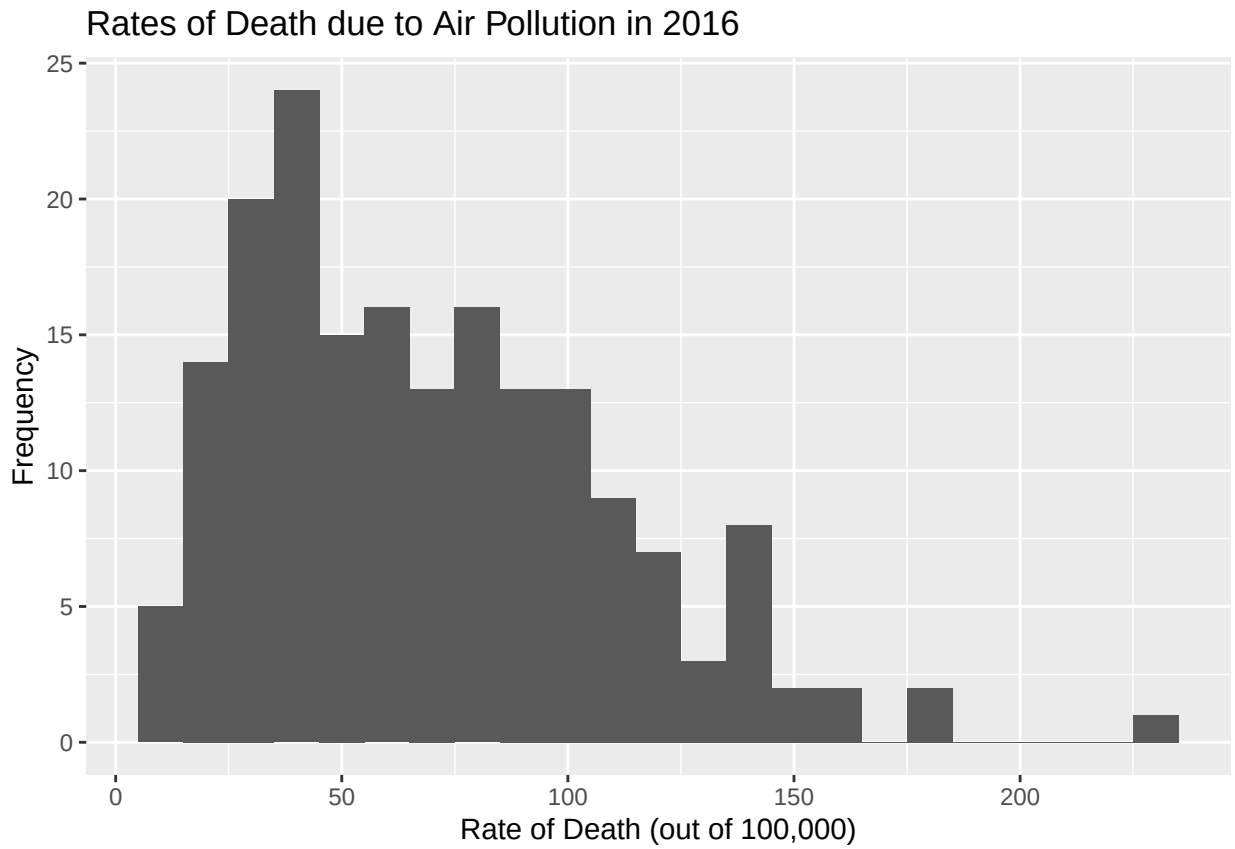
We downloaded this data from the World Health Organization Data Collections. There are several variables in this data including year, country, deaths by each type of air pollution, and deaths by other risk factors. We are focusing on the year 2016, which was the latest year data was collected in relation to air pollutants. We are also not looking at how they died. There are several different types of air pollution, but we will be looking at household and ambient matter pollution. We will compare these two variables to see which type of air pollution is the most fatal. We would also like to create a model to predict the amount of deaths per country or the likelihood of dying due to air concentration and other predictors. The main predictors we will look at are gender, amount of ambient air pollution (PM2.5), and the type of residence they live in.

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".

Exploratory Data Analysis


Rates of Death due to Ambient Air Pollution in 2016
Frequency
40
30
20
10
0
0
25
50
75
100
125
Rate of Death (out of 100,000)


Rates of Death due to Household Air Pollution in 2016
Frequency
40
20
0
0
50
100
150
Rate of Death (out of 100,000)



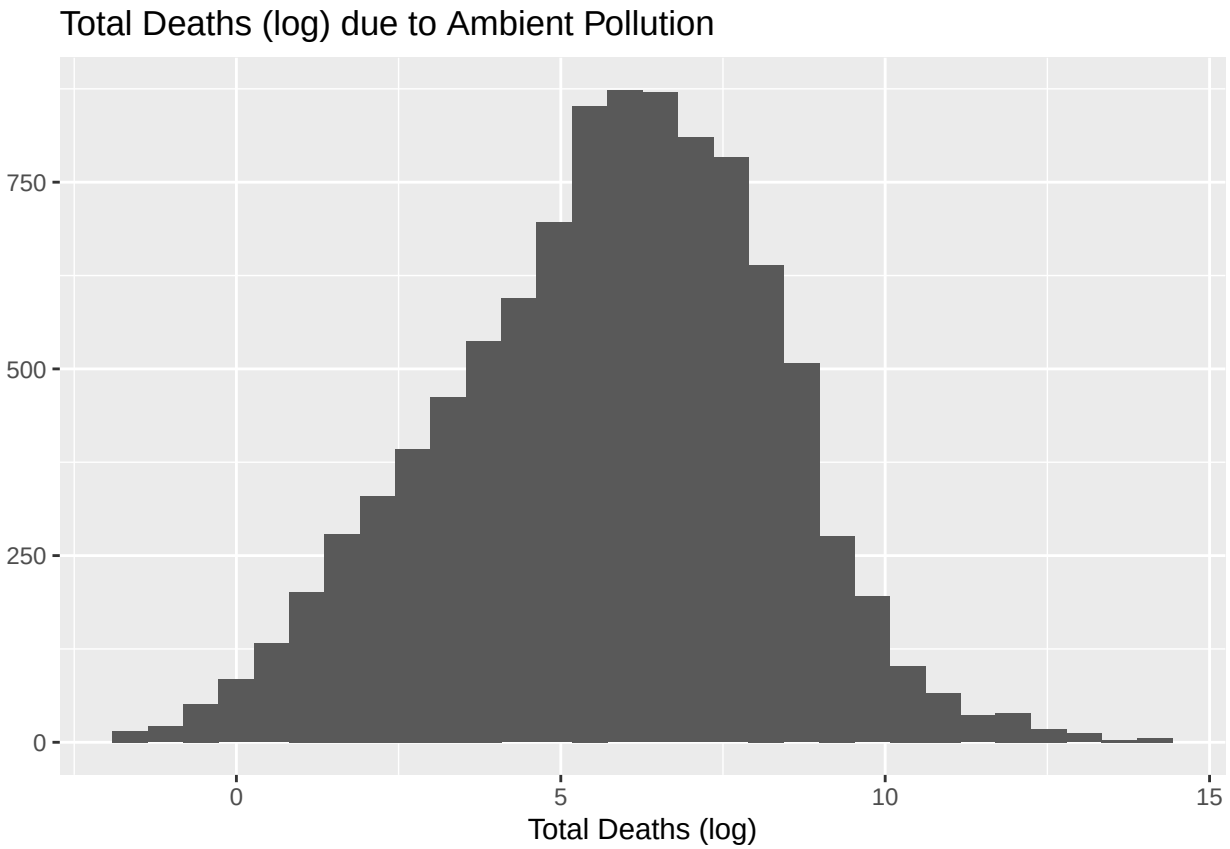
The first thing we wanted to look at was the frequency of higher death rates due to the ambient and household air pollution. As seen in the first visualization showing the rates of death due to Ambient Air Pollution, there tends to be a greater amount of countries that have death rates of around 25-30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

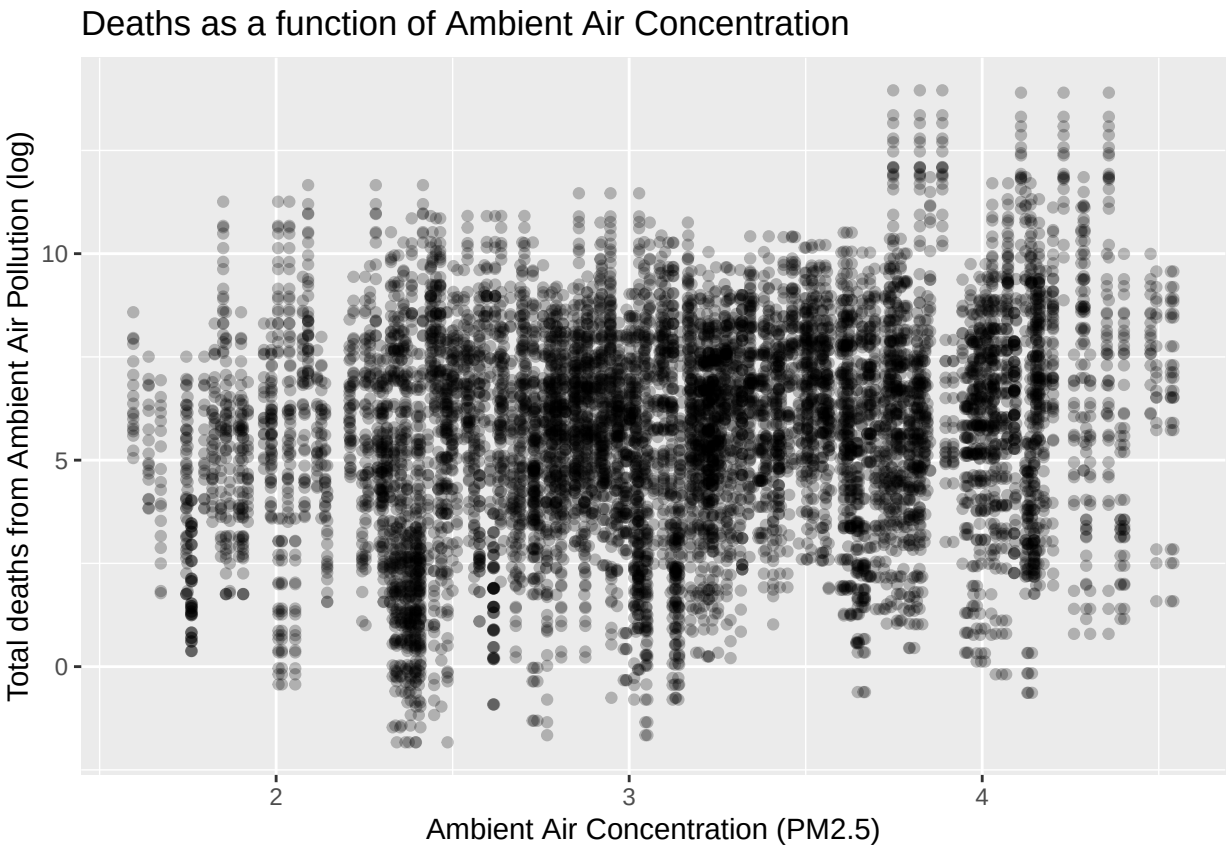
Out of these three plots, there are four countries that should be noted: Democratic People's Republic of Korea, Georgia, Chad, and Bosnia and Herzegovina. These countries are outliers and show much higher death rates due to air pollution than other countries.

Analytic Methods At first we wanted to try to create a linear model to predict how many individuals will die due to ambient air pollution. To do this we first had to check if the data was normal and if the data was independent.

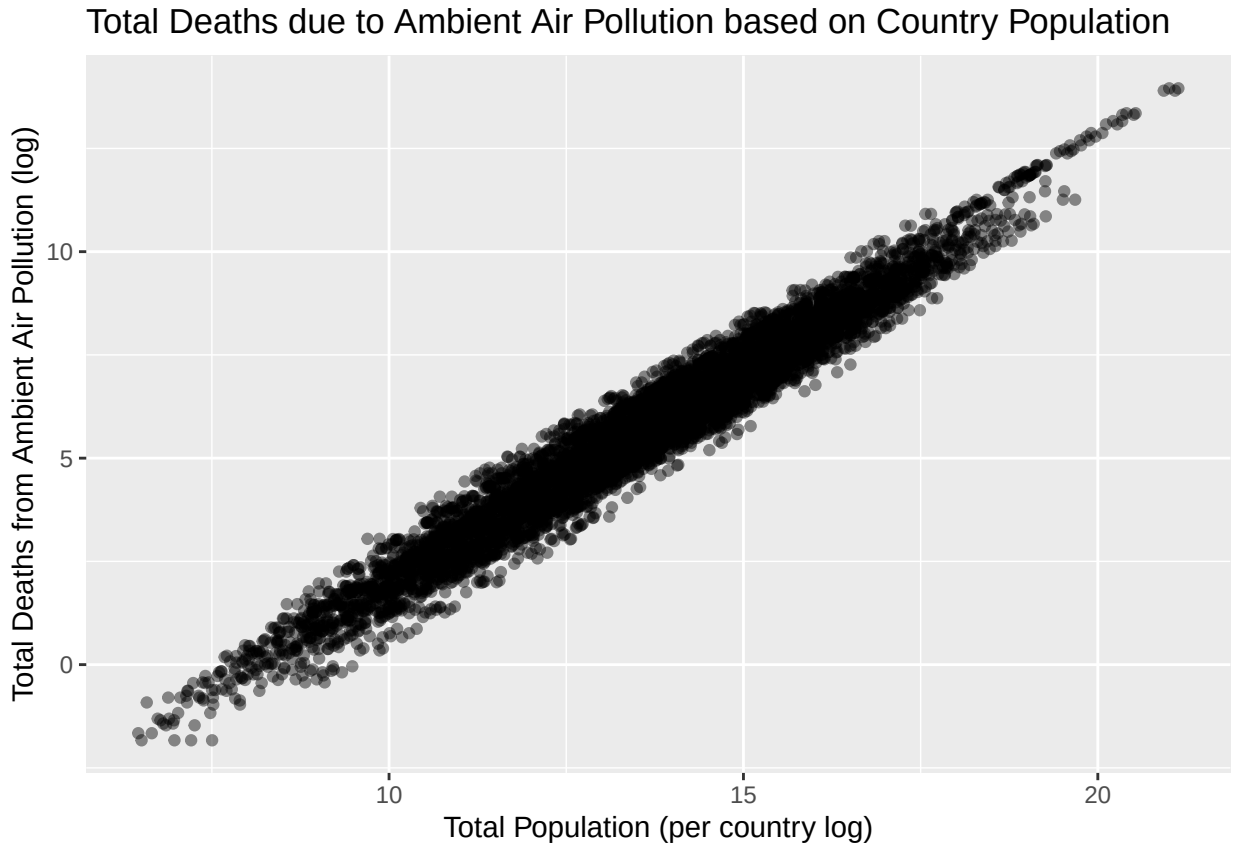
Within this graph, we normalized the total deaths due to ambient air pollution using the log function. We did this to create a normal distribution, so we could do a linear regression. Before doing this our deaths were skewed towards the left.



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.



In continuing our exploration of which factors influence deaths and death rates from ambient air pollution, we compared the total deaths from ambient air pollution to the total population of the countries we selected to examine. Per the linear arrangement of points in the plot, there seems to be a strong relationship between the ambient air pollution deaths and the country population. This is likely because an increased population will naturally lead to a greater chance of citizens facing global health crisis. This also could be related to an increased amount and spread of ambient air pollution in larger countries which rely on greater usage of fossil fuels.



For the next 2 sections we created two different models for a linear regression to predict the total number of deaths in a certain population. The first one we did just based off the outdoor air concentration. While the p-value (p-value<0.05) indicates that the air concentration variable is significant the r squared value indicates that the model is not a good fit for the data as the R-squared value is 0.039, meaning the model only explains 3.9% variance in the model.

```
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
```

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

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

We then tried to improve the linear regression model by adding two predictors, residence type and gender. The residence types could either be urben or rural, whild gender was either female or male.

Unfortunately, this linear regression model was also a bad fit for the data. The r-squared value was 0.038 so it only explains 3.8% of the data variance. Additionally the variables, gender and residence type, are not statistically significant as both of their p-values are greater then 0.05.

```
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
```

Instead of continuing with linear regression models, we chose to create a binomial regression model to predict the likelihood of a person dying based on the level of ambient air concentration. We did this because the linear model was not predicting the data well as seen by r squared value of both linear models.

Here we created a binomial regression model to predict the likelihood of a person dying based on the level of ambient air concentration. The resulting equation likelihood of dying = 0.0003 + 1.0124(Ambient Air Concentration). Because the p-value is less than alpha = 0.05 and thus statistically significant and the odds ratio is within the confidence interval, we reject the null hypothesis and have enough evidence to say that ambient air concentration is related to dying. The odds ratio confirms that it is 1.0124 times more likely for individuals that experience ambient air concentration, specifically 1 PM2.5, to die than those who do not.

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

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
```

```
tidy(modelagg1, 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)             0.000334 0.000276     -28975.       0 0.000333  0.000334
## 2 AmbientAirConcentration 1.01     0.00000542     2282.       0 1.01      1.01
```

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. The resulting equation is likelihood of dying = 0.0003 + 1.0124(Ambient Air Concentration) + 1.1822(Sex) + 0.9270(Residence Type). Because the p-values for all 3 of these variables are less than alpha = 0.05 and thus statistically significant and the odds ratio is within the confidence interval, we reject the null hypothesis and have enough evidence to say that ambient air concentration, sex and residential type are all related to dying. The odds ratio confirms that it is 1.1822 times more likely for males to die than females when the ambient air pollution concentration is 0 and they live in a rural community. The odds ratio confirms that it is 0.9270 times more likely for someone living in a urban community to die than someone living in an rural community when the ambient air pollution concentration is 0 and they are a female.

```
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
```

```
tidy(modelagg2, 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)              0.000317 0.000313     -25743.       0 0.000316  0.000317
## 2 AmbientAirConcentration 1.01     0.00000551     2289.       0 1.01      1.01
## 3 SexMale                  1.18     0.000243       690.       0 1.18      1.18
## 4 ResidencetypeUrban       0.927    0.000246       -308.       0 0.927     0.927
```

While we cannot directly compare the models we are concluding that the last model we made is the best one given the data we have. We are concluding this because the the two additional predictors we added are significant which likely means they are needed in addition to the air concentration graph. Also the AIC value in the second graph is smaller than the one in the first binomial regression, which indicates it is a better fit than the first model.

Lastly we wanted to explore the difference in effects between the two different types of air pollution within our datasets, household and ambient air pollution. So we conducted a paired t-test because the two types of pollution are coming from the same country and measuring the same outcome (deaths). We wanted to see if the two different types of air pollution had the same impact on mortality. We saw that they did not through the two sided paired t-test, but wanted to do a one sided t-test to see which type of air pollution had a greater effect on mortality. When we conducted the one sided t-test we saw that household air pollution had less of an effect on mortality than ambient air pollution. This was significant because the p-value was less than 0.05 and the mean was within the confidence interval for both t-tests.This goes along with our earlier exploratory graphs where we noticed the alarming amount of deaths due to household air pollution. It is important to note that it is likely that these two predictors are highly correlated as it is impossible to ethically separate and test the effects of different air pollutants

```
t.test(household$Totaldeathshoushold, household$Totaldeathsambient, paired = TRUE, alternative = "two.si
```

```
##
##  Paired t-test
##
## data:  household$Totaldeathshoushold and household$Totaldeathsambient
## t = -5.8255, df = 174959, p-value = 5.704e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -734.9047 -364.8837
## sample estimates:
## mean of the differences
##               -549.8942
```

```
t.test(household$Totaldeathshoushold, household$Totaldeathsambient, paired = TRUE, alternative = "less")
```

```
##
##  Paired t-test
##
## data:  household$Totaldeathshoushold and household$Totaldeathsambient
## t = -5.8255, df = 174959, p-value = 2.852e-09
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -394.6288
## sample estimates:
## mean of the differences
##               -549.8942
```

Conclusions

Through our analysis we have learned that ambient air pollution has an impact on mortality, furthermore we learned that ambient air pollution has a greater effect on mortality than household air pollution. This went along with our hypothesis made earlier. We also made a binomial regression model to predict the likelihood of a person dying due to ambient air pollution. This was a bit different than the original plan as we originally sought to create a model that would predict for the whole population. Through this regression model we found that it is more likely that an individual in an rural community to die due to air pollution than someone living in a urban community. We also found that a male is more likely to die than female.

We did have a big limitation with the t-tests, which was that household and ambient air pollution are likely highly correlated. We haven't learned how to test and adjust for this so this could be affecting the results. Another limitation we encountered was that we didn't have the data for multiple years which I think would have been a great addition. In terms of next steps, I think we should find a way to do a correlation test and perhaps gather data from multiple years. I think we should also gather some training data to test how well our model is predicting indivduals.