

# Midterm Project: Effects of Air Pollution on Countries

DataSci Warriors: Group 5

# Our Motivation

To explore how air pollution affect countries differently

Does air pollution affect underdeveloped countries disproportionately?

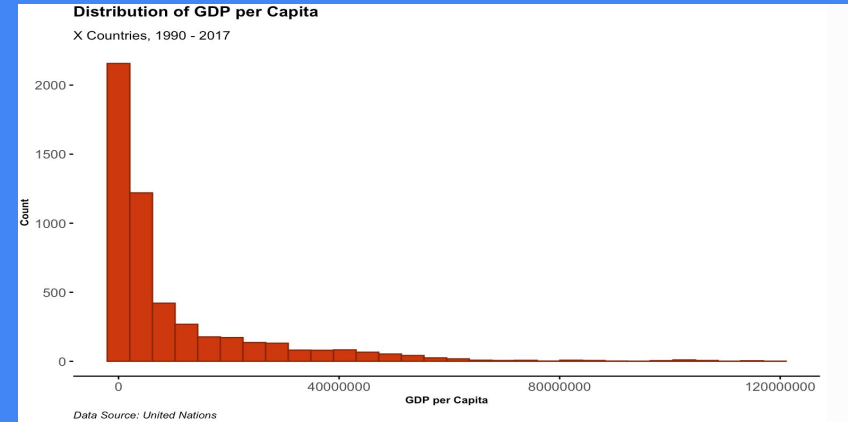
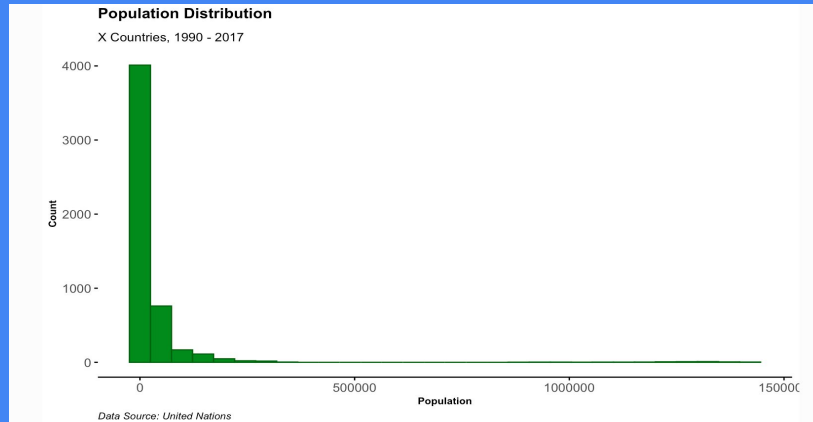
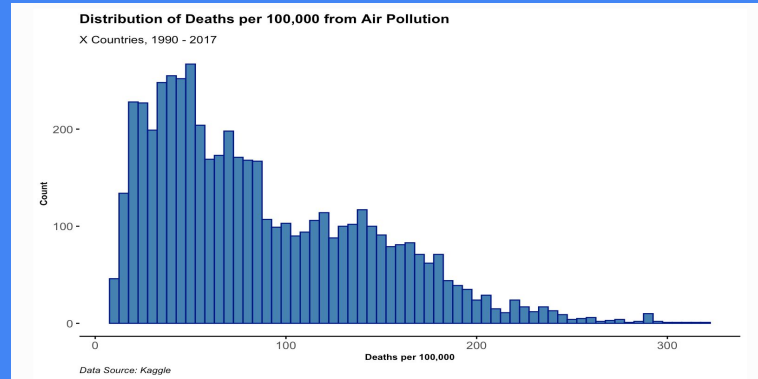
# Data Sources & Variables

Data	Source
Deaths Due to Air Pollution of Countries from 1990 - 2017	Kaggle
GDP Annual Growth of Countries from 1960 - 2020	Kaggle via WorldBank
United Nations Population and Region Data	United Nations
United Nations ISO-alpha3 code	United Nations
spData for Map Geometries	spData for Mapping

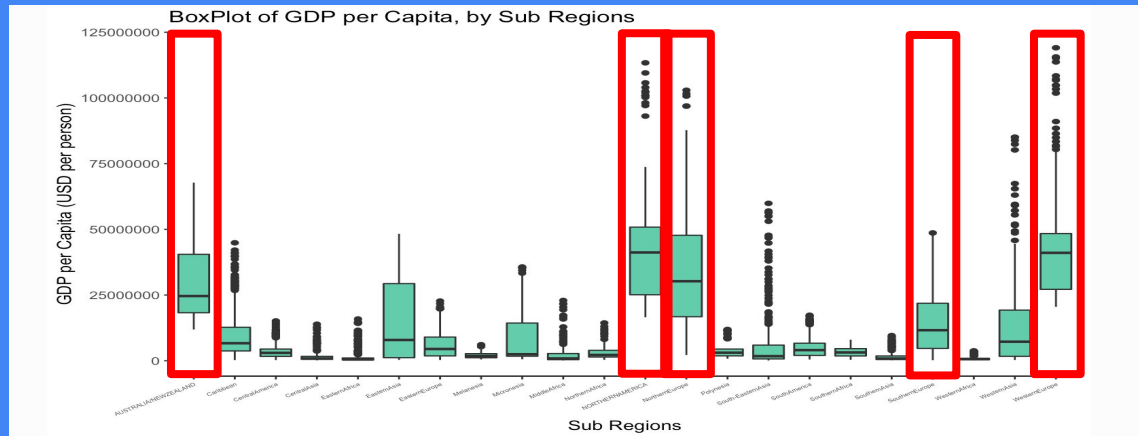
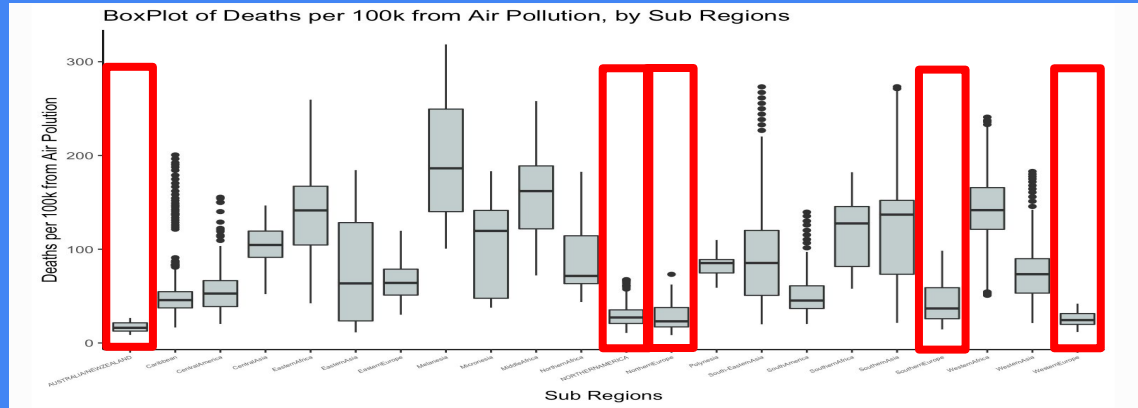
1. GDP : Numerical, Continuous
2. Population Size : Numerical, Continuous
3. Deaths due to Air Pollution : Numerical, Continuous
4. Country : Qualitative, Categorical
5. SDG Region : Qualitative, Categorical
6. Sub Region : Qualitative, Categorical
7. ISO- alpha3 Country Code : Qualitative, Categorical
8. ISO- alpha2 Country Code : Qualitative, Categorical
9. M49 Country Code : Numerical, Categorical
10. Year : Numerical, Categorical
11. GDP per Capita : Numerical, Continuous

E.D.A.

# EDA: Histogram of Air Pollution Induced Deaths

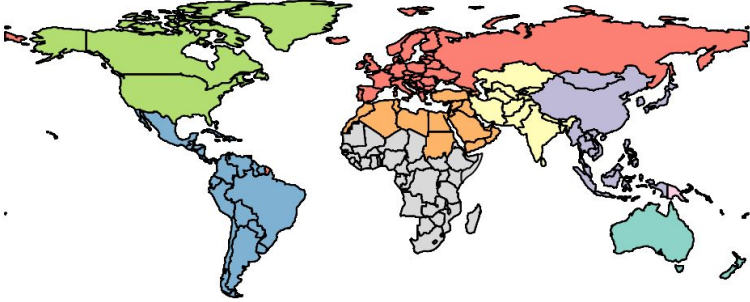


# EDA: Boxplots by Sub Region

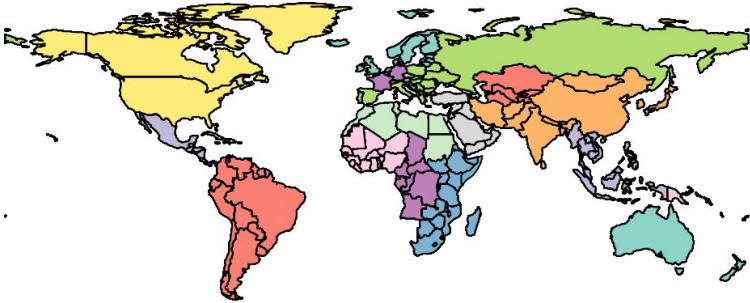


# E.D.A - Maps

**SDGRegion**



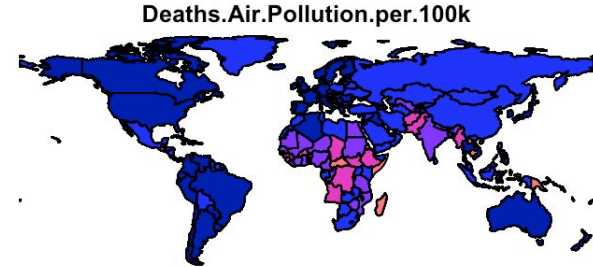
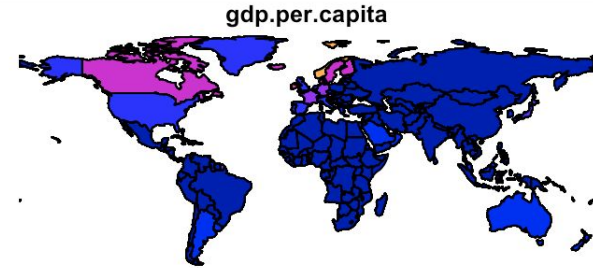
**SubRegion**



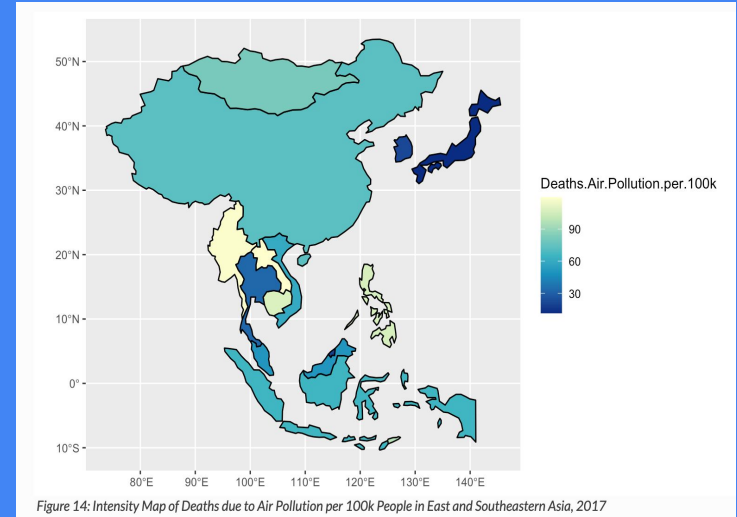
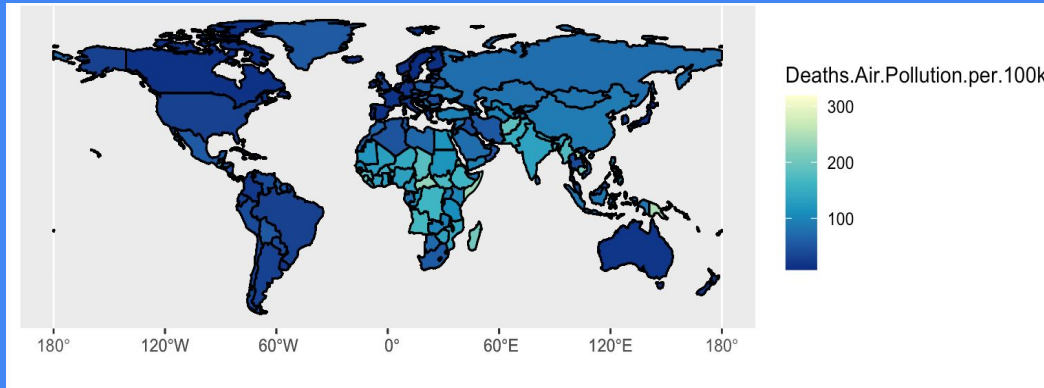
# Global Map of SDG Regions and SubRegions




# Global Intensity Map of Key Numerical Features, 1990 to 2017



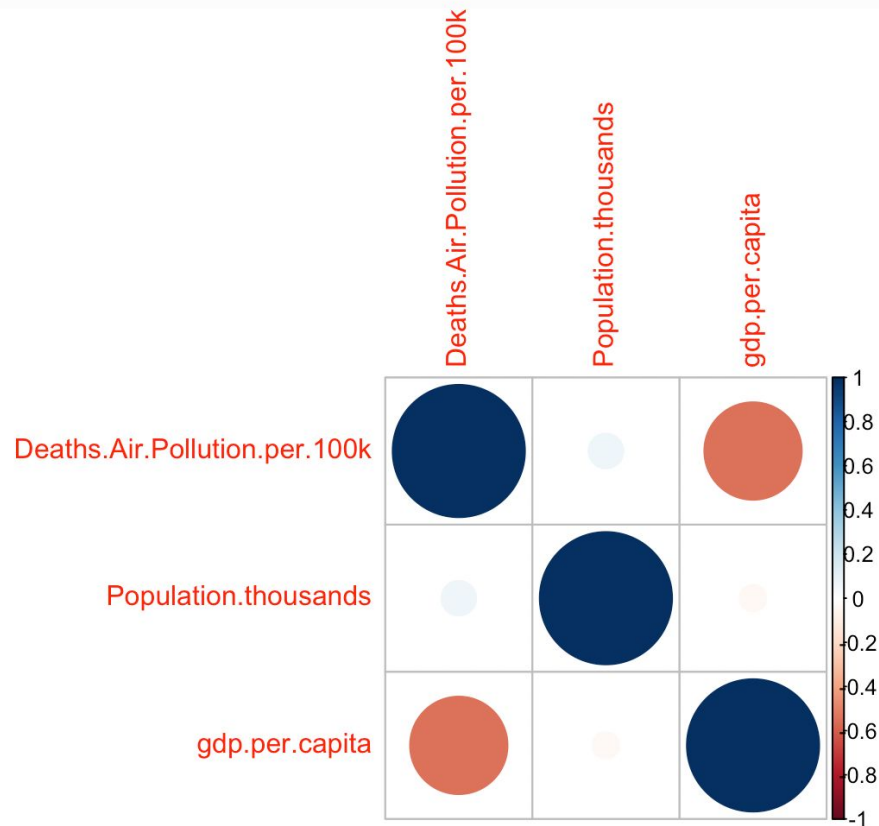
# Global Intensity Map of Deaths due to Air Pollution



# SMART Questions

1. Is there a relationship between population size and deaths due to air pollution?
  2. Which countries have the highest and lowest deaths due to air pollution? How is related to the region?
  3. Which years have the lowest and highest deaths due to air pollution?
- 

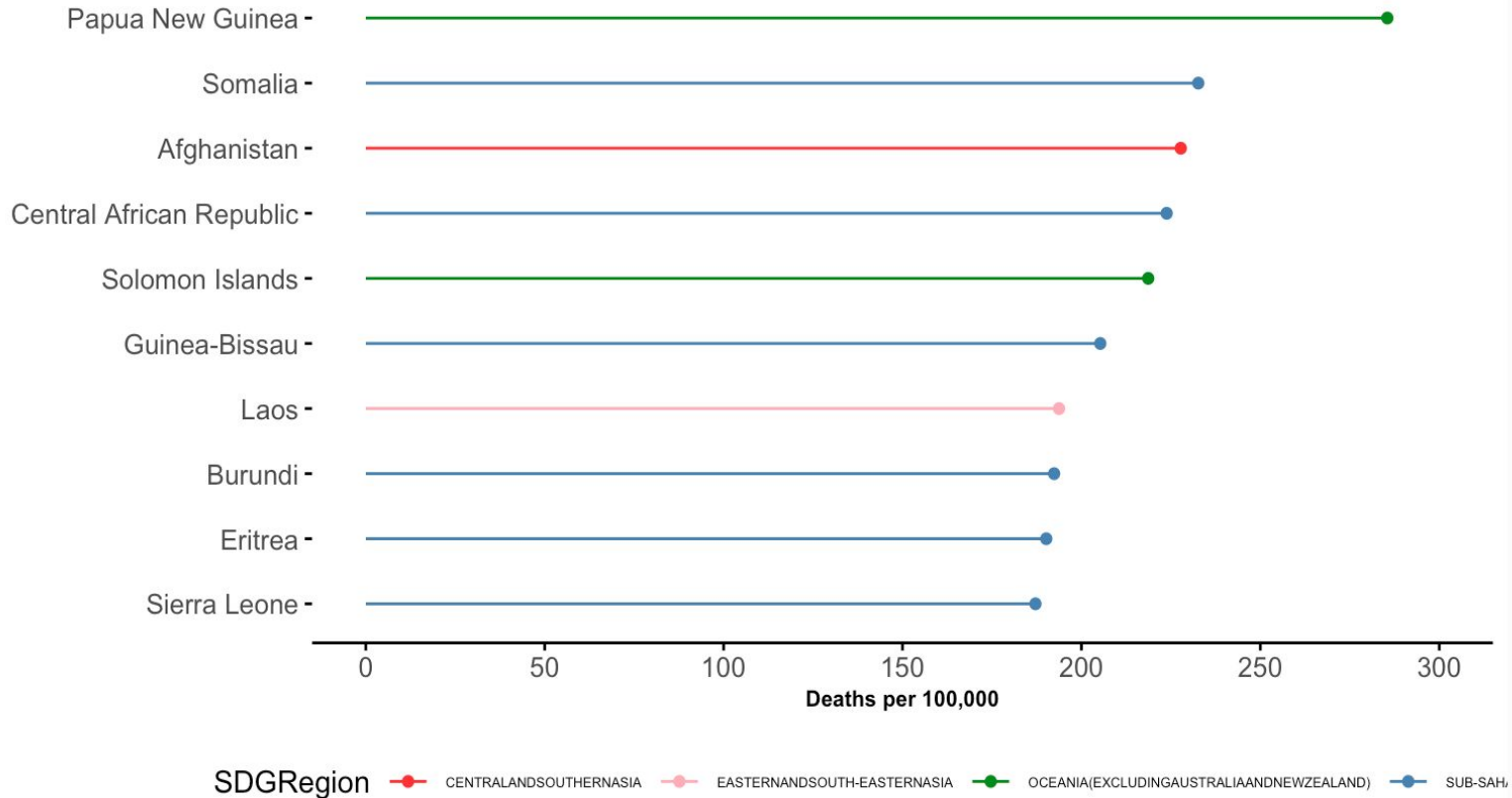
# Relationship between population size and deaths due to air pollution (1990-2017)



Which Countries have the highest and lowest deaths due to air pollution? How is it related to region?

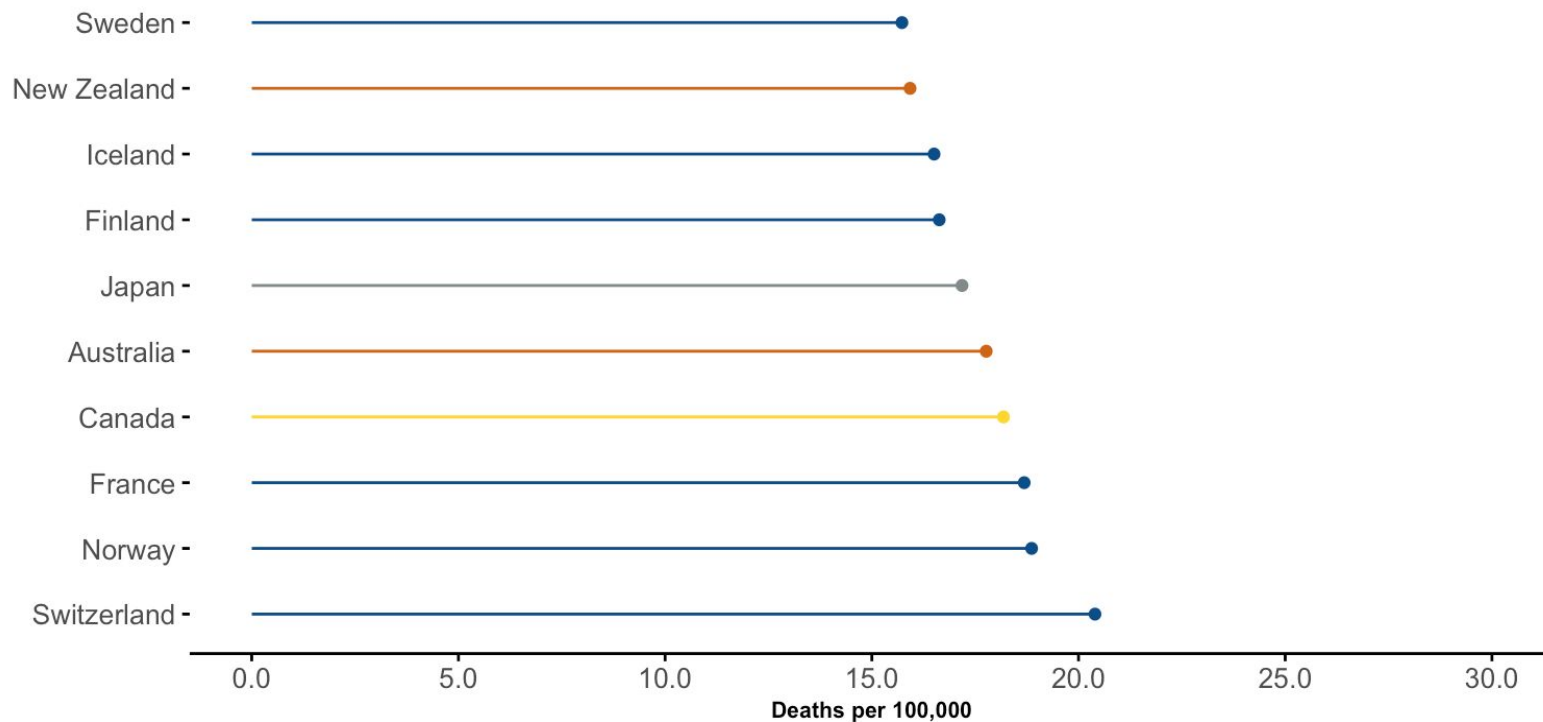
## 10 Countries with the Highest Average Deaths per 100,000

Deaths due to Air Pollution, between 1990 - 2017



## 10 Countries with the Lowest Average Deaths per 100,000

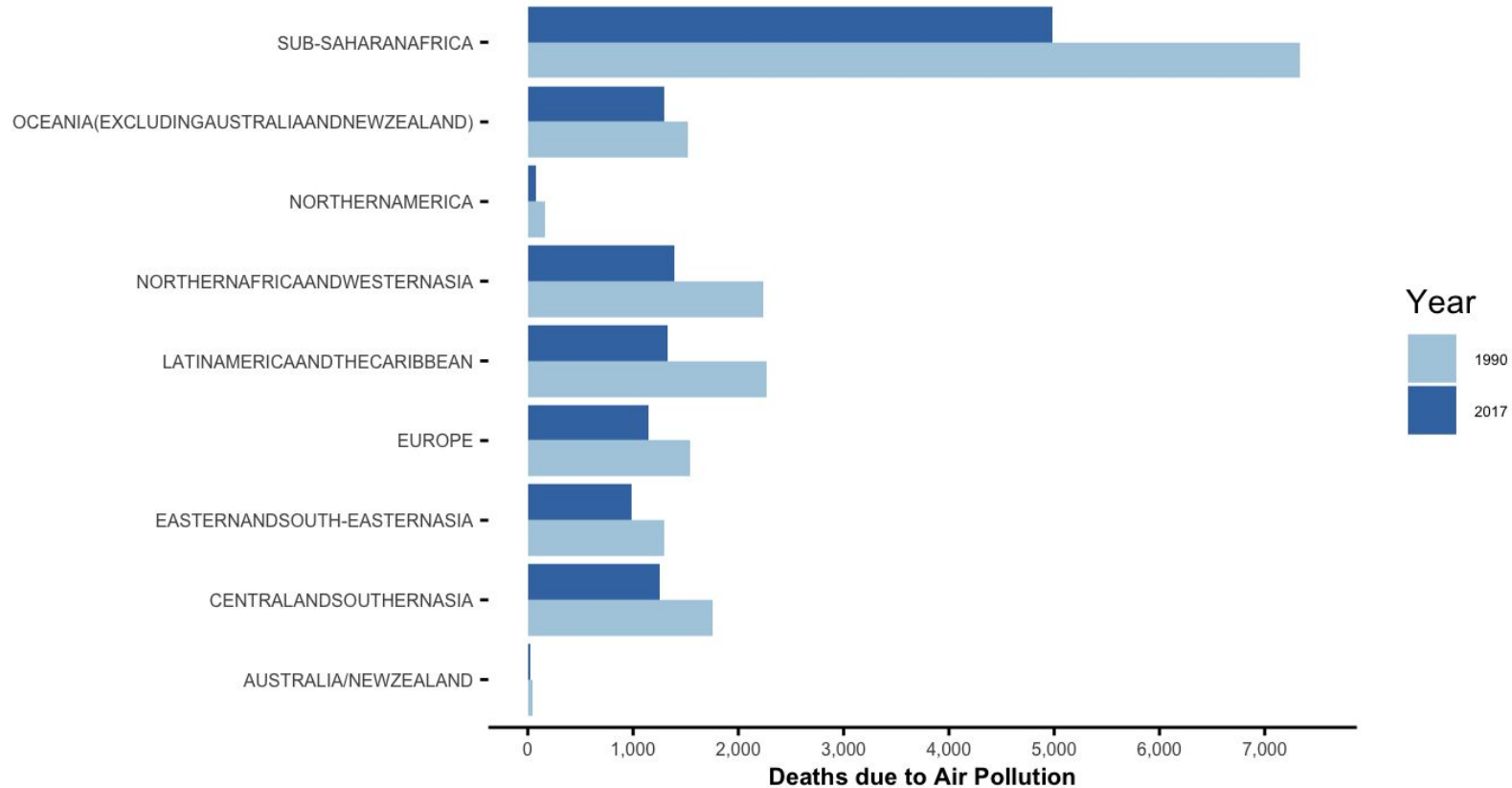
Deaths due to Air Pollution, between 1990 - 2017



SDGRegion — AUSTRALIA/NEWZEALAND — EASTERNANDSOUTH-EASTERNASIA — EUROPE — NORTHERNAMERICA

# Total Deaths per 100,000 by Air Pollution, by Region

1990 - 2017

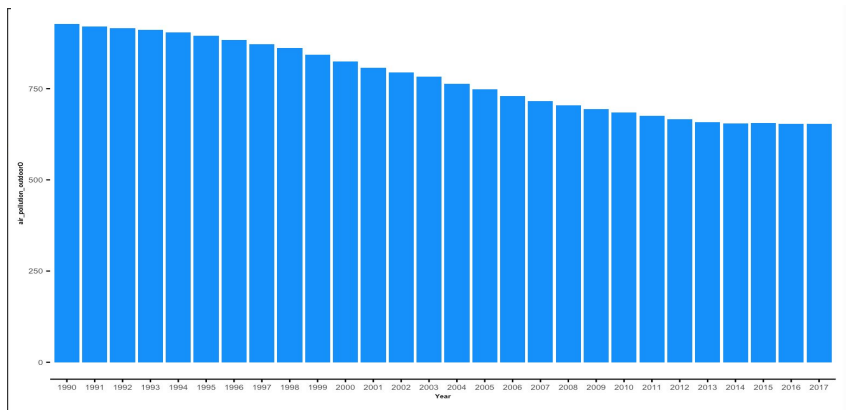
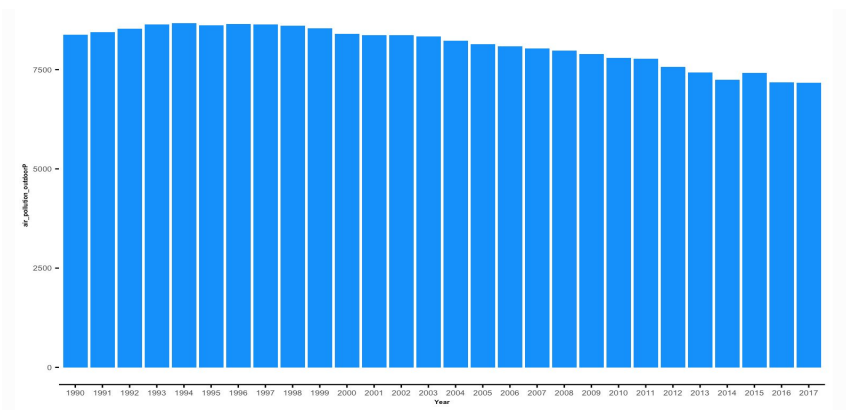
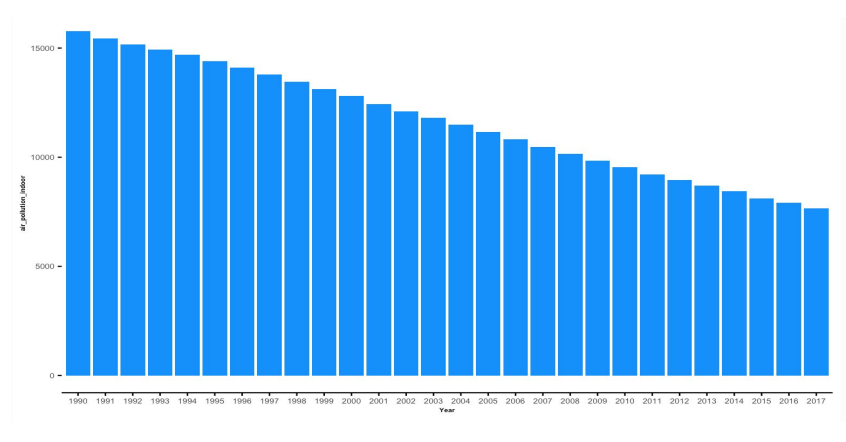
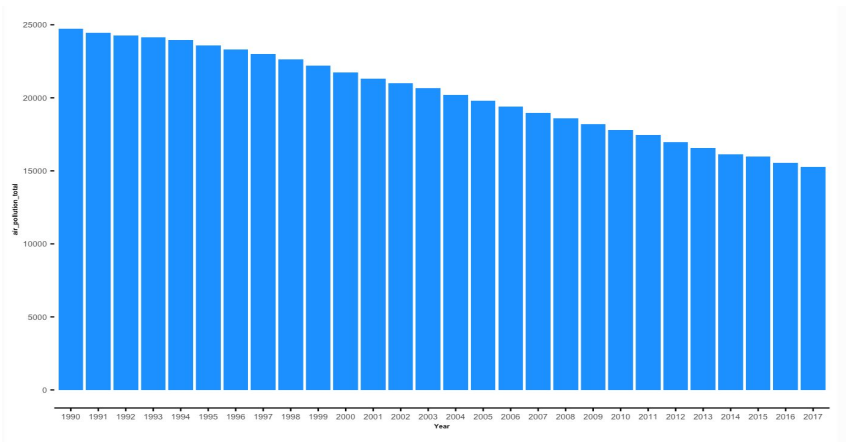


Data Source: Kaggle

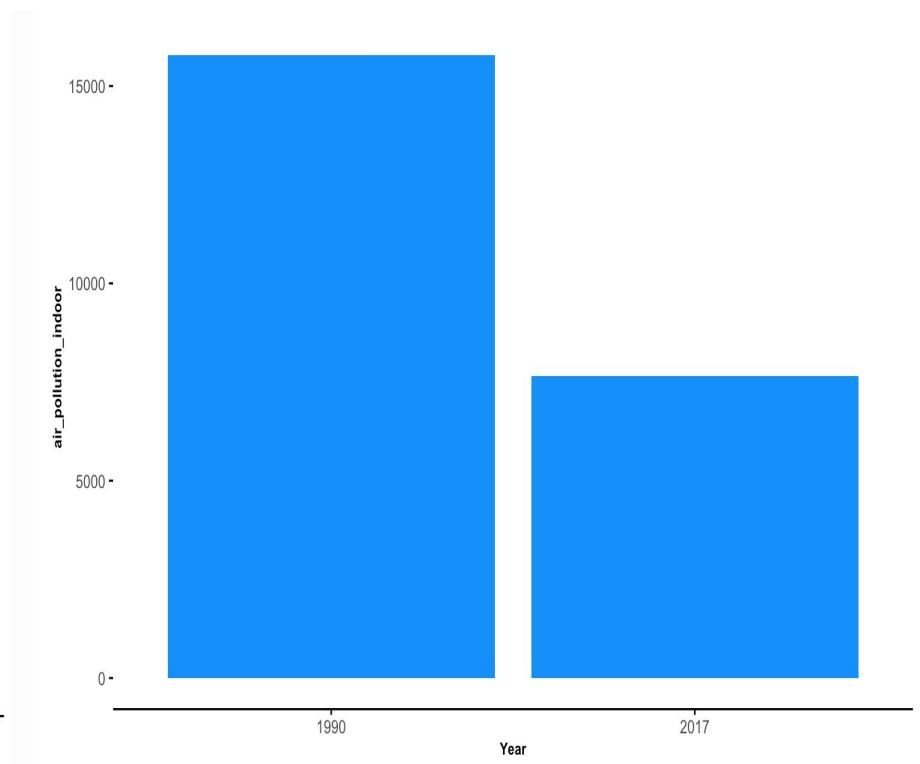
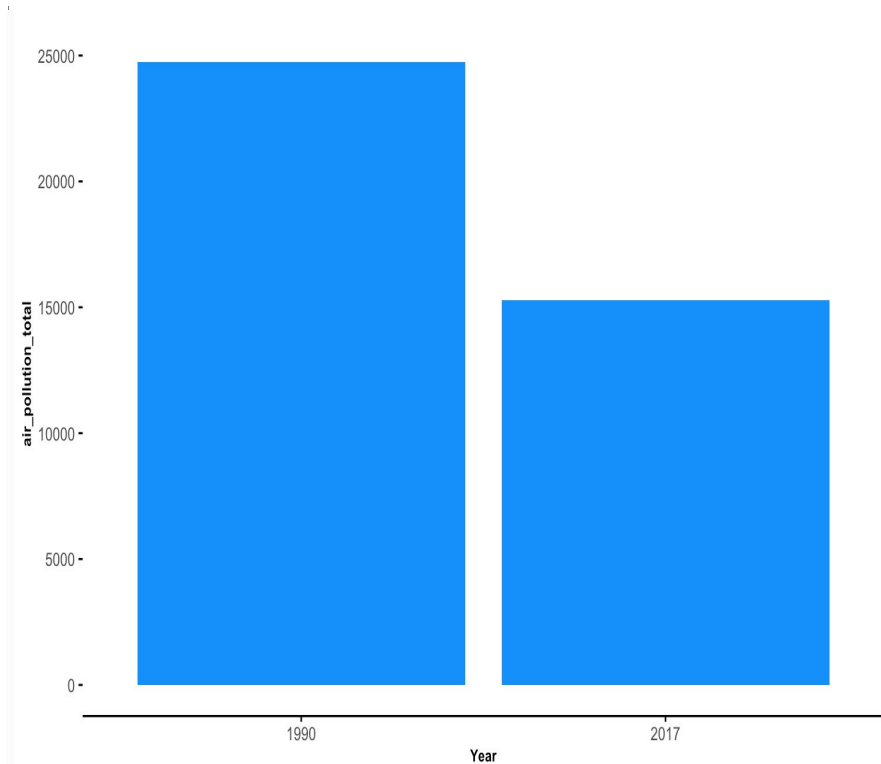


Which years have the lowest and highest deaths due to air pollution?

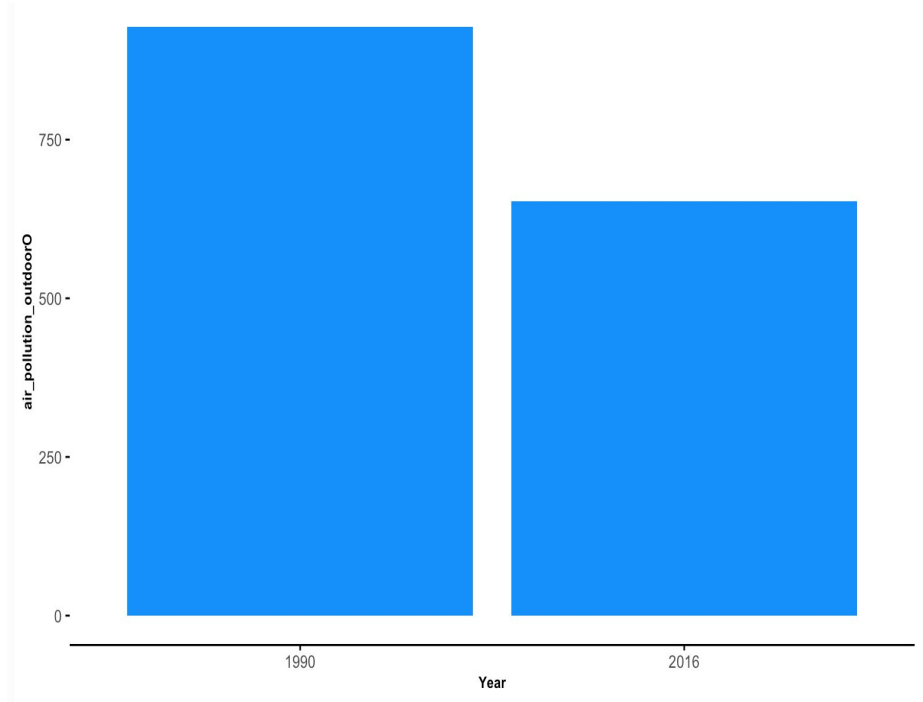
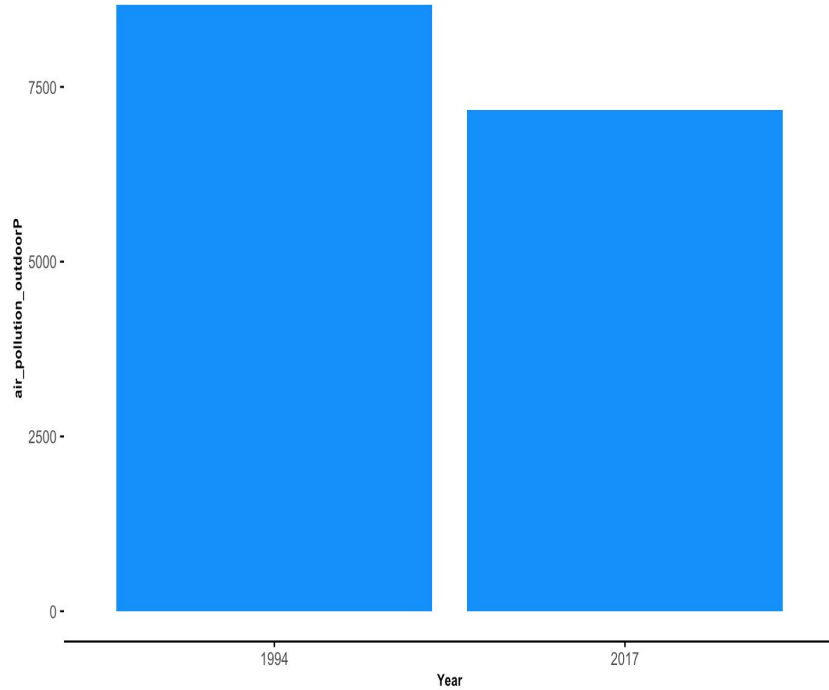
# Distributions of pollution deaths per year



# The Highest and Lowest



# The Highest and Lowest



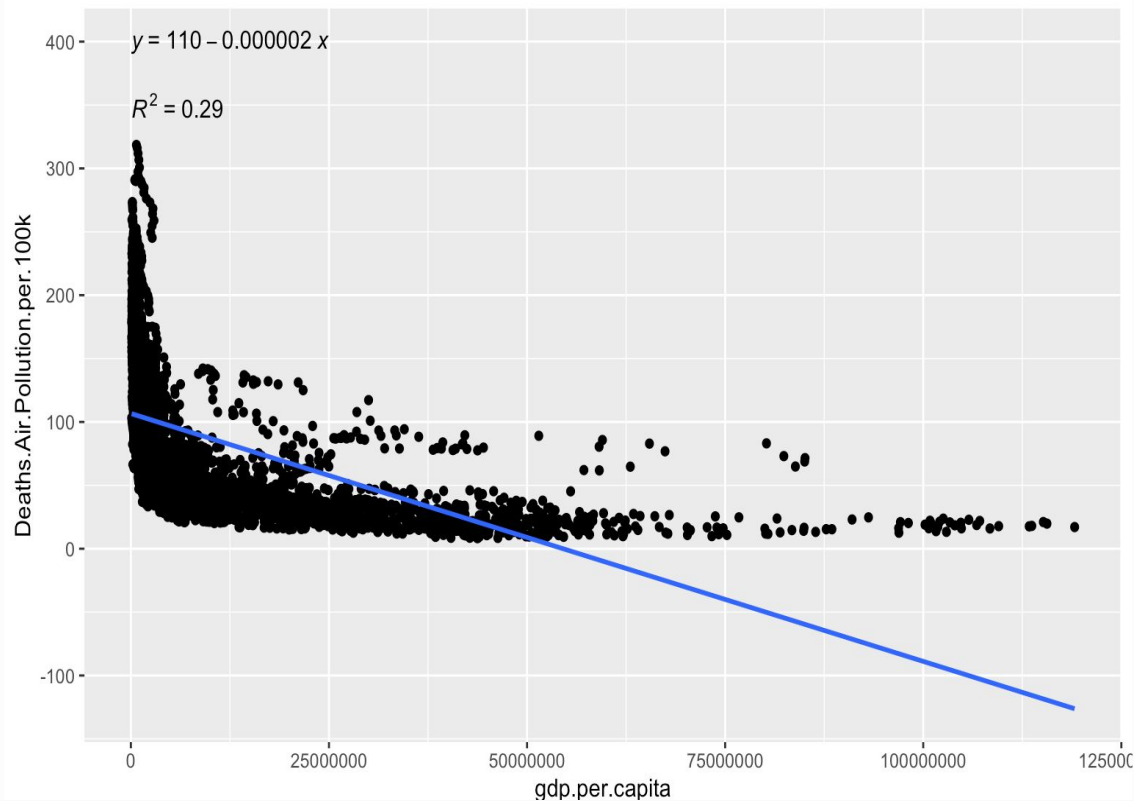
# Main Research Questions

- 1) Do lower GDP countries have more deaths per 100k due to air pollution?
- 2) Is there a correlation between GDP per capita and deaths caused by pollution? Is it linear? How strong is the correlation?

# Linear Fit

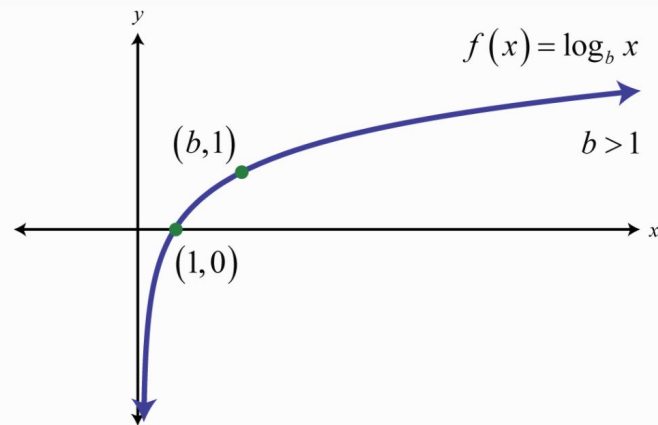
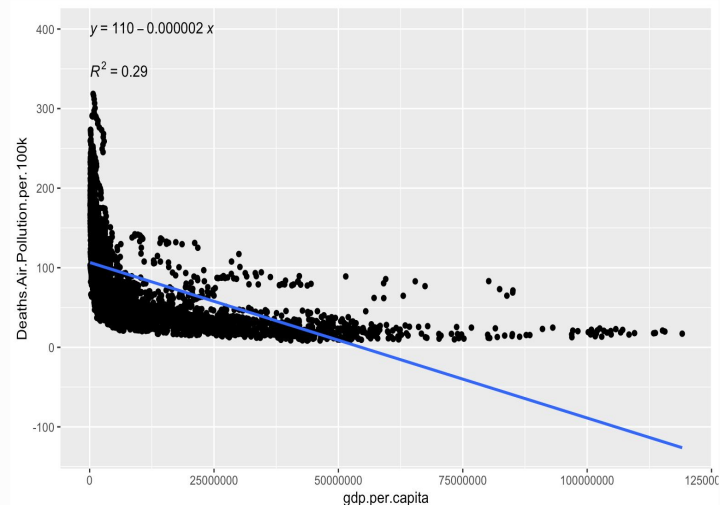
Negative correlation

But not a strong relationship,  
R squared is low at 0.295



# Transformed Log Scale - Linear Fit

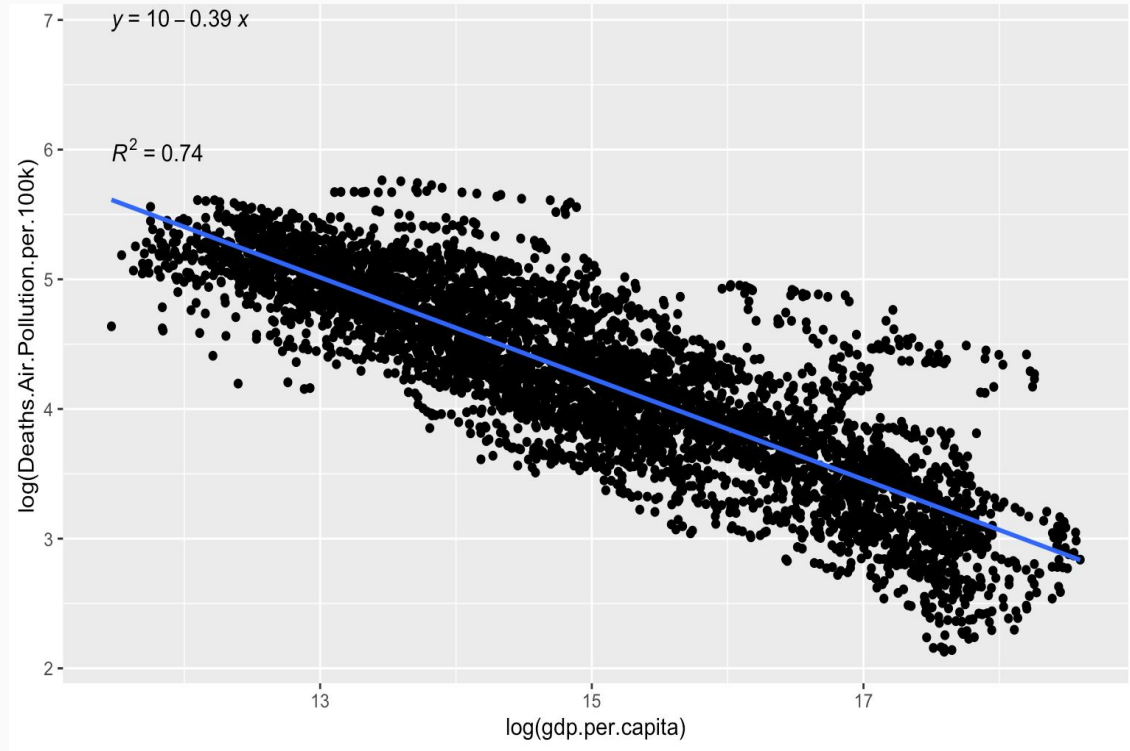
Our data looks like a (negative) logarithmic behavior



# Final Model

Transform features in model to log()

New R squared: 0.737



$$\log(\text{Deaths}_{\text{from air pollution|per year|per country}}/100,000) = 10.07849 - 0.38952 * \log(\text{GDP}_{\text{percapita}}) \quad \text{eqn(1)}$$

$$\text{Deaths}_{\text{from air pollution|per year|per country}} = 10^{10.07849 - 0.38952 * \log(\text{GDP}_{\text{percapita}})} * 100,000 \quad \text{eqn(2)}$$



# ANOVA Testing: Deaths Across Different GDP per Capita Levels are not Equal

$H_0: \mu_{\text{deaths\_lowest\_gdp}} = \mu_{\text{deaths\_low\_gdp}} = \mu_{\text{deaths\_medium\_gdp}} = \mu_{\text{deaths\_high\_gdp}}$

$H_1$ : At least one of  $\mu_{\text{deaths\_lowest\_gdp}}$ ,  $\mu_{\text{deaths\_low\_gdp}}$ ,  $\mu_{\text{deaths\_medium\_gdp}}$ ,  $\mu_{\text{deaths\_high\_gdp}}$  is not equal

**P-value =  $2e^{-16} \ll 0.05$ , reject null hypothesis**

Can confirm at least one of the means of deaths in different GDP per capita groupings are not the same

## Results: 2-Sample T-Tests

### Test 1

$$H_0: \mu_{\text{deaths\_lowest\_gdp}} = \mu_{\text{deaths\_low\_gdp}}$$

$$H_1: \mu_{\text{deaths\_lowest\_gdp}} \neq \mu_{\text{deaths\_low\_gdp}}$$

$$\text{p-value}_{\text{test1}}: 2.99\text{e-}203$$

$$\text{p-value}_{\text{test1}} < \alpha_{0.05} = \text{TRUE}$$

### Test 2

$$H_0: \mu_{\text{deaths\_low\_gdp}} = \mu_{\text{deaths\_medium\_gdp}}$$

$$H_1: \mu_{\text{deaths\_low\_gdp}} \neq \mu_{\text{deaths\_medium\_gdp}}$$

$$\text{p-value}_{\text{test2}}: 1.47\text{e-}13$$

$$\text{p-value}_{\text{test2}} < \alpha_{0.05} = \text{TRUE}$$

### Test 3

$$H_0: \mu_{\text{deaths\_medium\_gdp}} = \mu_{\text{deaths\_high\_gdp}}$$

$$H_1: \mu_{\text{deaths\_medium\_gdp}} \neq \mu_{\text{deaths\_high\_gdp}}$$

$$\text{p-value}_{\text{test3}}: 0\text{e+}00$$

$$\text{p-value}_{\text{test3}} < \alpha_{0.05} = \text{TRUE}$$

### Test 4

$$H_0: \mu_{\text{deaths\_lowest\_gdp}} = \mu_{\text{deaths\_high\_gdp}}$$

$$H_1: \mu_{\text{deaths\_lowest\_gdp}} \neq \mu_{\text{deaths\_high\_gdp}}$$

$$\text{p-value}_{\text{test4}}: 2.91\text{e-}06$$

$$\text{p-value}_{\text{test4}} < \alpha_{0.05} = \text{TRUE}$$

### Test 5

$$H_0: \mu_{\text{deaths\_lowest\_gdp}} = \mu_{\text{deaths\_medium\_gdp}}$$

$$H_1: \mu_{\text{deaths\_lowest\_gdp}} \neq \mu_{\text{deaths\_medium\_gdp}}$$

$$\text{p-value}_{\text{test5}}: 4.79\text{e-}48$$

$$\text{p-value}_{\text{test5}} < \alpha_{0.05} = \text{TRUE}$$

### Test 6

$$H_0: \mu_{\text{deaths\_low\_gdp}} = \mu_{\text{deaths\_high\_gdp}}$$

$$H_1: \mu_{\text{deaths\_low\_gdp}} \neq \mu_{\text{deaths\_high\_gdp}}$$

$$\text{p-value}_{\text{test6}}: 4.17\text{e-}70$$

$$\text{p-value}_{\text{test6}} < \alpha_{0.05} = \text{TRUE}$$

Conclusion of the tests: The p-values are less than an alpha value of 0.05, therefore we reject the null hypothesis and accept our alternative hypothesis

# Conclusion

- 1) The means of deaths caused by air pollution are statistically significant when grouped by different levels of GDP per capita
- 2) Findings reinforce the idea that deaths caused by air pollution has a significant relationship with GDP per capita and the strength and model can be quantified

$$Deaths_{from\ air\ pollution|per\ year|per\ country} = 10^{10.07849 - 0.38952 * \log(GDP_{per\ capita})} * 100,000 \quad eqn(2)$$