Predictive Insights into Air Pollution-Related Mortality

Eleanor, Owen, Clare, Kendall, Ryan, Kierany

November 2025

Presentation overview

- Problem statement
- Exploratory data analysis
- Modeling
- Results
- Evaluation
- Questions

Datasets

Death rate from air pollution per 100,000 people (1990-2019)

	A	B	C	D
1	Entity	Code	Year	Deaths - Cause: All causes - Risk: Air pollution - Sex: Both - Age: Age-standardized (Rate)
2	Afghanista	AFG	1990	402.18
3	Afghanista	AFG	1991	390.09
4	Afghanista	AFG	1992	383.2
5	Afghanista	AFG	1993	387.7
6	Afghanista	AFG	1994	394.02
7	Afghanista	AFG	1995	394.26
8	Afghanista	AFG	1996	395.64
9	Afghanista	AFG	1997	398.58
10	Afghanista	AFG	1998	401.16
11	Afghanista	AFG	1999	403.81
12	Afghanista	AFG	2000	403.5
13	Afghanista	AFG	2001	399.82
14	Afghanista	AFG	2002	386.45
15	Afghanista	AFG	2003	380.92
16	Afghanista	AFG	2004	372.72
17	Afghanista	AFG	2005	362.05
18	Afghanista	AFG	2006	351.97
19	Afghanista	AFG	2007	339.79
20	Afghanista	AFG	2008	327.88
21	Afghanista	AFG	2009	315.67
22	Afghanista	AFG	2010	304.63
23	Afghanista	AFG	2011	294.99
24	Afghanista	AFG	2012	286.2
25	Afghanista	AFG	2013	277.75
26	Afghanista	AFG	2014	270.26

Number of deaths from ambient particulate matter pollution (1990-2019)

	A	B	C	D	E	F	G	H	I	J	K
1	Entity	Code	Year	Deaths - C	Deaths - C	Deaths - C	Deaths - C	Deaths - C	Deaths - C	Deaths - C	Deaths - C
2	Afghanista	AFG	1990	3169	25633	1045	7077	356	3185	3702	4794
3	Afghanista	AFG	1991	3222	25872	1055	7149	364	3248	4309	4921
4	Afghanista	AFG	1992	3395	26309	1075	7297	376	3351	5356	5279
5	Afghanista	AFG	1993	3623	26961	1103	7499	389	3480	7152	5734
6	Afghanista	AFG	1994	3788	27658	1134	7698	399	3610	7192	6050
7	Afghanista	AFG	1995	3869	28090	1154	7807	406	3703	8378	6167
8	Afghanista	AFG	1996	3943	28587	1178	7943	413	3819	8487	6298
9	Afghanista	AFG	1997	4024	29021	1202	8075	420	3938	9348	6425
10	Afghanista	AFG	1998	4040	29349	1222	8173	425	4038	9788	6402
11	Afghanista	AFG	1999	4042	29712	1242	8265	426	4127	9931	6323
12	Afghanista	AFG	2000	4021	29999	1260	8328	427	4174	9942	6227
13	Afghanista	AFG	2001	4014	30421	1282	8440	432	4226	10052	6214
14	Afghanista	AFG	2002	3961	30189	1275	8383	432	4184	10004	6103
15	Afghanista	AFG	2003	4116	30157	1277	8398	437	4179	10841	6341
16	Afghanista	AFG	2004	4176	30225	1281	8433	445	4188	10761	6383
17	Afghanista	AFG	2005	4176	30089	1276	8415	452	4166	10118	6272
18	Afghanista	AFG	2006	4232	30075	1270	8418	456	4142	9081	6153
19	Afghanista	AFG	2007	4480	30080	1263	8426	465	4108	8168	6010
20	Afghanista	AFG	2008	4767	30219	1261	8474	478	4088	7245	5868
21	Afghanista	AFG	2009	5038	30280	1254	8488	485	4050	6437	5752
22	Afghanista	AFG	2010	5344	30352	1248	8512	492	4022	6021	5714
23	Afghanista	AFG	2011	5824	30684	1255	8620	496	4033	5600	5686
24	Afghanista	AFG	2012	6516	31090	1264	8753	502	4052	5243	5676
25	Afghanista	AFG	2013	7273	31462	1270	8854	508	4060	5220	5739
26	Afghanista	AFG	2014	7817	32002	1283	9001	514	4092	5010	5758

Source: Impact of Air Pollution on Human Health : Kaggle Data Set Compiled by Our World in Data

Problem Statement

Air pollution causes millions of deaths each year, underscoring the urgent need to reduce pollution and protect vulnerable communities.

We aimed to investigate how air pollution related death rates have changed across countries over time (1990-2019) and what risk factors best explain these differences.

Questions we're answering:

1. Can we model and predict air pollution-related mortality trends for a given country over the next five years?
2. Can we categorize the risk factors for air pollution-related deaths as having high, medium, or low significance in predicting death rates?

What we aim to accomplish:

1. Use EDA insights to forecast future air pollution deaths by country and region
2. Identify which factors matter most in predicting air pollution mortality

How we'll measure success:

1. Model accuracy for 5 year predictions using linear regression
2. Feature importance rankings that clearly separate high/medium/low significance factors

EDA

Structure of the Dataset

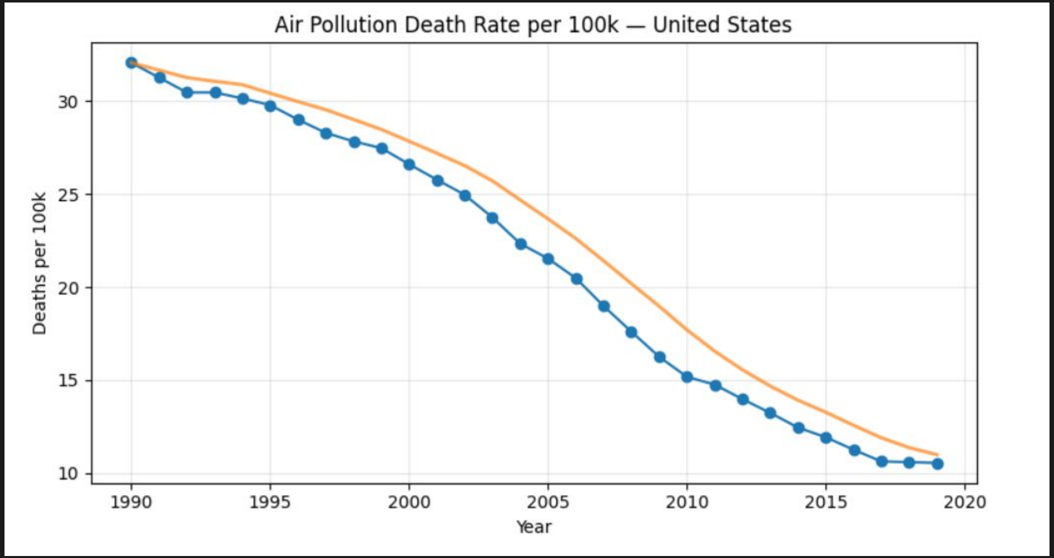
```
Rows: 5303
Countries: 177
Years: 30 1990 → 2019
Missing % in target: 0.0
```

- Panel data: Country x Year (1990 - 2019)
- Targets variables
  - **Death rate per 100k people (age standardized)**
  - **Absolute deaths**
- Features ex: **smoking, household air pollution, alcohol use, high blood pressure, and low physical activity etc.**
- Prep:
  - Standardized column names
  - Merged datasets on Country and Year
  - Handled missing values
  - Converted raw counter → per 100,000 people for cross-country comparability

Key Insights from EDA

1. Air pollution death rates have **varied significantly** across countries between 1990 and 2019
2. Countries like **Equatorial Guinea, Ethiopia, and Myanmar** have the most rapid reduction in air pollution death rates
3. **Uzbekistan, Lesotho, and Zimbabwe** have seen the most rapid rise in air pollution death rates

Trend Example: United States



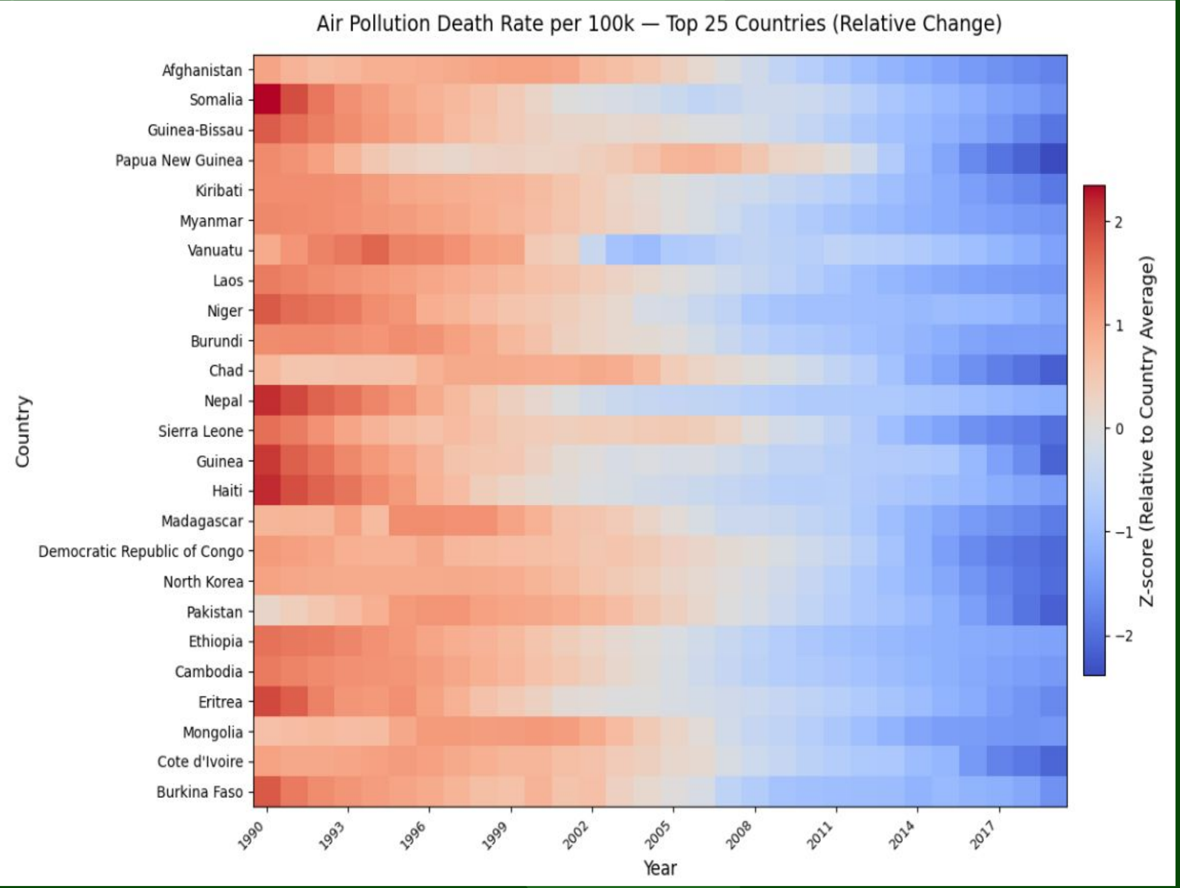
- **Blue:** yearly death rate per 100k
- **Orange:** 5 year rolling average

- Clear visual of decrease in death rates over time

- Suggests that time is an important predictor and we should include Year in our model

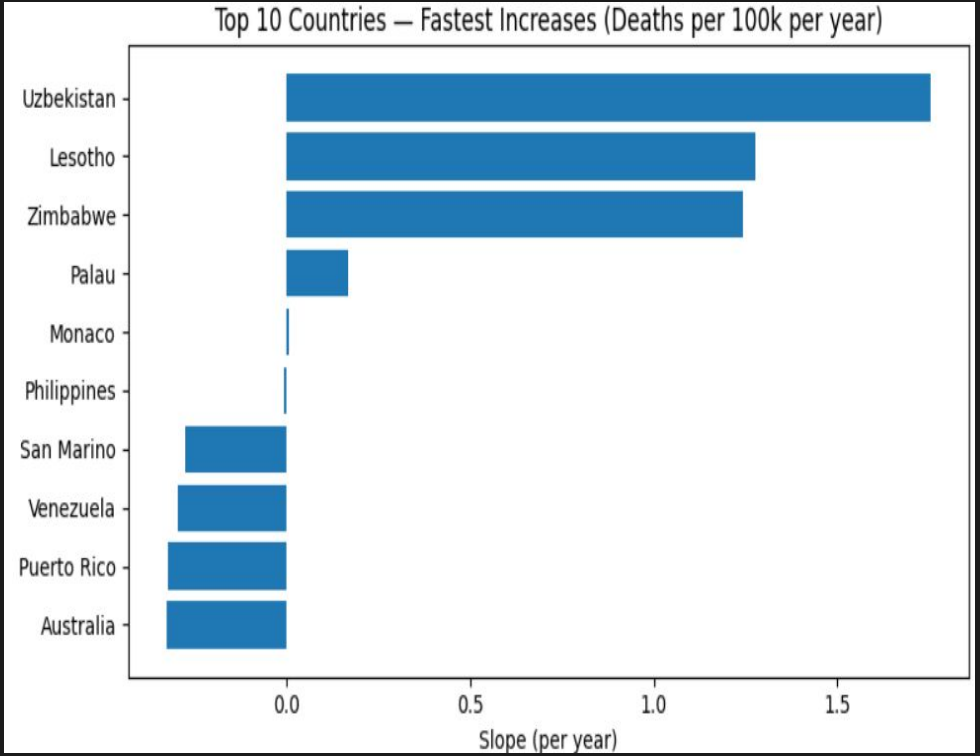
- Sets a baseline expectation for the next 5 year forecast

Global Change in Air Pollution Death Rates (Top 25 Countries)



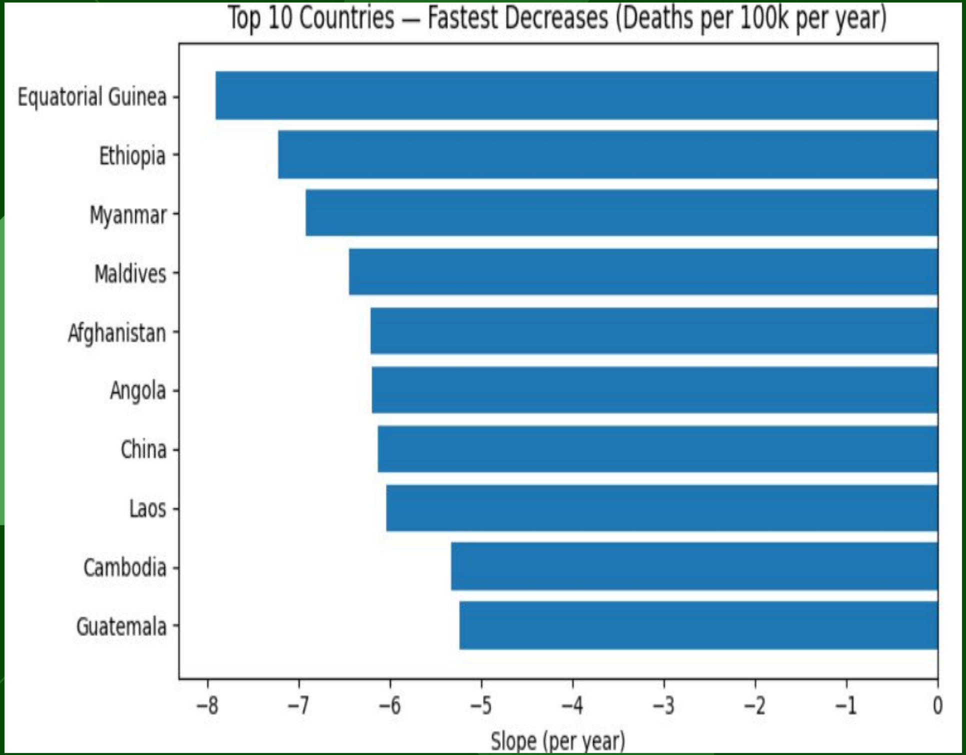
- **Row:** Country
- **Column:** Year (1990 – 2019)
- **Red**: above average years
- **Blue:** below average years (relative to each country's own average)
- Many countries shift from red → blue after early 2000s
- Shows global improvement, but uneven progress across regions

Top 10 Countries with the Fastest Increases in Air Pollution Death Rates



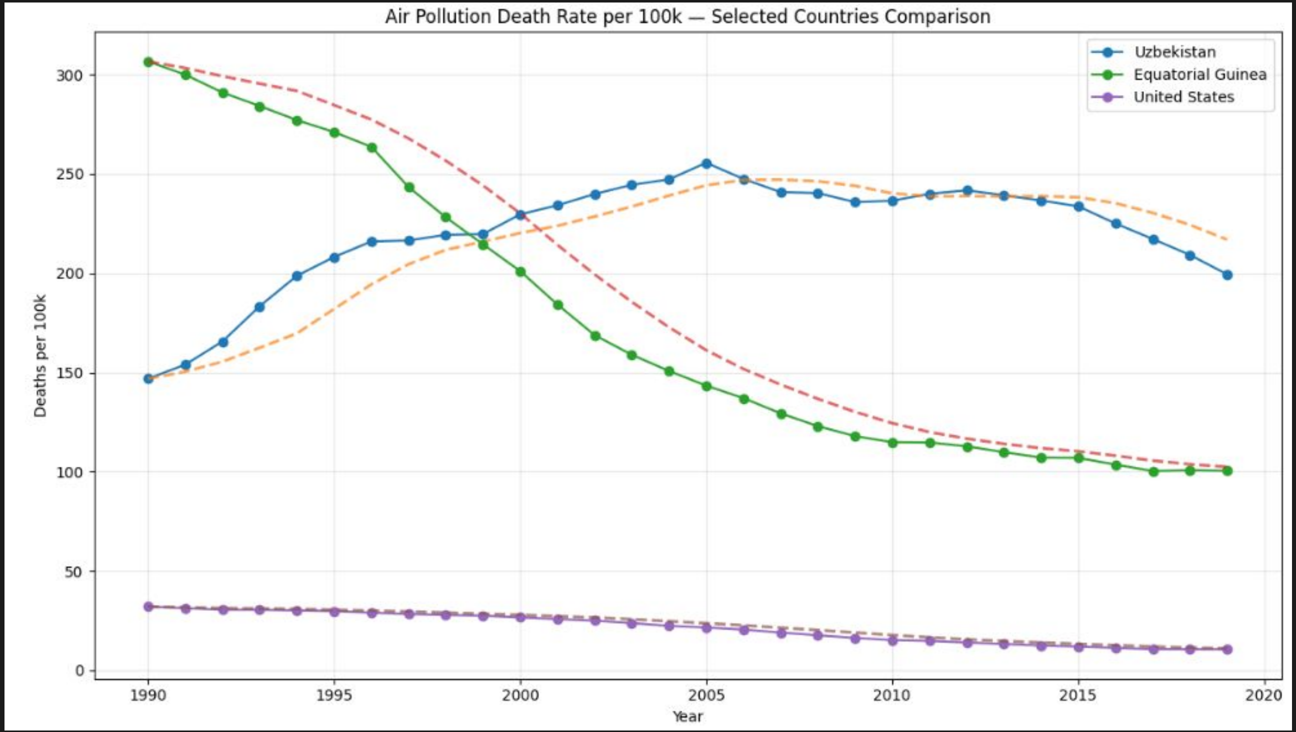
- Each bar shows the **rate of change in deaths per 100,000 people per year**
- **Only a handful of countries (Uzbekistan, Lesotho, Zimbabwe, and Palau) have death rates that are increasing over time**
- **Quantifies *how quickly* air pollution deaths are *increasing* by country**

Top 10 Countries - Fastest Decreases in Air Pollution Deaths



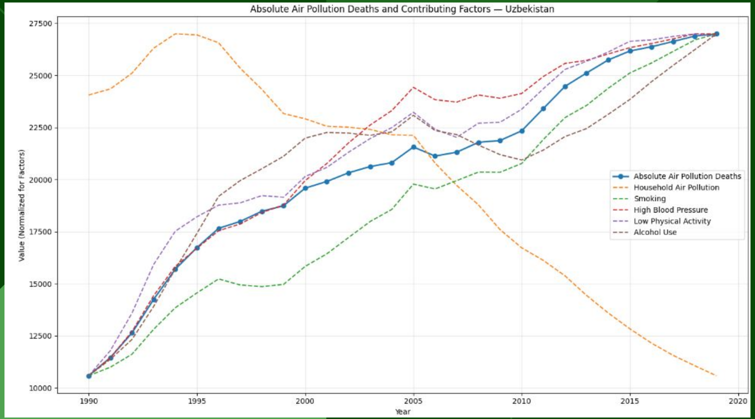
- Each bar shows the **rate of change in deaths per 100,000 people per year**
- All values are **negative slopes → faster declines over time**
- **Countries like Equatorial Guinea, Ethiopia, and Myanmar have seen the steepest improvements**
- **Quantifies *how quickly* air pollution deaths are *decreasing* by country**

Selected Countries Comparison

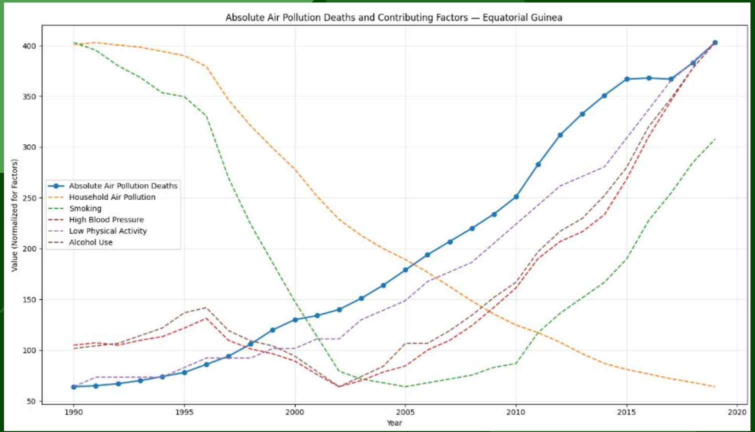


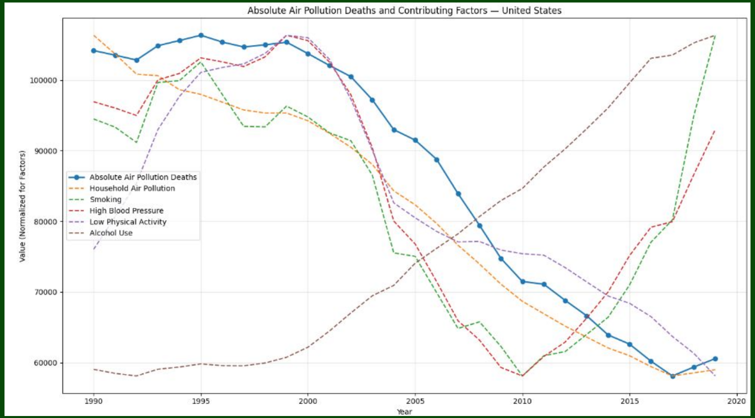
- **Uzbekistan**: Fastest increase
- **Equatorial Guinea**: Fastest decrease
- **United States:** Consistent and significant decreasing trend

Changes in Absolute Deaths Over Time

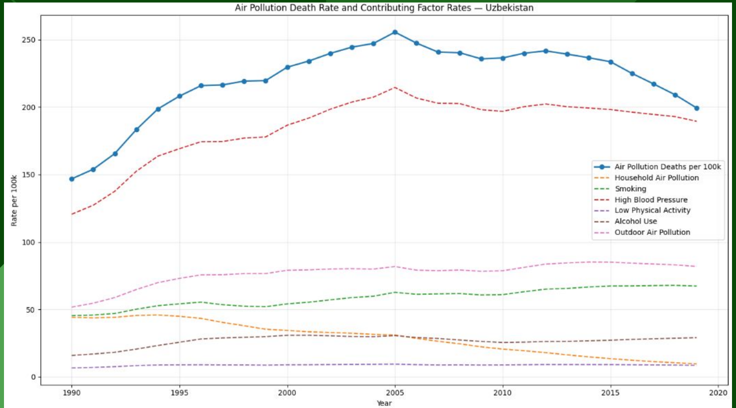


- **Reflects total burden & overall human cost of air pollution in selected countries**
- **Does not take into account population size**
- Highlights **global severity** of the issue & the need to allocate resources to reduce air pollution deaths, **especially in developing countries**

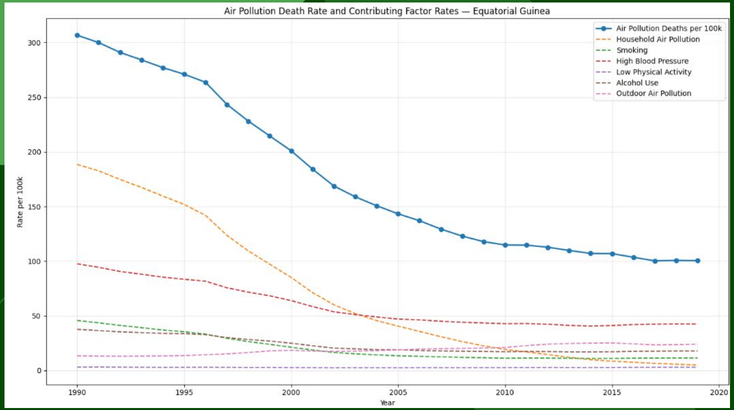


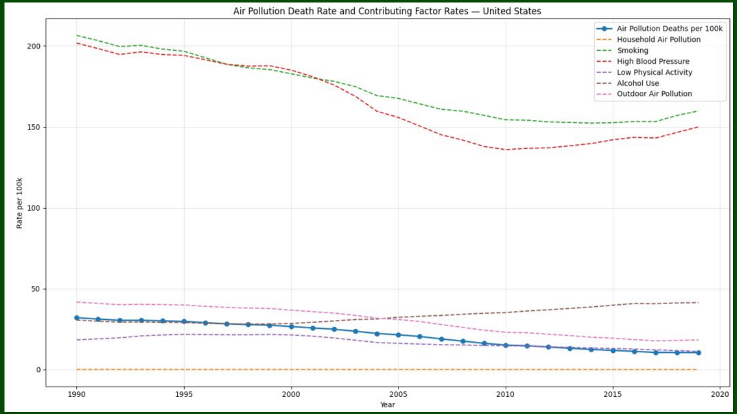


Changes in Absolute Death Rates Over Time



- **United States:** Outdoor air pollution (brown dashed line) appears strongly correlated with death rates
- **Uzbekistan**: High blood pressure rates track almost perfectly with the death rate trajectory
- **Equatorial Guinea**: Household air pollution closely mirrors death rate decline





Suggests air pollution death rates are driven by different primary factors depending on the country's development level and predominant pollution sources

Pearson correlations between air pollution death rate and each risk factor

	driver	pearson_r
0	Household Air Pollution	0.202640
1	Smoking	0.015831
2	High Blood Pressure	0.010369
4	Alcohol Use	-0.005801
3	Low Physical Activity	-0.062344

driver	mean_r
Household Air Pollution	0.742026
Smoking	-0.258573
Alcohol Use	-0.378161
High Blood Pressure	-0.421195
Low Physical Activity	-0.643935

- Pearson r shows how 2 variables move together
- **Household Air Pollution** → strongest positive link
- Other factors show **weak or negative relationships**

**correlation does not mean causation

Justify project choice based on insights from EDA

- EDA shows **clear linear time trends** → fits **multivariable linear regression**
- **Limited data points per country** → larger models (e.g., neural nets) not feasible
- **Small number of features** → simpler, interpretable model preferred
- **Linear regression** captures gradual yearly changes well
- Tested other options (**logistic, k-means)** for classification of risk-factors strength

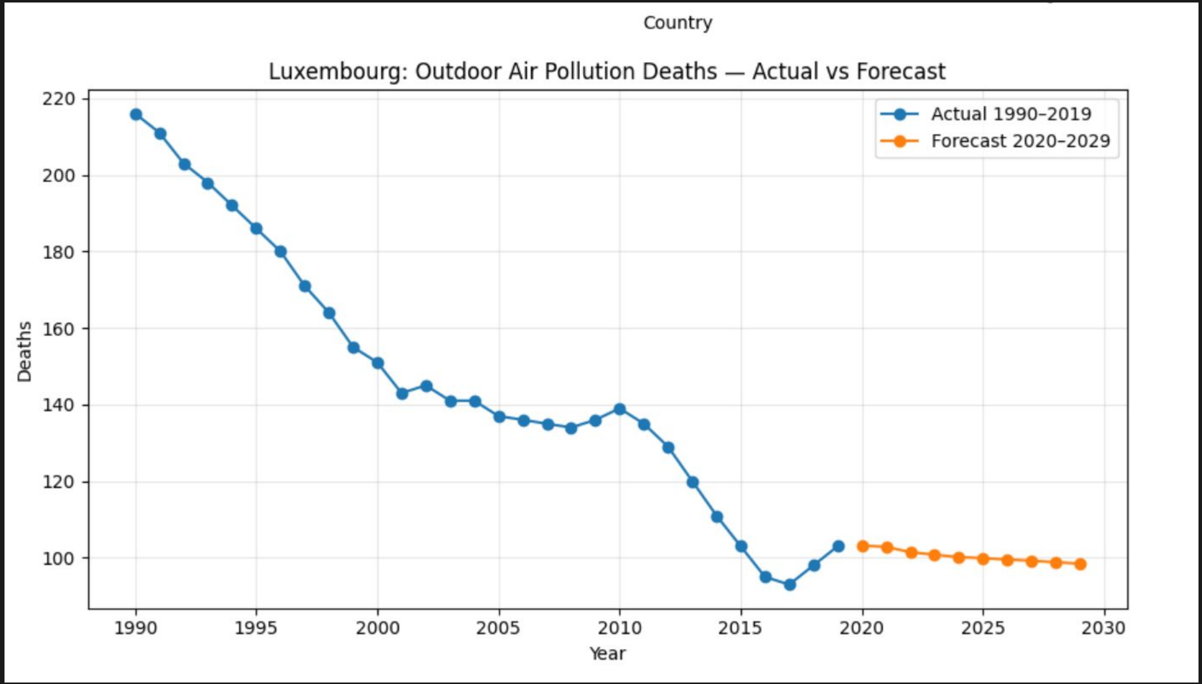
Modeling

Modeling Approaches Applied:

- Linear regression
- Lasso regression fit with 5-fold cross-validation

Forecasting Model

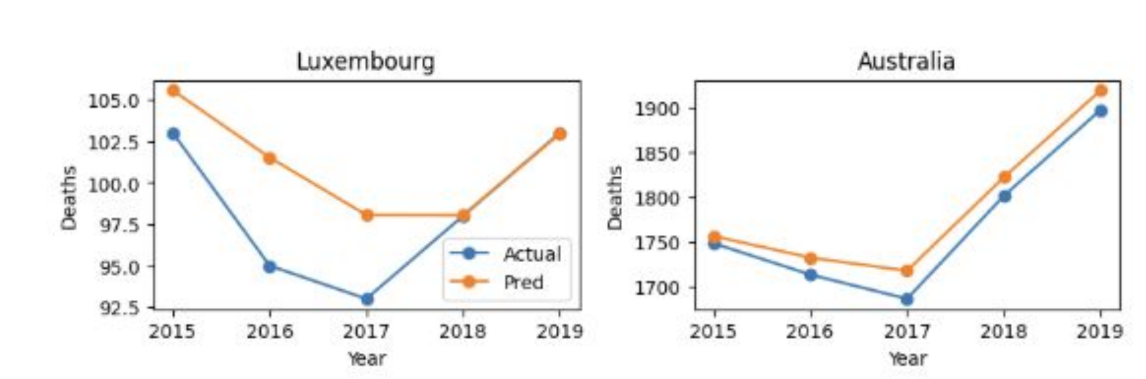
- Given historical mortality and related risk-factor deaths (1990–2019), can we build a ML model to forecast country-level deaths due to outdoor air pollution for 2020–2029?

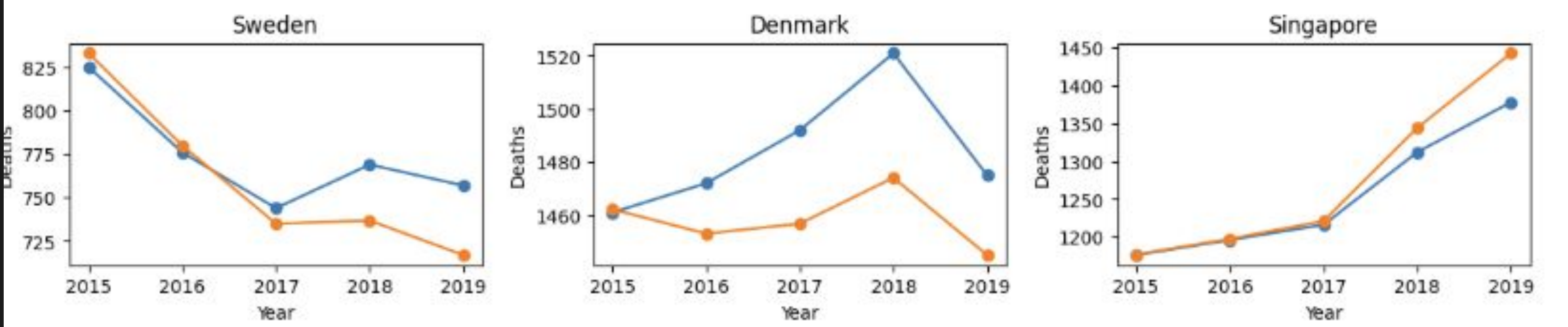


Methodology

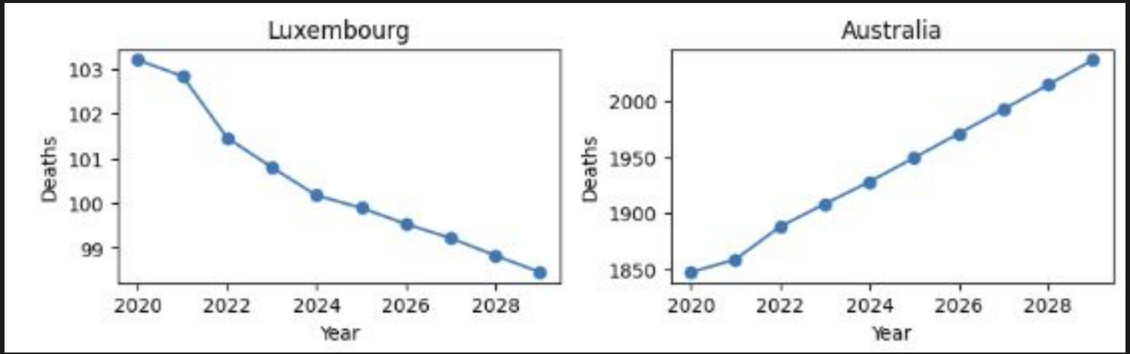
Multiple Linear Regression
Scope: Selected List of Developed Countries
Split: Train 1990–2014, Test 2015–2019
Methodology:

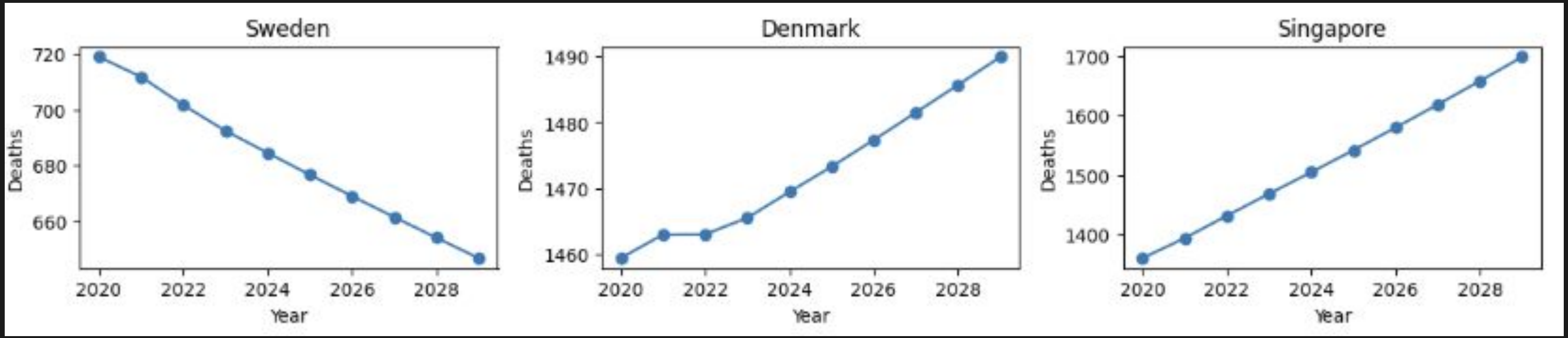
- **Ridge Regularization**
- **Log Transformation**
- **Choosing Features with Highest Correlation**



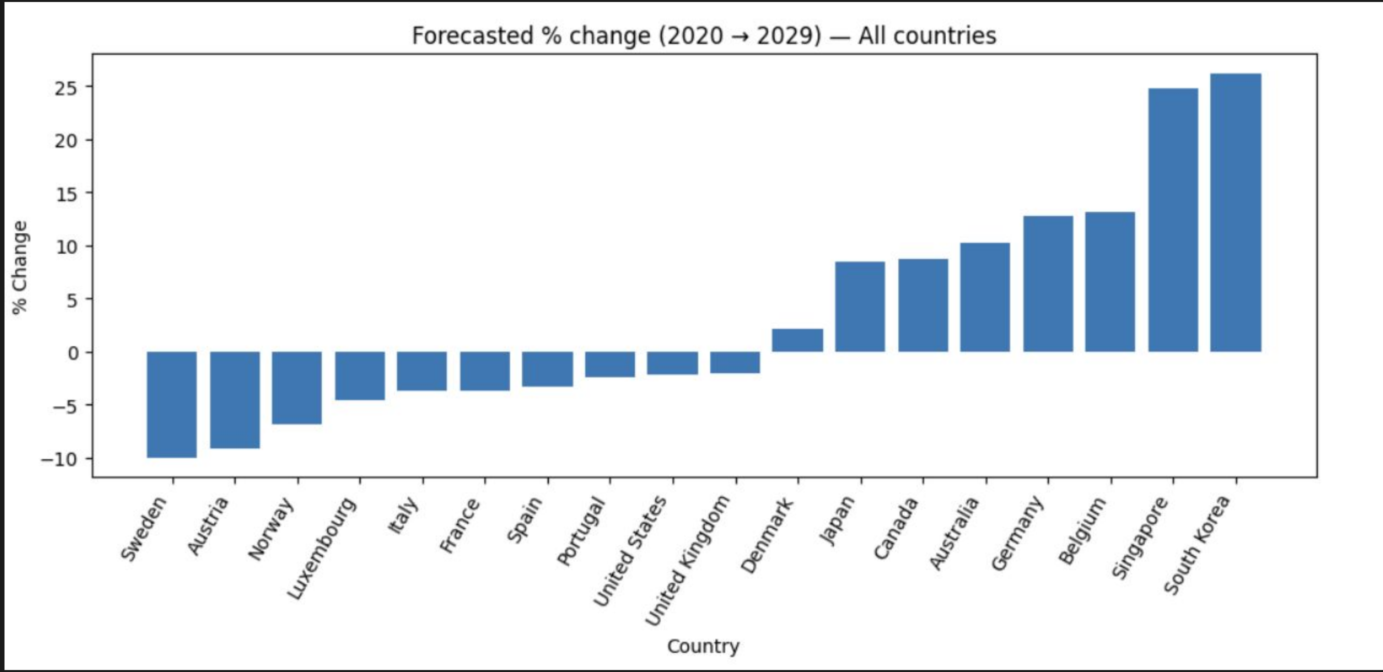


Outdoor Air Pollution Deaths Forecast (2020-2029)





What this Means



Take Away from our Predictions

Mixed Forecasts: several countries flat/declining; a few modest increases.

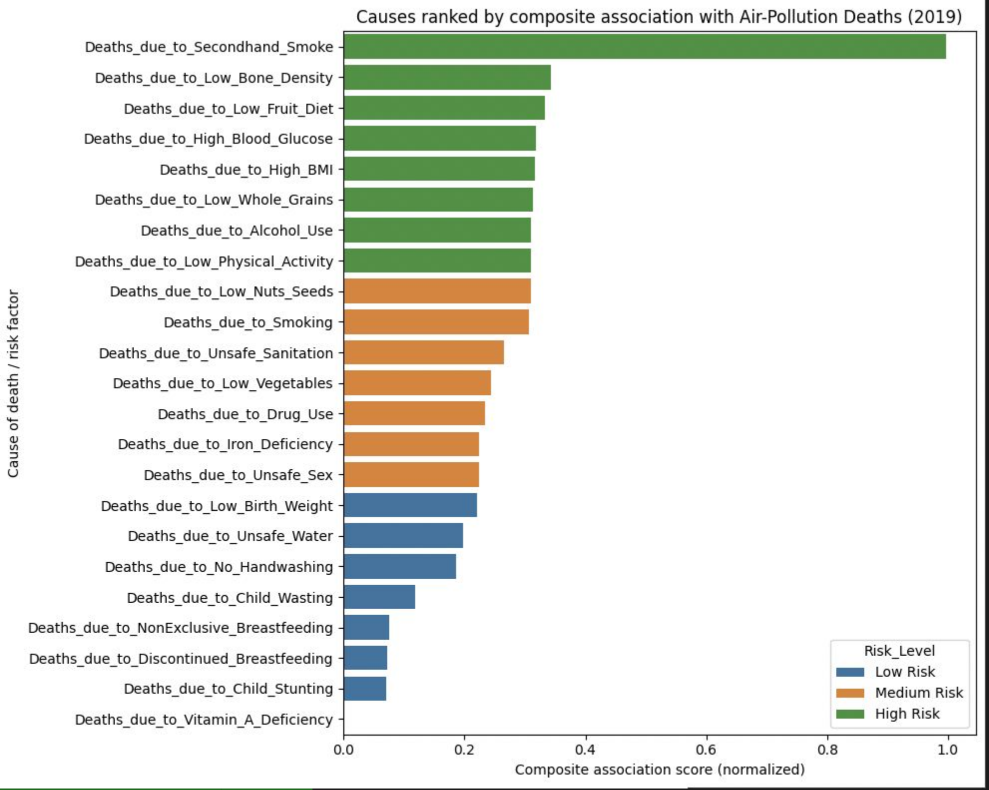
Declines: Sweden, Austria, Norway, Luxembourg.

Modest growth: Japan, Canada, Australia, Germany, Belgium.

Stand-outs: Singapore and **South Korea** show the largest projected increases.

Implications: prioritize monitoring/mitigation where growth is strongest; maintain/optimize resources where declining.

Lasso regression fit with 5-fold cross-validation



- Deaths due to secondhand smoke, deaths due to low bone density, and deaths due to low fruit diet indicate the highest risk of air pollution death
- Model depicts correlations between health factors and and associated risk levels (low, medium, high)

**correlation does not mean causation

Lasso Regression

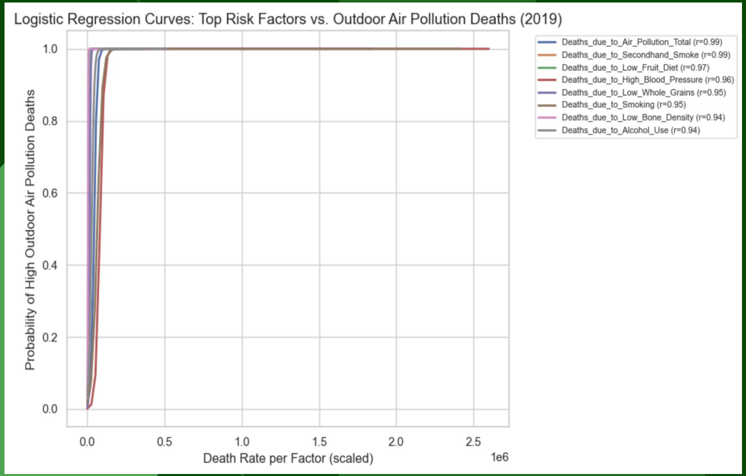
```
Model accuracy: 0.785
Weighted precision: 0.797
Weighted recall: 0.785

Classification Report:
              precision    recall  f1-score

        High       0.96      0.81      0.88
         Low       0.72      0.95      0.82
      Medium       0.71      0.59      0.65

    accuracy                           0.79
   macro avg       0.80      0.79      0.78
weighted avg       0.80      0.79      0.78
```

Performance Metrics

- Overall Accuracy: **78.5%**
  - Model performs reasonably well
- Class-by-Class Interpretation
  - **High Risk**
    - Precision (0.96): Almost always correct
    - Recall (0.81): Successfully identifies 81% of "High" cases
  - **Low Risk**
    - Precision (0.72): Moderately accurate when predicting "low"
    - Recall (0.95): Slightly less precise
  - **Medium Risk**
    - Precision (0.71): Some misclassifications
    - Recall (0.59): Lowest recall
- Final Interpretation:
  - Model performs best on "high" and "low" cases
  - "Medium" needs the most improvement

Challenges

Challenges faced



- **Choosing the right model** - balancing interpretability with limited data
- **Limited data availability** - missing key variables like GDP or pollution exposure
- **Dataset merging issues** - tried combining sources, but structure didn't align
- **Feature standardization** - had to convert values to **per 100 000 people** for fair comparison
- **Small sample size per country** - restricted use of advanced models (e.g., neural networks)
- **Refining research questions** - adjusted focus to match what our data could answer
- **Choosing the correct model -** ran several models and had to pick the best one (Logistic Regression is example of a failed model)

Suggestions For Future Work or Enhancements

- **Expand dataset**
  - Add more historical years and countries for stronger trends
- **Include broader indicators**
  - GDP, population, carbon emissions, energy use and policy data
- **Test advanced models**
  - Try random forest or time series forecasting once more data available
- **Add regional grouping**
  - Compare trends by continent or income level
- **Improve data quality**
  - Use consistent sources and fill missing years

Thank you!
Questions?