Emerging Talent

Milestone 3: Data Analysis

Air Pollution ($PM_{2.5}$) Health Impact Analysis: A Global Perspective

Group 9

Data analysis process

Data Collection

Ambient particulate matter pollution **Concentration**

Ambient particulate matter pollution **Exposure**

Data Preprocessing

Data Cleaning

Data Merging

Feature Engineering

Exploratory Data Analysis

Correlation Analysis

Distribution of key variables

Time series analysis

Model Building

Split Data

Train Linear Regression and Random Forest models

Model Evaluation

R^2

RMSE

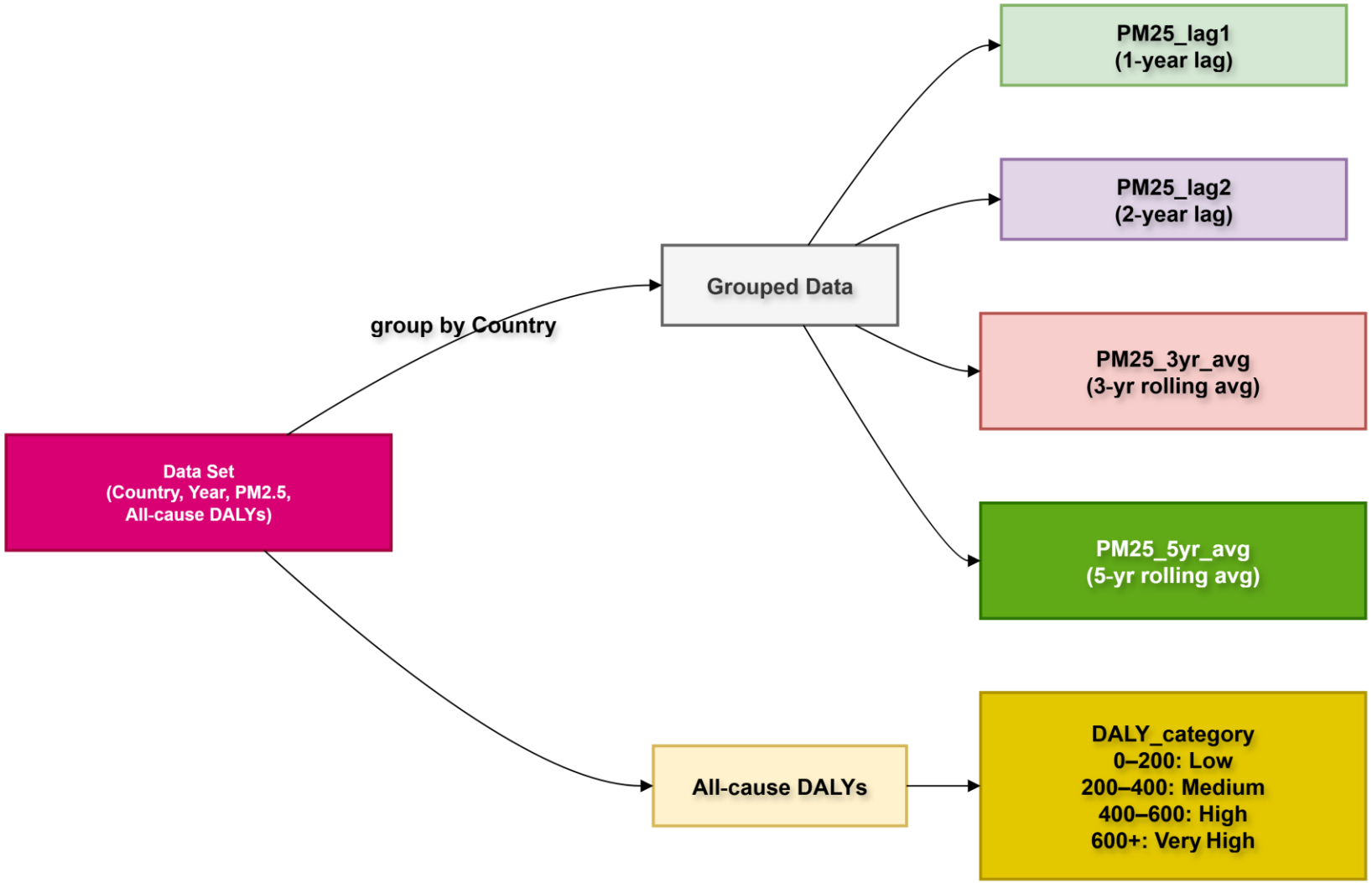
Correlation Analysis

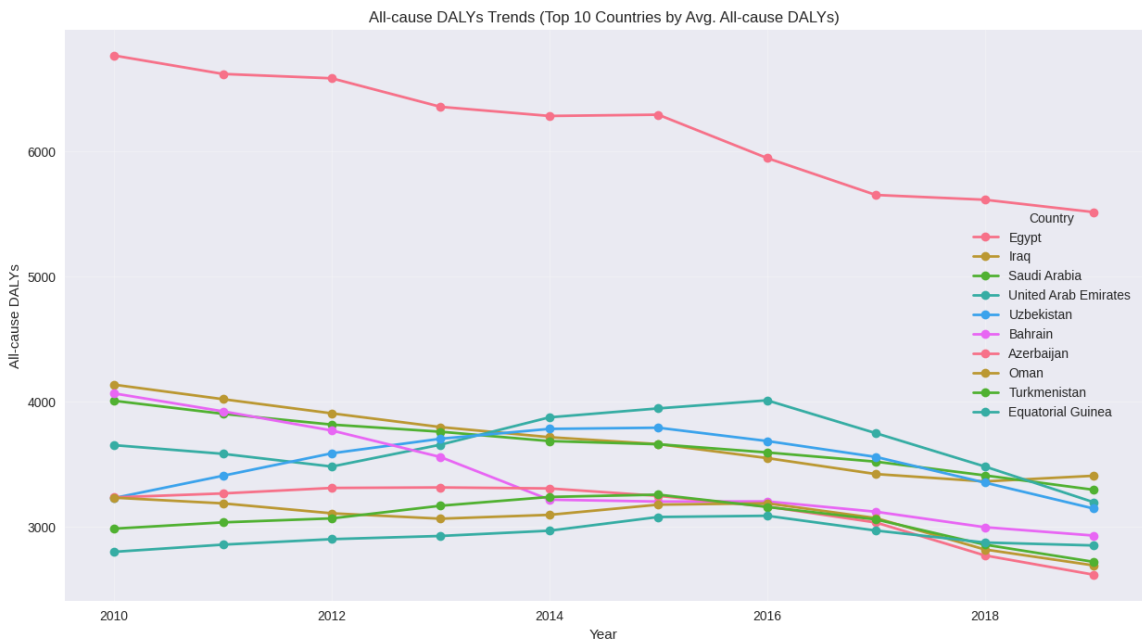
Datasets

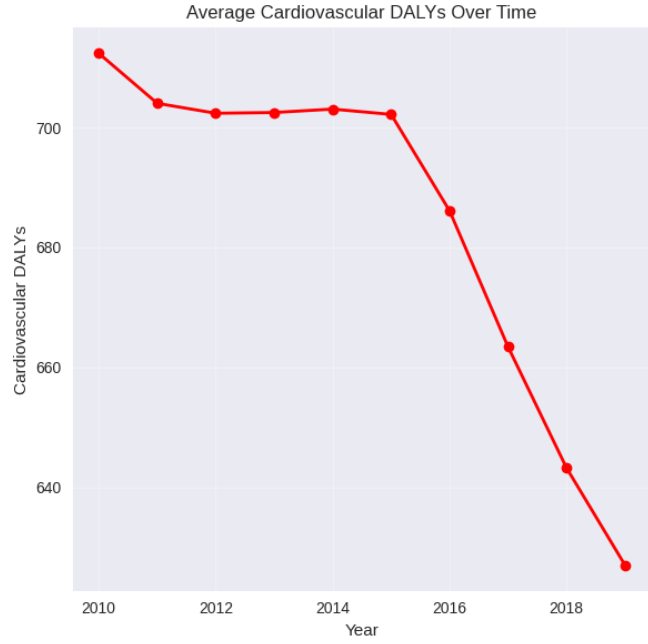
	Country	Year	SDI	PM2.5	All-cause DALYs	Cardiovascular DALYs	Stroke DALYs	Respiratory DALYs
0	Georgia	2010	0.805882	22.84	1312.882621	977.414421	407.617311	43.139413
1	Georgia	2011	0.811503	25.98	1270.511194	894.030277	363.367273	48.459765
2	Georgia	2012	0.815983	25.60	1327.455940	912.853130	364.818925	51.909633
3	Georgia	2013	0.820060	23.92	1384.849753	946.309316	387.791719	55.317720
4	Georgia	2014	0.823730	22.76	1419.667517	972.114994	410.655051	55.280263

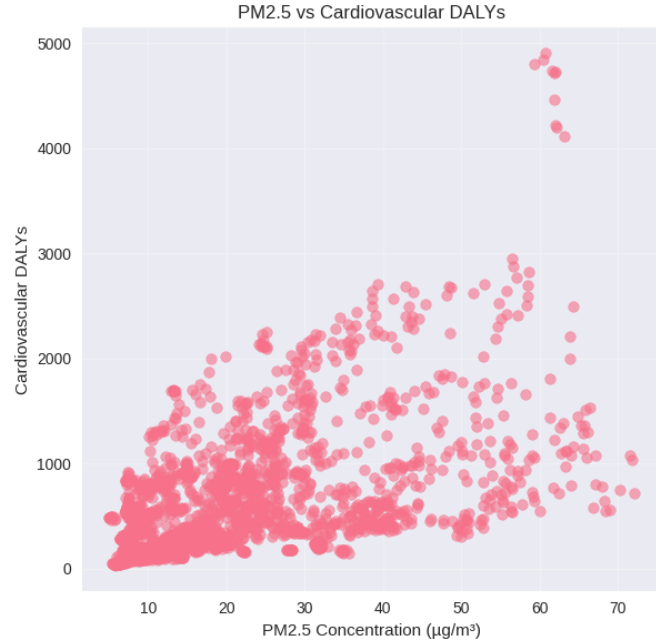

```
Total observations: 1950
Countries analyzed: 195
Time period: 2010 - 2019
```

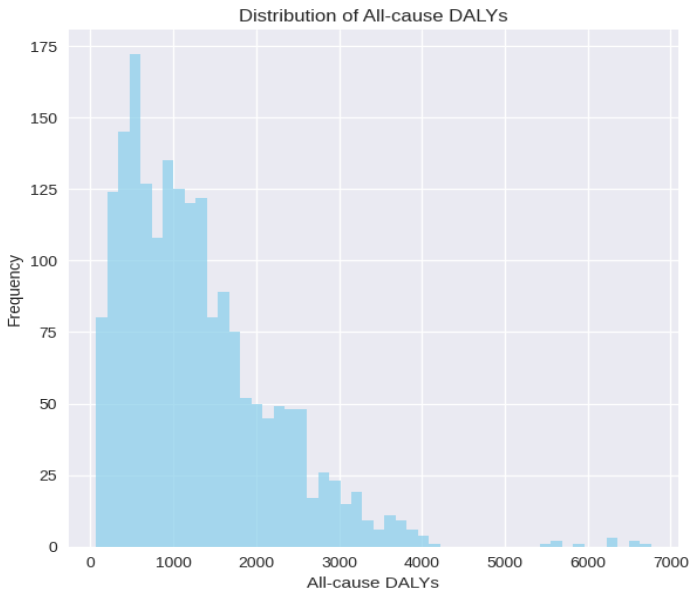
Data Preprocessing and Feature Engineering

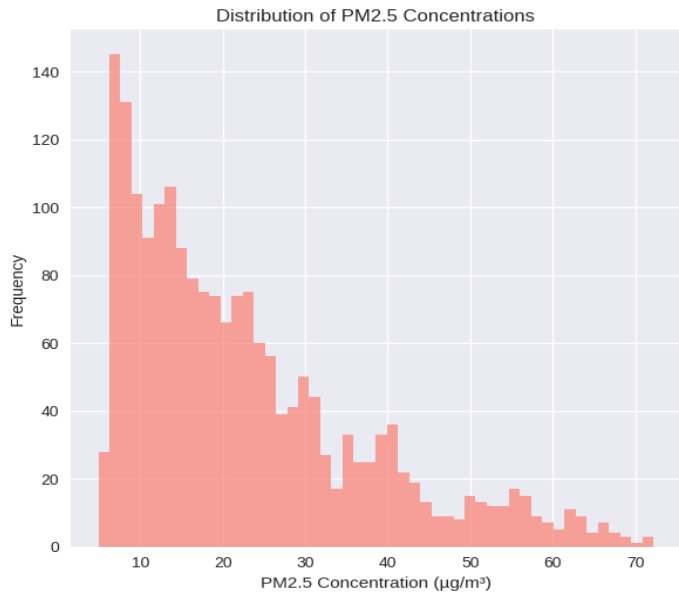


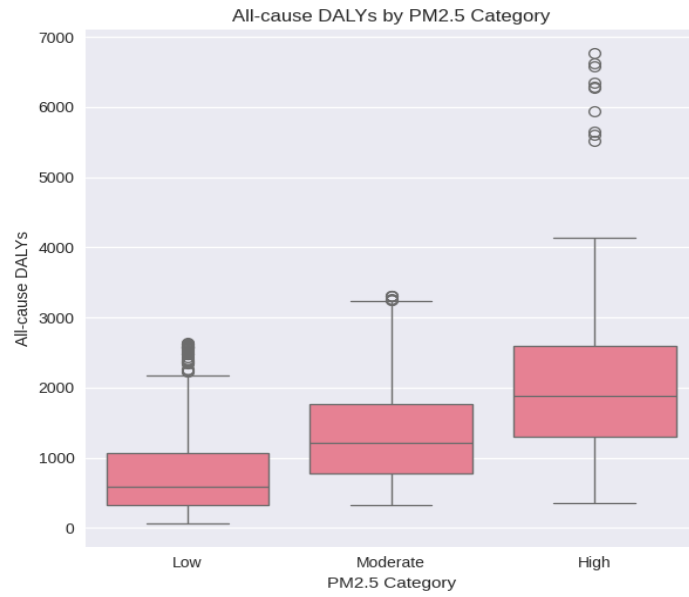

All-cause DALYs Trends (Top 10 Countries by Avg. All-cause DALYs)
All-cause DALYs
Year
Country
Egypt
Iraq
Saudi Arabia
United Arab Emirates
Uzbekistan
Bahrain
Azerbaijan
Oman
Turkmenistan
Equatorial Guinea

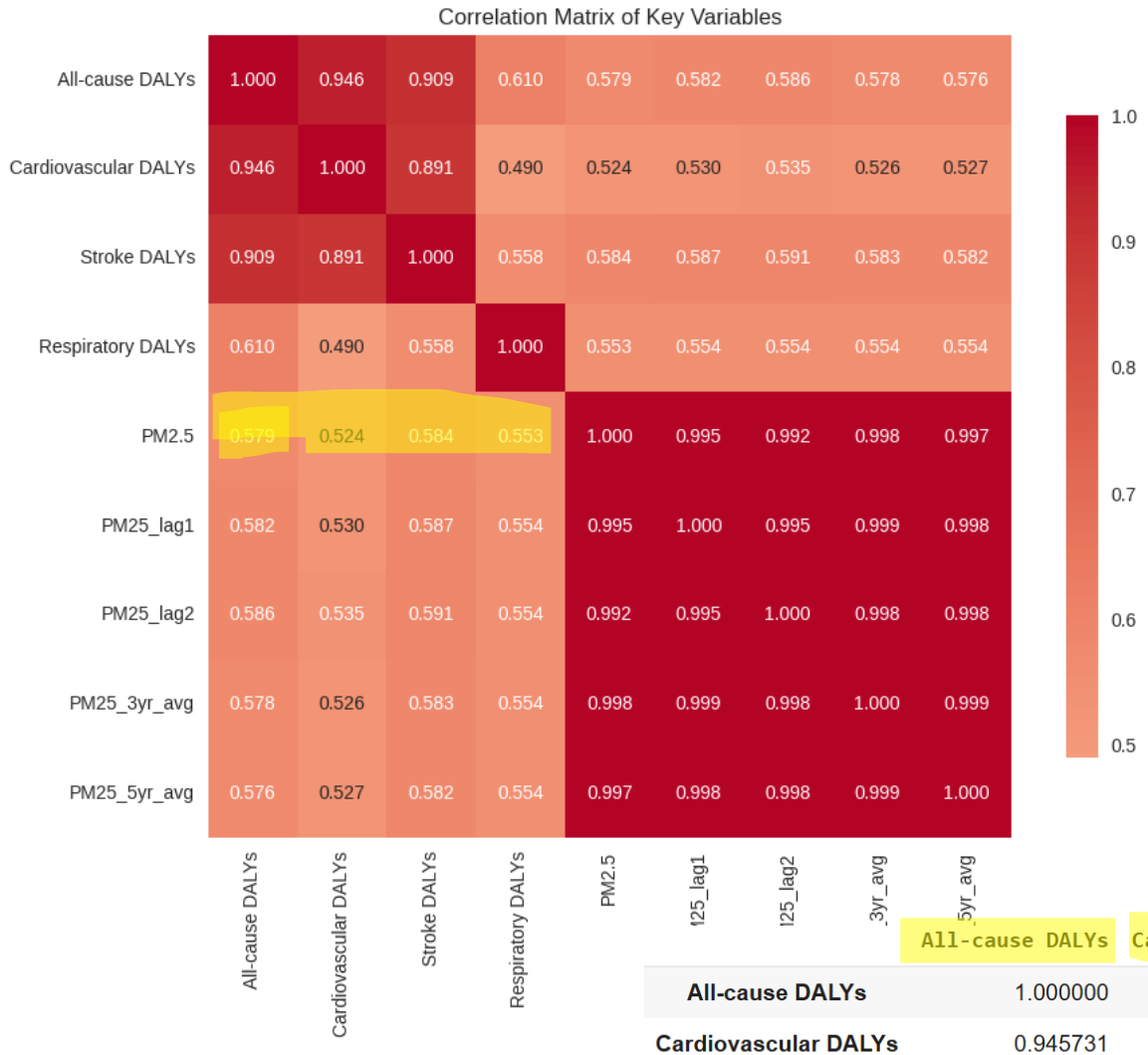

Average Cardiovascular DALYs Over Time
Cardiovascular DALYs
Year


PM2.5 vs Cardiovascular DALYs
Cardiovascular DALYs
PM2.5 Concentration (µg/m³)


Distribution of All-cause DALYs
Frequency
All-cause DALYs

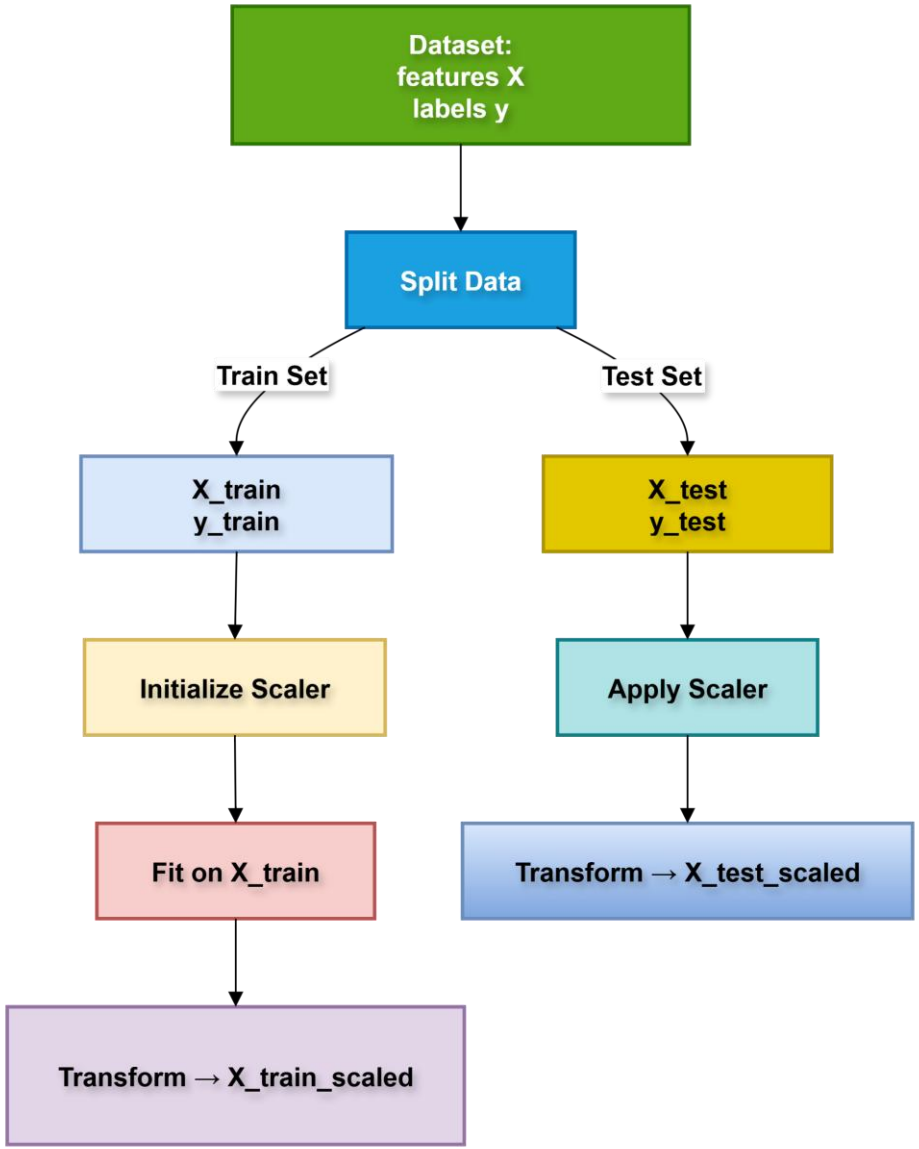

Distribution of PM2.5 Concentrations
Frequency
PM2.5 Concentration (µg/m³)

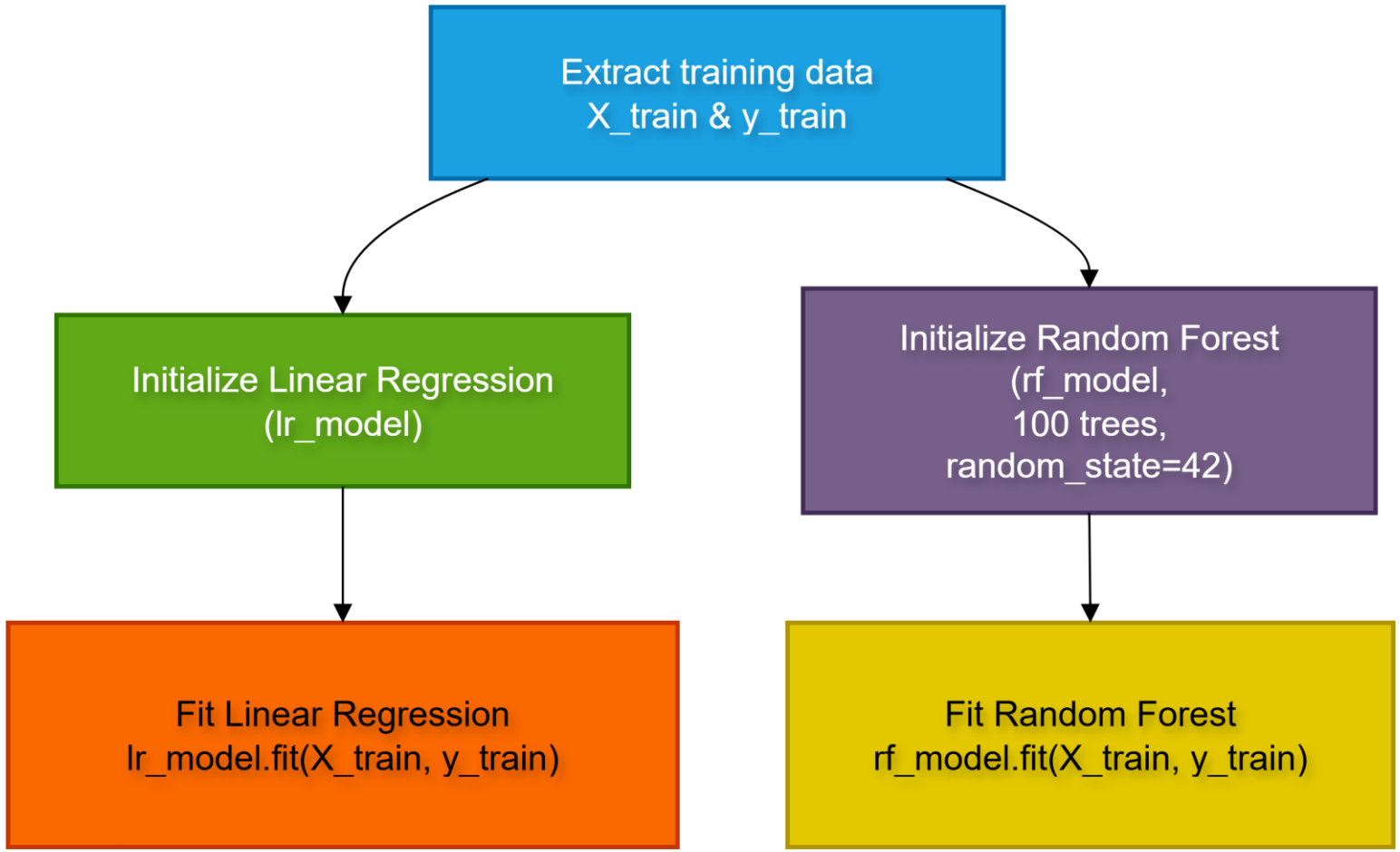

All-cause DALYs by PM2.5 Category
All-cause DALYs
Low
Moderate
High
PM2.5 Category



	All-cause DALYs	Cardiovascular DALYs	Stroke DALYs	Respiratory DALYs	PM2.5	PM25_lag1	PM25_lag2	PM25_3yr_avg	PM25_5yr_avg
All-cause DALYs	1.000000	0.945731	0.908569	0.609804	0.578767	0.582220	0.585576	0.577853	0.576342
Cardiovascular DALYs	0.945731	1.000000	0.891132	0.490146	0.524498	0.530011	0.535299	0.525898	0.526582
Stroke DALYs	0.908569	0.891132	1.000000	0.558182	0.584248	0.587427	0.590877	0.583413	0.581893
Respiratory DALYs	0.609804	0.490146	0.558182	1.000000	0.553190	0.554438	0.554295	0.553873	0.553592
PM2.5	0.578767	0.524498	0.584248	0.553190	1.000000	0.995100	0.992430	0.998042	0.996825
PM25_lag1	0.582220	0.530011	0.587427	0.554438	0.995100	1.000000	0.995315	0.998717	0.997892
PM25_lag2	0.585576	0.535299	0.590877	0.554295	0.992430	0.995315	1.000000	0.997798	0.997999
PM25_3yr_avg	0.577853	0.525898	0.583413	0.553873	0.998042	0.998717	0.997798	1.000000	0.999401
PM25_5yr_avg	0.576342	0.526582	0.581893	0.553592	0.996825	0.997892	0.997999	0.999401	1.000000

Split each dataset into training and testing sets.




Extract training data
X_train & y_train
Initialize Linear Regression
(lr_model)
Initialize Random Forest
(rf_model,
100 trees,
random_state=42)
Fit Linear Regression
lr_model.fit(X_train, y_train)
Fit Random Forest
rf_model.fit(X_train, y_train)

--- Summary of Model Performance ---

DALY Outcome	Model	R^2	RMSE
All-cause DALYs	Linear Regression	0.486215	646.716375
	Random Forest	0.788944	414.497636
Cardiovascular DALYs	Linear Regression	0.459761	460.639213
	Random Forest	0.792943	285.175787
Stroke DALYs	Linear Regression	0.459133	149.298449
	Random Forest	0.734233	104.655153
Respiratory DALYs	Linear Regression	0.348486	65.789872
	Random Forest	0.652087	48.076433

=== KEY FINDINGS AND INSIGHTS ===

1. DATASET OVERVIEW:
   - Total observations: 1950
   - Countries analyzed: 195
   - Time period: 2010 - 2019

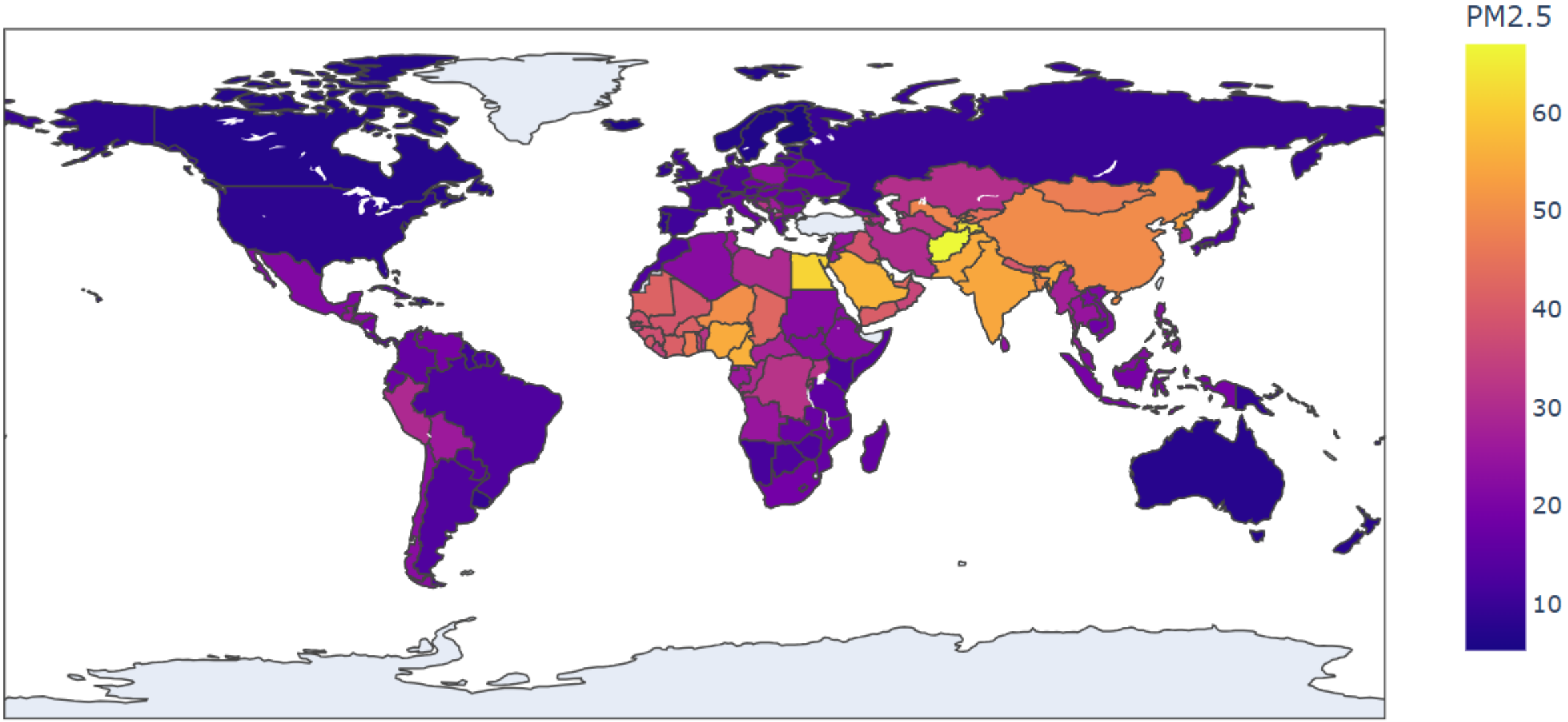
2. PM2.5 EXPOSURE PATTERNS:
   - Average PM2.5 concentration: 22.67 µg/m³
   - Highest PM2.5 country: Afghanistan (67.01 µg/m³)
   - PM2.5 trend over time: Decreasing

3. HEALTH BURDEN PATTERNS (All-cause DALYs):
   - Average All-cause DALY value: 1280.52
   - Highest All-cause DALY country: Egypt (6160.37)
   - All-cause DALY trend over time: Decreasing

4. MODEL PERFORMANCE (Summary):
   - Best performing model overall (based on R²): Linear Regression for Cardiovascular DALYs
   - Highest R² achieved: 0.991

5. LAG EFFECTS (All-cause DALYs):
   - Strongest PM2.5 feature correlation with All-cause DALYs: PM2.5 (r = 0.591)

Geographic Distribution of Average $PM_{2.5}$ Concentrations (2010–2019)



Workflow Diagram: $PM_{2.5}$ Exposure and DALY Analysis Framework