Air Quality Measures Analysis

This project provides an in-depth analysis of air quality data sourced from the National Environmental Health Tracking Network. The main goals are to clean and analyse the dataset, explore trends across regions and years, and build a simple Linear Regression model to predict average air quality measures over time.

Dataset Overview

1. Original Source:
Air_Quality_Measures_on_the_National_Environmental_Health_Tracking_Network.csv

2. Cleaned Dataset:
Air_Quality_Measures_on_the_National_Environmental_Health_Tracking_Network (1).csv

3. Selected Features for Analysis:

MeasureId
MeasureName
MeasureType
StateName
CountyName
ReportYear
Value

Data Cleaning

The initial dataset was cleaned by: -

Removing irrelevant columns

Handling missing values

Ensuring consistent datatypes

Retaining only essential fields for analysis and modelling

Exploratory Data Analysis (EDA)

1. Descriptive Statistics for Value

Statistic	Value
Count	404,394
Mean	20.89
Standard Deviation	31.06

Statistic	Value
Max	275.0
Skewness	2.47
Kurtosis	8.79

The distribution of Value (air quality measure score) is highly skewed, indicating potential outliers or metrics with large variances.

Temporal Trend

A time-series plot was created to visualise the Average Air Quality Measure Value Over Years, revealing how air quality trends have evolved.
Plot saved as:* average_value_over_years.png

Top 10 States by Average Value

StateName	Average Value
Oklahoma	35.98
West Virginia	34.50
New Jersey	34.18
Florida	33.67
North Carolina	32.40
South Carolina	31.57
Georgia	31.03
Tennessee	29.98
California	29.21
Illinois	28.32

Visualisation saved as: top_10_states_by_avg_value.png

These results highlight regional differences, with certain southern and midwestern states showing higher average air quality measure values.

Time-Series Modelling (Linear Regression)

A simple Linear Regression model was implemented using sklearn to predict the aggregated average air quality values based on the reporting year.

Modelling Workflow

Data Aggregation: Grouped data by ReportYear and computed the mean AverageValue

Split: 80% training and 20% testing

Model: sklearn.linear_model.LinearRegression

Evaluation: Model performance was assessed using MSE and R^2 metrics

Model Performance

Metric	Result
Mean Squared Error (MSE)	1.96
R-squared (R^2) Score	0.66

An R^2 score of 0.66 indicates that the linear trend of ReportYear explains approximately 66% of the variance in annual average air quality measure values.

Prediction Output

```
ReportYear |AverageValue |PredictedValue |2000 |
30.83 |   30.60 |
2001 |28.24 |   29.83 |
2002 |31.32 |   29.07 |
2003 |27.91 |   28.30 |
2004 |25.92 |   27.53 |
```

Required Libraries:

pandas

matplotlib

seaborn

sklearn

Author

Air Quality Measures Analysis Developed for educational and analytical pur- poses.
by pushkal thakre