Laporan Mata Kuliah Pembelajaran Mesin

Project 2 : Linear dan Polynomial Regression

disusun untuk memenuhi
tugas mata kuliah pembelajaran mesin
oleh :

Kelompok : 1

Anggota :

M. Syahidal Akbar Zas	**(2208107010045)**
Muhammad Raihan	**(2208107010021)**
Ammar Qurthuby	**(2208107010031)**
Azri Harniza	**(2208107010034)**

DEPARTEMEN INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SYIAH KUALA

TAHUN 2025

DAFTAR ISI

BAB 1

PENDAHULUAN

A. Latar Belakang

Polusi udara telah menjadi salah satu isu lingkungan paling kritis di abad ke-21, dengan dampak yang signifikan terhadap kesehatan manusia, lingkungan, dan ekonomi global. Menurut Organisasi Kesehatan Dunia (WHO), sekitar 7 juta kematian prematur setiap tahun dikaitkan dengan paparan polusi udara, menjadikannya salah satu ancaman kesehatan lingkungan terbesar yang dihadapi manusia saat ini. Peningkatan urbanisasi, industrialisasi, dan pertumbuhan populasi telah menyebabkan ekskalasi emisi polutan udara di berbagai belahan dunia.

Indeks Kualitas Udara (Air Quality Index - AQI) merupakan parameter utama yang digunakan secara global untuk mengukur dan mengkomunikasikan tingkat polusi udara kepada masyarakat umum. AQI mempertimbangkan konsentrasi berbagai polutan berbahaya, termasuk partikulat (PM2.5 dan PM10), ozon (O_3), karbon monoksida (CO), nitrogen dioksida (NO_2), dan sulfur dioksida (SO_2). Namun, identifikasi faktor-faktor yang berkontribusi terhadap fluktuasi AQI masih memerlukan penelitian lebih lanjut, terutama dengan mempertimbangkan variasi geografis dan pengaruh faktor meteorologi.

Memahami hubungan antara AQI dan berbagai faktor lingkungan seperti kadar polutan spesifik (CO, NO_2, SO_2) serta parameter meteorologi (suhu, kelembaban, kecepatan angin) menjadi sangat penting untuk mengembangkan strategi mitigasi yang efektif. Analisis hubungan ini dapat memberikan wawasan berharga mengenai bagaimana berbagai faktor berinteraksi dan berdampak pada kualitas udara secara keseluruhan.

Penelitian ini menggunakan dataset polusi udara global untuk menganalisis korelasi dan hubungan antara AQI sebagai variabel target dengan berbagai faktor lingkungan sebagai variabel independen. Dengan memanfaatkan teknik analisis data dan pemodelan statistik, penelitian ini bertujuan untuk mengidentifikasi variabel-variabel

yang memiliki pengaruh paling signifikan terhadap AQI, serta memprediksi tingkat polusi udara berdasarkan parameter-parameter tersebut

B. Rumusan Masalah

1. Bagaimana hubungan antara konsentrasi polutan udara tertentu (CO, NO_2, SO_2) dengan Indeks Kualitas Udara (AQI) pada dataset polusi udara global?
2. Seberapa besar pengaruh faktor meteorologi seperti suhu, kelembaban, dan kecepatan angin terhadap variasi nilai Indeks Kualitas Udara (AQI)?
3. Di antara variabel independen yang dianalisis (kadar CO, NO_2, SO_2, suhu, kelembaban, kecepatan angin), manakah yang memiliki kontribusi paling signifikan terhadap perubahan nilai AQI?
4. Bagaimana model prediktif dapat dikembangkan untuk memperkirakan nilai AQI berdasarkan kombinasi faktor polutan dan meteorologi, serta seberapa akurat model tersebut?
5. Apakah terdapat perbedaan signifikan pada pola hubungan antara variabel independen dan AQI di berbagai wilayah geografis yang tercakup dalam dataset global?

C. Tujuan

1. Menganalisis dan mengukur kekuatan hubungan antara konsentrasi polutan udara tertentu (CO, NO_2, SO_2) dengan Indeks Kualitas Udara (AQI) menggunakan dataset polusi udara global.
2. Mengevaluasi besarnya pengaruh faktor meteorologi seperti suhu, kelembaban, dan kecepatan angin terhadap variasi nilai Indeks Kualitas Udara (AQI).
3. Mengidentifikasi variabel independen yang memiliki kontribusi paling signifikan terhadap perubahan nilai AQI melalui analisis statistik dan pemodelan.
4. Mengembangkan model prediktif untuk memperkirakan nilai AQI berdasarkan kombinasi faktor polutan dan meteorologi, serta mengevaluasi tingkat akurasi model tersebut.
5. Membandingkan pola hubungan antara variabel independen dan AQI di berbagai wilayah geografis untuk mengidentifikasi variasi regional dalam faktor-faktor yang mempengaruhi kualitas udara.

BAB 2

METODOLOGI

1. Pengumpulan Data

Data yang digunakan dalam penelitian ini berasal dari dataset kualitas udara dengan beberapa fitur seperti konsentrasi polutan (misalnya PM10, SO2, CO, O3, dan NO2), serta label kualitas udara yang dikategorikan (seperti Baik, Sedang, Tidak Sehat, dsb). Dataset diimpor dalam format CSV dan dibaca menggunakan library *pandas*.

2. Pra-pemrosesan Data dan Eksplorasi Data

Tahap ini bertujuan untuk menyiapkan data agar layak untuk dianalisis dan digunakan dalam pelatihan model. Langkah-langkahnya adalah sebagai berikut:

a. Mengecek informasi awal dataset dengan menggunakan air.info() dan air.describe(), untuk memahami tipe data dan distribusi nilai.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23463 entries, 0 to 23462
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Country             23036 non-null  object
 1   City                23462 non-null  object
 2   AQI Value           23463 non-null  int64 
 3   AQI Category        23463 non-null  object
 4   CO AQI Value        23463 non-null  int64 
 5   CO AQI Category     23463 non-null  object
 6   Ozone AQI Value     23463 non-null  int64 
 7   Ozone AQI Category  23463 non-null  object
 8   NO2 AQI Value       23463 non-null  int64 
 9   NO2 AQI Category    23463 non-null  object
 10  PM2.5 AQI Value     23463 non-null  int64 
 11  PM2.5 AQI Category  23463 non-null  object
dtypes: int64(5), object(7)
memory usage: 2.1+ MB
```

b. Pembersihan Data

i. Menghapus nilai-nilai yang kosong atau *missing values* menggunakan fungsi dropna() agar tidak mengganggu proses pelatihan model.

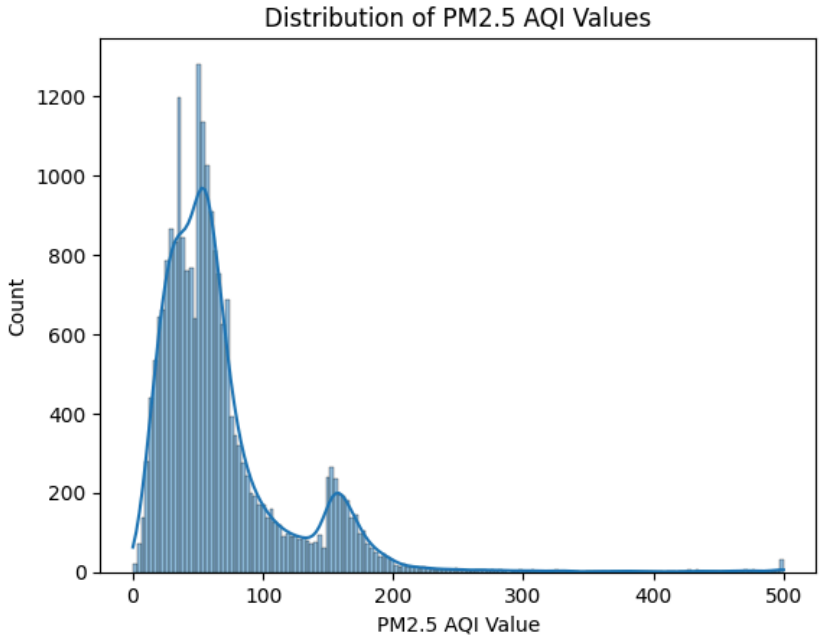
```
[ ] air.dropna(inplace=True)

    air.isnull().sum()
```

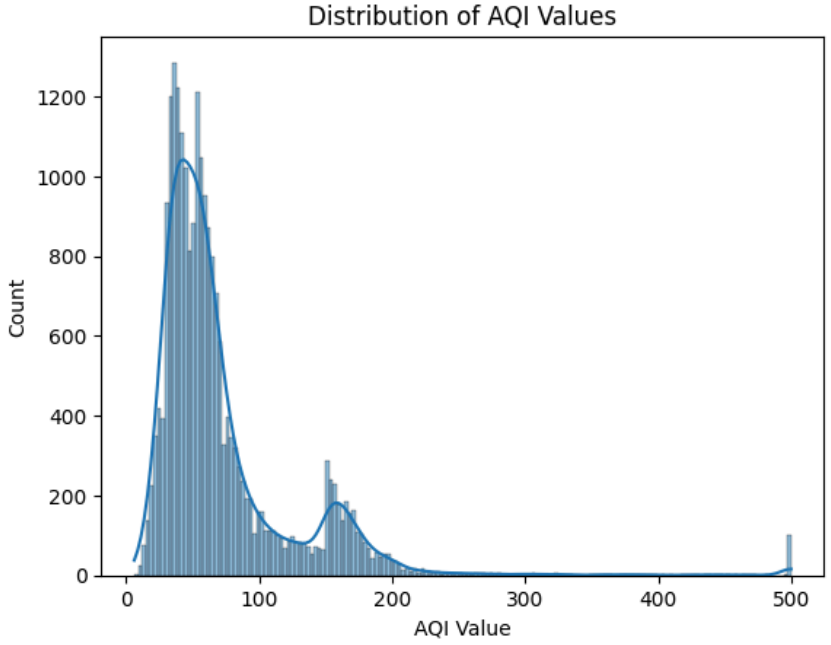
	0
Country	0
City	0
AQI Value	0
AQI Category	0
CO AQI Value	0
CO AQI Category	0
Ozone AQI Value	0
Ozone AQI Category	0
NO2 AQI Value	0
NO2 AQI Category	0
PM2.5 AQI Value	0
PM2.5 AQI Category	0

dtype: int64

c. Menampilkan dua histogram dengan kurva KDE (Kernel Density Estimation)

memvisualisasikan distribusi nilai kualitas udara (AQI Value) dan distribusi nilai partikel halus (PM2.5 AQI Value) dalam dataset, sehingga dapat dianalisis pola sebaran dan frekuensi nilai-nilai tersebut untuk memahami karakteristik polusi udara yang terjadi.

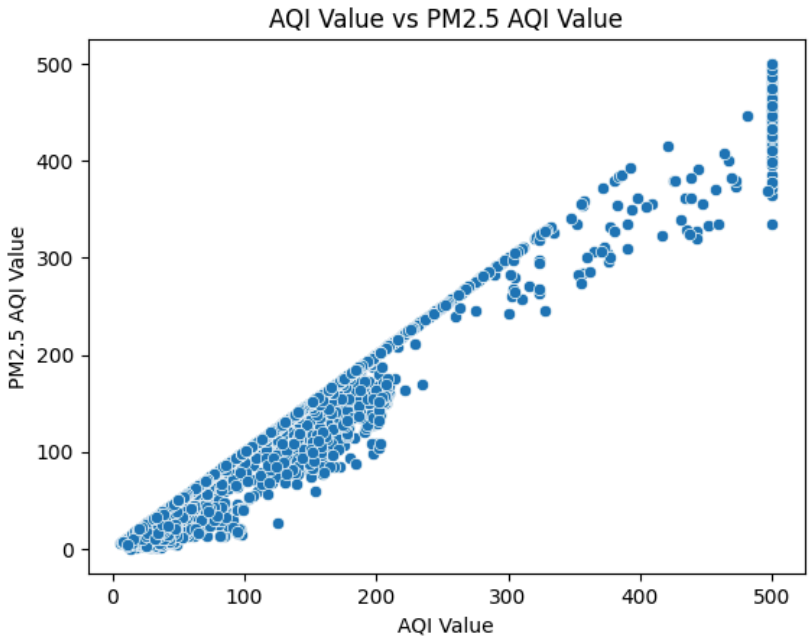


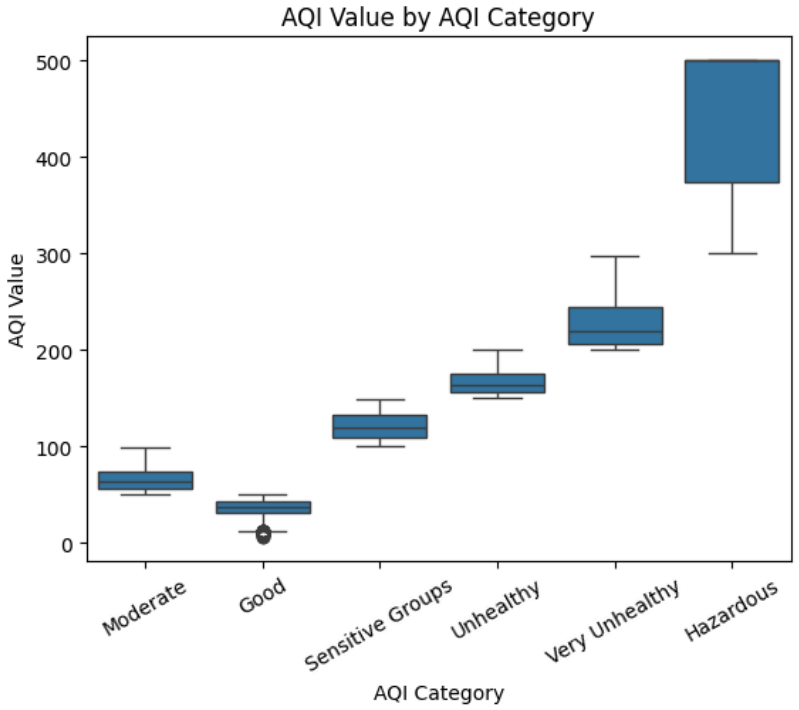


d. Distribusi Nilai AQI Berdasarkan Kategori Kualitas Udara

Scatter plot digunakan untuk menunjukkan hubungan dan korelasi antara nilai AQI secara umum dengan nilai AQI PM2.5, sehingga memudahkan identifikasi pola atau keterkaitan di antara keduanya. Sementara itu, box plot digunakan untuk membandingkan distribusi nilai AQI pada berbagai kategori

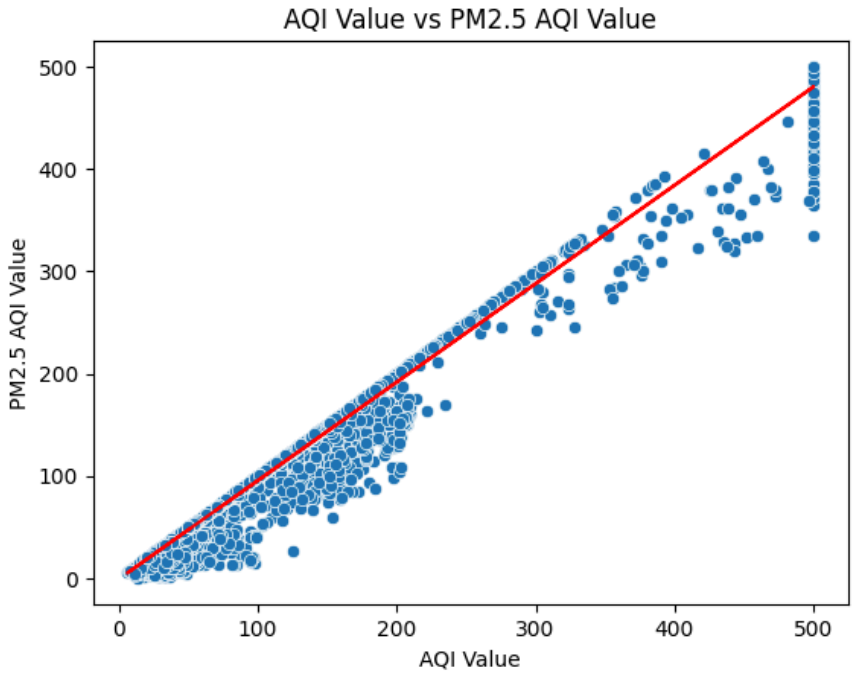
kualitas udara (seperti Good, Moderate, Unhealthy, dan lainnya), guna memvisualisasikan sebaran, median, serta outlier dalam setiap kategori.





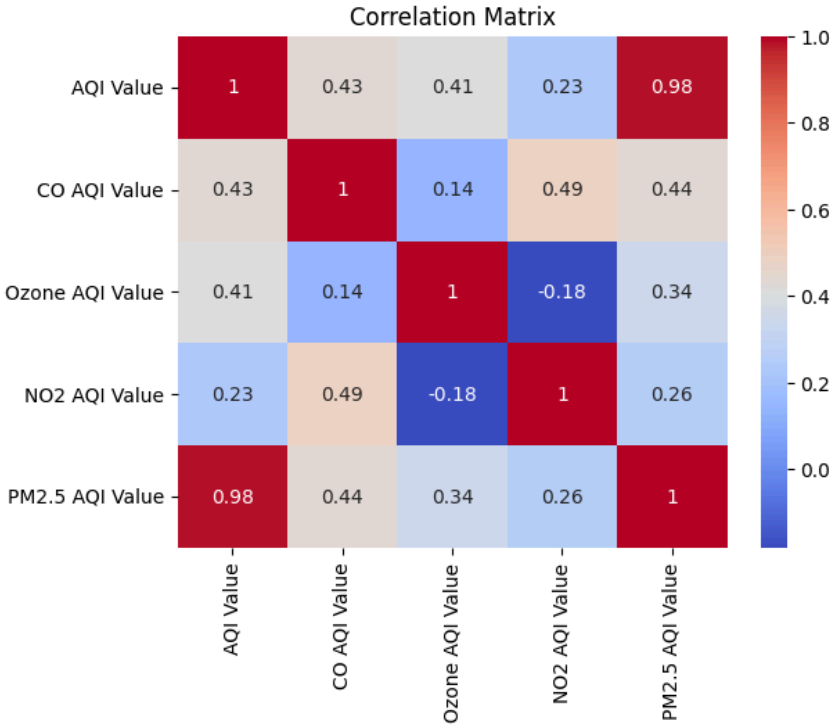
e. Hubungan Antara AQI Umum dan AQI PM2.5 dengan Regresi Linear

menampilkan scatter plot antara nilai AQI umum dan AQI PM2.5, dilengkapi dengan garis regresi linear berwarna merah. Visualisasi ini membantu menunjukkan adanya korelasi antara kedua variabel, serta memberikan gambaran pola hubungan linear antara tingkat polusi udara secara umum dengan konsentrasi PM2.5.



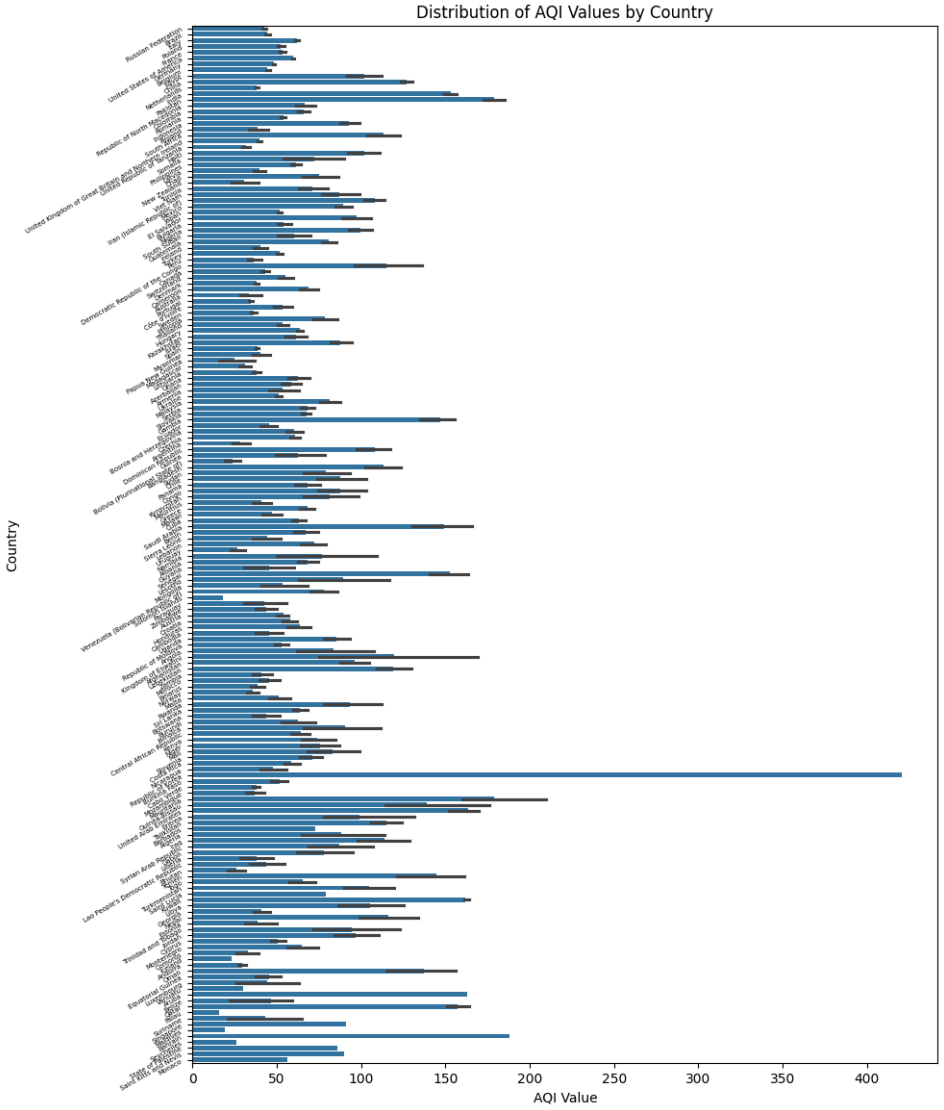
f. Matriks Korelasi AQI Berdasarkan Polutan

menunjukkan korelasi antara nilai AQI secara keseluruhan dengan nilai AQI masing-masing polutan (CO, Ozone, NO2, dan PM2.5), membantu mengidentifikasi hubungan antar parameter kualitas udara.


Correlation Matrix
AQI Value
CO AQI Value
Ozone AQI Value
NO2 AQI Value
PM2.5 AQI Value
1
0.43
0.41
0.23
0.98
0.43
1
0.14
0.49
0.44
0.41
0.14
1
-0.18
0.34
0.23
0.49
-0.18
1
0.26
0.98
0.44
0.34
0.26
1
AQI Value
CO AQI Value
Ozone AQI Value
NO2 AQI Value
PM2.5 AQI Value
1.0
0.8
0.6
0.4
0.2
0.0

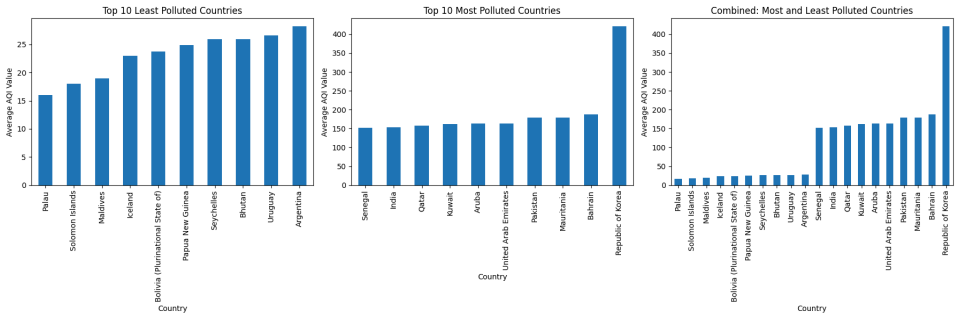
g. Distribusi Nilai AQI per Negara

menampilkan distribusi nilai AQI di berbagai negara, memberikan gambaran perbandingan tingkat polusi udara antar negara berdasarkan nilai AQI. Visualisasi ini memudahkan identifikasi negara dengan kualitas udara lebih tinggi atau rendah.



h. Perbandingan Rata-rata AQI Negara Paling Bersih dan Paling Tercemar

terdiri dari tiga barplot yang membandingkan rata-rata nilai AQI per negara. Plot pertama menampilkan 10 negara dengan kualitas udara terbaik (AQI terendah), plot kedua menunjukkan 10 negara dengan polusi tertinggi (AQI tertinggi), dan plot ketiga menggabungkan keduanya untuk memberikan perbandingan yang lebih jelas.



i. Encoding Label

i. Kolom Kualitas_Udara yang berupa data kategorikal diubah menjadi data numerik menggunakan LabelEncoder dari sklearn.preprocessing.

```
# categorical values into numerical format
from sklearn.preprocessing import LabelEncoder
def label(i) :
    air[i] = LabelEncoder().fit_transform(air[i])
    return air[i]

for i in air.iloc[:,:4].columns:
    label(i)

# 'object' to 'int'
for column in air.select_dtypes(include=['object']).columns:
    air[column] = LabelEncoder().fit_transform(air[column])

x = air[['Country','City','AQI Value', 'CO AQI Value', 'Ozone AQI Value', 'NO2 AQI Value','AQI Category']]

# Select target variable (output variable)
y = air['PM2.5 AQI Value']
```

3. Pembagian Dataset dan Hasil

a. Dataset yang telah dibersihkan kemudian dibagi menjadi dua bagian:
   i. **Data latih (training set)** sebanyak 80% dari total data.
   ii. **Data uji (testing set)** sebanyak 20% dari total data.

   Pembagian dilakukan menggunakan fungsi train_test_split dari library sklearn.model_selection dengan parameter random_state untuk menjaga konsistensi hasil eksperimen.

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)
```

b. Regresi Linear untuk Prediksi AQI PM2.5

   Model regresi linear dilatih menggunakan data latih (x_train, y_train) untuk memprediksi nilai AQI PM2.5 berdasarkan data uji (x_test). Hasil prediksi disimpan dalam y_pred, yang nantinya dapat digunakan untuk evaluasi performa model atau visualisasi hasil prediksi.

```
from sklearn.linear_model import LinearRegression

regression = LinearRegression()
regression.fit(x_train, y_train)
y_pred = regression.predict(x_test)
```

c. Ringkasan Statistik AQI dan Polutan Utama

   merangkum nilai tertinggi, terendah, dan rata-rata dari AQI serta empat jenis polutan utama (CO, Ozone, NO2, dan PM2.5). Selain itu, ditampilkan juga negara dengan rata-rata polusi tertinggi dan terendah berdasarkan nilai AQI keseluruhan, memberikan gambaran umum kualitas udara global dalam dataset.

	Statistic	AQI Value	CO AQI Value	Ozone AQI Value	NO2 AQI Value	PM2.5 AQI Value	Most Polluted Country	Least Polluted Country
0	Highest	346.00	133.00	235.00	91.00	500.00	126	117
1	Lowest	0.00	0.00	0.00	0.00	0.00	126	117
2	Average	65.45	1.38	35.23	3.08	68.88	126	117

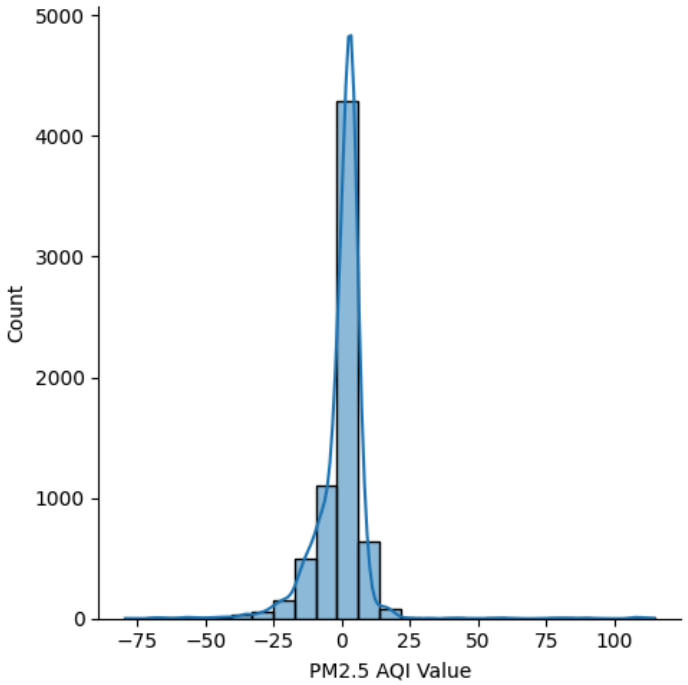
d. Evaluasi Performa Model Regresi Linear

Evaluasi performa model regresi linear dilakukan menggunakan empat metrik: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan Mean Absolute Percentage Error (MAPE). MAE menunjukkan rata-rata selisih absolut antara nilai prediksi dan nilai aktual, sedangkan MSE menghitung rata-rata kuadrat dari selisih tersebut. RMSE memberikan interpretasi kesalahan dalam satuan aslinya, dan MAPE menyajikan tingkat kesalahan dalam bentuk persentase. Keempat metrik ini digunakan untuk menilai sejauh mana model mampu memprediksi nilai AQI PM2.5 secara akurat.

```
Mean Absolute Error 5.7897
Mean Squared Error 102.0117
Root Mean Squared Error 10.1001
Mean Absolute Percentage Error 0.1613
```

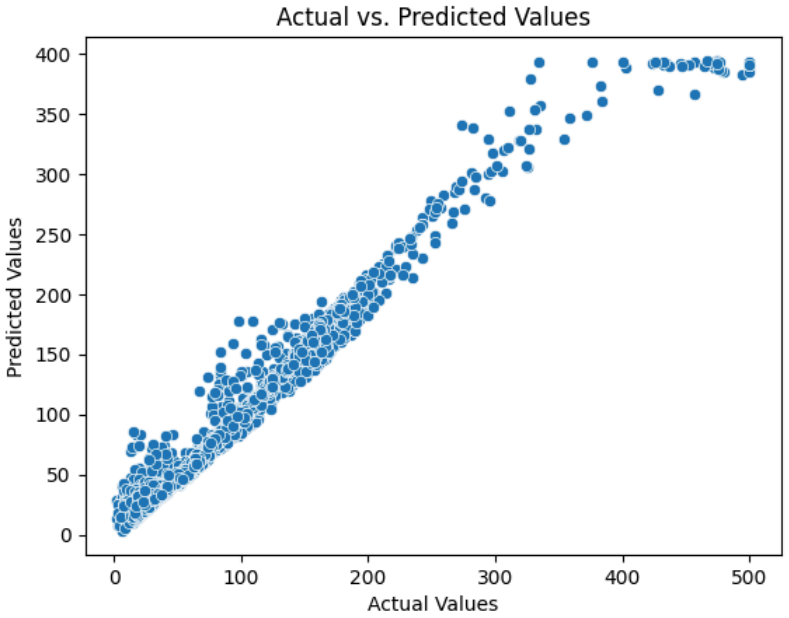
e. Distribusi Residual pada Model Regresi

menunjukkan distribusi residual (selisih antara nilai aktual dan prediksi) dari model regresi. Dengan menggunakan histogram dan kurva KDE (Kernel Density Estimation), visualisasi ini membantu mengidentifikasi apakah residual tersebar normal, yang menjadi salah satu indikator bahwa model regresi bekerja dengan baik tanpa bias sistematis.



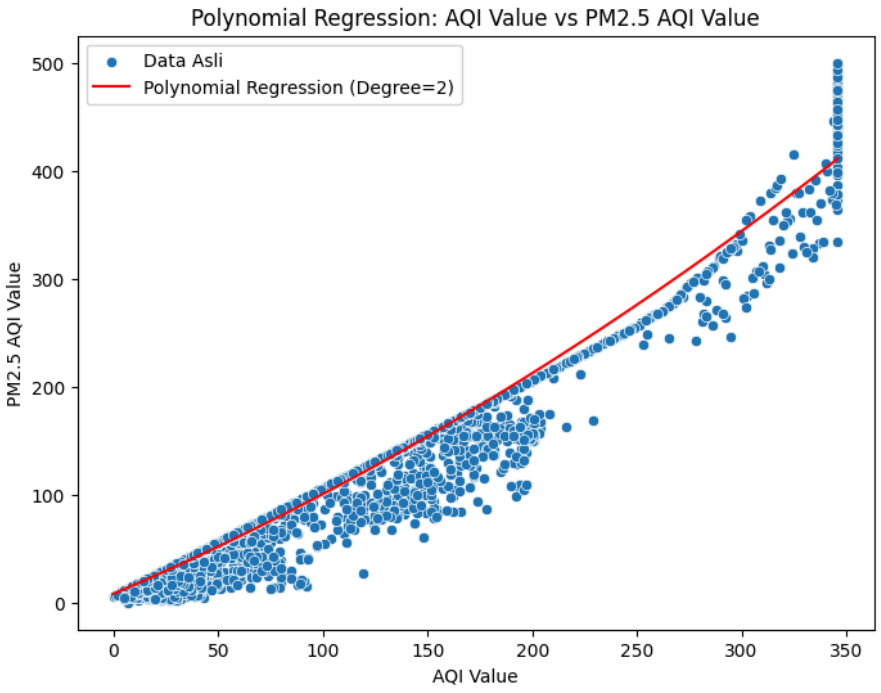
f. Perbandingan Nilai Aktual dan Prediksi

membandingkan nilai aktual dengan nilai prediksi dari model regresi. Titik-titik yang mendekati garis diagonal mengindikasikan prediksi yang akurat, sedangkan penyebaran yang jauh dari garis tersebut menunjukkan adanya deviasi atau kesalahan prediksi. Visualisasi ini berguna untuk mengevaluasi seberapa baik model merepresentasikan data.



g. Regresi Polinomial antara AQI Umum dan AQI PM2.5

menunjukkan hubungan antara nilai AQI umum dan nilai AQI PM2.5 menggunakan regresi polinomial derajat 2. Scatter plot merepresentasikan data asli, sementara kurva merah menggambarkan hasil prediksi model polinomial. Model ini digunakan untuk menangkap pola non-linear yang mungkin tidak terdeteksi oleh regresi linear sederhana.



h. Evaluasi Model Regresi Polinomial

Evaluasi performa regresi polinomial dilakukan dengan menghitung MAE, MSE, RMSE, dan MAPE. MAE mengukur rata-rata selisih absolut antara prediksi dan nilai aktual, MSE mengkuadratkan selisih untuk penalti kesalahan besar, RMSE memberikan interpretasi dalam satuan asli data, dan MAPE menunjukkan rata-rata kesalahan dalam persentase. Metrik-metrik ini digunakan untuk menilai seberapa baik model polinomial dalam memetakan hubungan antara AQI umum dan PM2.5.

BAB 3
KESIMPULAN

Berdasarkan hasil evaluasi, model Linear Regression memberikan performa yang lebih baik dibandingkan Polynomial Regression derajat 2 dalam memprediksi nilai AQI PM2.5. Hal ini ditunjukkan oleh nilai error yang lebih rendah pada Linear Regression, termasuk MAE, MSE, dan RMSE. Selain itu, nilai MAPE pada Polynomial Regression menghasilkan "inf", yang kemungkinan disebabkan oleh adanya nilai nol pada data aktual sehingga menyebabkan pembagian dengan nol. Ini menunjukkan bahwa regresi linier lebih stabil dan akurat untuk dataset ini, serta lebih sesuai dalam memodelkan hubungan antara AQI umum dan AQI PM2.5.

Link:

GitHub: twosecondz/Kelompok_05_Tugas02_Linear_Regression: Tugas Kelompok – Kelas B Linear dan Polynomial Regression

Dataset: https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset

Link Video: bit.ly/ML_05_Project_2