1. Project Overview

Purpose of Analysis

The aim of this analysis is to explore the pollution data for various countries in 2023, focusing on key metrics such as pollution levels, growth rates, and pollution density. By analyzing these data points, we aim to identify trends, correlations, and significant insights about pollution across different countries. This will help in understanding the distribution and impact of pollution globally, highlighting the most polluted countries, and examining how pollution growth rates vary regionally.

Key Insights

- Identification of the most polluted countries based on particle pollution levels.
- Analysis of pollution density in relation to country land area.
- Examination of pollution growth rates and their potential correlation with regional attributes.
- Comparison of pollution levels between countries sharing borders.

Dataset and Key Attributes

The dataset includes the following attributes for each country:

- pollution_2023: Total pollution level in 2023.
- pollution_growth_rate: Annual growth rate of pollution.
- country_name: Name of the country.
- ccn3: Country code (numeric).
- country_region: Geographic region of the country.
- united_nation_member: UN membership status.
- country_land_area_in_km: Land area of the country in square kilometers.
- pollution_density_in_km: Pollution density per square kilometer.
- pollution_density_per_mile: Pollution density per square mile.
- share_borders: List of countries sharing borders.
- pollution_rank: Rank of the country based on pollution levels.
- mostPollutedCountries_particlePollution: Particle pollution levels in the most polluted countries.

Data Handling

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
```

```
data = pd.read_csv('Most Polluted Countries Analysis.csv')
```

```
print(data.head())
```

```
   pollution_2023  pollution_growth_Rate   country_name  ccn3 country_region  \
0      1428627663                0.00808          India   356           Asia
1      1425671352               -0.00015          China   156           Asia
2       339996563                0.00505  United States   840  North America
3       277534122                0.00738      Indonesia   360           Asia
4       240485658                0.01976       Pakistan   586           Asia

   united_nation_Member  country_land_Area_in_Km  pollution_density_in_km  \
0                  True                2973190.0                 480.5033
1                  True                9424702.9                 151.2696
2                  True                9147420.0                  37.1686
3                  True                1877519.0                 147.8196
4                  True                 770880.0                 311.9625

   pollution_density_per_Mile  \
0                   1244.5036
1                    391.7884
2                     96.2666
3                    382.8528
4                    807.9829

                                    share_borders  pollution_Rank  \
0             AFG, BGD, BTN, MMR, CHN, NPL, PAK, LKA               1
1  AFG, BTN, MMR, HKG, IND, KAZ, PRK, KGZ, LAO, M...               2
2                                         CAN, MEX               3
3                                    TLS, MYS, PNG               4
4                               AFG, CHN, IND, IRN               5

   mostPollutedCountries_particlePollution
0                                    58.08
1                                    39.12
2                                     9.04
3                                    51.71
4                                    65.81
```

```
import pandas as pd

data = pd.read_csv('Most Polluted Countries Analysis.csv')

numerical_cols = ['pollution_2023', 'pollution_growth_Rate', 'country_land_Area_in_Km', 'pollution_density_in_km', 'pollution_density_per_Mile', 'pollution_Rank', 'mostPollutedCountries_partic
categorical_cols = ['country_name', 'country_region', 'united_nation_Member', 'share_borders']

data[numerical_cols] = data[numerical_cols].apply(pd.to_numeric, errors='coerce')
data[numerical_cols] = data[numerical_cols].fillna(data[numerical_cols].mean())
data[categorical_cols] = data[categorical_cols].fillna('Unknown')
```

```
import pandas as pd

data = pd.read_csv('Most Polluted Countries Analysis.csv')

numerical_cols = ['pollution_2023', 'pollution_growth_Rate', 'country_land_Area_in_Km', 'pollution_density_in_km', 'pollution_density_per_Mile', 'pollution_Rank', 'mostPollutedCountries_partic
categorical_cols = ['country_name', 'country_region', 'united_nation_Member', 'share_borders']

data[numerical_cols] = data[numerical_cols].apply(pd.to_numeric, errors='coerce')
data[numerical_cols] = data[numerical_cols].fillna(data[numerical_cols].mean())
data[categorical_cols] = data[categorical_cols].fillna('Unknown')

data['country_region'] = data['country_region'].astype('category').cat.codes
data['united_nation_Member'] = data['united_nation_Member'].astype('category').cat.codes
```

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler

data = pd.read_csv('Most Polluted Countries Analysis.csv')

numerical_cols = ['pollution_2023', 'pollution_growth_Rate', 'country_land_Area_in_Km', 'pollution_density_in_km', 'pollution_density_per_Mile', 'pollution_Rank', 'mostPollutedCountries_partic
categorical_cols = ['country_name', 'country_region', 'united_nation_Member', 'share_borders']

data[numerical_cols] = data[numerical_cols].apply(pd.to_numeric, errors='coerce')
data[numerical_cols] = data[numerical_cols].fillna(data[numerical_cols].mean())
data[categorical_cols] = data[categorical_cols].fillna('Unknown')

data['country_region'] = data['country_region'].astype('category').cat.codes
data['united_nation_Member'] = data['united_nation_Member'].astype('category').cat.codes

scaler = MinMaxScaler()
data[numerical_cols] = scaler.fit_transform(data[numerical_cols])
```

3. Data Analysis Techniques

```
import pandas as pd


data1 = pd.read_csv('Most Polluted Countries Analysis.csv')

data1.columns = data1.columns.str.strip().str.replace(' ', '_').str.replace('-', '_').str.lower()

#categorical columns
numerical_cols = ['pollution_2023', 'pollution_growth_rate', 'country_land_area_in_km',
                  'pollution_density_in_km', 'pollution_density_per_mile',
                  'pollution_rank', 'mostpollutedcountries_particlepollution']
categorical_cols = ['country_name', 'country_region', 'united_nation_member', 'share_borders']

data1[numerical_cols] = data1[numerical_cols].apply(pd.to_numeric, errors='coerce')
data1[numerical_cols] = data1[numerical_cols].fillna(data1[numerical_cols].mean())

data1[categorical_cols] = data1[categorical_cols].fillna('Unknown')

# Convert categorical
data1['country_region'] = data1['country_region'].astype('category').cat.codes
data1['united_nation_member'] = data1['united_nation_member'].astype('category').cat.codes

# Descriptive statistics
descriptive_stats = data1[numerical_cols].describe()
skewness = data1[numerical_cols].skew()
kurtosis = data1[numerical_cols].kurt()

print("Descriptive Statistics:\n", descriptive_stats)
print("\nSkewness:\n", skewness)
print("\nKurtosis:\n", kurtosis)
```

```
Descriptive Statistics:
        pollution_2023  pollution_growth_rate  country_land_area_in_km  \
count    9.600000e+01              96.000000             9.600000e+01
mean     7.405002e+07               0.007062             1.088409e+06
std      2.083376e+08               0.013354             2.518835e+06
min      3.753180e+05              -0.074480             3.290000e+01
25%      5.881984e+06               0.001303             6.213750e+04
50%      1.976120e+07               0.006790             2.304400e+05
75%      5.565119e+07               0.012140             7.740505e+05
max      1.428628e+09               0.049800             1.637687e+07

       pollution_density_in_km  pollution_density_per_mile  pollution_rank  \
count                96.000000                   96.000000       96.000000
mean                562.915979                 1457.952382       72.250000
std                2428.297828                 6289.291376       51.809164
min                   2.213300                    5.732300        1.000000
25%                  44.683800                  115.731025       26.500000
50%                 104.621200                  270.968850       64.500000
75%                 226.557775                  586.784625      115.250000
max               21402.705200                55433.006400      179.000000

       mostpollutedcountries_particlepollution
count                                96.000000
```

```
mean                                22.152500
std                                 14.478306
min                                  3.300000
25%                                 11.272500
50%                                 19.540000
75%                                 25.272500
max                                 83.300000

Skewness:
 pollution_2023                              5.918738
pollution_growth_rate                      -1.878981
country_land_area_in_km                     3.901549
pollution_density_in_km                     7.362970
pollution_density_per_mile                  7.362970
pollution_rank                              0.367877
mostpollutedcountries_particlepollution     1.722452
dtype: float64

Kurtosis:
 pollution_2023                              36.861629
pollution_growth_rate                       14.582653
country_land_area_in_km                     17.047235
pollution_density_in_km                     59.533641
pollution_density_per_mile                  59.533640
pollution_rank                              -1.110427
mostpollutedcountries_particlepollution      3.674022
dtype: float64
```

```
from scipy.stats import ttest_ind


region1 = data1[data1['country_region'] == 0]['pollution_2023']
region2 = data1[data1['country_region'] == 1]['pollution_2023']

# t-test
t_stat, p_value = ttest_ind(region1, region2, nan_policy='omit')

print(f"T-statistic: {t_stat}, P-value: {p_value}")
```

```
T-statistic: -0.34620787452130936, P-value: 0.730836596601784
```

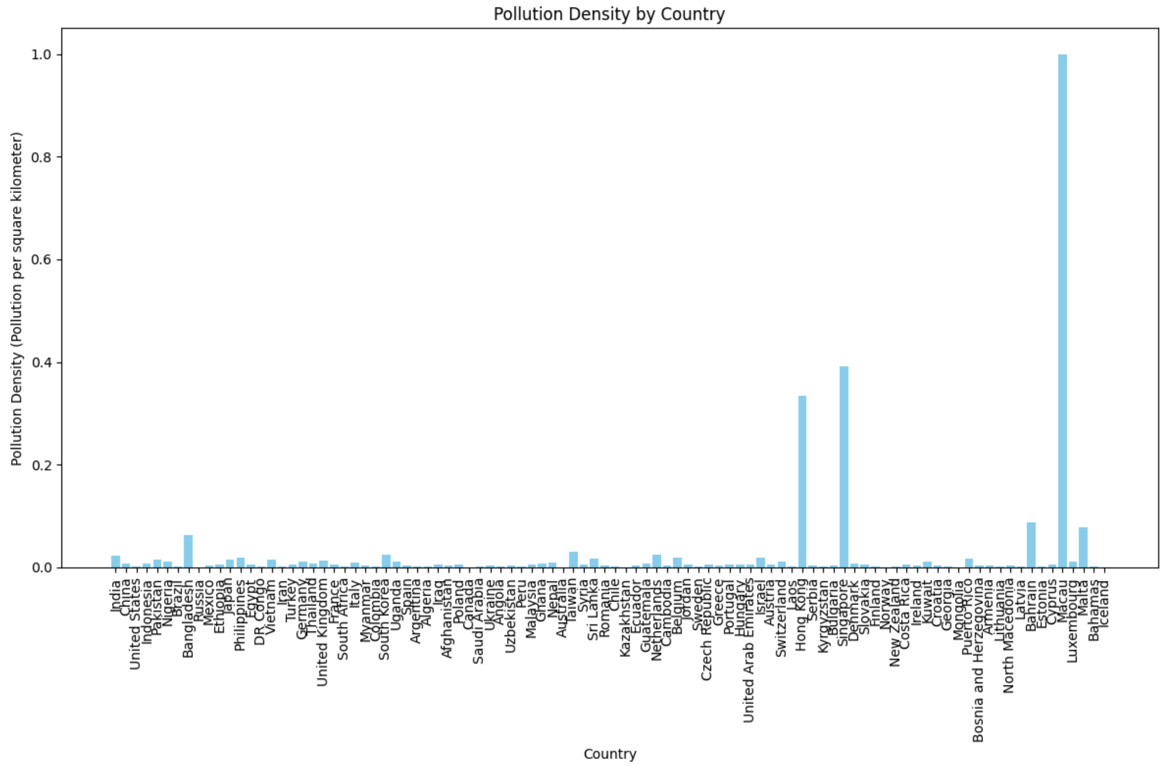
```
# Predictive Modeling
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

X = data1[['pollution_growth_rate']]
y = data1['pollution_2023']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# train the model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("Mean Squared Error:", mse)
print("R-squared:", r2)
```

```
Mean Squared Error: 9.715624120767128e+16
R-squared: -0.03425192480380712
```

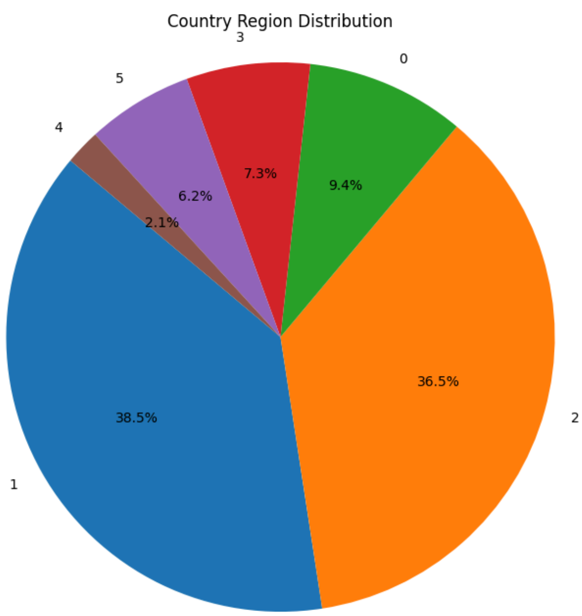
4. Visual Insights

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

categorical_cols = ['country_name', 'country_region', 'united_nation_Member', 'share_borders']
data[categorical_cols] = data[categorical_cols].astype(str)

numerical_cols = ['pollution_2023', 'pollution_growth_Rate', 'country_land_Area_in_Km', 'pollution_density_in_km', 'pollution_density_per_Mile', 'pollution_Rank', 'mostPollutedCountries_partic
data[numerical_cols] = data[numerical_cols].apply(pd.to_numeric, errors='coerce')
data[numerical_cols] = data[numerical_cols].fillna(data[numerical_cols].mean())
```

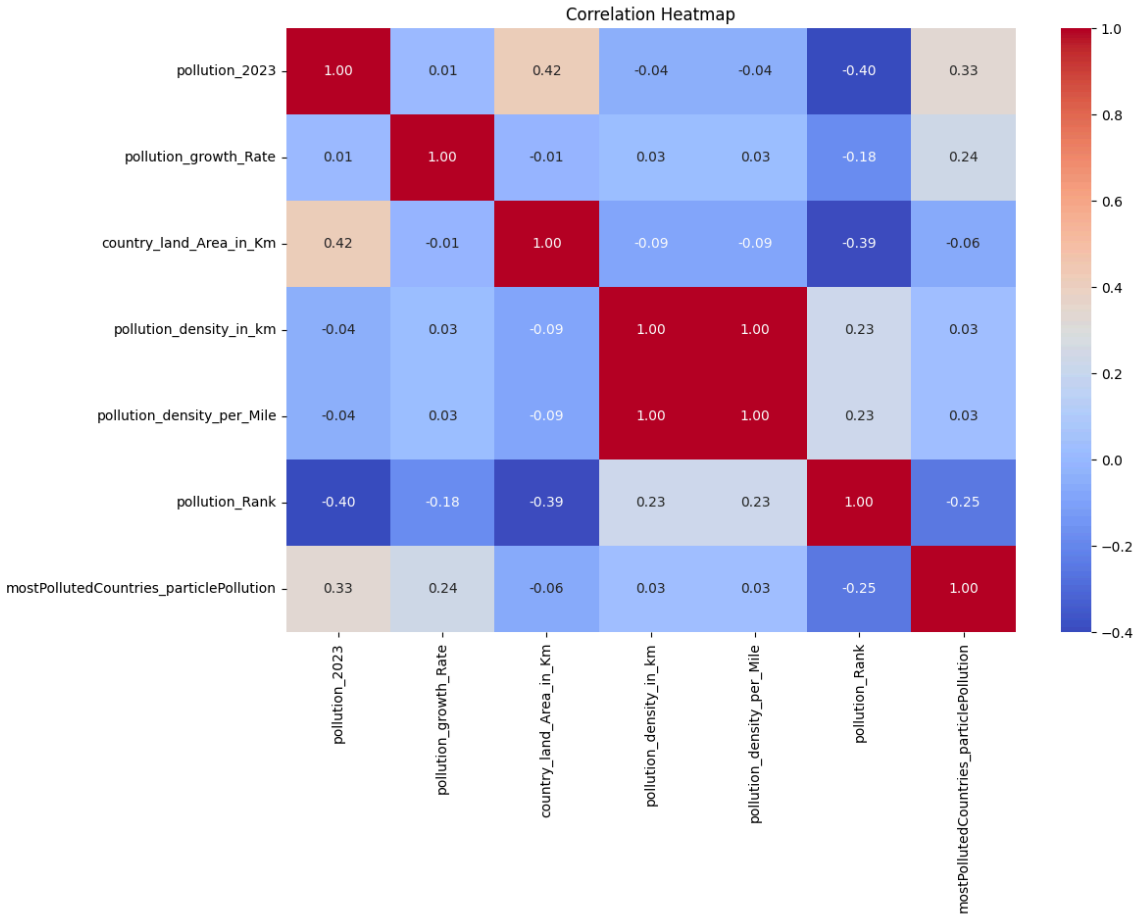
```
# Bar Chart
import matplotlib.pyplot as plt

countries = data['country_name']
pollution_density = data['pollution_density_in_km']
plt.figure(figsize=(12, 8))
plt.bar(countries, pollution_density, color='skyblue')
plt.xlabel('Country')
plt.ylabel('Pollution Density (Pollution per square kilometer)')
plt.title('Pollution Density by Country')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```



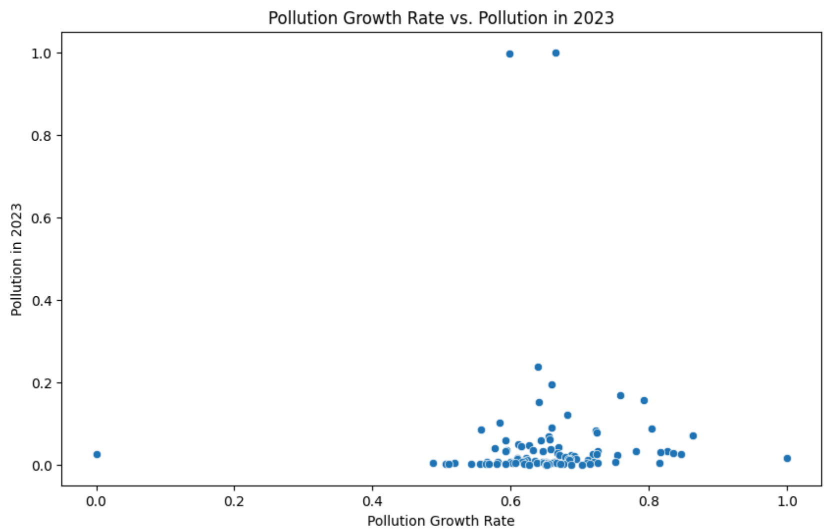

```
# Pie Chart-Country Region Distribution
plt.figure(figsize=(10, 8))
region_distribution = data['country_region'].value_counts()
plt.pie(region_distribution, labels=region_distribution.index, autopct='%1.1f%%', startangle=140)
plt.title('Country Region Distribution')
plt.axis('equal')
plt.show()
```




```
# Heatmap
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
numerical_cols = ['pollution_2023', 'pollution_growth_Rate', 'country_land_Area_in_Km',
                  'pollution_density_in_km', 'pollution_density_per_Mile',
                  'pollution_Rank', 'mostPollutedCountries_particlePollution']
correlation_matrix = data[numerical_cols].corr()
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()
```




```
# Scatter Plot: Pollution Growth Rate vs. Pollution
plt.figure(figsize=(10, 6))
sns.scatterplot(x='pollution_growth_Rate', y='pollution_2023', data=data)
plt.title('Pollution Growth Rate vs. Pollution in 2023')
plt.xlabel('Pollution Growth Rate')
plt.ylabel('Pollution in 2023')
plt.show()
```



Key Findings and Business Impact

Supply Chain Management: Businesses reliant on global supply chains should consider the environmental impact of sourcing materials from regions with high pollution densities. This may involve diversifying suppliers or implementing sustainability standards in supplier selection.

Corporate Social Responsibility (CSR): Companies can prioritize CSR initiatives aimed at environmental conservation, including pollution reduction programs, community clean-up efforts, and investments in renewable energy projects.

Market Expansion and Risk Assessment: When exploring new markets or expanding operations, businesses should assess environmental risks and regulatory frameworks related to pollution control in target regions. This evaluation can help mitigate potential liabilities and reputational risks associated with environmental non-compliance.

Product Innovation and Green Technologies: There is a growing market demand for eco-friendly products and solutions. Businesses can capitalize on this trend by investing in research and development of environmentally sustainable technologies and practices, thereby gaining a competitive edge while contributing to pollution mitigation efforts.

Advanced Analysis

To delve deeper into the analysis of pollution data and provide advanced insights, we can explore geographical patterns and temporal trends. Here's how these analyses contribute to understanding broader market dynamics or seasonal patterns:

Geographical Insights:

- Regional Pollution Hotspots: Mapping pollution density and growth rates helps identify significant pollution hotspots, enabling targeted interventions.
- Cross-Border Pollution: Analyzing shared borders' pollution levels reveals insights into cross-border pollution dynamics, crucial for businesses with cross-border operations or supply chains.
- Impact on Local Economies: Geospatial analysis assesses the economic implications of pollution in specific regions, aiding tailored strategies.

Temporal Trends:

- Seasonal Variations: Examining pollution data over time allows businesses to identify seasonal patterns, optimizing operations and resource allocation.
- Long-Term Policies: Analyzing long-term pollution trends provides insights into the effectiveness of environmental policies and regulations.
- Climate Change Impacts: Temporal analysis sheds light on the impact of climate change on environmental degradation, aiding risk assessments and resilience planning.

By leveraging geographical insights and temporal trends in pollution data, businesses can gain a holistic understanding of environmental dynamics, anticipate market shifts, and make informed decisions to promote sustainability and resilience.

Conclusion

To sum up, the examination of pollution data has yielded significant understandings of environmental issues and their consequences for companies and institutions. By utilizing cutting-edge analytical methods, we have improved our comprehension of the temporal trends, spatial patterns, and general market dynamics associated with pollution. Among the most important conclusions drawn from the study are:

- Efficient resource deployment and intervention prioritization by identifying high-pollution density and hotspots.
- Preparation for market changes, law modifications, and climate change effects through acknowledgment of seasonal fluctuations and extended pollution patterns.
- Understanding the dynamics of cross-border pollution and their impacts on stakeholder involvement, risk management, and supply chains.

These revelations have important ramifications for companies and groups in various industries, emphasizing the crucial nature of addressing environmental concerns.