Air Quality Index & Water Pollution In Major Cities of Pakistan (2020)

👨💻 By: Irfan Ullah Khan

GITHUB PROFILE | KAGGLE PROFILE | LINKEDIN PROFILE

YOUTUBE PROFILE | EMAIL CONTACT ME | WEBSITE CONTACT ME

Air Quality Index & Water Pollution
In Major Cities of Pakistan (2020)
BY: Irfan Ullah Khan
Air Quality Index
Water Pollution

Install Required Libraries:

```
!pip install pandas matplotlib seaborn plotly
```

Show hidden output

Import Required Libraries:

```
#Libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from scipy import stats
```

Load the Dataset:

```
# Load the dataset
df = pd.read_csv('water-air-quality-big-cities-of-pakistan-2020.csv', encoding='latin-1')
```

Clean the Data:

```
df.columns = df.columns.str.strip().str.replace('"', '')
print(df.head())
```

```
       Country                 Region        City Postal code population  \
0   "Pakistan"   "Khyber Pakhtunkhwa"  Abbottabad       22010     148587   
1   "Pakistan"               "Punjab"      Attock       43600     91 475   
2   "Pakistan"                     ""  Bahawalpur       63071     681696   
3   "Pakistan"               "Punjab"     Chakwal       04882    105 977   
4   "Pakistan"                     ""        Dadu       76150     146179   

   Latitude  Longitude  AirQuality  WaterPollution  
0   34.1500    73.2167        62.5       80.357143  
1   33.7667    72.3598        75.0       50.000000  
2   29.3956    71.6722        75.0       60.000000  
3   32.9300    72.8500        50.0       50.000000  
4   26.7319    67.7750        50.0       50.000000  
```

```
df.tail()
```

	Country	Region	City	Postal code	population	Latitude	Longitude	AirQuali
27	"Pakistan"	"Punjab"	Sheikhupura	39060	411834	31.7083	74.00000	5
28	"Pakistan"	"Punjab"	Sialkot	5132	3893672	32.5000	74.53330	3
29	"Pakistan"	"Gilgit-Baltistan"	Skardu	16100	214,848	35.0000	75.63372	10
30	"Pakistan"	"Sindh"	Sukkur	65080	476776	27.6995	68.86730	5
[illegible]	[illegible]	[illegible]	Nankana	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
```

```
---  ------         --------------  -----
 0   Country        32 non-null     object
 1   Region         32 non-null     object
 2   City           32 non-null     object
 3   Postal code    32 non-null     object
 4   population     32 non-null     object
 5   Latitude       32 non-null     float64
 6   Longitude      32 non-null     float64
 7   AirQuality     32 non-null     float64
 8   WaterPollution 32 non-null     float64
dtypes: float64(4), object(5)
memory usage: 2.4+ KB
```

```
df.describe()
```

	Latitude	Longitude	AirQuality	WaterPollution
count	32.000000	32.000000	32.000000	32.000000
mean	31.487891	72.196551	51.414259	68.112060
std	2.847269	2.296013	23.656493	20.924851
min	24.860000	67.007000	0.000100	25.000000
25%	30.552400	71.492775	32.263514	50.000000
50%	31.895950	73.052400	50.000000	67.187500
75%	33.715850	73.717625	75.000000	83.289659
max	35.920800	75.633720	100.000000	100.000000

```
df.shape
```

```
(32, 9)
```

```
df.isnull().sum()
```

```
                0
Country         0
Region          0
City            0
Postal code     0
population      0
Latitude        0
Longitude       0
AirQuality      0
WaterPollution  0

dtype: int64
```

Visualize the Data:

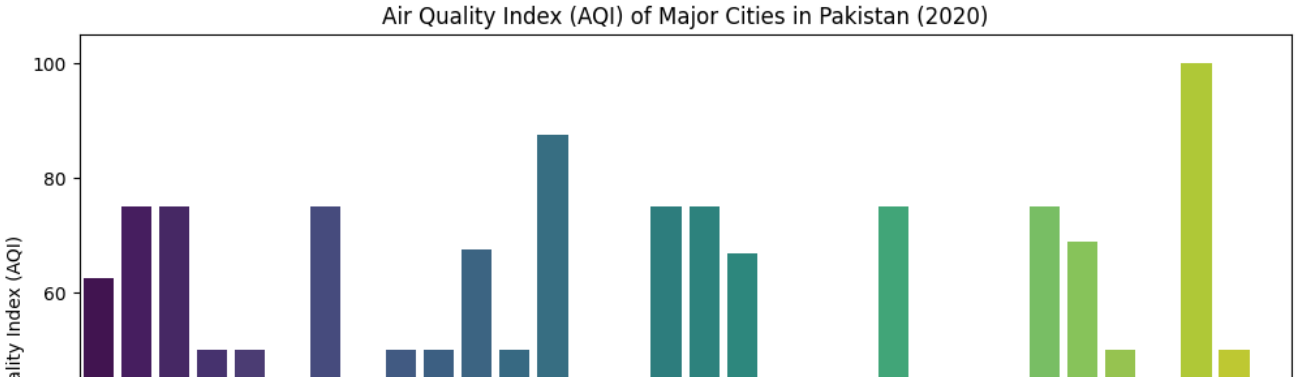
You can create various visualizations. Here’s an example of a line plot to visualize AQI over time:

Visualize Air Quality Index (AQI):

Create a bar plot to visualize the AQI for each city:

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

plt.figure(figsize=(12, 6))
sns.barplot(data=df, x='City', y='AirQuality', palette='viridis')
plt.title('Air Quality Index (AQI) of Major Cities in Pakistan (2020)')
plt.xticks(rotation=45)
plt.xlabel('City')
plt.ylabel('Air Quality Index (AQI)')
plt.show()
```

```
<ipython-input-5-23c566443f61>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v

  sns.barplot(data=df, x='City', y='AirQuality', palette='viridis')
```

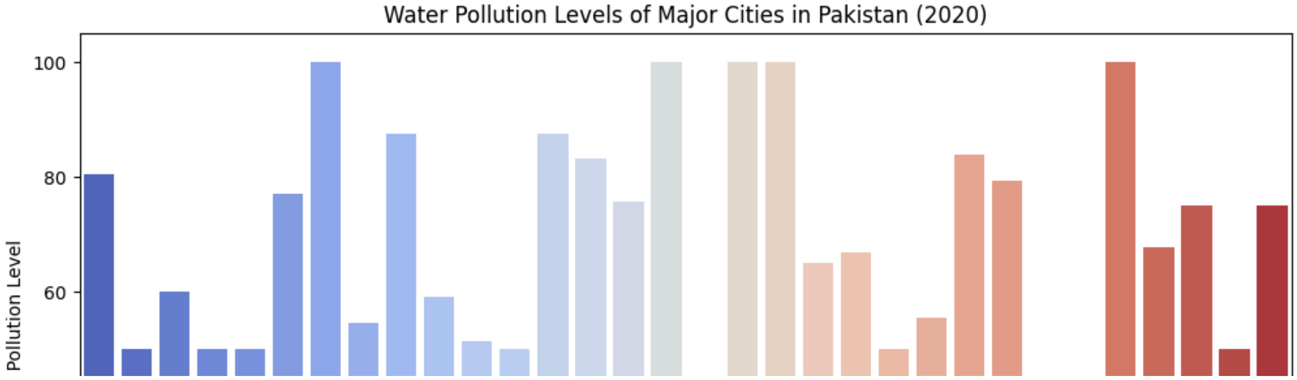


Visualize Water Pollution:

Create a similar bar plot for water pollution:

```
plt.figure(figsize=(12, 6))
sns.barplot(data=df, x='City', y='WaterPollution', palette='coolwarm')
plt.title('Water Pollution Levels of Major Cities in Pakistan (2020)')
plt.xticks(rotation=45)
plt.xlabel('City')
plt.ylabel('Water Pollution Level')
plt.show()
```

```
<ipython-input-6-54b8a81baf47>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v

  sns.barplot(data=df, x='City', y='WaterPollution', palette='coolwarm')
```

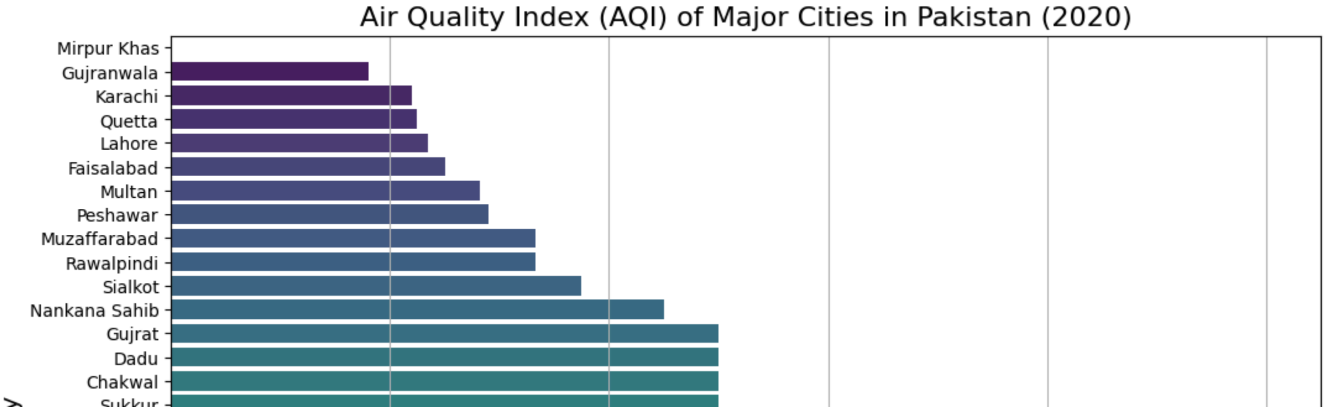


Visualizing Air Quality Index (AQI):

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

# Convert 'AirQuality' to numeric, if necessary
df['AirQuality'] = pd.to_numeric(df['AirQuality'], errors='coerce')

# Sort values by Air Quality
df_sorted_aqi = df.sort_values(by='AirQuality')

plt.figure(figsize=(12, 8))
```

```
sns.barplot(data=df_sorted_aqi, x='AirQuality', y='City', palette='viridis')
plt.title('Air Quality Index (AQI) of Major Cities in Pakistan (2020)', fontsize=16)
plt.xlabel('Air Quality Index (AQI)', fontsize=14)
plt.ylabel('City', fontsize=14)
plt.grid(axis='x')
plt.show()
```

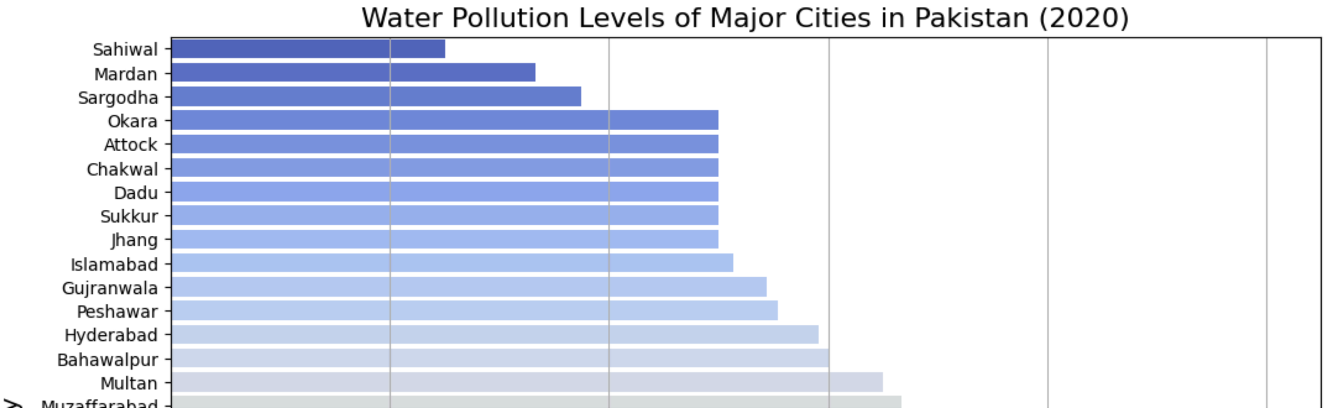
```
<ipython-input-7-a707b4c385e0>:11: FutureWarning: 

Passing `palette` without assigning `hue` is deprecated and will be removed in v

  sns.barplot(data=df_sorted_aqi, x='AirQuality', y='City', palette='viridis')
```



Visualizing Water Pollution:

```
# Convert 'WaterPollution' to numeric, if necessary
df['WaterPollution'] = pd.to_numeric(df['WaterPollution'], errors='coerce')

# Sort values by Water Pollution
df_sorted_wp = df.sort_values(by='WaterPollution')

plt.figure(figsize=(12, 8))
sns.barplot(data=df_sorted_wp, x='WaterPollution', y='City', palette='coolwarm')
plt.title('Water Pollution Levels of Major Cities in Pakistan (2020)', fontsize=16)
plt.xlabel('Water Pollution Level', fontsize=14)
plt.ylabel('City', fontsize=14)
plt.grid(axis='x')
plt.show()
```

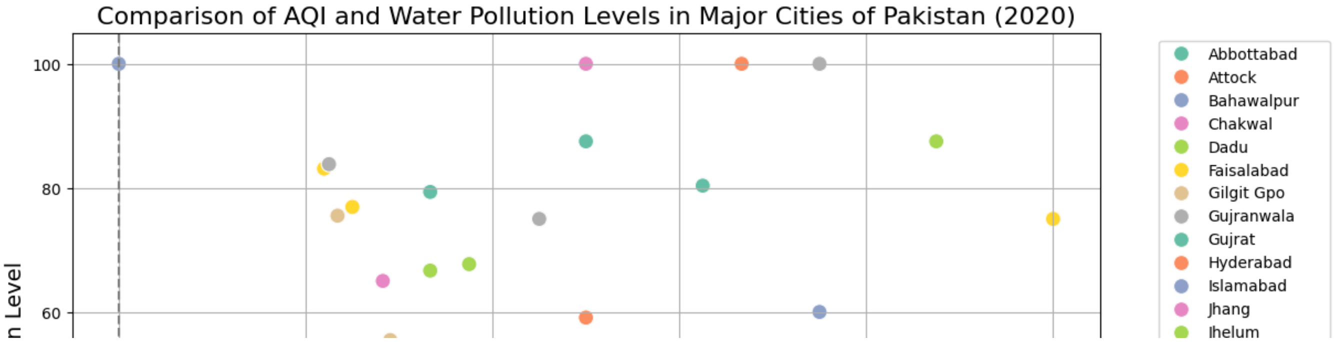
```
<ipython-input-8-cec9154d34f7>:8: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v

  sns.barplot(data=df_sorted_wp, x='WaterPollution', y='City', palette='coolwarm
```



Combined Visualization:

Create a combined scatter plot to compare AQI and Water Pollution levels for each city.

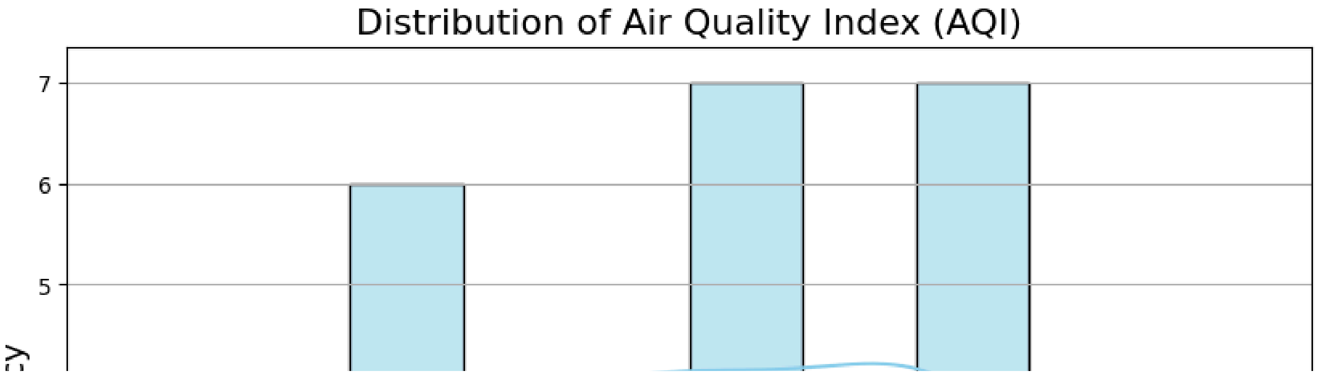
```
plt.figure(figsize=(12, 8))
sns.scatterplot(data=df, x='AirQuality', y='WaterPollution', hue='City', palette='Set2', s=
plt.title('Comparison of AQI and Water Pollution Levels in Major Cities of Pakistan (2020)'
plt.xlabel('Air Quality Index (AQI)', fontsize=14)
plt.ylabel('Water Pollution Level', fontsize=14)
plt.axhline(0, color='gray', linestyle='--')
plt.axvline(0, color='gray', linestyle='--')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.grid()
plt.show()
```



Histogram of Air Quality Index (AQI):

This will help you understand the distribution of AQI values across cities.

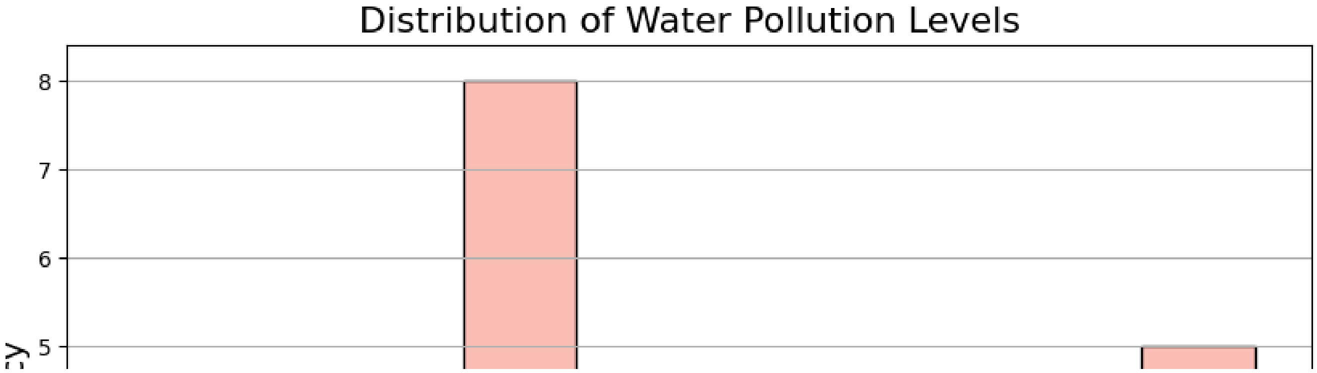
```
plt.figure(figsize=(10, 6))
sns.histplot(df['AirQuality'], bins=10, kde=True, color='skyblue')
plt.title('Distribution of Air Quality Index (AQI)', fontsize=16)
plt.xlabel('Air Quality Index (AQI)', fontsize=14)
plt.ylabel('Frequency', fontsize=14)
plt.grid(axis='y')
plt.show()
```



Histogram of Water Pollution Levels:

Similar to the AQI histogram, this will show the distribution of water pollution levels.

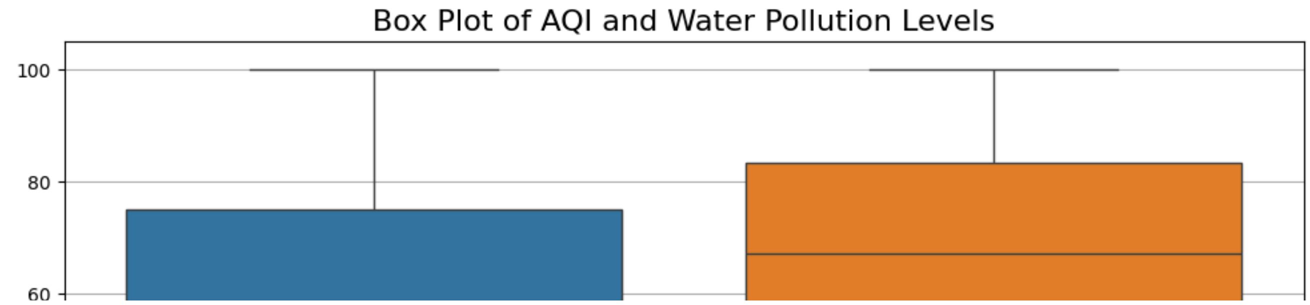
```
plt.figure(figsize=(10, 6))
sns.histplot(df['WaterPollution'], bins=10, kde=True, color='salmon')
plt.title('Distribution of Water Pollution Levels', fontsize=16)
plt.xlabel('Water Pollution Level', fontsize=14)
plt.ylabel('Frequency', fontsize=14)
plt.grid(axis='y')
plt.show()
```



Box Plot for AQI and Water Pollution:

Box plots can provide insights into the spread and potential outliers in the data.

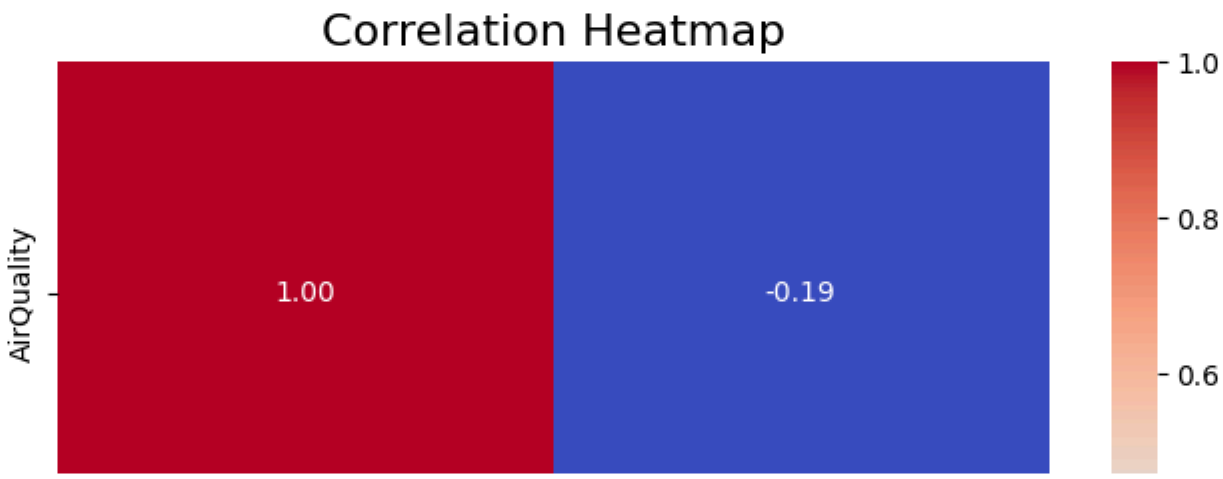
```
plt.figure(figsize=(12, 6))
sns.boxplot(data=df[['AirQuality', 'WaterPollution']])
plt.title('Box Plot of AQI and Water Pollution Levels', fontsize=16)
plt.ylabel('Values', fontsize=14)
plt.xticks([0, 1], ['Air Quality Index (AQI)', 'Water Pollution Level'])
plt.grid(axis='y')
plt.show()
```



Correlation Heatmap:

A heatmap will help visualize the correlation between AQI and water pollution.

```
plt.figure(figsize=(8, 6))
correlation_matrix = df[['AirQuality', 'WaterPollution']].corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap', fontsize=16)
plt.show()
```



City-wise Summary Statistics:

Generate summary statistics to get a clearer picture of the data.

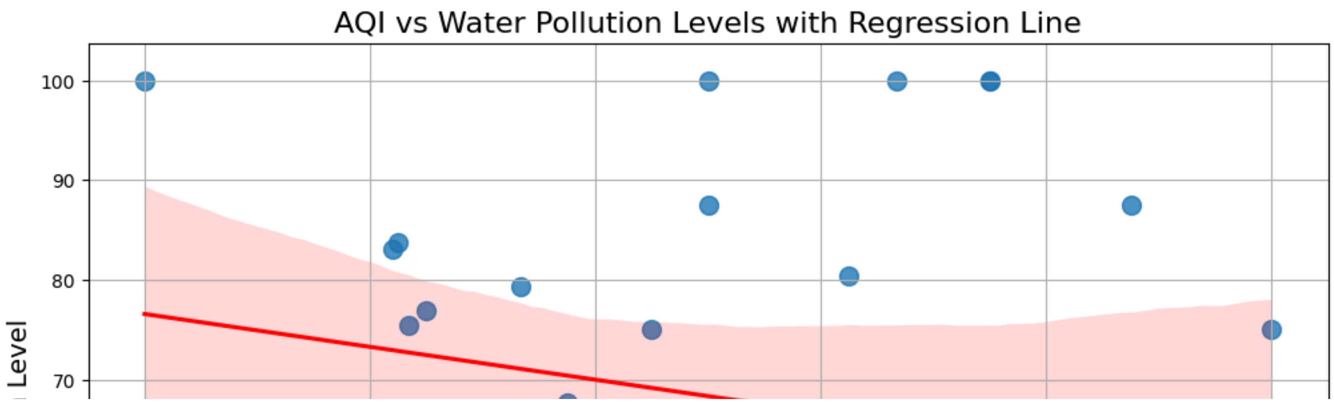
```
summary_stats = df[['AirQuality', 'WaterPollution']].describe()
print(summary_stats)
```

```
       AirQuality  WaterPollution
count   32.000000       32.000000
mean    51.414256       68.112060
std     23.656500       20.924851
min      0.000000       25.000000
25%     32.263514       50.000000
50%     50.000000       67.187500
75%     75.000000       83.289659
max    100.000000      100.000000
```

Scatter Plot with Trend Line:

Add a regression line to the scatter plot to visualize the relationship between AQI and water pollution.

```
plt.figure(figsize=(12, 8))
sns.regplot(data=df, x='AirQuality', y='WaterPollution', scatter_kws={'s':100}, line_kws={'
plt.title('AQI vs Water Pollution Levels with Regression Line', fontsize=16)
plt.xlabel('Air Quality Index (AQI)', fontsize=14)
plt.ylabel('Water Pollution Level', fontsize=14)
plt.grid()
plt.show()
```

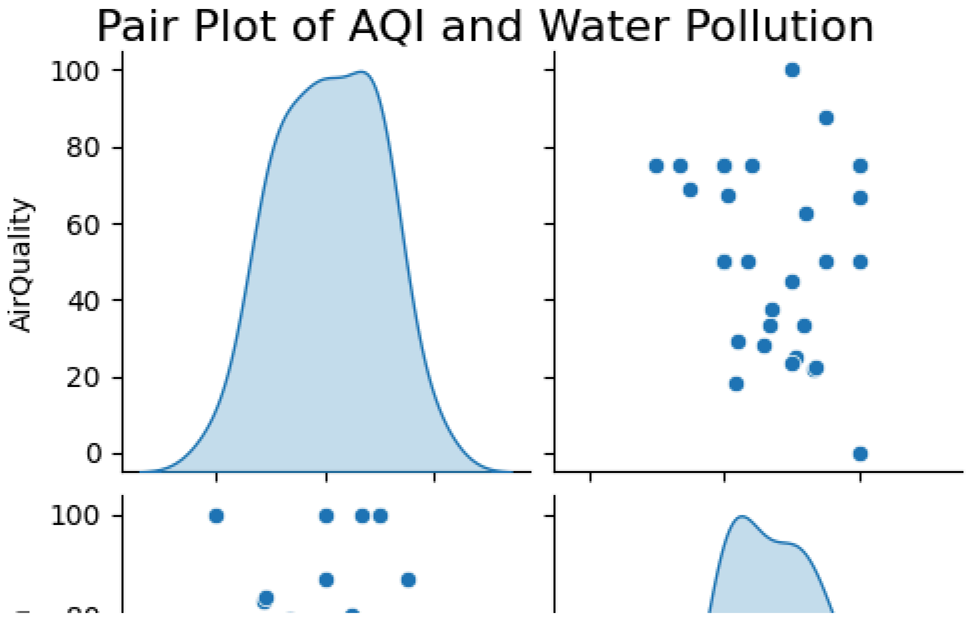


Pair Plot:

This allows you to visualize relationships between multiple variables in the dataset simultaneously.

```
sns.pairplot(df[['AirQuality', 'WaterPollution']], diag_kind='kde', palette='Set2')
plt.suptitle('Pair Plot of AQI and Water Pollution', y=1.02, fontsize=16)
plt.show()
```

```
/usr/local/lib/python3.10/dist-packages/seaborn/axisgrid.py:1513: UserWarning: I
  func(x=vector, **plot_kwargs)
/usr/local/lib/python3.10/dist-packages/seaborn/axisgrid.py:1513: UserWarning: I
  func(x=vector, **plot_kwargs)
/usr/local/lib/python3.10/dist-packages/seaborn/axisgrid.py:1615: UserWarning: I
  func(x=x, y=y, **kwargs)
/usr/local/lib/python3.10/dist-packages/seaborn/axisgrid.py:1615: UserWarning: I
  func(x=x, y=y, **kwargs)
```

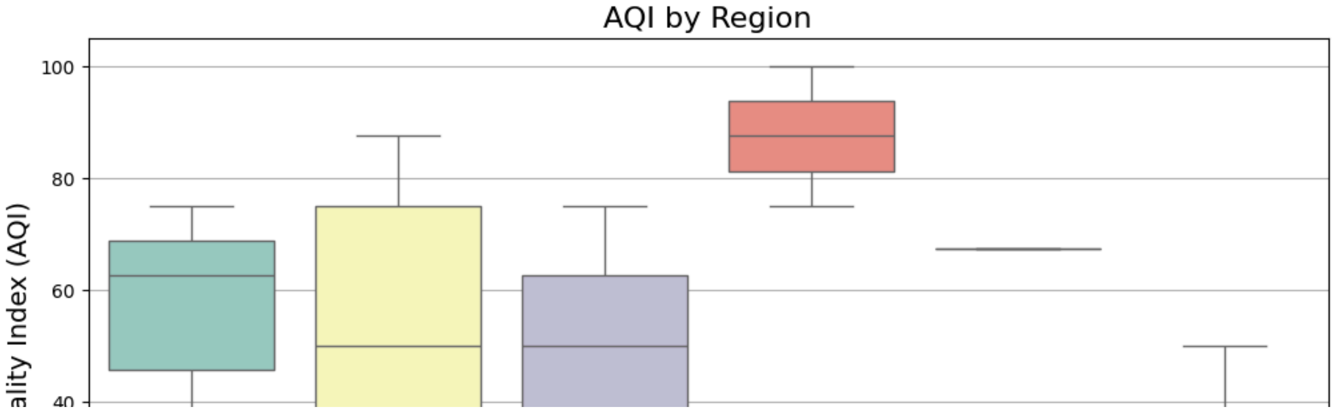


Categorical Analysis:

If you have categorical data (like region), you can analyze AQI and water pollution by region.

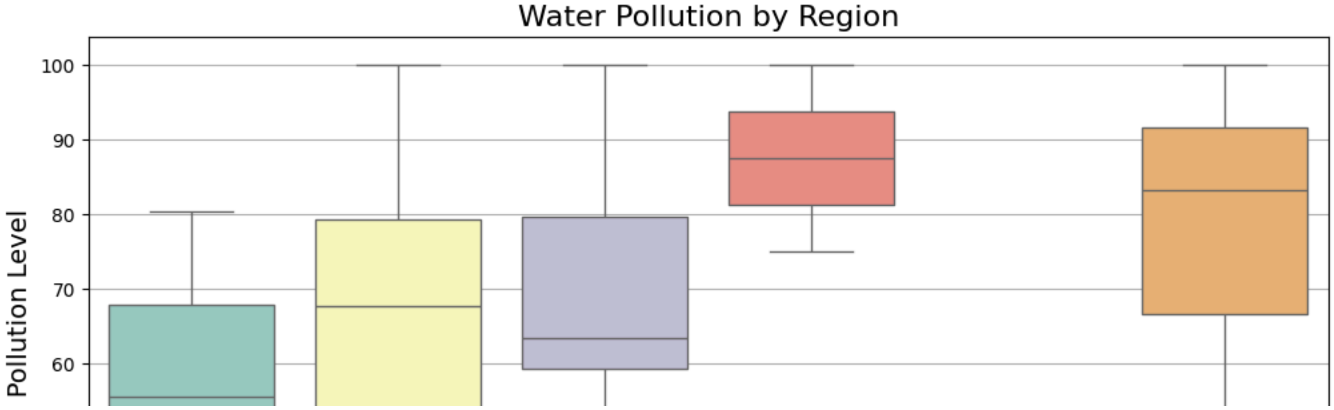
```
plt.figure(figsize=(12, 6))
sns.boxplot(data=df, x='Region', y='AirQuality', palette='Set3')
plt.title('AQI by Region', fontsize=16)
plt.xticks(rotation=45)
plt.ylabel('Air Quality Index (AQI)', fontsize=14)
plt.grid(axis='y')
plt.show()
```

```
<ipython-input-17-62863a7d982f>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v

  sns.boxplot(data=df, x='Region', y='AirQuality', palette='Set3')
```


AQI by Region

```
plt.figure(figsize=(12, 6))
sns.boxplot(data=df, x='Region', y='WaterPollution', palette='Set3')
plt.title('Water Pollution by Region', fontsize=16)
plt.xticks(rotation=45)
plt.ylabel('Water Pollution Level', fontsize=14)
plt.grid(axis='y')
plt.show()
```

```
<ipython-input-18-18470e63c7b4>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v

  sns.boxplot(data=df, x='Region', y='WaterPollution', palette='Set3')
```

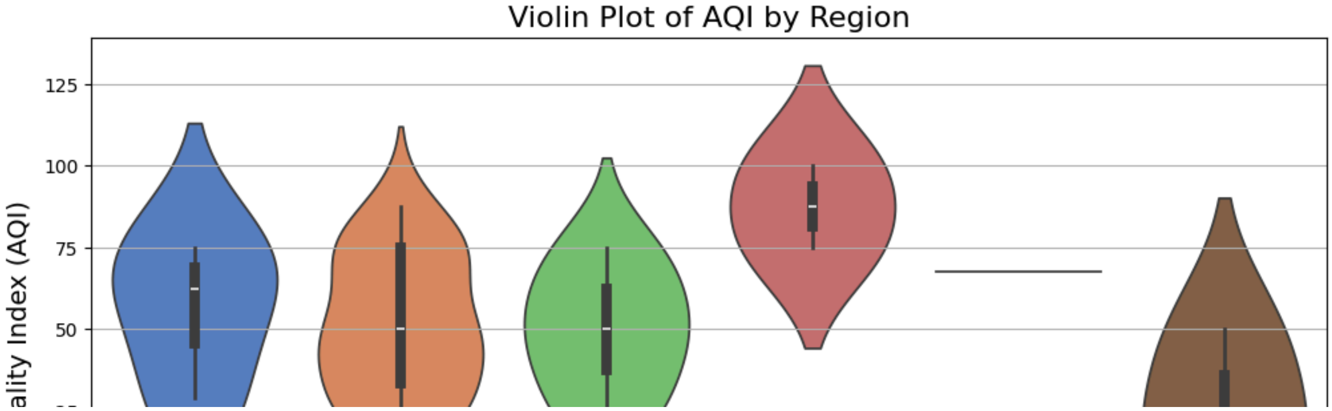


Violin Plot:

A violin plot combines a box plot with a KDE plot to show the distribution of the data.

```
plt.figure(figsize=(12, 6))
sns.violinplot(data=df, x='Region', y='AirQuality', palette='muted')
plt.title('Violin Plot of AQI by Region', fontsize=16)
plt.xticks(rotation=45)
plt.ylabel('Air Quality Index (AQI)', fontsize=14)
plt.grid(axis='y')
plt.show()
```

```
<ipython-input-19-480d2d3d80c2>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v

  sns.violinplot(data=df, x='Region', y='AirQuality', palette='muted')
```

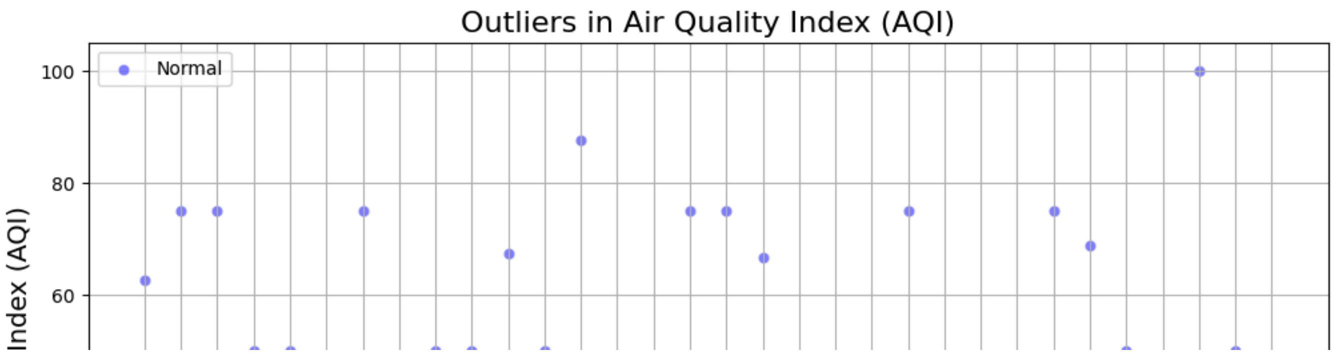


Outlier Detection:

Identify and visualize outliers in AQI and water pollution levels using z-scores.

```
from scipy import stats
```

```
df['AQI_z'] = stats.zscore(df['AirQuality'])
df['WP_z'] = stats.zscore(df['WaterPollution'])

outliers_aqi = df[(df['AQI_z'] > 3) | (df['AQI_z'] < -3)]
outliers_wp = df[(df['WP_z'] > 3) | (df['WP_z'] < -3)]

plt.figure(figsize=(12, 6))
sns.scatterplot(data=outliers_aqi, x='City', y='AirQuality', color='red', label='Outliers (
sns.scatterplot(data=df, x='City', y='AirQuality', color='blue', label='Normal', alpha=0.5)
plt.title('Outliers in Air Quality Index (AQI)', fontsize=16)
plt.xticks(rotation=45)
plt.ylabel('Air Quality Index (AQI)', fontsize=14)
plt.legend()
plt.grid()
plt.show()
```



Correlation Coefficients:

Calculate and display the correlation coefficients.

```
correlation_aqi_wp = df[['AirQuality', 'WaterPollution']].corr().iloc[0, 1]
print(f'Correlation between AQI and Water Pollution: {correlation_aqi_wp:.2f}')
```

```
Correlation between AQI and Water Pollution: -0.19
```

Group Summary Statistics:

Summarize AQI and water pollution by region.

```
region_summary = df.groupby('Region')[['AirQuality', 'WaterPollution']].mean()
print(region_summary)
```

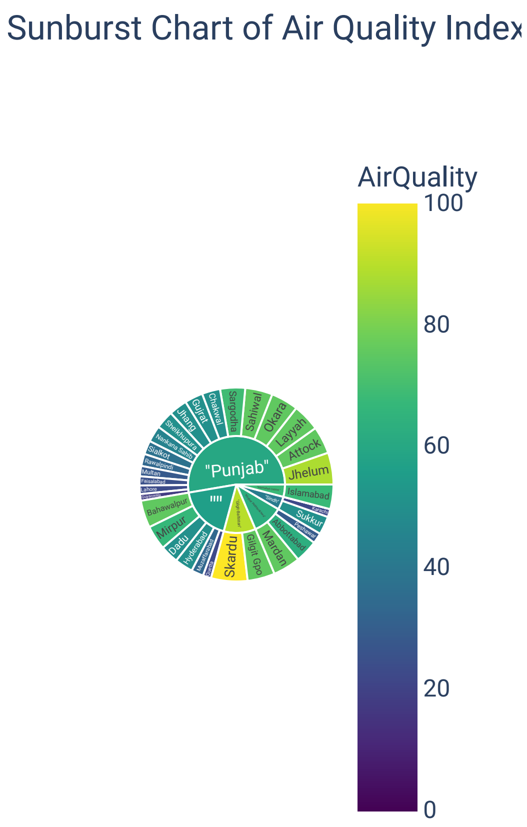
```
                        AirQuality  WaterPollution
Region
""                       49.583333       69.930184
"Gilgit-Baltistan"       87.500000       87.500000
"Islamabad Capital "     67.418033       51.327434
"Khyber Pakhtunkhwa"     55.518018       56.386409
"Punjab"                 50.989082       66.553321
"Sindh"                  23.989899       77.703901
```

Sunburst Chart

a sunburst chart, we can use the plotly library, which provides a straightforward way to create interactive visualizations, including sunburst charts. Sunburst charts are useful for visualizing hierarchical data, and in this case, we can visualize the AQI and water pollution data by region and city.

```
import pandas as pd
import plotly.express as px

# Load the dataset
df = pd.read_csv('water-air-quality-big-cities-of-pakistan-2020.csv', encoding='latin-1')
# Clean the column names
df.columns = df.columns.str.strip().str.replace('"', '')

# Ensure 'AirQuality' is numeric and replace zeros/negatives with small positive values
df['AirQuality'] = pd.to_numeric(df['AirQuality'], errors='coerce')
df['AirQuality'] = df['AirQuality'].replace(0, 0.0001)  # Replace zeros with a tiny value
df.loc[df['AirQuality'] < 0, 'AirQuality'] = 0.0001  # Replace negative values too

# Create a sunburst chart for Air Quality Index (AQI)
fig_aqi = px.sunburst(
    df,
    path=['Region', 'City'],  # Hierarchical path
    values='AirQuality',       # Values to be represented
    title='Sunburst Chart of Air Quality Index (AQI) by Region and City',
    color='AirQuality',        # Color by AQI values
    color_continuous_scale='viridis'
)
```

```
fig_aqi.show()
```



```
import pandas as pd
import plotly.express as px

# Create a sunburst chart for Water Pollution
fig_wp = px.sunburst(
    df,
    path=['Region', 'City'],
    values='WaterPollution',
    title='Sunburst Chart of Water Pollution by Region and City',
    color='WaterPollution',
    color_continuous_scale=px.colors.sequential.RdBu  # Changed to a Plotly colorscale
)

fig_wp.show()
```

Sunburst Chart of Water Pollution

WaterPollution