**Descrição do Projeto:**
O projeto visa realizar uma análise abrangente dos dados fornecidos pelo banco de dados "Water Quality", que contém informações sobre países, cidades e regiões em relação à qualidade do ar e à poluição da água. A análise busca identificar padrões, valores nulos e tendências significativas nos dados, além de gerar visualizações gráficas para facilitar a compreensão dos resultados. Outrossim, o projeto inclui a realização de inferências estatísticas para explorar as relações entre a qualidade do ar e a poluição da água, fornecendo insights interessantes para a compreensão dos impactos ambientais.

**Descrição das Variáveis:**
City: Nome da cidade onde os dados foram coletados.
Region: Região da cidade onde os dados foram coletados.
Country: Nome do país onde a cidade está localizada.
AirQuality: Índice de qualidade do ar na cidade.
WaterPollution: Índice de poluição da água na cidade.

**Métricas Básicas:**
O banco de dados possui 5 colunas com 3963 registros cada, com exceção apenas da variável “Region”, que apresenta 425 valores nulos.

City: 3963 valores não nulos, tipo categórica (Object).
Region: 3538 valores não nulos, tipo categórica (Object).
Country: 3963 valores não nulos, tipo categórica (Object).
AirQuality: 3963 valores não nulos, tipo contínua (Float64).
WaterPollution: 3963 valores não nulos, tipo contínua (Float64).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3963 entries, 0 to 3962
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   City            3963 non-null   object 
 1   Region          3538 non-null   object 
 2   Country         3963 non-null   object 
 3   AirQuality      3963 non-null   float64
 4   WaterPollution  3963 non-null   float64
dtypes: float64(2), object(3)
memory usage: 154.9+ KB
None
```

```
Valores nulos em cada coluna:
City                0
Region            425
Country             0
AirQuality          0
WaterPollution      0
dtype: int64
```

**Estatísticas Descritivas:**
Focando nos atributos: Qualidade do Ar (AirQuality) e Poluição da Água (WaterPollution). Podemos entender melhor as características dos dados, obtendo uma visão clara de como esses atributos se comportam em diferentes áreas e regiões.

Qualidade do Ar (AirQuality):
A qualidade do ar varia bastante de país para país. Por exemplo, no Afeganistão, os valores da qualidade do ar estão em média em torno de 37, o que indica uma condição moderada. No entanto, há uma ampla variação, com algumas medições alcançando o valor máximo de 100, e outras caindo para 0.
Isso mostra que a qualidade do ar no Afeganistão é bastante inconsistente, com grandes flutuações.

Já na Albânia, a qualidade do ar é ligeiramente melhor, com uma média de aproximadamente 52. As medições na Albânia também variam bastante, mas a mediana de 58 sugere que a maioria das cidades tende a ter uma qualidade do ar razoável.

Poluição da Água (WaterPollution):
A poluição da água também apresenta variações interessantes entre os países. No Afeganistão, a média da poluição da água é em torno de 53, o que aponta para um nível de poluição moderado. Assim como a qualidade do ar, a poluição da água também varia bastante, chegando a um mínimo de 0 e a um máximo de 83.
Por outro lado, na Argélia, a média da poluição da água é cerca de 51, mas com um desvio considerável, o que indica que em algumas áreas a poluição pode ser muito baixa, enquanto em outras pode ser extremamente alta.

Comparação por país:
Observando os países, vemos que a qualidade do ar e a poluição da água variam significativamente. Por exemplo, a qualidade do ar no Brasil é aprazível, com uma média em torno de 55 e uma variabilidade menor, indicando que a maioria das cidades tem uma qualidade de ar razoavelmente estável. Em contrapartida, na Índia, a qualidade do ar é mais baixa, com uma média de 65 e uma alta variabilidade, mostrando que algumas áreas têm ar muito poluído.

Em termos de poluição da água, a Venezuela apresenta altos níveis de poluição, com uma média de quase 79, e valores variando de 25 a 100. Isso sugere que a poluição da água é um problema grave e generalizado. Comparativamente, na Alemanha, a poluição da água é muito menor, com uma média de cerca de 40, e menos variabilidade, indicando que a água é geralmente mais limpa.

```
Digite o número de linhas que deseja exibir:  24
Estatísticas Descritivas por País:
                  Country AirQuality                                    \
                                mean     median        std        min
0             Afghanistan  37.213694  31.085526  36.242793   0.000000
1                 Albania  51.873625  58.333333  40.051927   0.000000
2                 Algeria  57.607466  50.000000  27.295630   0.000000
3                 Andorra  43.750000  43.750000   8.838835  37.500000
4                  Angola  15.000000  15.000000        NaN  15.000000
5               Argentina  68.147781  75.000000  25.615390   0.000000
6                 Armenia  22.270115  25.000000  21.038428   0.000000
7               Australia  80.598013  84.375000  21.001346   0.000000
8                 Austria  83.490955  87.500000  13.504922  50.000000
9              Azerbaijan  29.896907  29.896907        NaN  29.896907
10                Bahrain  32.656695  32.692308   7.638951  25.000000
11             Bangladesh  37.141523  25.000000  28.125403   6.250000
12               Barbados  88.333333  90.000000  12.583057  75.000000
13                Belarus  50.506834  53.125000  12.877226  31.250000
14                Belgium  67.730261  75.000000  22.958873   0.000000
15                 Belize  78.125000  81.250000  25.769410  50.000000
16                  Benin  25.000000  25.000000        NaN  25.000000
17                 Bhutan  59.821429  59.821429  21.465742  44.642857
18                Bolivia  44.832516  38.970588  22.820055  26.388889
19  Bosnia and Herzegovina  57.296646  45.588235  34.117366  10.937500
20               Botswana  64.802632  64.802632   3.256413  62.500000
21                 Brazil  63.291876  62.500000  24.842414   0.000000
22                 Brunei  47.727273  47.727273  49.818887  12.500000
23               Bulgaria  43.882159  37.500000  24.692733  12.500000
```

Comparação por região
As regiões também mostram variações notáveis. Por exemplo, na região de Aceh, a qualidade do ar é surpreendentemente alta, com uma média de 85. Isso sugere que a região possui ar excepcionalmente limpo,

apesar de alguma variabilidade. Em contraste, a região de Abruzzo na Itália tem uma qualidade do ar mais baixa, com uma média de 74, mas ainda assim acima da média global.

Quando se trata de poluição da água, a província de Ad Daqahliyah tem níveis extremamente altos, com valores fixos em 100, indicando uma poluição severa e uniforme. Em contrapartida, a região de Abruzzo apresenta níveis moderados de poluição da água, com uma média de 64, sugerindo que, embora a poluição exista, ela não é tão crítica quanto em outras regiões.

Essas métricas básicas nos dão uma visão rica e detalhada dos dados de qualidade do ar e poluição da água em diferentes países e regiões. Ao examinar a média, os extremos e a variabilidade, podemos identificar padrões e anomalias que são de grande importância para estudos ambientais e políticas públicas.

**Estatísticas descritivas:**

1. Qualidade do Ar (AirQuality)

```
Estatísticas Descritivas - Qualidade do Ar:
count    3963.000000
mean       62.253452
std        30.944753
min         0.000000
25%        37.686567
50%        69.444444
75%        87.500000
max       100.000000
Name: AirQuality, dtype: float64
```

2. Poluição da água (WaterPollution)

```
Estatísticas Descritivas - Poluição da Água:
count    3963.000000
mean       44.635372
std        25.663910
min         0.000000
25%        25.000000
50%        50.000000
75%        57.719393
max       100.000000
Name: WaterPollution, dtype: float64
```

Count: Registros de qualidade do ar disponíveis.
Mean: Média da qualidade do ar.
Std: Desvio do padrão.
Min: Valor mínimo.
Max: Valor máximo.
25% : 25% dos dados têm valores de determinada variável iguais ou menores que x.
50%: 50%  dos dados têm valores de determinada variável iguais ou menores que x.
75%: 75% dos dados têm valores de determinada variável iguais ou menores que x.

**Correlação:**
O valor calculado da correlação entre qualidade do ar e poluição da água foi de -0.454, o que indica uma correlação negativa moderada entre as variáveis.

A correlação negativa de -0.454 demonstra que, de forma moderada, regiões com melhor qualidade do ar tendem a ter menor poluição da água e vice-versa.
A inferência realizada é importante pois sugere que fatores e intervenções que melhoram a qualidade do ar podem estar associadas a reduções na poluição da água.

```
# calcular a correlação entre AirQuality e WaterPollution
correlation = data['AirQuality'].corr(data['WaterPollution'])
print(f'Correlação entre Qualidade do Ar e Poluição da Água: {correlation}')
```

```
Correlação entre Qualidade do Ar e Poluição da Água: -0.45417262259393154
```

**Teste t para comparação entre países:**
Foi realizado um teste t manual para comparar a qualidade do ar entre os Estados Unidos da América e a Alemanha, e depois, Brasil e Índia, que foram os países escolhidos pelos integrantes do grupo.

1. O objetivo do teste t é determinar se há uma diferença estatisticamente significativa na média da qualidade do ar entre os países.
2. Os dados utilizados foram retirados das colunas 'AirQuality' fornecidos pelo banco de dados do site kaggle.

**Estatísticas Calculadas:**

1. Média da Qualidade do Ar (EUA e Alemanha)
2. Desvio Padrão da Qualidade do Ar (EUA e Alemanha)
3. Média da Qualidade do Ar (Brasil e Índia)
4. Desvio Padrão da Qualidade do Ar (Brasil e Índia)
5. Estatística t: Medida que indica a diferença entre as duas médias padronizada pela variabilidade dos dados.
6. Grau de Liberdade (df): Estimado para o teste t.

**Resultados:**

1. Para EUA e Alemanha:
   Estatística t (t-stat): 3.748
   Grau de Liberdade (df): 427.86
2. Para Brasil e Índia
   Estatística t (t-stat): 6.346
   Grau de Liberdade (df): 200.03
3. Ainda para EUA e Alemanha, realizamos um comparativo da poluição da água.
   Estatística t(t-stat): 3.079
   Grau de Liberdade (df): 174.25

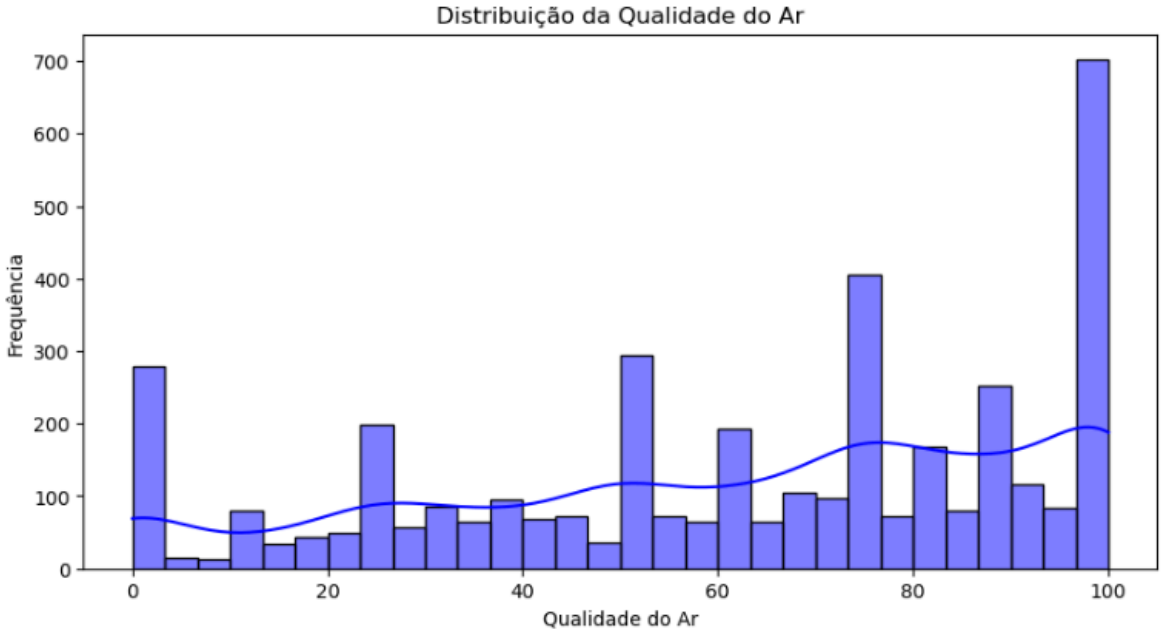
A estatística t de 3.748 indica que a diferença entre as médias da qualidade do ar nos Estados Unidos e na Alemanha é aproximadamente 3.748 vezes o desvio padrão da diferença das médias. Para determinar se essa diferença é significativa, comparamos o valor absoluto da estatística t com um valor crítico de uma distribuição t com 427.86 graus de liberdade. O mesmo é realizado com os valores entre Brasil e Índia.

Sobre os dois primeiros países escolhidos, realizamos também o test t para comparar a poluição da água onde encontramos uma discrepância significativa de aproximadamente 3.079 vezes o desvio padrão da diferença das médias.

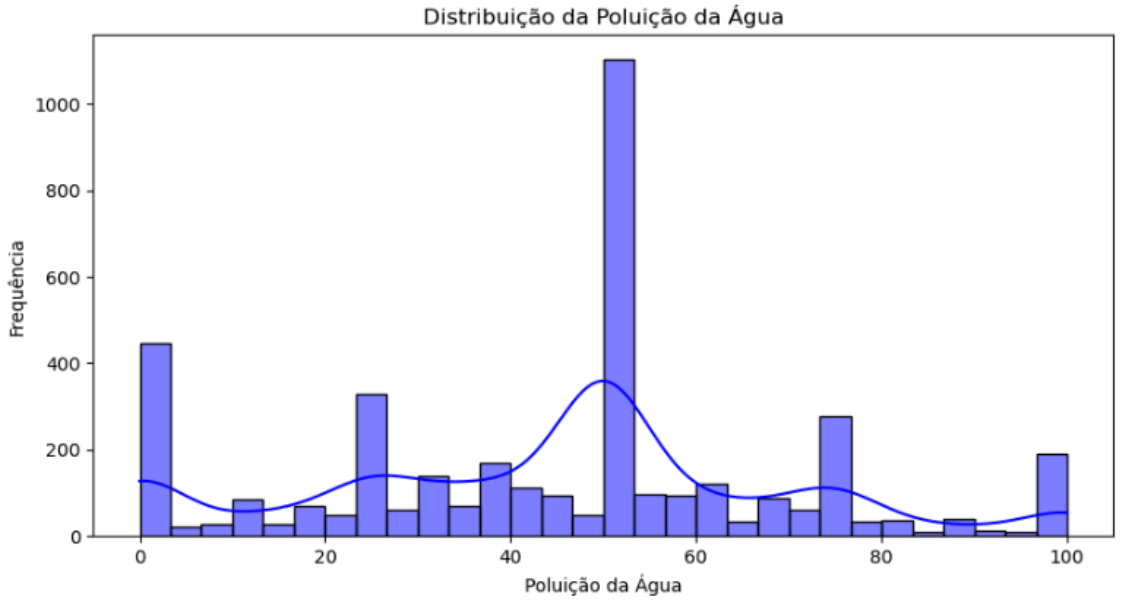
```
Teste t para Poluição da Água entre EUA e Alemanha: t-stat=3.0792990314303323, df=174.25598088242202
Diferença significativa na Poluição da Água entre EUA e Alemanha.
```

**Gráficos:**

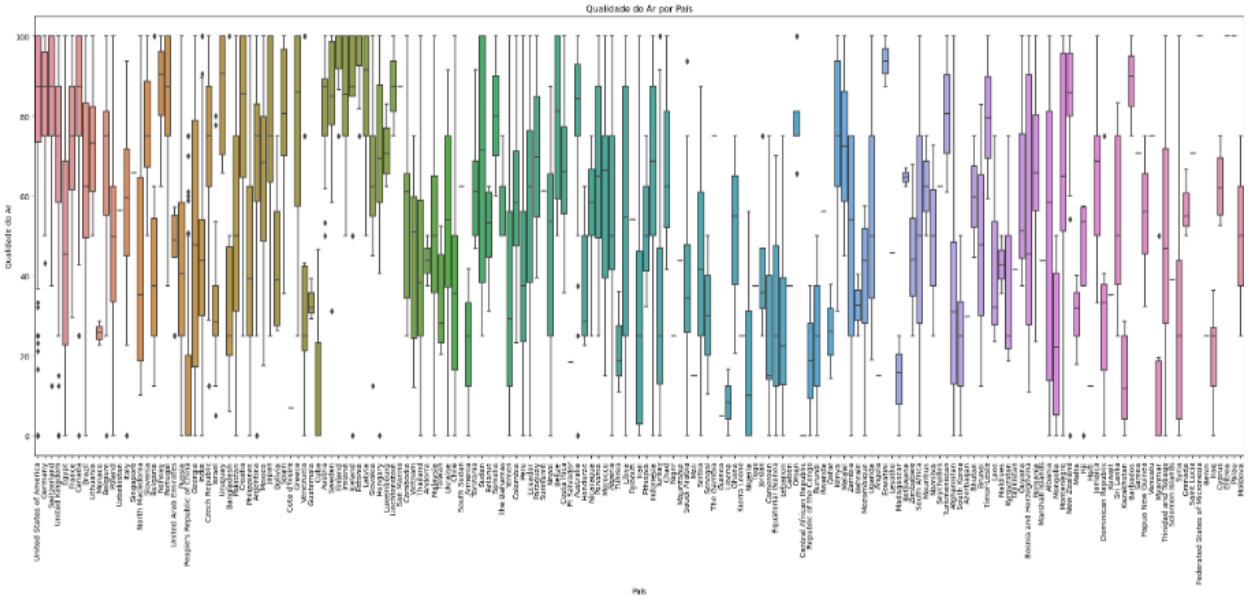
1. Histograma que mostra a distribuição da qualidade do ar nos dados, com barras representando a frequência de diferentes faixas de valores da qualidade do ar.



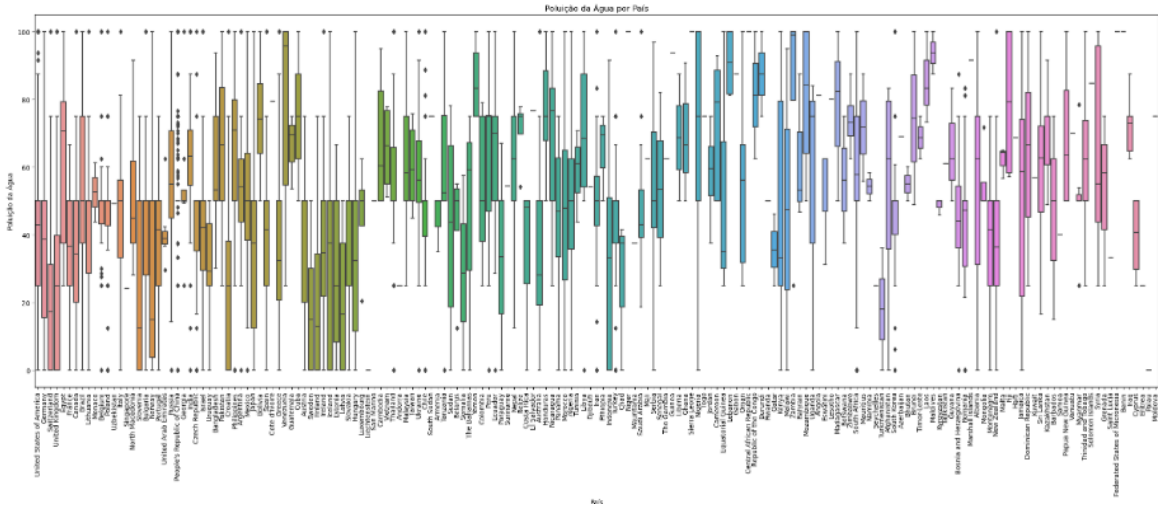

2. Gráfico, em barras, que mostra a distribuição da poluição da água nos dados fornecidos. A distribuição é representada por barras, onde cada barra indica a frequência de observações em diferentes faixas de valores da poluição da água.



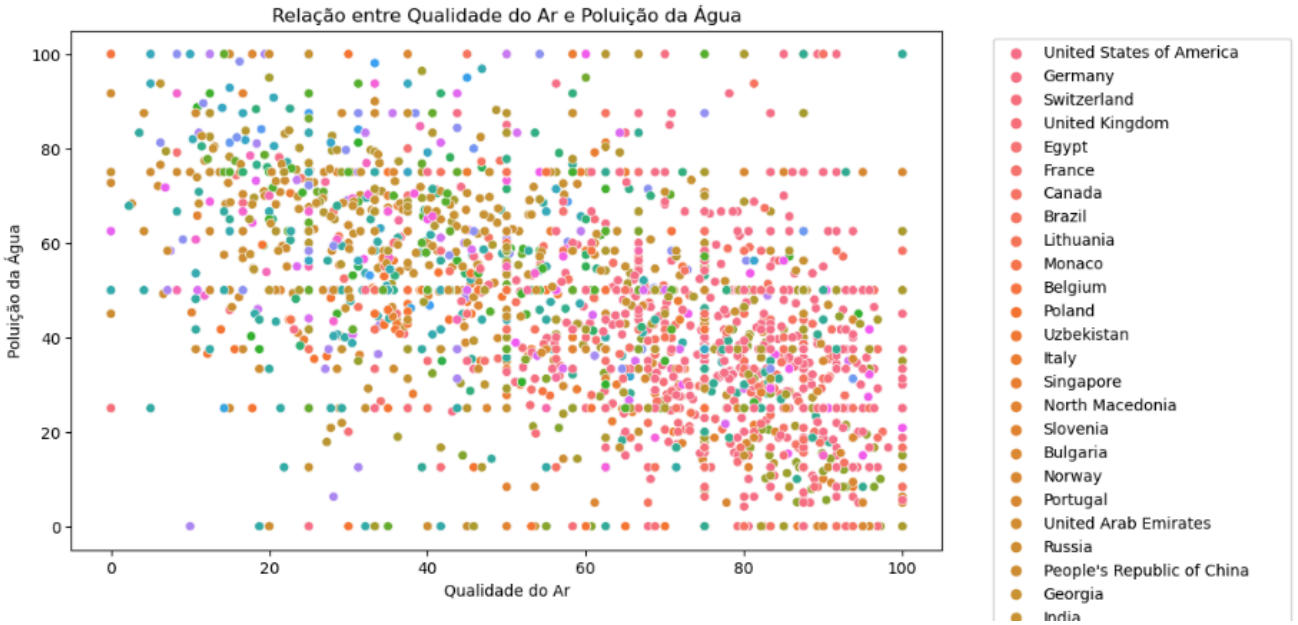
3. Gráfico, em box plot(caixa), mostrando a distribuição da qualidade do ar em diferentes países. Cada caixa representa a distribuição dos valores da qualidade do ar em um país específico.O eixo x mostra os países e o eixo y mostra os valores da qualidade do ar. Essa visualização permite comparar a distribuição da qualidade do ar entre diferentes países e identificar discrepâncias ou padrões distintos



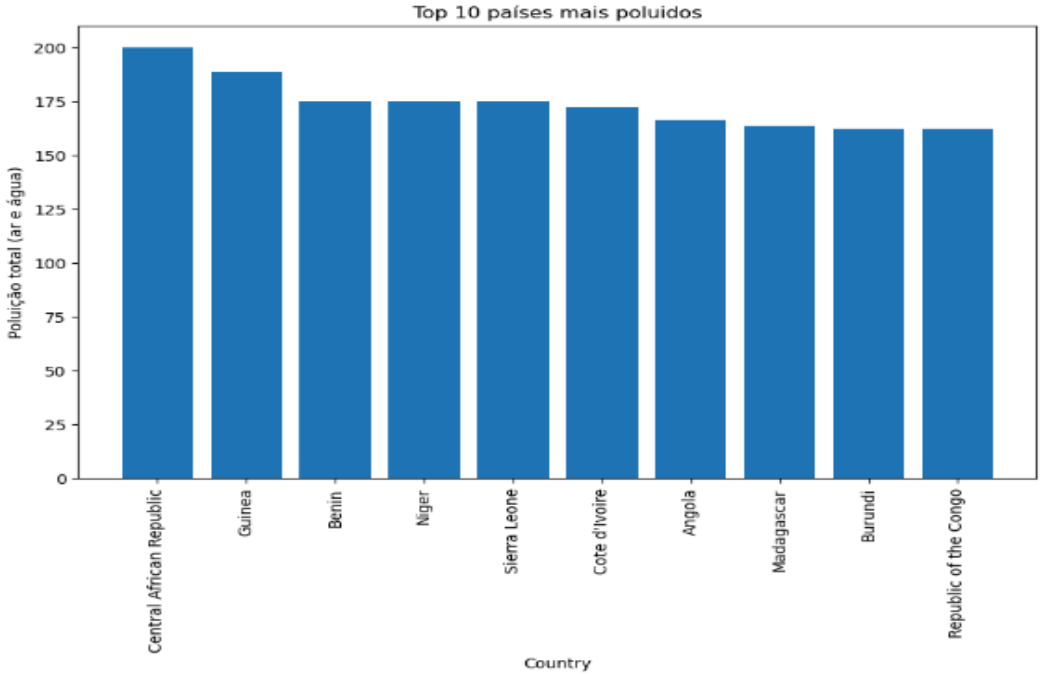

4. Também em box plot, o gráfico a seguir apresenta a distribuição da poluição da água em diferentes países. O eixo x mostra os países e o eixo y mostra os valores da poluição da água. Essa visualização permite comparar a distribuição da poluição da água entre diferentes países e identificar discrepâncias ou padrões distintos.



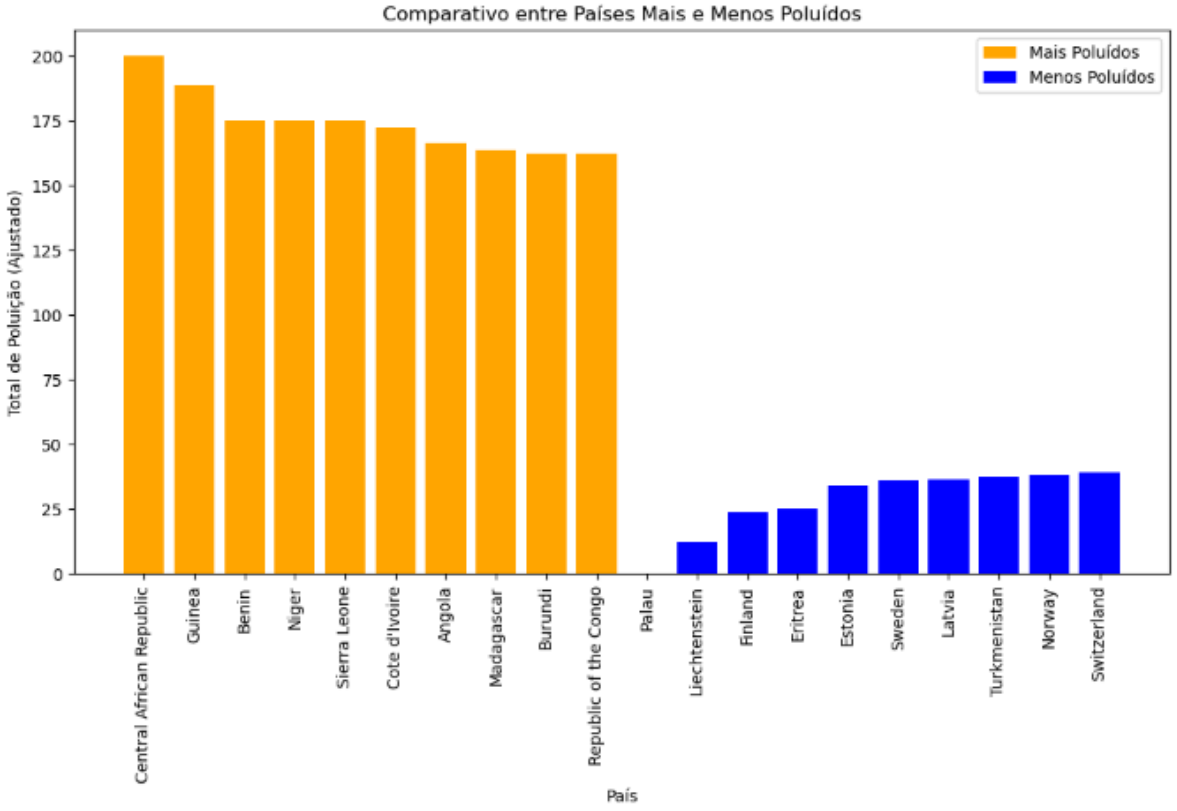
5. O gráfico mostra a relação entre a qualidade do ar e a poluição da água. Cada ponto no gráfico representa uma observação, onde o eixo x representa a qualidade do ar e o eixo y a poluição da água.




6. Este gráfico mostra os 10 países com maior poluição combinada. Esta visualização nos permite identificar os países que enfrentam os maiores desafios em termos de poluição ambiental, combinando dados de qualidade do ar e poluição da água.



7. O gráfico apresenta uma comparação entre os 10 países mais poluídos e os 10 países menos poluídos, considerando a poluição total ajustada, que combina a qualidade do ar e a poluição da água.





Alunos
Debora da Silva Amaral RM 550412
Levy Nascimento Junior RM 98655
Lívia Namba Seraphim RM 97819