Global Solution: DATA SCIENCE

- Rm: 99805
  João Pedro de Souza Vieira
- Rm:550923
  Felipe de Campos Mello Arnus
- Rm:97857
  João Pedro Oliveira Chambrone

1.Descrição do projeto e das variáveis:

Para esse projeto será utilizada a planilha *Cities.csv,* obtido no link fornecido no enunciado do pdf. O projeto consiste em analisar dados relativos à poluição da água e a qualidade do ar com o objetivo de entender padrões de poluição, impactos em diferentes regiões, locais mais afetados e outras tendências relevantes. Utilizaremos uma planilha de dados que contém informações sobre a qualidade do ar e a poluição da água, segmentados por cidade, região, país, qualidade do ar e poluição da água.

1.1Variáveis a serem utilizadas:

1. Cities:

Variável do tipo string, relativo a cidade que está sendo analisada. Devido a presença de muitas cidades, porém poucas repetições das mesmas, não mostra-se viável fazer histogramas dessa variável.

```
Quantidade de cidades e suas aparições
City
Albany          4
Rochester       4
Alexandria      4
Jackson         4
Cambridge       4
               ..
Murfreesboro    1
Debrecen        1
Sierra Vista    1
Seward          1
Zamora city     1
Name: count, Length: 3796, dtype: int64

Total de cidades repetidas: 139
Total de cidades não repetidas: 3657
```

Com base nisso, observa-se que mesmo existindo mais de 3700 cidades, pouco menos de 100 tem mais de uma aparição no banco de dados e o máximo de aparições não passa de 4.

2. Region:

Variável do tipo string, refere-se às regiões onde esses dados foram coletados. Ao analisar a região, assim como a variável anterior, percebe-se uma maior incidência de regiões não repetidas, mesmo tendo uma diferença menor.

```
Quantidade de regiões e suas aparições
Region
England                  142
California               122
Texas                     51
Florida                   48
Ontario                   47
                         ...
Mwanza Region              1
Durango                    1
Montevideo Department      1
Elbasan County             1
Los Rios Region            1
Name: count, Length: 1152, dtype: int64

Total de regiões repetidas: 436
Total de regiões não repetidas: 716
```

3. Country:

Tipo de variável do tipo string que refere se ao país onde esses dados foram coletados. Ao analisar os países, percebe-se que opostamente as outras variáveis, ele

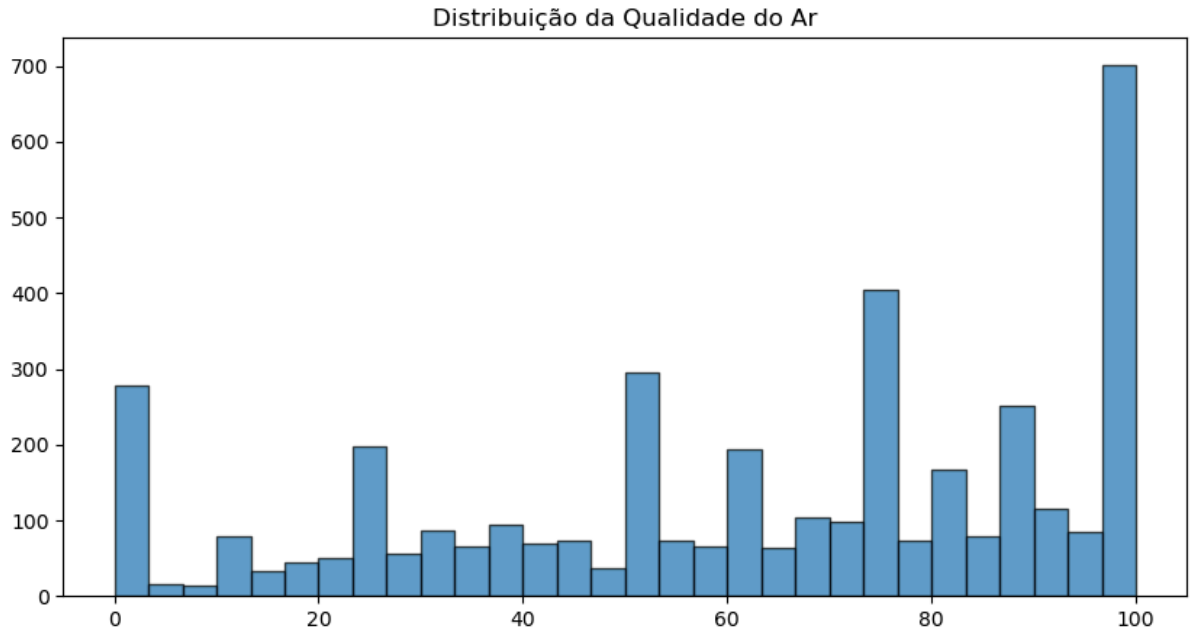
possui mais valores repetidos, dessa forma mostrando-se uma variável melhor para se trabalhar, e fazer medições como a média da poluição por país, como será mostrado a frente.

```
Quantidade de países e suas aparições:
Country
United States of America      842
People's Republic of China    238
United Kingdom                170
Canada                        157
India                         154
                             ...
El Salvador                     1
Suriname                        1
Haiti                           1
Togo                            1
Lesotho                         1
Name: count, Length: 177, dtype: int64

Total de países repetidos: 143
Total de países não repetidos: 34
```

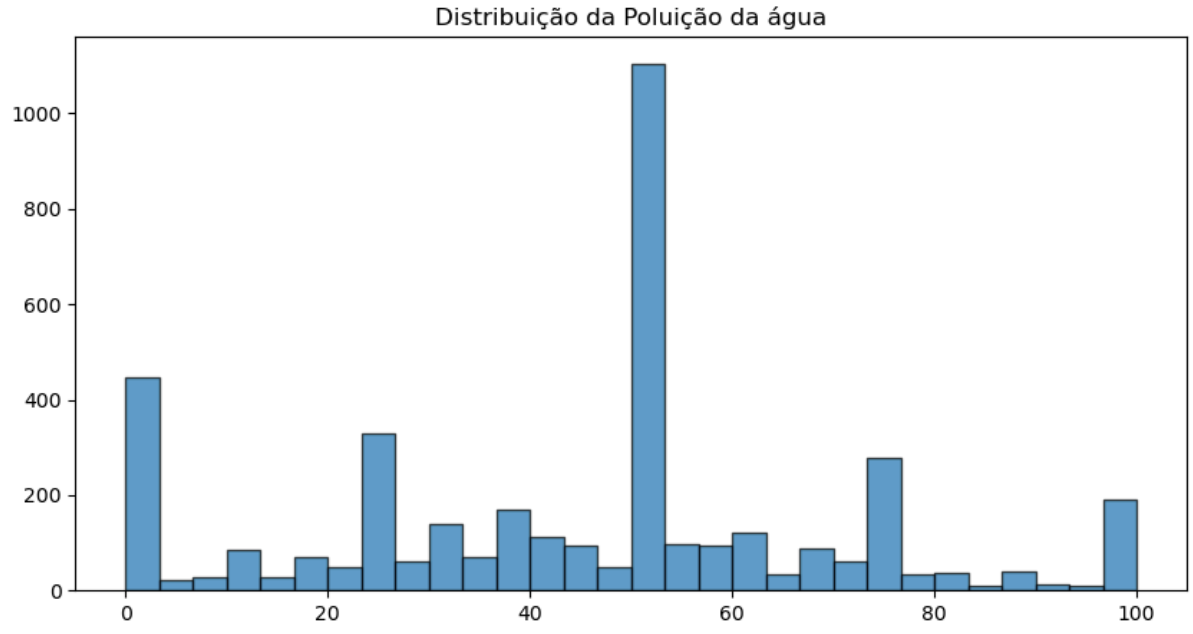
4. AirQuality:

Variável de tipo float, que refere-se à qualidade do ar do local onde foi coletado o dado. Com uma distribuição aparentemente equilibrada, ou seja, sem a presença de outliers.



5. WaterPollution:

Variável de tipo float, que refere-se a poluição da água do local onde foi coletado o dado. Com uma distribuição aparentemente equilibrada, ou seja, sem a presença de outliers.



2.Valores nulos e estatísticas descritivas:

Para uma melhor organização do documento e evitar repetição, estará sendo colocada abaixo deste parágrafo uma imagem com o retorno de dois comandos que trazem informações importantes e que serão utilizadas nesse tópico como um todo.

```
print(base.describe())
print(base.isnull().sum())
```

```
         AirQuality  WaterPollution
count  3963.000000     3963.000000
mean     62.253452       44.635372
std      30.944753       25.663910
min       0.000000        0.000000
25%      37.686567       25.000000
50%      69.444444       50.000000
75%      87.500000       57.719393
max     100.000000      100.000000
City                0
Region            425
Country             0
AirQuality          0
WaterPollution      0
dtype: int64
```

2.1 Média, mediana e moda:

Para o cálculo desses índices, foi levada em consideração todos os valores contidos nas variáveis AirQuality e WaterPollution, sem ter uma separação por cidade ou região, para a obtenção de uma visão geral.

1.Média:

Para obter o valor da média, foi utilizado o comando “.describe”, o qual tem como um dos retornos o indicador mean (que representa a média), o qual deixa em vitrine que a média da variável AirQuality é 62.253462, já a da variável WaterPollution é 44.635372.

2.Moda:

Para obter o valor da moda, foi utilizado o comando “.mode”, onde retornou o valor 100.00 para AirQuality e 50.00 para WaterPollution.

```
Coluna: AirQuality
Mediana: 69.44
Moda: 100.0

Coluna: WaterPollution
Mediana: 50.00
Moda: 50.0
```

3.Mediana:

Para obter o valor da mediana, foi utilizado o comando “.median”, onde retornou o valor 69.44 para AirQuality e 50.00 para WaterPollution.

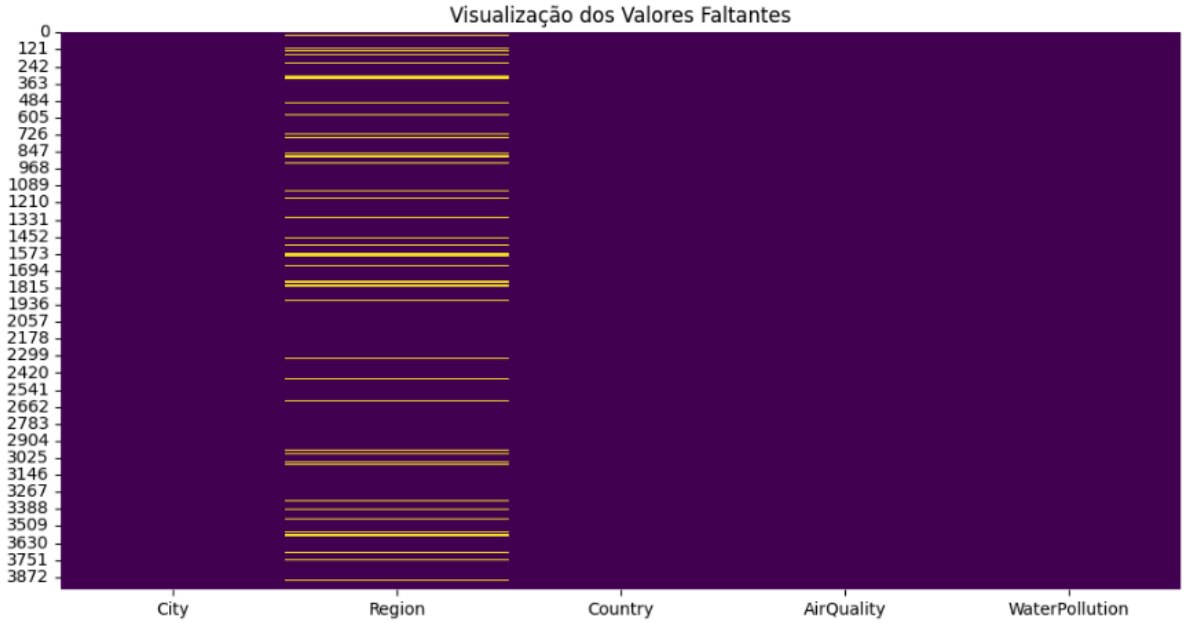

```
Coluna: AirQuality
Mediana: 69.44
Moda: 100.0

Coluna: WaterPollution
Mediana: 50.00
Moda: 50.0
```


2.2 Valores nulos:

A obtenção desses índices foi feita a partir do comando “.isnull”, referenciado acima, para saber se existem valores nulos e em quais variáveis eles se encontram.

Como revelado pelo retorno do comando, apenas uma variável possui valores nulos, a variável “Region”, além disso observa-se que são 425 valores nulos, que representa aproximadamente 10.7% do tamanho da base de dados, uma quantidade relevante de dados que possuem valores nulos. Como é melhor observado abaixo.



Na imagem acima, todos os traços amarelos representam valores nulos em uma determinada variável.

Felizmente, tais valores nulos estão restritos exclusivamente a variável relativa às regiões, sendo assim, nenhum dado de variável numérica (AirQuality e WaterPollution) possui valores nulos, sendo assim, os dados numéricos até agora não possuem nenhuma análise enviesada.

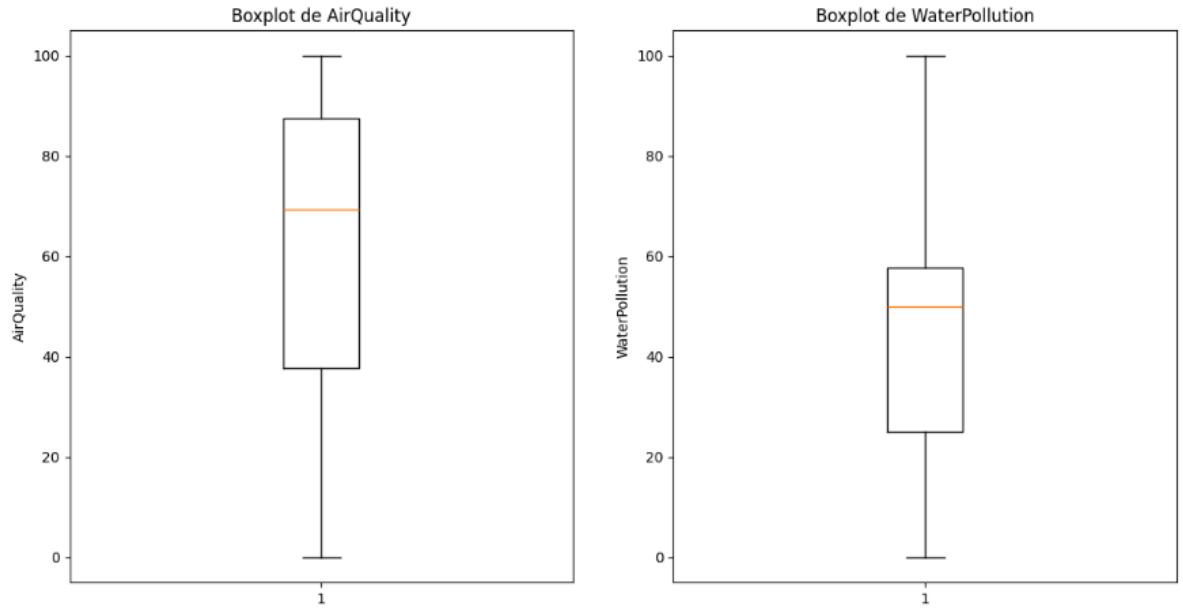
2.3 Desvio padrão, coeficiente de variação e outliers:

Para a obtenção do desvio padrão, é analisado o retorno do comando “.describe”, onde o índice std evidencia o desvio padrão, sendo 30.944753 o desvio para AirQuality e 25.663910 o desvio para WaterPollution.

A partir desse retorno é possível calcular o coeficiente de variação fazendo a divisão do desvio padrão pela média, ao observar o resultado, pode ser prevista uma chance de existirem outliers a partir dele, tendo como métrica a proximidade do número 1, sendo que de 25% para baixo é uma chance pequena de existires outliers.

```
Coeficiente de Variação para AirQuality: 49.71%
Coeficiente de Variação para WaterPollution: 57.50%
```

Tendo em vista o retorno obtido, a chance de ter outliers mostra-se ser não baixa. Mas para obter a certeza de que esses realmente não existam, precisa ser feita a análise de outros parâmetros, como a análise dos histogramas e a geração de boxplot.



Com base na análise dos parâmetros já citados, é improvável a existência de outliers, mesmo com os coeficientes mostrando valores alterados, uma vez que na visualização dos boxplots não há nenhuma marcação além dos limites superiores ou inferiores das duas variáveis.

3.Análise relativa aos valores nulos presentes no banco de dados:

Nesse ponto, será feita a análise de parâmetros a partir da presença de valores nulos. Tendo em vista que os valores nulos estão restritos às regiões, mas mesmo assim possuem o valor do país, dessa forma sendo possível saber a quantidade de ocorrência de valores nulos por país e sendo assim pode-se fazer um levantamento de quais países possuem valores nulos e ordenando-os em ordem crescente. Para uma melhor visualização dos retornos foram criados gráficos de barras para ilustrar melhor, sendo que o primeiro mostra todos os países, enquanto o segundo agrega os países com a quantidade de valores nulos abaixo do limite definido como “outros” para uma melhor visualização. Dessa forma, obtivemos o seguinte:

Retorno da quantidade de valores nulos por cada país:

```
Países com valores nulos ordenados em ordem decrescente:
                City  Region  Country  AirQuality  WaterPollution
Country
Czech Republic     0      57        0           0               0
France             0      42        0           0               0
Philippines        0      36        0           0               0
Finland            0      23        0           0               0
Greece             0      20        0           0               0
...              ...     ...      ...         ...             ...
Mozambique         0       1        0           0               0
Niger              0       1        0           0               0
Nigeria            0       1        0           0               0
Norway             0       1        0           0               0
Marshall Islands   0       1        0           0               0

[99 rows x 5 columns]
```

Gráfico relativo a quantidade de valores nulos por país:

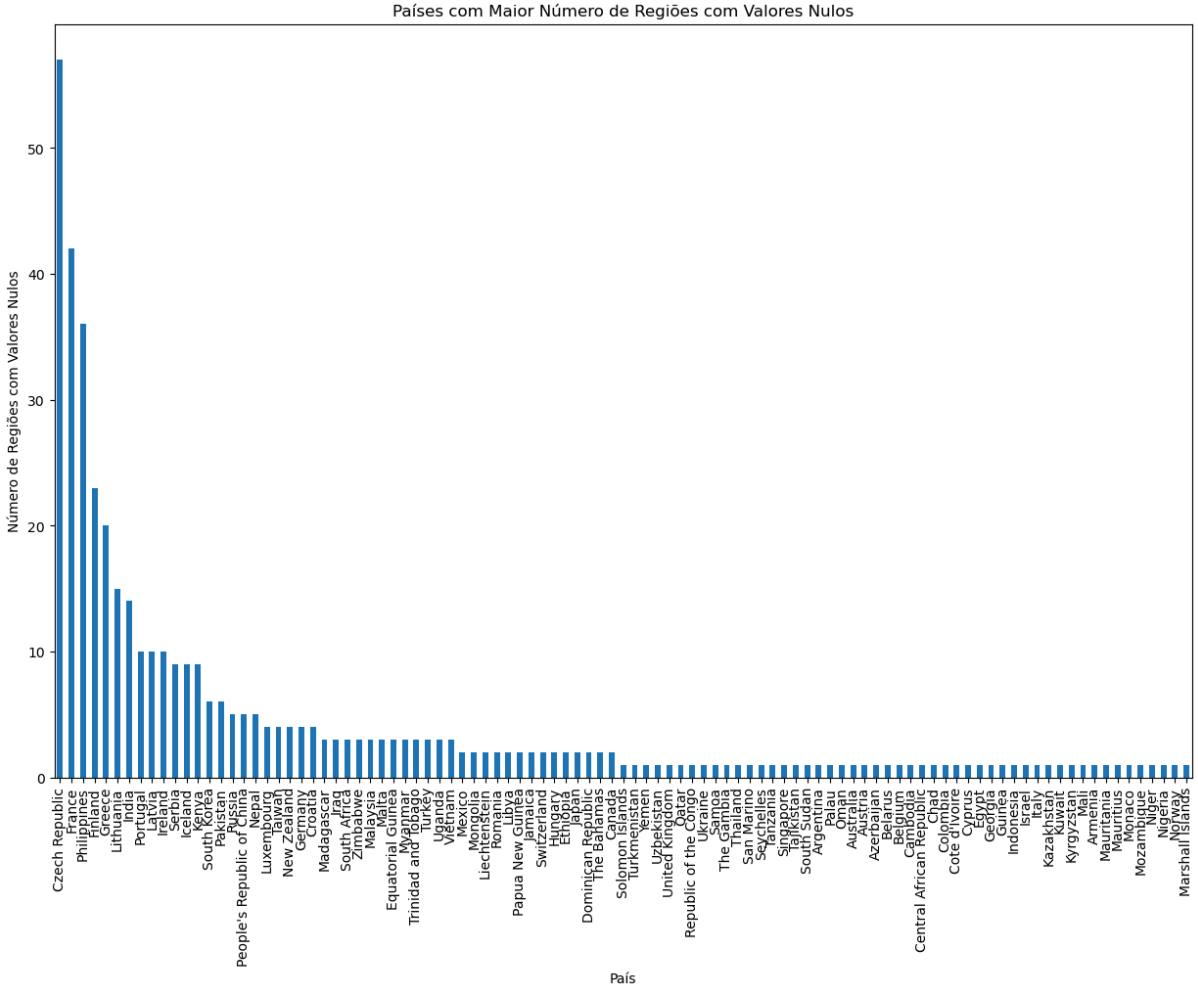
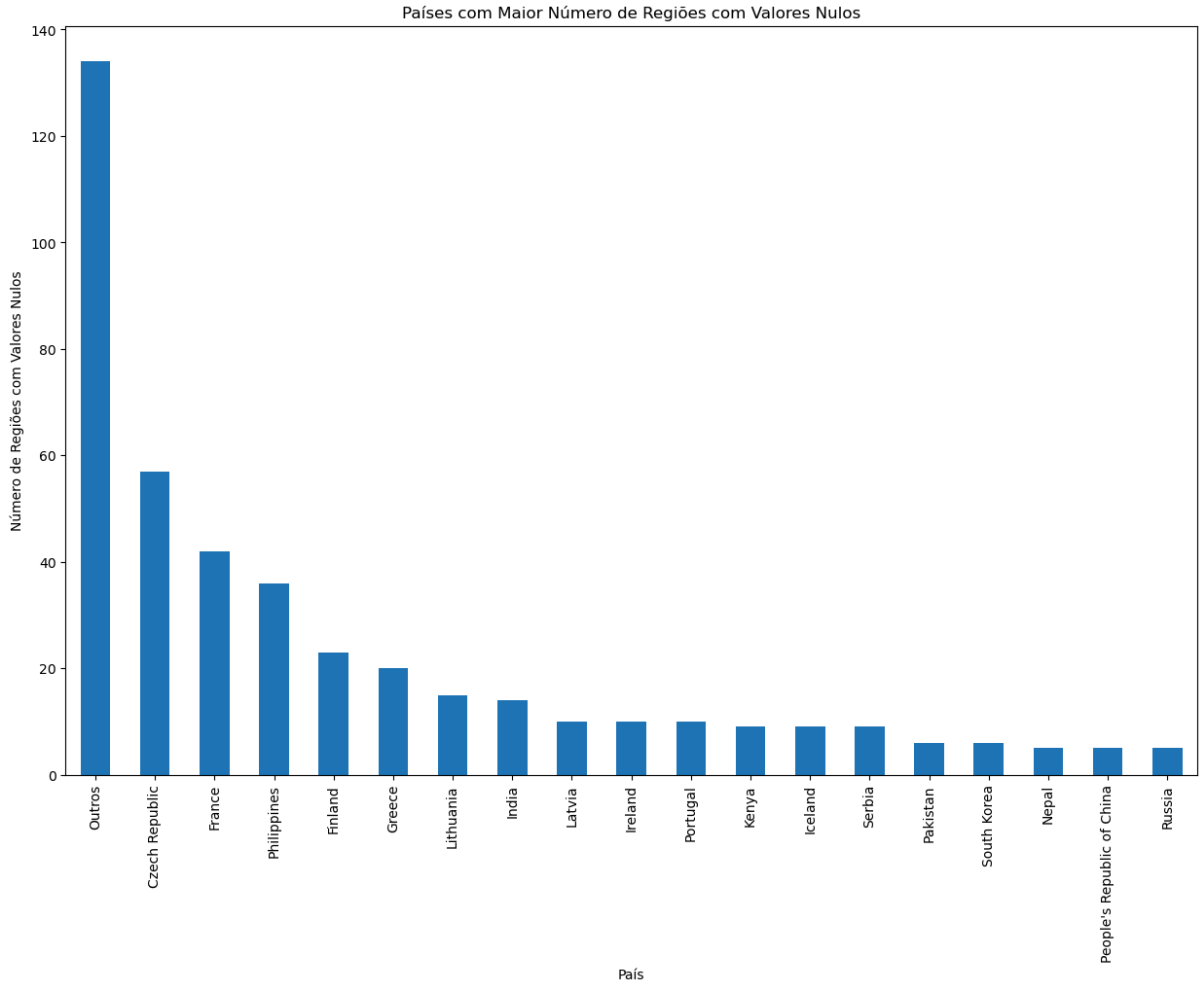


Gráfico reduzido relativo a quantidade de valores nulos por país:



Conclusão relativa aos gráficos:

Com isso, evidencia se que há uma grande diferença na distribuição dos valores nulos para cada país, havendo uma grande concentração em alguns países e poucas em outros, como por exemplo a *"Czerch Republic"* que possui 57 valores nulos enquanto Marshall Islands possui apenas 1, sendo uma diferença de 57 vezes do maior número para o menor, que mostra-se ser um valor muito elevado quando comparado com a maioria dos outros países, como aponta o primeiro gráfico, algo que indica tal disparidade é a comparação dos dois gráficos, uma vez que no segundo a variável *"Outros"* agrega todos os países com menos de 5 valores nulos relacionados a eles, o que reduziu consideravelmente a quantidade de países presentes no gráfico, uma vez que como observado no primeiro gráfico, a maioria dos valores encontrasse abaixo de 10.

Além disso, após calcular a média de valores nulos por país, ou seja, somando a quantidade de vezes que cada país possui de valores nulos e dividir esse valor pela quantidade de valores nulos, percebe-se que há uma grande variação, pois a média é aproximadamente 2.4 e o país com a maior quantidade de incidência de valores nulos é 57, uma diferença de 23.75 vezes.

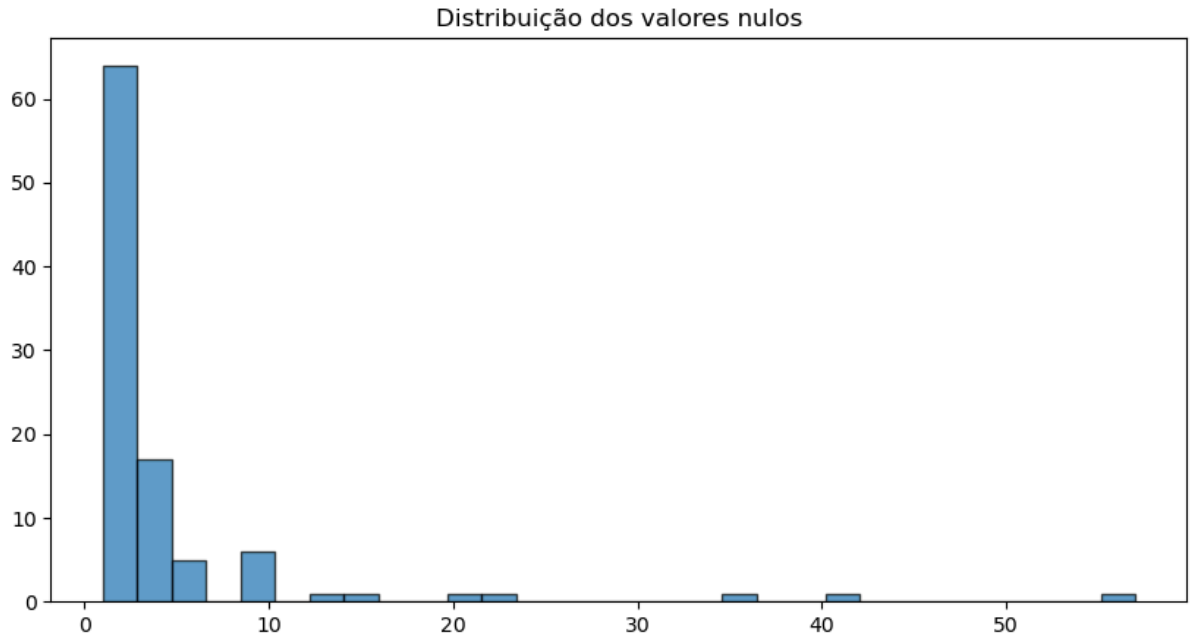
```
Média total de valores nulos por país:
2.401129943502825
```

Análise de variação:

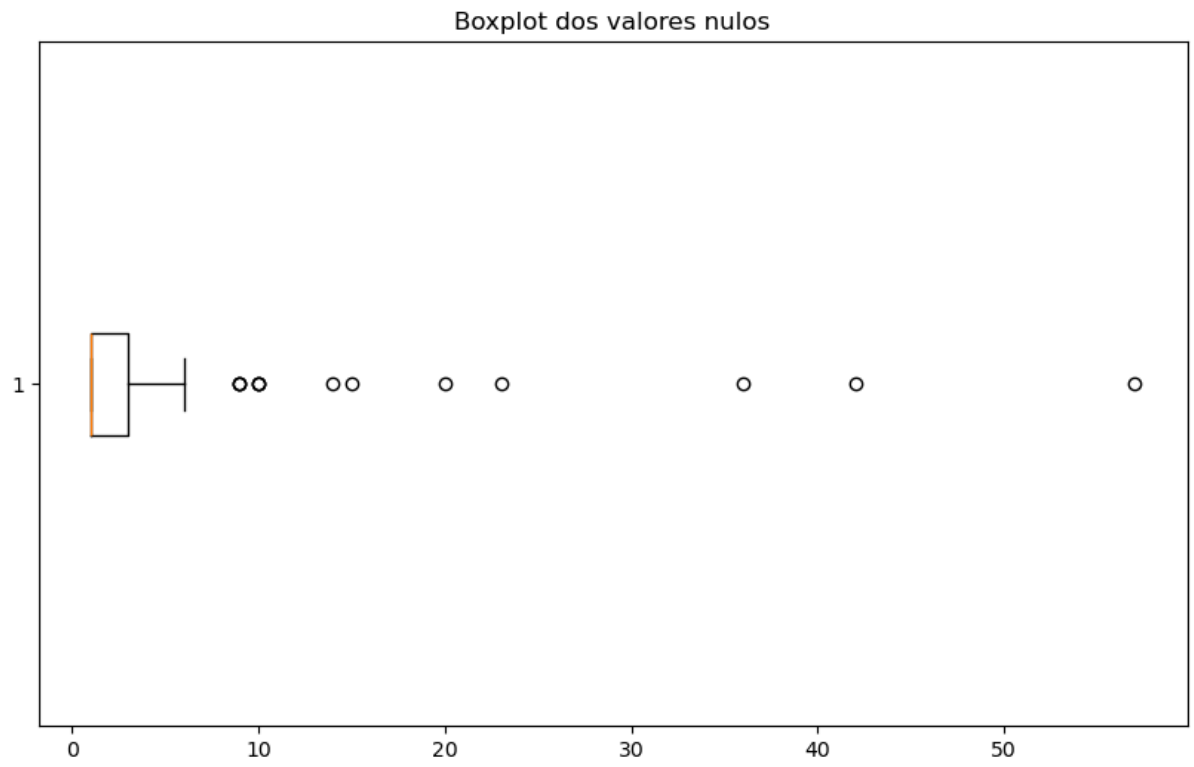
Para isso, será calculado o coeficiente de variação, e será gerado um histograma, além de um boxplot para a verificação da existência de outliers relativos a incidência de valores nulos por país. Com o cálculo do coeficiente de variação, obtivemos como retorno um coeficiente de 194.30%, sendo assim um valor que indica uma grande probabilidade para a presença de outliers.

```
O desvio pdrão dos valores nulos é de:
8.341257827690 65%
O coeficiente de variação dos valores nulos é de:
194.30%
```

Porém, para ter certeza da existência de outliers, foi gerado um histograma que mostrou uma grande variação, o que é mais um indicativo da presença de outliers.

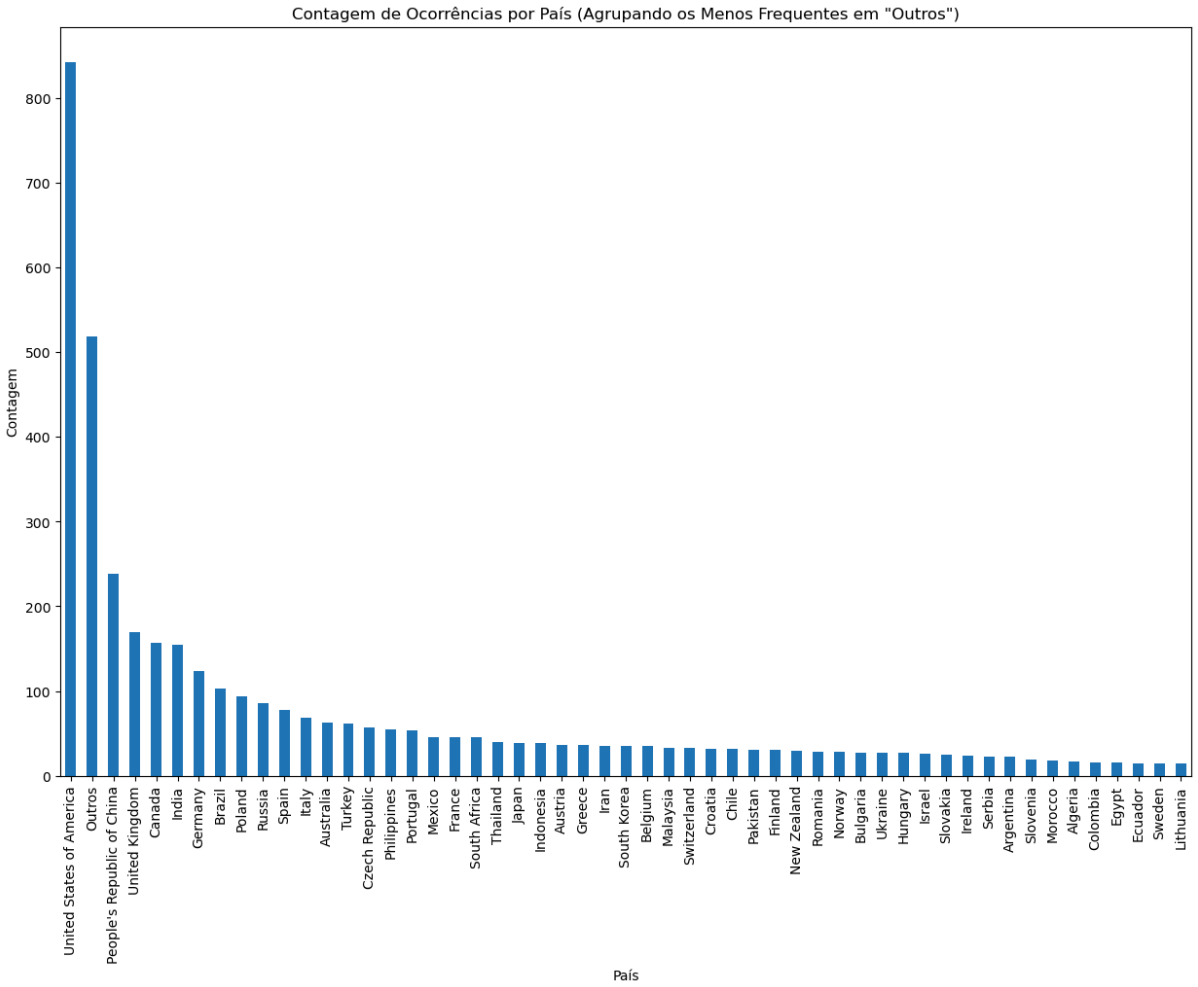


Por fim, para uma verificação final foi gerado um boxplot, que confirmou a presença de outliers devido às marcações circulares além dos limites indicados na imagem. Sendo assim, há “valores fora da curva”, que são gerados por um grupo de dados que se diferencia significativamente das outras observações.



4.Análise relativa a quantidade de aparições de cada país presentes no banco de dados:

Com essa análise, será mensurada a quantidade de vezes em que cada país aparece no banco de dados, dessa forma podendo-se saber quais deles possuem mais informações coletadas. Para essa análise, mostra-se viável possuir apenas um gráfico com a variável *"Outros"* já inserida, essa por si tendo os mesmo princípio da presente na análise anterior, mas com a diferença de que países com 15 aparições ou menos fosse para essa. Tal diferença existe pois, este gráfico possui muito mais países que o anterior, então um gráfico sem esse parâmetro de agrupamento seria impossível de compreender, uma vez que mesmo que as barras pudessem ser avistadas, os países aos quais elas representam não seriam visíveis.



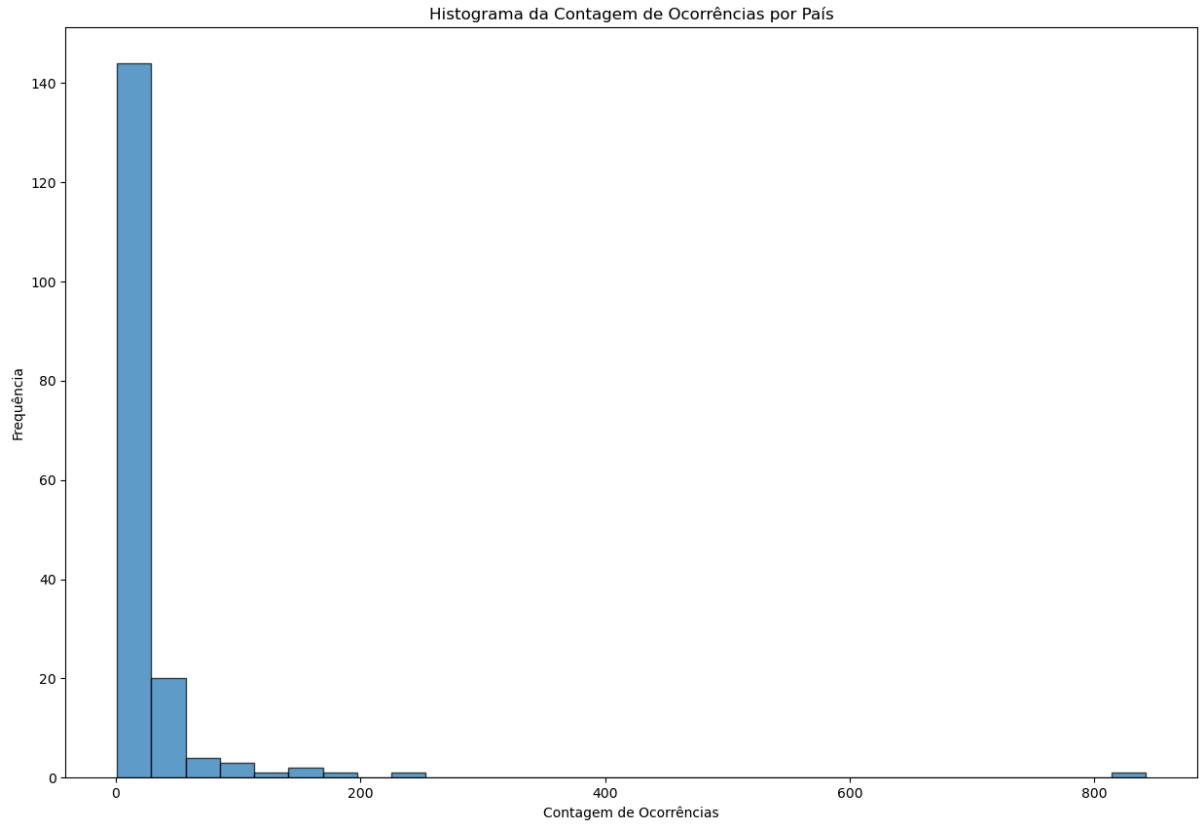
Com esse gráfico, é possível perceber que há também uma grande variação dos valores, uma vez que mesmo agregando aqueles com menos de 15 aparições em uma mesma variável, os *"United States of America"* ainda sim possuem um valor maior que todos os agregados em *"Outros"*, pois como observado ele possui mais de 800 aparições, enquanto a outra possui pouco mais de 500.

Sendo assim, essa mostra ser uma variação muito grande principalmente quando comparada com a média, uma vez que a média total da quantidade de aparições de cada país no banco de dados é de cerca de 22.39, enquanto a maior quantidade de aparições é de 842, sendo assim uma diferença de 842 vezes para o país com a menor quantidade de aparições, *"Lesotho"* com apenas 1 aparição, e de 37.6 vezes para a média geral.

```
Média total de aparições por país:
22.389830508474578
```

```
Country
United States of America      842
People's Republic of China    238
United Kingdom                170
Canada                        157
India                         154
                              ...
El Salvador                     1
Suriname                        1
Haiti                           1
Togo                            1
Lesotho                         1
Name: count, Length: 177, dtype: int64
```

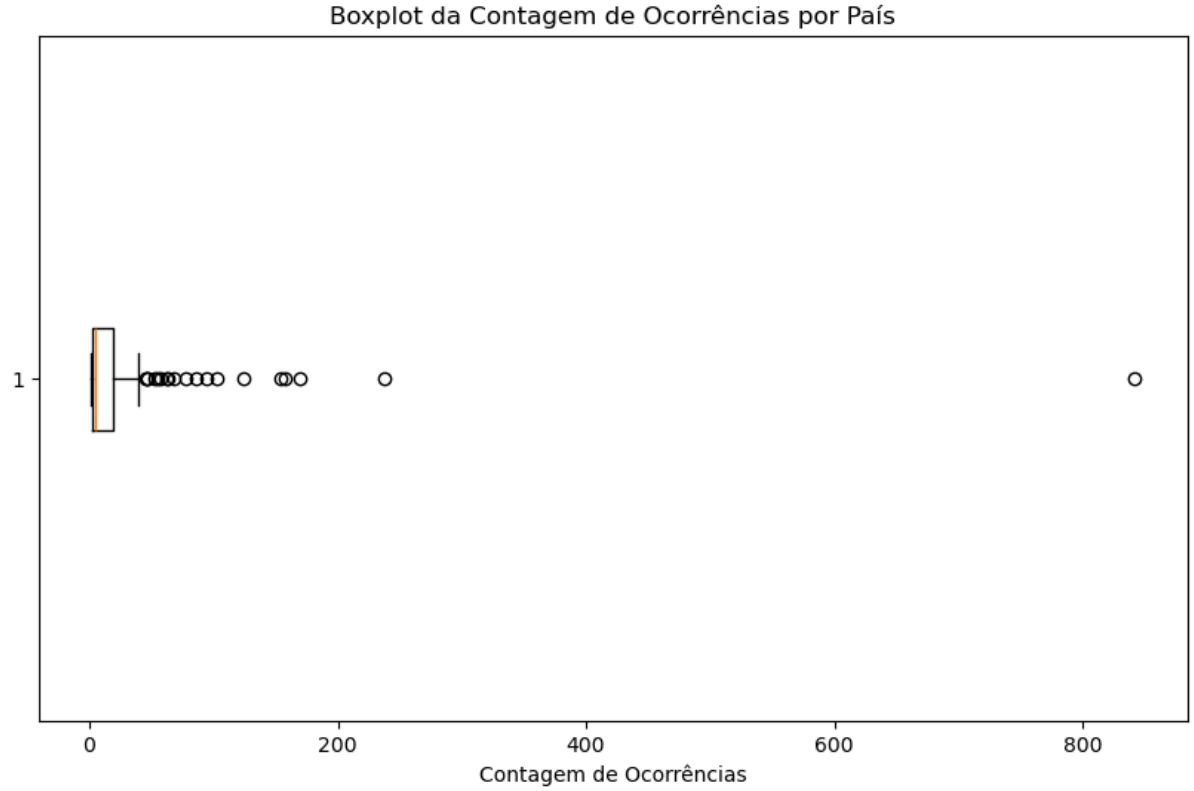
Ainda na análise desses dados, pode ser feita a análise relativa à presença de outliers na quantidade de aparições de cada país. Para isso foi gerado o histograma a seguir:



O qual ilustra uma grande variação dos dados, com muitas ocorrências com números próximos a 0 e com poucas para outros, com 800, o que é um indício da presença de outliers. Para uma verificação inicial, foi calculado o Coeficiente de Variação, o qual foi de 312%, sendo assim uma probabilidade alta da presença de outlier.

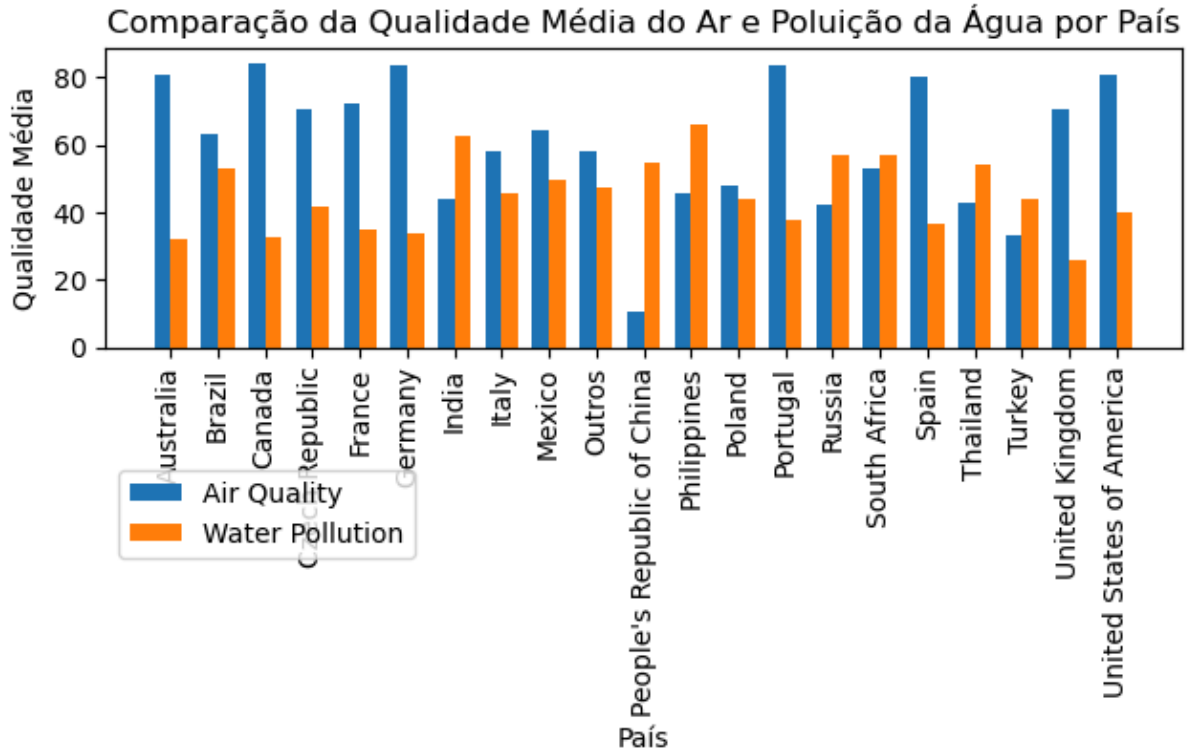
```
O coeficiente de variação da quantidade de aparições é de:
312.31%
```

Por fim, para a confirmação da presença de outliers, foi gerado um boxplot, o qual revelou haver marcações além dos limites, confirmando assim a presença de outliers na análise da quantidade de aparições de cada país no banco de dados.



5.Análise relativa a qualidade media do ar e da poluição da água de cada país presentes no banco de dados:

Para essa análise, foi calculada a média desses dois valores para cada país, ainda usando o método de agrupamento da variável *“Outros”*, para a melhor visualização de gráficos, o critério de agrupamento foi selecionado ainda com base na quantidade de vezes que o país apareceu no banco de dados, e não com base em alguma das duas médias, pois acreditamos que assim pode ser feita uma análise mais precisa relativa ao país como um todo devido a quantidade de dados. Com essa análise espera-se visualizar a diferença desse valores a depender de cada país. Primordialmente foi gerado um gráfico para a análise desses valores.



Tendo esse gráfico em vitrine, pode-se chegar a algumas conclusões iniciais, como que dentre os países que não foram agregados em *"Outros"*, a *"People's Republic of China"*, possui a menor media de AirQuality, enquanto países como *"Portugal"* e *"Germany"* estão entre médias que estão entre as maiores desse mesmo indicador. Além disso, percebe-se também que para a maioria dos países a qualidade do ar supera a poluição da água, porém é claro existem exceções como a *"India"* e a *"People's Republic of China"*.

6.Conclusão:

Após todas as análises realizadas com base no banco de dados relativo a Qualidade do Ar e a Poluição da Água, e estabelecendo parâmetros de agrupamento para melhor visualização de gráfico que também pode ser utilizado para a evidenciação da má distribuição de valores, além da geração de gráficos, histogramas, boxplot, cálculos de média, desvio padrão e coeficiente de variação, pode-se ter as seguinte conclusões:

6.1 Valores nulos e estatísticas descritivas:

Para essa parte, pode-se concluir que há uma significativa presença de valores nulos presentes no banco de dados, e como apontado tanto pelo retorno do comando ".isnull()", tanto pela imagem para melhor visualização, todos esses valores nulos estão restritos exclusivamente a variável "Region", o que tornaria análises tendo essa variável como um parâmetro poderiam ser enviesada, uma vez que esses valores representam cerca de 10.7% do banco como um todo.

Tendo isso em vista, as análises feitas neste projeto têm como base outras variáveis ou análise dos próprios valores nulos, para evitar resultados enviesados. Porém, mesmo com isso, o projeto não foi limitado ou prejudicado, devido a substituição da variável com valores nulos por outra equivalente, mas sem a presença deles.

Por fim, para o entendimento de estatísticas descritivas, comandos como ".describe()" e ".median()" foram usados para calcular os valores relativos às estatísticas descritivas, com isso, obtivemos os valores da média, mediana, moda e desvio padrão, que foram usadas para de forma extensiva na análise tanto em uma visão geral, quanto nas análises mais específicas, como as de valores nulos e das aparições de cada país.

6.2 Presença de Outliers:

Quanto a essa parte do trabalho, ela foi aplicada tanto na análise geral, quanto nas análises de valores nulos e da quantidade de aparições de cada país no banco. Para todas foram usados os mesmos métodos, primeiro a análise dos histogramas relativos a cada uma dessas partes, depois o cálculo do coeficiente de variação, dividindo a desvio padrão pela média, e por fim a geração de boxplot na tentativa de identificar marcações circulares além dos limites estabelecidos.

Tendo como parâmetro 25% para baixa probabilidade da existência de outliers detectados por meio dos valores do coeficiente de variação, foi percebido que todas as análises que contavam com esse método, apontaram alterações, como os valores de AirQuality e WaterPollution que deram próximo a 50%, ou o valor relativo a quantidade de aparições de cada país tendo como retorno 312.31%, sendo assim um valor muito alto.

O passo seguinte foi a geração dos histogramas, já nesse método não teve um resultado unânime como o outro, uma vez que nem todos os histogramas apontavam a existência de outliers, que pode ser percebida por uma concentração anômala e extrema de valores representados por barras. Nesse cenário, o histograma destoante foi o da análise geral relativa aos índices de AirQuality e WaterPollution, uma vez que tiveram valores melhor distribuídos que outros histogramas, como o caso do relativo a ocorrência de cada país.

Por fim, a última etapa desse processo é a geração do boxplot, que aponta outliers com marcações circulares além de seus limites. Com essa parte, afirmou-se que não existiam outliers para os valores gerais de AirQuality e WaterPollution, e foi confirmada a presença de outliers nas demais análises, valores nulos por país, e a quantidade de aparições de cada país no banco de dados.

Sendo assim, a conclusão geral que pode ser tirada é que para os valores de AirQuality e WaterPollution, os valores estão melhor distribuídos e sem "pontos fora da curva", mesmo que o coeficiente de variação tenha sido um valor acima do limite para ser considerada uma baixa probabilidade de ocorrência desses valores. Já para as outras análises, existe uma "concentração" que pode ser constatada por todos os 3 métodos, confirmando assim a existência desse "pontos fora da curva".

6.3 Análise dos valores nulos:

Para essa análise, teve como objetivo observar qual foi a distribuição dos valores nulos presentes na variável “Region”, utilizando o país em que eles se encontram como parâmetro para ver sua distribuição e quantidade de ocorrência de valores nulos por cada país. Inicialmente, foi averiguado que haviam 99 com pelo menos uma ocorrência de valores nulos, e considerando que o total de países é de 177, sendo assim 55.93% dos países possuem pelo menos um valor nulo na variável “Region”, além disso constatou-se que *“Czech Republic”* é o país que possui a maior quantidade de valores nulos, com 57, e *“Marshall Islands”* está entre os países com a menor quantidade de aparições (1), além disso a média de valores nulos por país é de 2.4, sendo assim o país com a maior incidência de valores nulos tem uma quantidade 57 vezes maior que o que possui menos, e 23.75 vezes mais que a média desses valores por país, sendo pode-se concluir que existe uma maior ocorrência de valores mais baixos. Algo que reforça isso é a variável “Outros”, criada para agrupar países com menos de 5 valores nulos, uma vez que ela agrupa 81 países, com isso essa variável equivale a 81.81% da quantidade total de países que possuem valores nulos.

Já sobre a presença de outliers relativa a essa análise, além de uma concentração de valores mais baixos que já foi percebida com a variável “Outros”, há também o coeficiente de variação medindo 194.3%, que é um número muito alto, o histograma que evidencia ainda mais a concentração de valores, e por fim existem marcações além dos limites do boxplot, confirmando assim a presença de outliers.

Sendo assim, conclui-se que há uma grande quantidade de países com pelo menos uma aparição de valores nulos, ao ponto desses países representarem mais de 50% da quantidade total, e há uma discrepância na distribuição da quantidade de valores nulos de cada país, pois a maioria dos países possui uma concentração baixa desses valores, mas existem uns poucos que tem valores muito altos, principalmente quando observa-se que a diferença do maior valor para a média é de mais de 20 vezes.

6.4 Análise das aparições de cada país:

Com essa análise, foi levado em conta a quantidade de vezes que cada país aparecia no banco de dados, sem se deixar enviesar por valores nulos, uma vez que no variável “Country” não há valores desse tipo.Constatou-se que haviam 177 países diferentes, sendo que desses o que conta com mais aparições no banco é o *“United States of America”*, tendo 842 dados coletados relacionados a esse, enquanto entre os países que possuem apenas uma aparição há o *“Lesotho”*. Além dessa colossal diferença de mais de 800 vezes da quantidade mínima e máxima de aparições, há também a diferença do máximo para a quantidade média, considerando que essa é 22.38, a diferença existente é de cerca de 37.6 vezes, o que evidencia uma diferença na quantidade de aparições de cada país. Outro ponto que auxilia a perceber isso, é a variável “Outros”, que segue o mesmo conceito de agrupar os países menos frequentes, uma vez que essa agrupa 125 países com menos de 15 aparições, sendo assim ela nos revela que foi agrupado 70.62% da quantidade total dos países nessa variável, mesmo ela possuindo um parâmetro abaixo da média, ou seja, a maioria dos países possui uma quantidade de aparições abaixo da média.

Em adição ao que foi apontado no parágrafo anterior, a detecção de outliers para essa análise foi feita seguindo os mesmos passos de todas as análises presentes nesse projeto, obtendo como resultado um coeficiente de variação de 312.31%, um histograma que aponta valores concentrados, e por fim um boxplot que apontou marcações além dos limites, ou seja, cada análise apontou e confirmou o que já havia sido observado anteriormente, que há valores “fora da curva” relacionados a quantidade de aparição de cada país no banco de dados.

Em conclusão, essa diferença da quantidade de aparições dos países pode influenciar na visão geral desses países, uma vez que a área territorial de um país é grande, e com certeza sempre vai além de apenas um único local, sendo assim isso implica na precisão dos dados medidos quando considerados como a medição relativa ao país como um todo, pois quanto mais vezes ele aparecer mais medições diferentes ele apresenta, tendo assim mais dados sobre o país como um todo, tendo assim uma medição menos enviesada, já que nem todos os locais têm o mesmo nível de poluição de água ou qualidade do ar, isso mesmo dentro do mesmo país. Tendo isso em vista, percebe-se que alguns países apresentaram dados médios que equivalham mais a real métrica média desses valores nesse país, pois há vários valores diferentes coletados em diferentes lugares, enquanto isso existem também países que mostraram dados médios enviesados devido a pouca quantidade de dados coletados sobre eles, sendo esses últimos os que estarão em maior quantidade, uma vez que já foi estabelecido que há a presença de outliers e que a maioria dos países está bem abaixo da média total de aparições devido a presença desses valores.

6.5 Análise das média dos índices por país:

Em última instância, essa foi a análise final desse projeto. Está tem como objetivo mostrar as médias dos valores da qualidade do ar e da poluição da água de cada país simultaneamente por meio de um gráfico. Para isso foi usada mais uma vez a técnica de agrupamento de países menos frequentes no banco de dados, com a diferença de que o limite definido dessa vez foi de menos de 40 aparições, uma vez que como está sendo analisado duas médias ao menos tempo, colocar qualquer uma delas como métrica dessas medidas poderia enviesar o gráfico com relação a outra, então foi escolhida essa forma pois, como já estabelecido anteriormente, os países que tiverem mais aparições apresentaram uma média que melhor condiz com a real média geral desse país.

Sendo assim, ao gerar o gráfico foram mostrados os 20 países com maior frequência no banco, uma vez que a variável “Outro” agrupou 157 países, e desses que foram postos em vitrine percebeu-se que, pelo menos entre os países mais frequentes, geralmente a qualidade do ar se sobressai sobre a poluição da água de cada país, mesmo existindo exceções como a *“People’s Republic of China”* e a *“India”*, ainda sobre essas medias, percebeu-se que dos países mostrados no gráfico a já citada *“People’s Republic of China”* possui a menor qualidade de ar e está entre os que possuem a maior poluição da água. Já analisando a maior média relativa a qualidade do ar, percebeu-se que além de geralmente se sobressair sobre a poluição da água, países como *“Portugal”* e *“Germany”* possuem a qualidade do ar entre as maiores, em antônimo quando analisado a média da poluição da água, percebeu-se que entre os países que possuem uma maior poluição da água é *“Philippines”* e o que possui a menor média desse mesmo valor é *“United Kingdom”*. Tendo esses pontos como exemplo, foi observado que os países que têm a poluição média da água que supera a qualidade média do ar, são os que estão ou próximo ou entre os

países com a maior poluição da água entre todos os que apareceram nesse gráfico, ou seja os mais frequentes no banco.

Por fim, a conclusão que obtivemos é que, entre os países mais frequentes no banco, geralmente há um maior nível da qualidade do ar do que a poluição da água, e aqueles que quebram esse padrão, estão entre os maiores níveis de poluição da água observados. A título de reforço, essas análise tem como base os países mais frequentes no banco de dados, uma vez que isso evidencia melhor a média geral real desses países, pois eles possuem mais dados coletados, então essa análise não evidencia a realidade de todas as médias, pois nenhuma das duas médias foi usada como parâmetro para o critério de agrupamento, o qual é necessário ser utilizado, pois se não torna-se impossível analisar o gráfico.