Data Science and Statistical Computing

Global Solution – FIAP – 2ESPV

Nome: Mateus Iago Sousa Conceição

1. Introdução

A poluição da água é um problema ambiental crítico que afeta ecossistemas, saúde pública e a disponibilidade de recursos hídricos. Neste trabalho de ciência de dados, utilizamos um banco de dados sobre a qualidade da água para investigar os níveis de poluição e suas principais fontes. A análise dos dados permitiu identificar padrões e tendências que podem ser fundamentais para a formulação de estratégias de mitigação e políticas de conservação.

A abordagem metodológica envolveu a análise dos dados fornecidos, que incluem a região e o nível de poluição da água, medido numa escala de 0 (sem poluentes) a 100 (extremamente poluída). Embora o banco de dados não seja muito completo, esses indicadores foram suficientes para identificar as regiões mais críticas e entender a distribuição da poluição hídrica. A análise estatística básica e a visualização de dados nos permitiram mapear os hotspots de poluição e inferir possíveis causas e consequências a partir dos padrões observados.

Os resultados obtidos destacam pontos com níveis significativos de poluição, permitindo uma análise mais detalhada das possíveis causas e consequências. Essa investigação inicial possibilitará estudos futuros que poderão explorar fatores culturais, econômicos e outros que contribuem para a poluição da água. Com base nesses insights, será possível desenvolver estratégias mais eficazes para mitigar o problema e promover a sustentabilidade dos recursos hídricos.

1.1 Descrição da base de dados

Variáveis Categóricas

- City (Cidade)
- Region (Região)
- Country (País)

Variáveis numéricas contínuas

- WaterPollution (Poluição da Água)
- AirQuality (Qualidade do Ar)

Na base de dados foi atribuído um valor de 0 (sem poluição) a 100 (extremamente poluído) para qualificar a qualidade da água de determinada região em um país

Logo de início retiramos do nosso data frame principal a coluna AirQuality pois ela não teria utilidade nas análises referente a poluição da água

```
df_cities = df_cities = df_cities.drop(columns=['AirQuality'])
display(df_cities.head(5), df_cities.shape)
```

	City	Region	Country	WaterPollution
0	New York City	New York	United States of America	49.504950
1	Washington, D.C.	District of Columbia	United States of America	49.107143
2	San Francisco	California	United States of America	43.000000
3	Berlin	NaN	Germany	28.612717
4	Los Angeles	California	United States of America	61.299435

```
(3963, 4)
```

O data frame é composto por 3963 linhas e 4 colunas sendo elas:

City – Essa coluna diz respeito a qual cidade foram extraídos os dados (str)

Region – Essa coluna mostra em qual região do país foram extraídos os dados (str)

Country – Essa coluna mostra em qual país foram extraídos os dados (str)

WaterPollution – Essa coluna qualifica a qualidade da água do lugar (float)

Identificando valores nulos e os tipos das colunas

Pudemos identificar a ocorrência de 425 valores nulos na coluna “Region” além disso verificamos os tipos e algumas informações básicas de cada coluna

```python
# Contar valores nulos em cada coluna
null_counts = df_cities.isnull().sum()
print(null_counts)
```

```
City                0
Region            425
Country             0
WaterPollution      0
dtype: int64
```

```python
print(df_cities.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3963 entries, 0 to 3962
Data columns (total 4 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   WaterPollution  3963 non-null   float64
dtypes: float64(1), object(3)
memory usage: 124.0+ KB
None
```

Métricas básicas do data frame no geral

	count	mean	std	min	25%	50%	75%	max
WaterPollution	3963.0	44.635372	25.66391	0.0	25.0	50.0	57.719393	100.0

```python
#Descobrindo em quantos paises foram colhidos dados para o dataset. No caso, 177 paises diferentes
print(df_cities['Country'].nunique())
#Descobrindo em quantas regioes foram colhidos dados para o dataset. No caso, 1152 regiões distintas
print(df_cities['Region'].nunique())
##Descobrindo em quantas cidades foram colhidos dados para o dataset. No caso, 3796 regiões distintas
print(df_cities['City'].nunique())
```

```
177
1152
3796
```

Número de observações de cada país

No presente trabalho, a análise de um conjunto de dados com 3.963 linhas revelou uma concentração significativa de informações em determinados países. Aproximadamente 51% das linhas se referem a apenas 10 países, enquanto 81% estão concentradas em 40 dos 172 países registrados. Essa disparidade na distribuição dos dados por país representa um desafio para a análise precisa das médias de poluição.

A comparação direta entre países com centenas de observações e aqueles com apenas algumas pode levar a interpretações enviesadas e distorcer a visão geral da poluição global. No entanto, essa concentração de dados também oferece oportunidades para análises mais aprofundadas em países com maior número de observações, permitindo identificar padrões e tendências com maior granularidade.

Top 10 paises com maiores observações

```
df_cities_counts = pd.DataFrame(df_cities['Country'].value_counts())
df_cities_counts.head(10)
```

Country	count
United States of America	842
People's Republic of China	238
United Kingdom	170
Canada	157
India	154
Germany	124
Brazil	103
Poland	94
Russia	86
Spain	78

```
porcentagem_country = (df_cities_counts.head(10).sum()/df_cities_counts.sum())*100
porcentagem_country
```

```
count    51.627555
dtype: float64
```

2. Análise dos dados dos paises

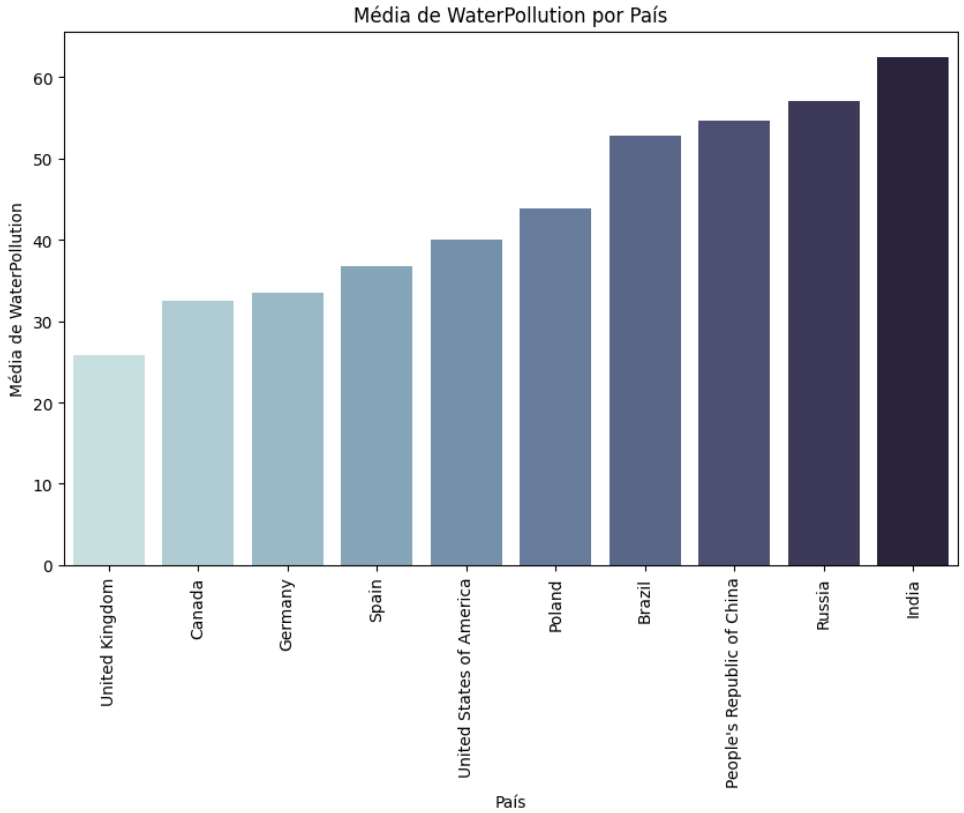
Para a primeira analise do nosso dataframe, escolhemos os 10 paises com maior

Numero de observações para compararmos, para isso foi necessarior criar um outro dataframe com os paises filtrados e a partir dai tirar as médias.

```
country_top_10 = ["United States of America", "People's Republic of China", "United Kingdom", "Canada", "India", "Germany", "Brazil", "Poland", "Russia",
df_filtrado = df_cities[df_cities['Country'].isin(country_top_10)]
df_filtrado
```

```
media_water_pollution = df_filtrado.groupby('Country')['WaterPollution'].mean().reset_index()
media_water_pollution = media_water_pollution.sort_values(by='WaterPollution')
media_water_pollution
```

	Country	WaterPollution
8	United Kingdom	25.791133
1	Canada	32.551674
2	Germany	33.495474
7	Spain	36.747945
9	United States of America	40.052397
5	Poland	43.860232
0	Brazil	52.747595
4	People's Republic of China	54.617488
6	Russia	57.089550
3	India	62.491712



O presente gráfico ilustra a média da poluição da água em dez países com maior incidência de registros na base de dados (média de WaterPollution por País). Observa-se que a Índia, com 154 registros (quinta maior quantidade), apresenta média superior a 60, indicando níveis elevados de poluição da água. Em contrapartida, o Reino Unido ostenta a menor média entre os dez países, além de apresentar o maior número de registros, com 170 (terceira maior quantidade). O número elevado e próximo de registros em cada país sugere a representatividade dos dados e a ausência de vieses.

Essa disparidade significativa na qualidade da água entre os dois países provavelmente se origina de fatores culturais e socioeconômicos. O Reino Unido, país desenvolvido com cerca de 70 milhões de habitantes, contrasta com a Índia, país emergente considerado de baixa renda, com cerca de 1,5 bilhão de habitantes, vinte vezes mais. Essa diferença nos perfis socioeconômicos dos países pode explicar, ao menos em parte, a disparidade na qualidade da água.

Nota-se, ainda, que o Brasil ocupa a quarta posição em termos de média de poluição da água, superando, inclusive, países comumente considerados mais poluídos, como os Estados Unidos da América. Essa realidade preocupante se encontra pouco abaixo da China, apesar de o Brasil ser um país com menor extensão territorial, população significativamente inferior e menor grau de concentração urbana.

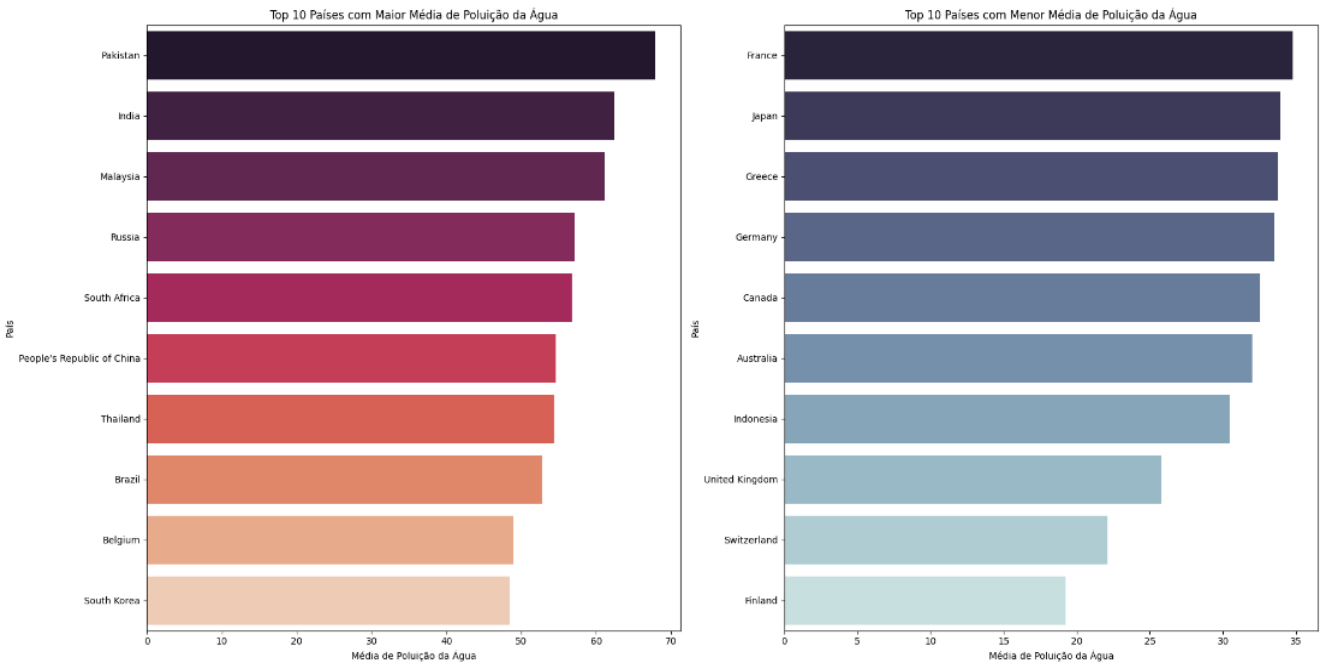
Essa realidade reflete, em parte, a cultura do país e o descaso com práticas de reciclagem e manejo adequado de resíduos. No Brasil, é comum observarmos lixo nas ruas e lançamento irregular em cursos d'água, como o caso emblemático do Rio Tietê, que outrora ostentava águas cristalinas e hoje é considerado o mais poluído do país.

É fundamental ressaltar que a análise da poluição da água no Brasil deve considerar diversos fatores interligados, como aspectos culturais, socioeconômicos, políticos e de gestão ambiental. Abordar essa questão de forma abrangente e propositiva é crucial para a busca de soluções eficazes que garantam a qualidade da água para as futuras gerações.

Top 40

Na próxima análise, selecionamos os 40 países com maior número de observações registradas no conjunto de dados. O critério de seleção para o Top 40 foi a presença de pelo menos 25 observações por país**.** A escolha de um mínimo de 25 observações visa garantir a representatividade dos dados e minimizar vieses nos resultados. Essa seleção garante maior confiabilidade nos resultados e permite uma análise mais aprofundada dos dados.

```
country_top_40 = ["United States of America", "People's Republic of China", "United Kingdom", "Canada", "India", "Germany", "Brazil", "Poland", "Russia",
df_filtrado_top_40 = df_cities[df_cities['Country'].isin(country_top_40)]
df_mean_Country = df_filtrado_top_40.pivot_table(values=['WaterPollution'], index=['Country'], aggfunc=['mean','median'])
df_mean_Country = df_mean_Country.sort_values(by=('mean', 'WaterPollution'), ascending=False)
df_mean_Country
```



Dos 40 principais países analisados, podemos dividi-los em dois grupos: os 10 países com os maiores índices de poluição da água e os 10 com os menores índices. É evidente uma divisão baseada no nível de desenvolvimento dos países. No grupo dos mais poluídos, predominam os países emergentes, incluindo todos os cinco países originais do BRICS (Brasil, Rússia, Índia, China e África do sul). Em contraste, no grupo dos menos poluídos, encontramos predominantemente países de primeiro mundo, que há muito tempo são considerados potências econômicas.

2.1 Análise dos dados do Brasil

Para começar a analisar os dados do Brasil foi necessario reunir todas as observações em um data frame só

```
df_brasil = df_cities[df_cities['Country'].isin(['Brazil'])]
display(df_brasil.head(5), df_brasil.shape)
```

	City	Region	Country	WaterPollution
16	Sao Paulo	Sao Paulo	Brazil	73.717949
246	Brasilia	Federal District	Brazil	38.043478
390	Osasco	Sao Paulo	Brazil	66.666667
399	Curitiba	Parana	Brazil	48.913043
511	Rio de Janeiro	Rio de Janeiro	Brazil	77.241379

```
(103, 4)
```

Podemos já perceber que o Brasil tem 103 linhas das 3963 totais da base de dados original. E essas observações foram dividias em varios estados e varias regiões do Brasil.

```
df_brazil_counts = pd.DataFrame(df_brasil['Region'].value_counts())
df_brazil_counts.head(22)
```

Region	count
Sao Paulo	32
Rio de Janeiro	14
Rio Grande do Sul	10
Parana	8
Minas Gerais	7
Santa Catarina	6
Pernambuco	4
Espirito Santo	4
Bahia	3
Mato Grosso do Sul	2
Piaui	2
Paraiba	1
Roraima	1
Rondonia	1
Mato Grosso	1
Goias	1
Rio Grande do Norte	1
Federal District	1
Ceara	1
Amazonas	1
Maranhao	1
Para	1

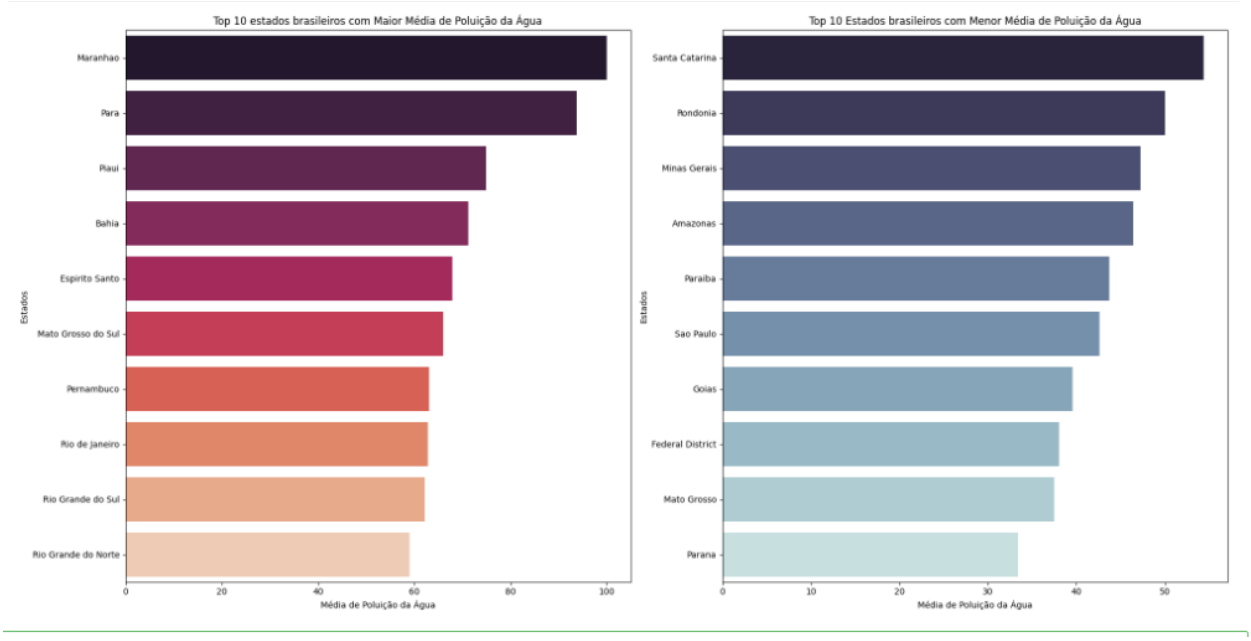
A partir dessa tabela, podemos observar que as 103 observações referentes ao Brasil estão mal distribuídas entre os estados, resultando em uma amostragem enviesada. Estados como Maranhão, Pará e Piauí possuem apenas uma observação cada, enquanto São Paulo conta com 34 observações e o Rio de Janeiro com 14. Essa

disparidade pode comprometer a representatividade dos dados e afetar a precisão das análises, evidenciando a necessidade de uma distribuição mais equilibrada das observações entre os estados.

Region	mean WaterPollution
Maranhao	100.000000
Para	93.750000
Piaui	75.000000
Bahia	71.271930
Espirito Santo	67.916667
Mato Grosso do Sul	66.071429
Pernambuco	63.104839
Rio de Janeiro	62.822436
Rio Grande do Sul	62.175926
Rio Grande do Norte	59.090909
Roraima	58.333333
Ceara	56.944444
Santa Catarina	54.369139
Rondonia	50.000000
Minas Gerais	47.255851
Amazonas	46.428571
Paraiba	43.750000
Sao Paulo	42.625249
Goias	39.583333
Federal District	38.043478
Mato Grosso	37.500000
Parana	33.405797

Considerando essa tabela, poderíamos concluir que as águas do Maranhão são completamente insalubres, enquanto as do Mato Grosso são mais limpas do que as de países altamente desenvolvidos. Embora essas informações possam ser verdadeiras, é essencial ter mais de uma observação para obter resultados realmente representativos

e confiáveis. Uma única observação não é suficiente para fazer generalizações precisas sobre a qualidade da água em uma região.



Dado o baixo número de observações na maioria dos estados do Brasil, é difícil tirar conclusões precisas e evitar análises enviesadas. No entanto, algo notável neste gráfico é a presença de São Paulo, que possui o maior número de observações no Brasil, entre os estados menos poluídos. São Paulo, sendo a capital econômica do país e o local com maior concentração de pessoas, reflete uma diversidade cultural significativa. Essa observação sugere que, mesmo em um ambiente densamente povoado e diversificado, pode haver iniciativas eficazes de controle de poluição e gestão ambiental.

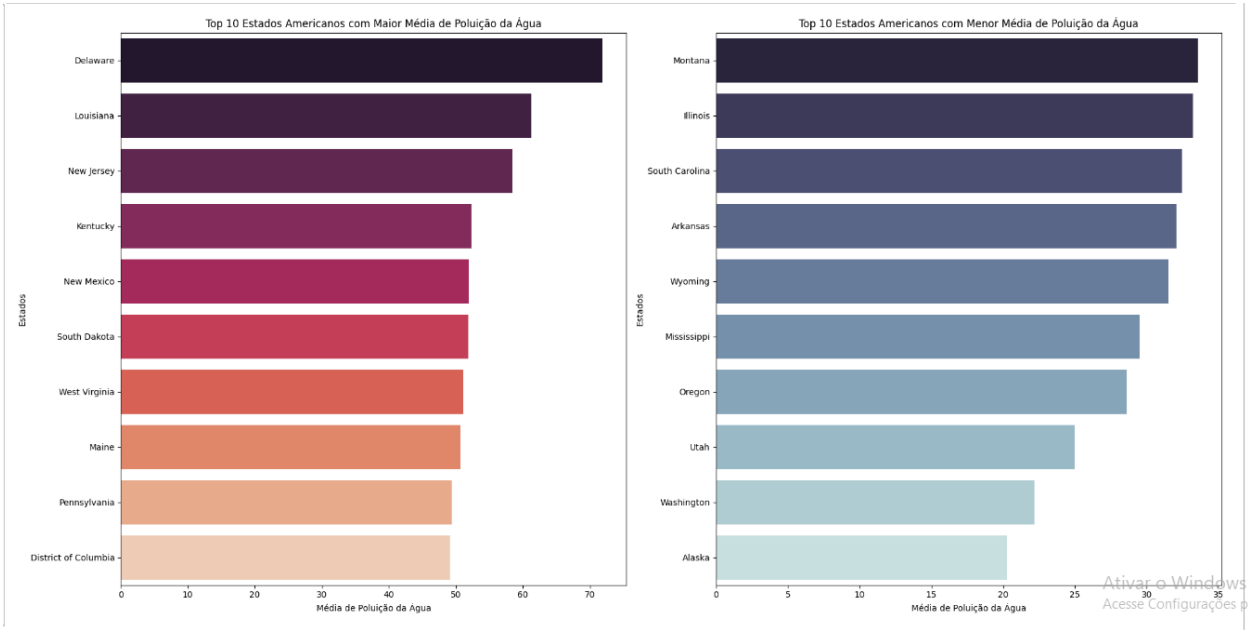
2.2 Análise dos dados dos Estados Unidos

Para a análise dos dados referentes aos Estados Unidos, utilizamos o mesmo processo de separação dos dados. Os Estados Unidos foram escolhidos para esta análise por serem o país com o maior número de observações (842) e por representarem a maior economia global. Além disso, a cultura consumista predominante no país gera uma quantidade significativa de resíduos. Isso levanta a questão: as políticas de limpeza e gestão de resíduos são eficientes o suficiente para reduzir os níveis de poluição da água?

	Region
California	122
Texas	51
Florida	48
Georgia	31
New York	31
Washington	29
North Carolina	27
Indiana	26
Illinois	26
Michigan	24
Massachusetts	22
Ohio	19
Tennessee	18
Oregon	17
Pennsylvania	17
Arizona	16
South Carolina	16
Virginia	16
Wisconsin	15
Minnesota	14
Connecticut	14
Colorado	13
Kentucky	13
Missouri	13
New Jersey	13
Arkansas	13
West Virginia	11
Iowa	11
Maryland	11
Kansas	10
Mississippi	10
Utah	9
New Hampshire	9
Louisiana	8
Maine	8
Alaska	8
Montana	8
Oklahoma	7
Wyoming	7
Idaho	7
North Dakota	7
Nevada	6
New Mexico	6
Vermont	6
Alabama	6
Nebraska	6
South Dakota	6
Rhode Island	5
Delaware	4
District of Columbia	1

Podemos observar que, entre os 50 estados americanos, apenas o “District of Columbia” possui menos de quatro observações. Isso confere uma maior confiabilidade aos dados analisados, tornando os resultados mais precisos e representativos.

Region	WaterPollution
Delaware	71.875000
Louisiana	61.268601
New Jersey	58.461538
Kentucky	52.332938
New Mexico	51.945930
South Dakota	51.884921
West Virginia	51.136364
Maine	50.669643
Pennsylvania	49.378071
District of Columbia	49.107143
North Dakota	49.107143
Alabama	48.482143
Tennessee	47.567061
Rhode Island	46.923077
Idaho	46.309524
Oklahoma	45.766733
Iowa	45.324675
Nevada	43.922919
Indiana	43.738674
Ohio	42.727135
Michigan	42.361524
Massachusetts	42.263018
Texas	42.240131
Virginia	42.219762
North Carolina	41.965326
Hawaii	41.517857
New York	40.971462
Florida	39.886222
Missouri	39.851530
California	39.792043
Nebraska	39.142157
Georgia	39.096516
Maryland	38.532060
Minnesota	38.171769
Connecticut	38.005952
Kansas	36.715686
Vermont	36.250000
Wisconsin	34.463240
Colorado	34.247940
Arizona	34.223560
New Hampshire	33.765432
Montana	33.559028
Illinois	33.216180
South Carolina	32.477783
Arkansas	32.081448
Wyoming	31.547619
Mississippi	29.523810
Oregon	28.629396
Utah	25.000000
Washington	22.184259



Analisando os resultados dos gráficos, podemos concluir que, apesar de sua forte cultura consumista, os Estados Unidos conseguem manter os níveis de poluição da água em um patamar aceitável, com uma média geral de 40.

3. Testes de hipótese

Analisando os dados, é possível levantar questionamentos que nos conduzem à formulação de testes de hipótese. Testes de hipótese são procedimentos estatísticos utilizados para avaliar a veracidade de afirmações sobre características de uma população com base em amostras dos dados observados. Em outras palavras, os testes de hipótese nos ajudam a tomar decisões sobre uma população com base em informações de uma amostra.

Ao realizar um teste de hipótese, formulamos duas hipóteses: a hipótese nula (H0) e a hipótese alternativa (H1). A hipótese nula representa a posição inicial ou a crença padrão, afirmando que não há efeito ou diferença significativa. Por outro lado, a hipótese alternativa desafia essa posição, sugerindo que há uma diferença ou efeito significativo.

Por meio do teste de hipótese, aplicamos técnicas estatísticas para calcular uma estatística de teste apropriada (por exemplo, teste t, teste de Kruskal-Wallis, teste de qui-quadrado, etc.) com base nos dados amostrais disponíveis. Comparamos então o valor da estatística de teste com um valor crítico (ou intervalo crítico) determinado pela distribuição de probabilidade subjacente (por exemplo, distribuição t de Student, distribuição qui-quadrado, etc.). Com base nessa comparação, podemos decidir se rejeitamos ou não a hipótese nula.

Em resumo, os testes de hipótese fornecem um método sistemático e objetivo para fazer inferências sobre populações com base em dados amostrais, permitindo-nos tirar conclusões sobre o mundo com base em evidências estatísticas.

Hipótese sobre a Poluição da Água e Desenvolvimento Socioeconômico:

H0 (Hipótese Nula): Não há diferença significativa nos níveis de poluição da água entre países desenvolvidos (norte do mapa) e países emergentes (sul do mapa).

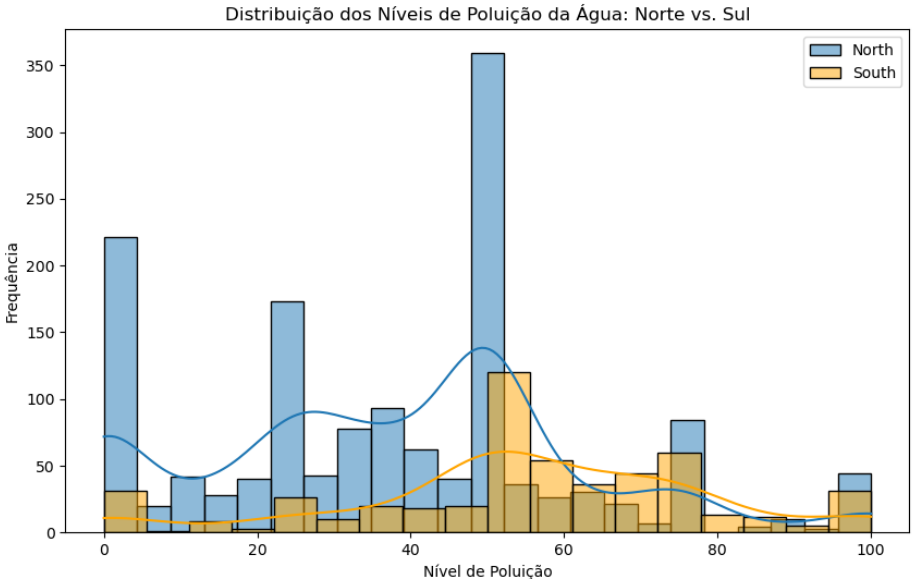
H1 (Hipótese Alternativa): Há uma diferença significativa nos níveis de poluição da água entre países desenvolvidos (norte do mapa) e países emergentes (sul do mapa)

Método: Teste t de Student para comparar as médias dos níveis de poluição da água entre os grupos de países.

Para conduzir este teste, inicialmente, dividimos os países listados no top 40, com um mínimo de 25 observações, com base em seus hemisférios geográficos, resultando em dois grupos distintos: norte sul.

In [29]:
```
#Carregando o dataset novamente
df = pd.read_csv('C:/Users/Mateus Iago/Desktop/GLOBAL SOLUTION/GS - DATA_SCIENCE/Cities1.csv')

#Separando os paises do norte e do sul do mundo
northern_countries = ['Canada', 'United Kingdom', 'Germany', 'France', 'United States of America','Japan','Russia']
southern_countries = ['Brazil', 'India', 'South Africa', 'Indonesia', 'Malaysia', 'Pakistan','Thailand','Chile','Iran']

#Criando a coluna 'hemisphere' no  dataset original
df['hemisphere'] = df['Country'].apply(lambda x: 'North' if x in northern_countries else ('South' if x in southern_countries else

print("Distribuição dos dados após classificação:")
print(df['hemisphere'].value_counts())

#Filtando o dataset a partir de um data frame somente com os paises do hemisferio sul e norte selecionados
df_hemisphere = df[df['hemisphere'].isin(['North', 'South'])]

#Separando em dois grupos
north_data = df_hemisphere[df_hemisphere['hemisphere'] == 'North']['WaterPollution']
south_data = df_hemisphere[df_hemisphere['hemisphere'] == 'South']['WaterPollution']
```

```
Distribuição dos dados após classificação:
Other    1987
North    1464
South     512
Name: hemisphere, dtype: int64
```



após a separação devemos realizar o teste t para então saber se devemos ou não rejeitar a hipótese nula.

```
# Realizar o teste t de Student
t_stat, p_value = ttest_ind(north_data, south_data, equal_var=False)

# Interpretação dos resultados
alpha = 0.05
if p_value < alpha:
    print("Rejeitamos a hipótese nula: Há uma diferença significativa nos níveis de poluição da água entre os países do Norte e d
else:
    print("Falhamos em rejeitar a hipótese nula: Não há uma diferença significativa nos níveis de poluição da água entre os países
```

```
Rejeitamos a hipótese nula: Há uma diferença significativa nos níveis de poluição da água entre os países do Norte e do Sul.
```

Para esse teste devemos rejeitar a hipótese nula já que existe uma diferença significativa nos níveis de poluição nas águas dos hemisférios.

Uma outra observação relevante foi a disparidade no número de observações entre os países. Esse aspecto nos motiva a desenvolver mais testes estatísticos para explorar essa diferença.

Hipótese sobre Desigualdade na Distribuição das Observações:

H0 (Hipótese Nula): A distribuição das observações entre os estados não é significativamente diferente.

H1 (Hipótese Alternativa): A distribuição das observações entre os estados é significativamente diferente.

Método: Para testar a hipótese sobre a desigualdade na distribuição das observações, foi utilizado o teste estatístico de Kruskal-Wallis

```
#Hipótese sobre Desigualdade na Distribuição das Observações

#H0 (Hipótese Nula): A distribuição das observações entre os Paises não é significativamente diferente.
#H1 (Hipótese Alternativa): A distribuição das observações entre os estados é significativamente diferente.

from scipy.stats import kruskal

#Carregando o dataset novamente
df = pd.read_csv('C:/Users/Mateus Iago/Desktop/GLOBAL SOLUTION/GS - DATA_SCIENCE/Cities1.csv')

# Realizar o teste de Kruskal-Wallis para comparar as distribuições das observações de poluição da água entre os diferentes paise
statistic, p_value = kruskal(*[group['WaterPollution'] for name, group in df.groupby('Country')])

# Exibir o resultado do teste
print("Resultado do Teste de Kruskal-Wallis:")
print(f"Estatística do Teste: {statistic}")
print(f"Valor-p (p-value): {p_value}")

# Interpretar os resultados
alpha = 0.05  # Nível de significância
if p_value < alpha:
    print("Rejeitamos a hipótese nula.")
    print("Há evidências estatísticas suficientes para suportar a hipótese alternativa.")
else:
    print("Não rejeitamos a hipótese nula.")
    print("Não há evidências estatísticas suficientes para suportar a hipótese alternativa.")
```

```
Resultado do Teste de Kruskal-Wallis:
Estatística do Teste: 1035.2945103979266
Valor-p (p-value): 1.1627683913927116e-121
Rejeitamos a hipótese nula.
Há evidências estatísticas suficientes para suportar a hipótese alternativa.
```

Com a análise dos resultados, concluímos que devemos rejeitar a hipótese nula, pois existem evidências suficientes para sustentar a hipótese alternativa.

4. Conclusão

É impressionante como a análise de uma única coluna sobre poluição da água em uma base de dados pode revelar tanto sobre o mundo em que vivemos. A partir dessa análise, torna-se evidente a desigualdade social que permeia o planeta, confirmando o que vemos nos noticiários e estudamos nas escolas. Através do estudo deste database, foi possível compreender melhor o famoso mapa da desigualdade social, onde a parte norte concentra os países mais desenvolvidos e a parte sul, os países emergentes. Embora o fator socioeconômico seja um dos principais contribuintes para a poluição das águas, inúmeros outros fatores também desempenham um papel significativo.

Fatores culturais e geográficos também influenciam significativamente nos índices de poluição da água. Espero que este documento permita uma melhor compreensão de alguns dos motivos por trás da poluição e que possa fornecer insights valiosos para a criação de medidas eficazes no combate a esse problema.