DATA CLEANING

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

water_pollution = pd.read_csv("/Users/tonydao/Documents/PythonProjects/PollutionProject/water_pollution_disease.csv")

water_pollution.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 24 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   Country                                     3000 non-null   object
 1   Region                                      3000 non-null   object
 2   Year                                        3000 non-null   int64
 3   Water Source Type                           3000 non-null   object
 4   Contaminant Level (ppm)                     3000 non-null   float64
 5   pH Level                                    3000 non-null   float64
 6   Turbidity (NTU)                             3000 non-null   float64
 7   Dissolved Oxygen (mg/L)                     3000 non-null   float64
 8   Nitrate Level (mg/L)                        3000 non-null   float64
 9   Lead Concentration (µg/L)                   3000 non-null   float64
 10  Bacteria Count (CFU/mL)                     3000 non-null   int64
 11  Water Treatment Method                      2253 non-null   object
 12  Access to Clean Water (% of Population)     3000 non-null   float64
 13  Diarrheal Cases per 100,000 people          3000 non-null   int64
 14  Cholera Cases per 100,000 people            3000 non-null   int64
 15  Typhoid Cases per 100,000 people            3000 non-null   int64
 16  Infant Mortality Rate (per 1,000 live births)  3000 non-null   float64
 17  GDP per Capita (USD)                        3000 non-null   int64
 18  Healthcare Access Index (0-100)             3000 non-null   float64
 19  Urbanization Rate (%)                       3000 non-null   float64
 20  Sanitation Coverage (% of Population)       3000 non-null   float64
 21  Rainfall (mm per year)                      3000 non-null   int64
 22  Temperature (°C)                            3000 non-null   float64
 23  Population Density (people per km²)         3000 non-null   int64
dtypes: float64(12), int64(8), object(4)
memory usage: 562.6+ KB
```

```python
print(water_pollution.isnull().sum())
```

```
Country                                        0
Region                                         0
Year                                           0
Water Source Type                              0
Contaminant Level (ppm)                        0
pH Level                                       0
Turbidity (NTU)                                0
Dissolved Oxygen (mg/L)                        0
Nitrate Level (mg/L)                           0
Lead Concentration (µg/L)                      0
Bacteria Count (CFU/mL)                        0
Water Treatment Method                       747
Access to Clean Water (% of Population)        0
Diarrheal Cases per 100,000 people             0
Cholera Cases per 100,000 people               0
Typhoid Cases per 100,000 people               0
Infant Mortality Rate (per 1,000 live births)  0
GDP per Capita (USD)                           0
Healthcare Access Index (0-100)                0
Urbanization Rate (%)                          0
Sanitation Coverage (% of Population)          0
Rainfall (mm per year)                         0
Temperature (°C)                               0
Population Density (people per km²)            0
dtype: int64
```

```python
print(water_pollution.duplicated().sum())
```

```
0
```

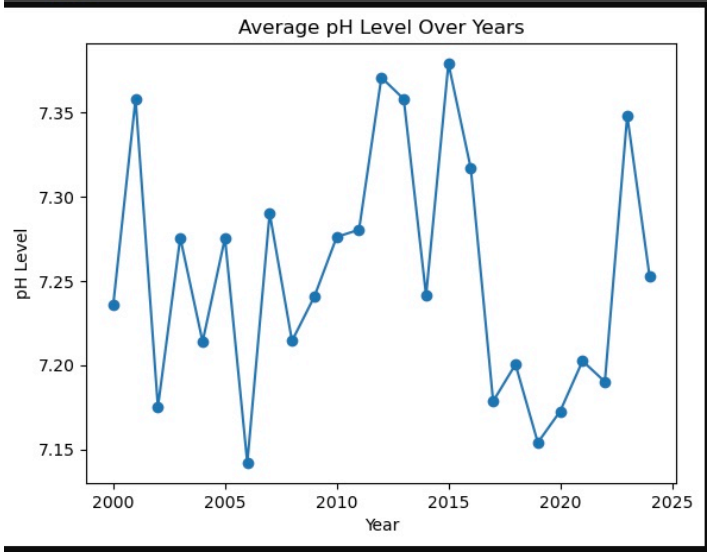
```python
#Clean column names (remove spaces, lowercase)
water_pollution["Country"] = water_pollution["Country"].str.strip().str.title()
water_pollution["Region"] = water_pollution["Region"].str.strip().str.title()
water_pollution["Water Source Type"] = water_pollution["Water Source Type"].str.strip().str.title()
```

```python
#Remove rows/columns with missing values:
water_pollution = water_pollution.dropna()
```

UNIVARIATE ANALYSIS

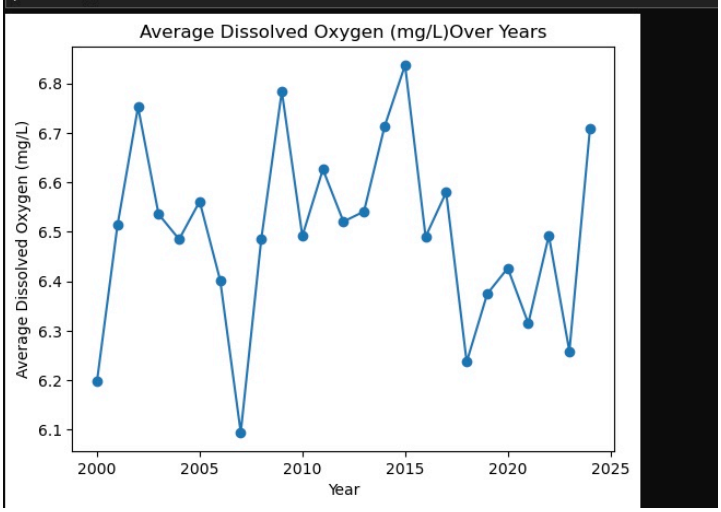
- Overall pH mean over the years.

```
water_pollution.groupby('Year')['pH Level'].mean().plot(marker='o')
plt.title('Average pH Level Over Years')
plt.ylabel('pH Level')
plt.show()
```



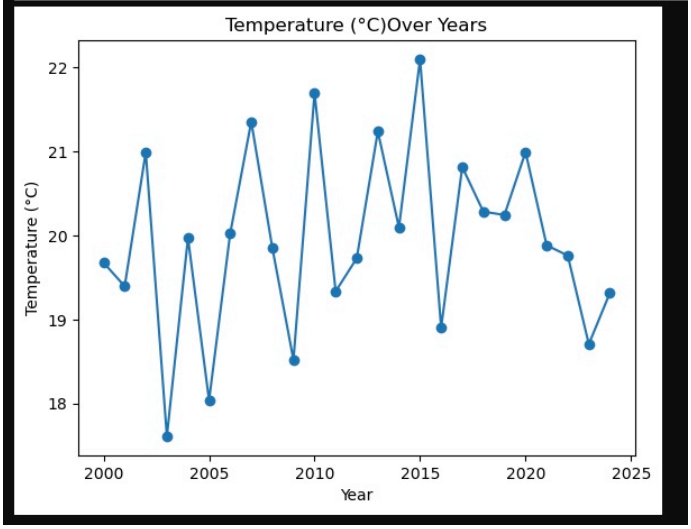
  - Slightly alkaline
- Average dissolved oxygen over the years

```
water_pollution.groupby('Year')['Dissolved Oxygen (mg/L)'].mean().plot(marker='o')
plt.title('Average Dissolved Oxygen (mg/L)Over Years')
plt.ylabel('Average Dissolved Oxygen (mg/L)')
plt.show()
```



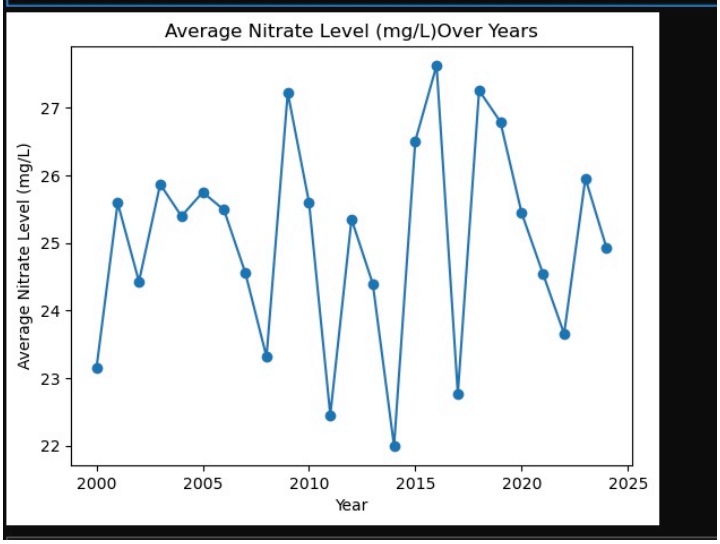
- Temperature (C) over the years

```
water_pollution.groupby('Year')['Temperature (°C)'].mean().plot(marker='o')
plt.title('Temperature (°C)Over Years')
plt.ylabel('Temperature (°C)')
plt.show()
```



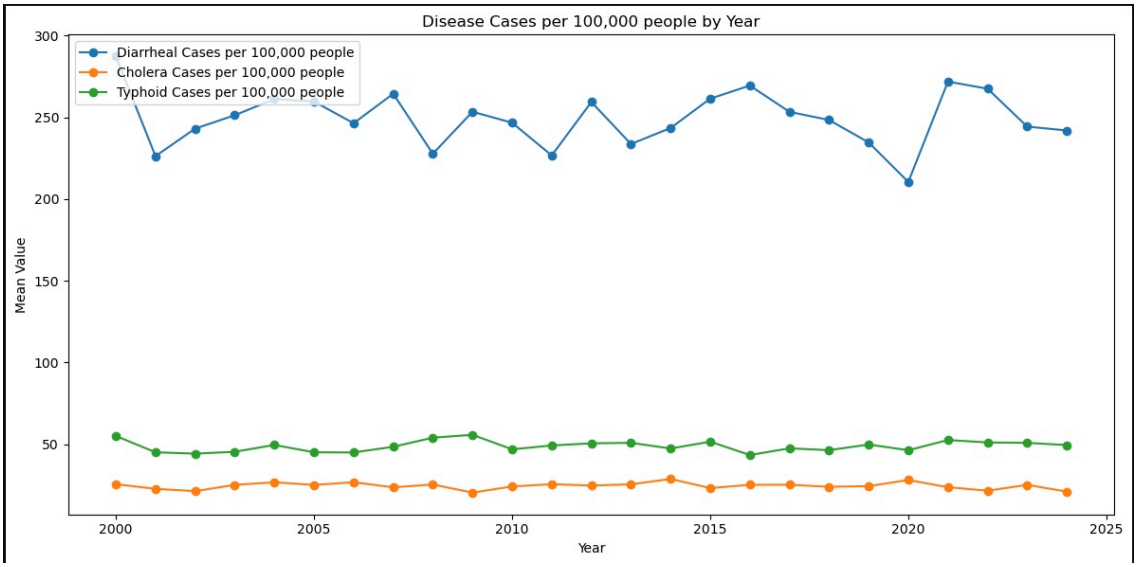
  - What could cause the temperature drop after 2015??
- Avg nitrate level (mg/L) over years

```
water_pollution.groupby('Year')['Nitrate Level (mg/L)'].mean().plot(marker='o')
plt.title('Average Nitrate Level (mg/L)Over Years')
plt.ylabel('Average Nitrate Level (mg/L)')
plt.show()
```



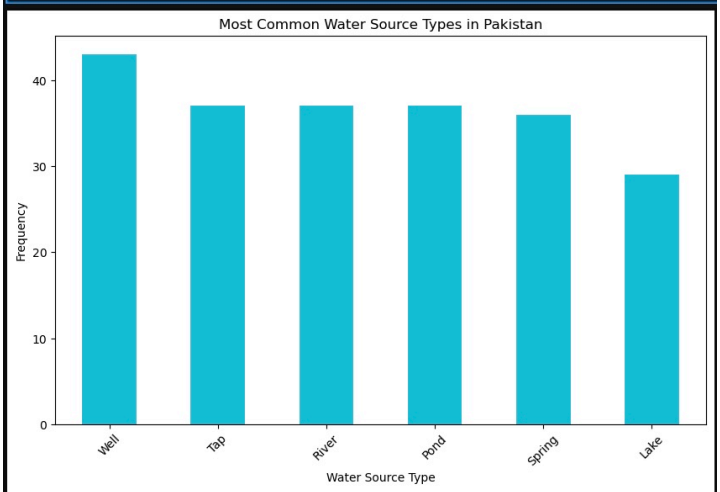
- Diseases cases per 100,000 over the years

```
cols = [
    'Diarrheal Cases per 100,000 people',
    'Cholera Cases per 100,000 people',
    'Typhoid Cases per 100,000 people',
]

water_pollution.groupby('Year')[cols].mean().plot(marker='o', figsize=(12, 6))
plt.title('Disease Cases per 100,000 people by Year')
plt.xlabel('Year')
plt.ylabel('Mean Value')

# Legend in upper left
plt.legend(loc='upper left')

plt.tight_layout()
plt.show()
```



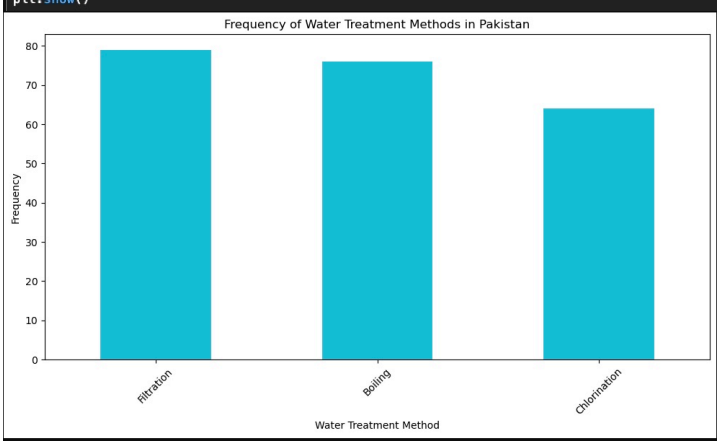
- Distribution of Water Source Types

```python
# Count the frequency of Water Source Types
water_source_counts = Pakistan_df['Water Source Type'].value_counts()

# Plot
plt.figure(figsize=(10,6))
water_source_counts.plot(kind='bar', color='#14bed6')
plt.title('Most Common Water Source Types in Pakistan')
plt.xlabel('Water Source Type')
plt.ylabel('Frequency')
plt.xticks(rotation=45)
plt.tight_layout
```



- Wells are the most frequently reported water type.
- Close counts for Tap, River, Pond, and Spring indicate that water is drawn from a variety of sources in Pakistan
- FUTURE ANALYSIS: Looking at water quality via indicators (ph, contaminant levels and more)

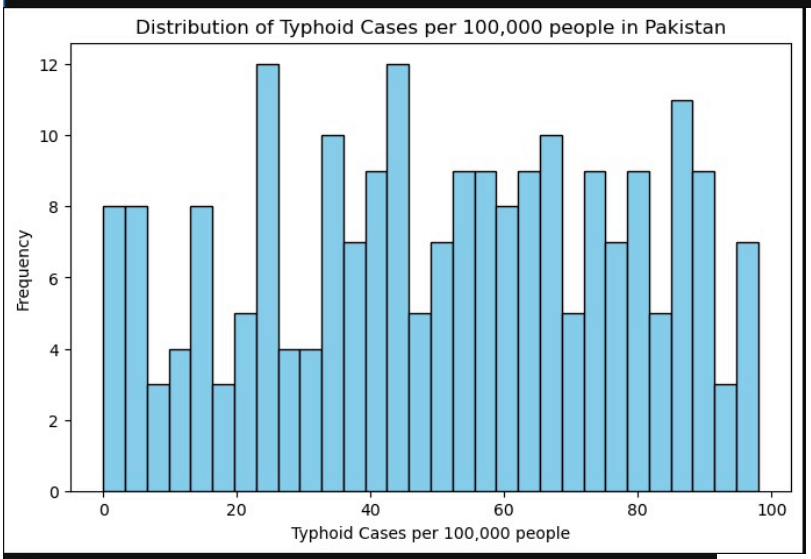
```python
# Count the frequency of Water Treatment Methods
water_treatment_counts = Pakistan_df['Water Treatment Method'].value_counts(dropna=False)

# Plot
plt.figure(figsize=(10,6))
water_treatment_counts.plot(kind='bar', color='#14bed6')
plt.title('Frequency of Water Treatment Methods in Pakistan')
plt.xlabel('Water Treatment Method')
plt.ylabel('Frequency')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



- Most common method would be filtration.
  - almost equally as common as boiling.

- Distribution of Typhoid Cases per 100,000 people in P

```
#Distribution of Typhoid Cases per 100,000 people in Pakistan
# Plot histogram
plt.figure(figsize=(8,5))
plt.hist(Pakistan_df['Typhoid Cases per 100,000 people'], bins=30, color='skyblue', edgecolor='black')
plt.title('Distribution of Typhoid Cases per 100,000 people in Pakistan')
plt.xlabel('Typhoid Cases per 100,000 people')
plt.ylabel('Frequency')
plt.show()
```



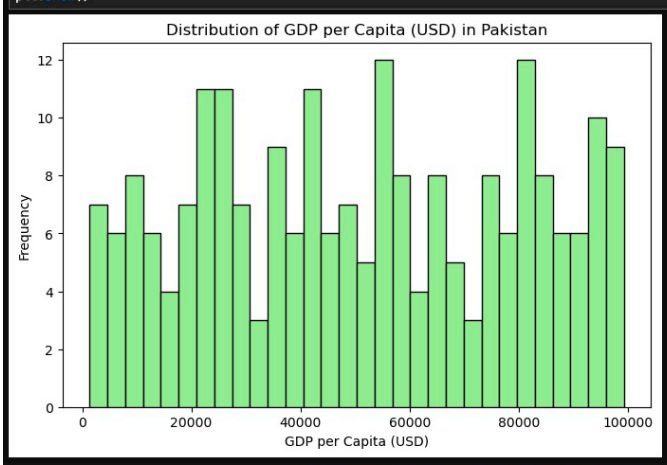

```
# Summary statistics
typhoid_stats = Pakistan_df['Typhoid Cases per 100,000 people'].describe()
print(typhoid_stats)
```

```
count    219.000000
mean      50.945205
std       27.152926
min        0.000000
25%       29.500000
50%       53.000000
75%       73.500000
max       98.000000
Name: Typhoid Cases per 100,000 people, dtype: float64
```

- Avg number of typhoid cases per 100,000 people is 51
- 27.15 STD indicates a wide spead of values around the mean.
- The highest recorded value is 98 cases per 100,000.
- 75% of the data is below the 73.5 cases per 100,000.
- mean and median are close (about 51 and 53), suggesting a roughly symmetrical distribution for central tendency.

- Reported values for GDP per capita (USD) in P

```
#Reported values for GDP per Capita (USD) in Pakistan
plt.figure(figsize=(8,5))
plt.hist(Pakistan_df['GDP per Capita (USD)'], bins=30, color='lightgreen', edgecolor='black')
plt.title('Distribution of GDP per Capita (USD) in Pakistan')
plt.xlabel('GDP per Capita (USD)')
plt.ylabel('Frequency')
plt.grid(False)
plt.show()
```



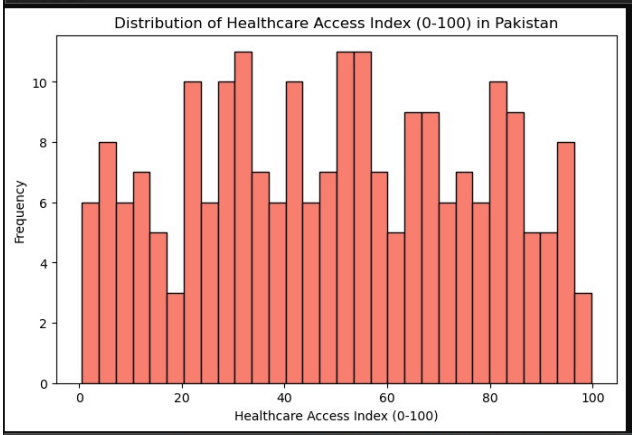

```
GDP_stats = Pakistan_df['GDP per Capita (USD)'].describe()
print(GDP_stats)
```

```
count      219.000000
mean     51138.926941
std      28484.647253
min       1327.000000
25%      25544.000000
50%      50759.000000
75%      77665.500000
max      99266.000000
Name: GDP per Capita (USD), dtype: float64
```

- Mean: $51,139
- STD: $28,485
  - large variability
- min: $1,327
- max: $99,266
- high standard deviation and range indicate significant inequality or diversity in the economic situation

- Values and distribution of healthcare access index HAI in P

```
#Values and distribution of the Healthcare Access Index (0-100) in Pakistan
plt.figure(figsize=(8,5))
plt.hist(Pakistan_df['Healthcare Access Index (0-100)'], bins=30, color='salmon', edgecolor='black')
plt.title('Distribution of Healthcare Access Index (0-100) in Pakistan')
plt.xlabel('Healthcare Access Index (0-100)')
plt.ylabel('Frequency')
plt.show()
```



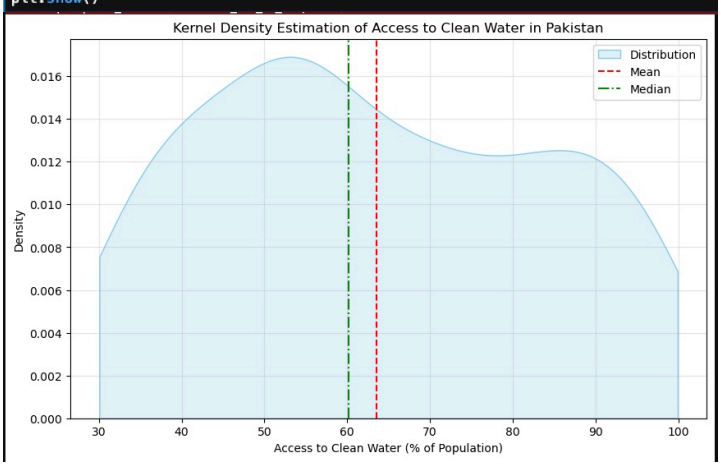

```
healthcare_stats = Pakistan_df['Healthcare Access Index (0-100)'].describe()
print(healthcare_stats)
```

```
count    219.000000
mean      49.778402
std       27.146189
min        0.550000
25%       28.230000
50%       50.880000
75%       72.320000
max       99.750000
Name: Healthcare Access Index (0-100), dtype: float64
```

- Mean: 49.8
  - Healthcare access is moderate
- STD: 27.1
  - High variability
- Min: 0.55
  - Almost no access!!!
- Max: 99.75
  - Extreme disparities
- Median is 50.88 and is near the mean thus the distribution is symmetric.
- **FUTURE Q: Does low healthcare access correlates with higher disease rates or lower GDP**.

- Kernal Density Estimation (KDE)

```
Kernel Density Estimation (KDE)
import seaborn as sns
plt.figure(figsize=(10, 6))
sns.kdeplot(Pakistan_df['Access to Clean Water (% of Population)'], fill=True,
            color='skyblue', cut=0, label='Distribution')
plt.axvline(Pakistan_df['Access to Clean Water (% of Population)'].mean(),
            color='red', linestyle='--', label='Mean')
plt.axvline(Pakistan_df['Access to Clean Water (% of Population)'].median(),
            color='green', linestyle='-.', label='Median')
plt.title('Kernel Density Estimation of Access to Clean Water in Pakistan')
plt.xlabel('Access to Clean Water (% of Population)')
plt.legend()
plt.grid(True, alpha=0.3)
plt.show()
```



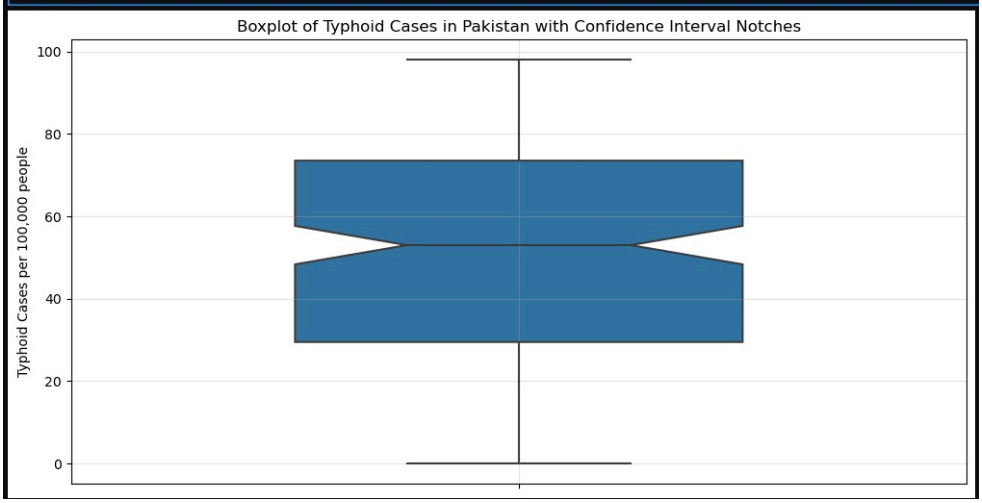

```
access_clean_water_stats =Pakistan_df['Access to Clean Water (% of Population)'].describe()
print(access_clean_water_stats)
```

```
count    219.000000
mean      63.498721
std       20.206259
min       30.060000
25%       46.720000
50%       60.120000
75%       81.710000
max       99.920000
```

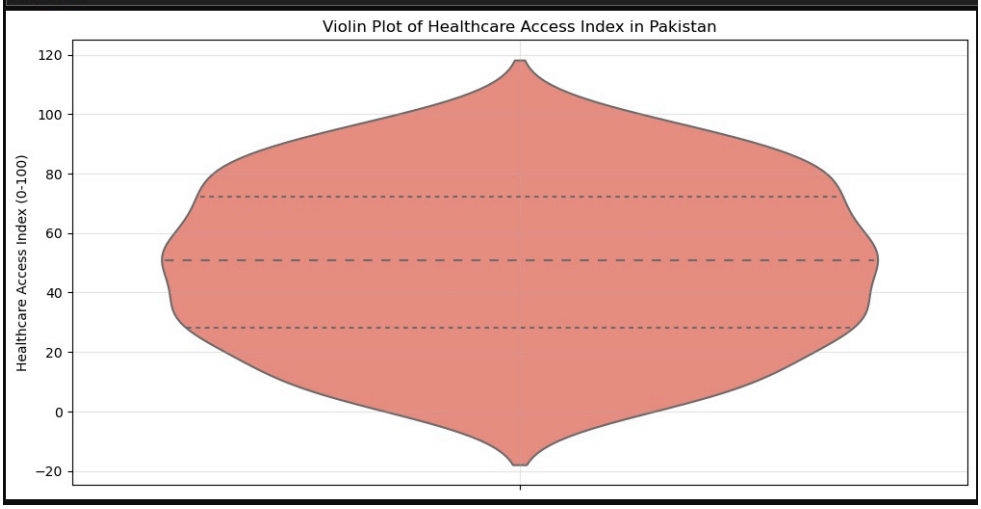
- Majorities of values between 30% and 100%
- Mean: 63.5%
- Median: 60.1%
  - Distribution not strongly skewed.
- STD: 20.2%
  - Wide spead
- Peak in KDE curve
  - where access to clean water is most common
- If the curve is high around 60-80%, that means many regions/years in Pakistan have access rates in that range.

- Using boxplot with CI notches.

```
#Box Plots with Enhanced Features
plt.figure(figsize=(12, 6))
sns.boxplot(y=Pakistan_df['Typhoid Cases per 100,000 people'],
            width=0.5, notch=True, showfliers=True)
plt.title('Boxplot of Typhoid Cases in Pakistan with Confidence Interval Notches')
plt.grid(True, alpha=0.3)
plt.show()
```




```
#Violin Plots
plt.figure(figsize=(12, 6))
sns.violinplot(y=Pakistan_df['Healthcare Access Index (0-100)'],
               inner='quartile', color='salmon')
plt.title('Violin Plot of Healthcare Access Index in Pakistan')
plt.grid(True, alpha=0.3)
plt.show()
```



  - Mulitmodal?