

Student Name: Abdulaziz Dafer Ali Alobud

Student Id : 253001366

Dataset link : <https://www.kaggle.com/datasets/imdevskp/corona-virus-report>

```

import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("covid_19_clean_complete.csv")

print("Original shape:", df.shape)

# 1) Handle missing values

num_cols = df.select_dtypes(include=[np.number]).columns
cat_cols = df.select_dtypes(exclude=[np.number]).columns

df[num_cols] = df[num_cols].fillna(df[num_cols].median())

for col in cat_cols:
    mode = df[col].mode()
    if not mode.empty:
        df[col] = df[col].fillna(mode[0])
    else:
        df[col] = df[col].fillna("Unknown")

print("Missing values after handling:")
print(df.isna().sum())

# Handle outliers (IQR)

def cap_outliers_iqr(series, factor=1.5):
    """IQR (Winsorization)."""
    q1 = series.quantile(0.25)
    q3 = series.quantile(0.75)
    iqr = q3 - q1
    lower = q1 - factor * iqr
    upper = q3 + factor * iqr
    return series.clip(lower, upper)

for col in ["Confirmed", "Deaths", "Recovered", "Active"]:
    if col in df.columns:
        df[col] = cap_outliers_iqr(df[col])

# 3) Data visualization

df["Date"] = pd.to_datetime(df["Date"], errors="coerce")

ts = (
    df.dropna(subset=["Date"])
    .groupby("Date")["Confirmed"]
    .sum()
    .reset_index()
    .sort_values("Date")
)

plt.figure(figsize=(10, 4))
plt.plot(ts["Date"], ts["Confirmed"])
plt.title("Global Confirmed COVID-19 Cases Over Time")
plt.xlabel("Date")
plt.ylabel("Confirmed Cases")
plt.tight_layout()
plt.show()

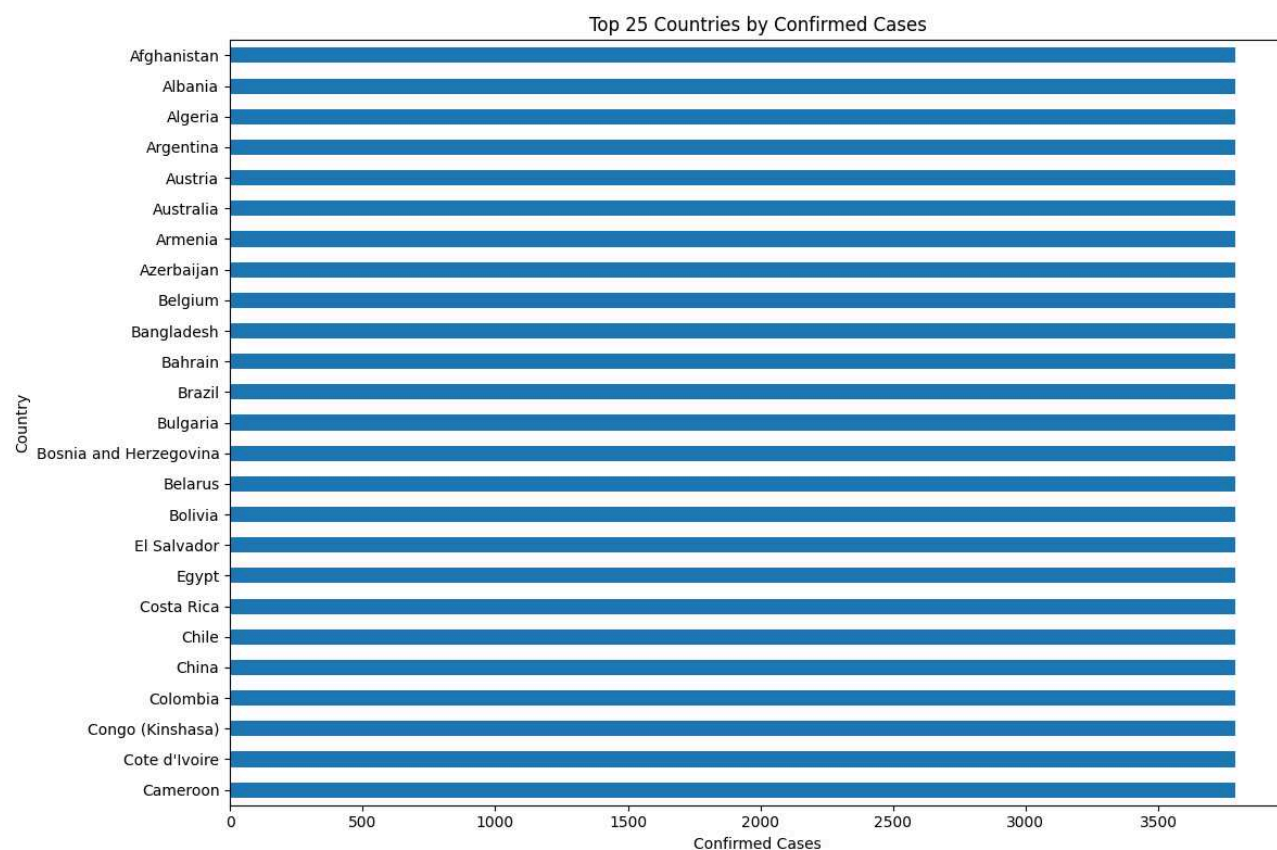
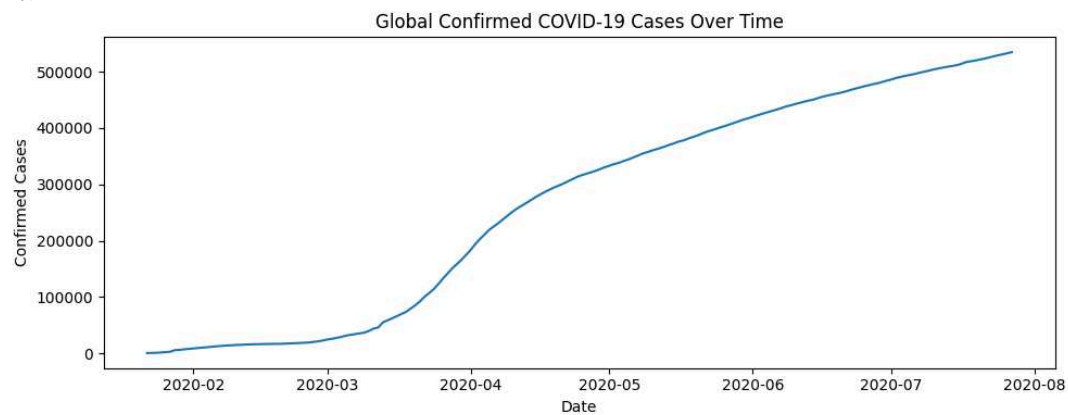
if "Country/Region" in df.columns:
    top25 = (
        df.groupby("Country/Region")["Confirmed"]
        .max()
        .sort_values(ascending=False)
        .head(25)
    )

    plt.figure(figsize=(12, 8))
    top25.plot(kind="barh")
    plt.title("Top 25 Countries by Confirmed Cases")
    plt.xlabel("Confirmed Cases")
    plt.ylabel("Country")
    plt.gca().invert_yaxis()
    plt.tight_layout()
    plt.show()

# outliers
plt.figure(figsize=(6, 4))
plt.boxplot(df["Confirmed"], vert=True, showfliers=True)
plt.title("Boxplot of Confirmed Cases After Outlier Handling")
plt.ylabel("Confirmed Cases")
plt.tight_layout()
plt.show()

```

```
Original shape: (49068, 10)
Missing values after handling:
Province/State    0
Country/Region    0
Lat               0
Long              0
Date              0
Confirmed         0
Deaths            0
Recovered         0
Active            0
WHO Region        0
dtype: int64
```



Double-click (or enter) to edit