covid-analysis-project

July 4, 2025

import numpy as np import pandas as pd

[19]:
```
df = pd.read_csv("Datasets/covid.csv")
df.head()
```

[19]:

	Country/Region	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	Deaths / 100 Recovered	Confirmed last week	1 week change	1 week % increase	WHO Region
0	Afghanistan	36263	1269	25198	9796	106	10	18	3.50	69.49	5.04	35526	737	2.07	Eastern Mediterranean
1	Albania	4880	144	2745	1991	117	6	63	2.95	56.25	5.25	4171	709	17.00	Europe
2	Algeria	27973	1163	18837	7973	616	8	749	4.16	67.34	6.17	23691	4282	18.07	Africa
3	Andorra	907	52	803	52	10	0	0	5.73	88.53	6.48	884	23	2.60	Europe
4	Angola	950	41	242	667	18	1	0	4.32	25.47	16.94	749	201	26.84	Africa

1 STEP 1: Data Understanding

[20]:
```
df.columns
```

[20]:
```
Index(['Country/Region', 'Confirmed', 'Deaths', 'Recovered', 'Active',
       'New cases', 'New deaths', 'New recovered', 'Deaths / 100 Cases',
       'Recovered / 100 Cases', 'Deaths / 100 Recovered',
       'Confirmed last week', '1 week change', '1 week % increase',
       'WHO Region'],
      dtype='object')
```

[21]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Country/Region          187 non-null    object
 1   Confirmed               187 non-null    int64
 2   Deaths                  187 non-null    int64
 3   Recovered               187 non-null    int64
 4   Active                  187 non-null    int64
 5   New cases               187 non-null    int64
 6   New deaths              187 non-null    int64
 7   New recovered           187 non-null    int64
 8   Deaths / 100 Cases      187 non-null    float64
 9   Recovered / 100 Cases   187 non-null    float64
 10  Deaths / 100 Recovered  187 non-null    float64
 11  Confirmed last week     187 non-null    int64
 12  1 week change           187 non-null    int64
 13  1 week % increase       187 non-null    float64
 14  WHO Region              187 non-null    object
dtypes: float64(4), int64(9), object(2)
memory usage: 22.0+ KB
```

2 STEP 2: Data Cleaning

Standardize Column Names

[23]:
```
#Remove spaces, make lowercase, use _ instead of spaces, Replace slashes with␣
 ↪underscores
df.columns = (
    df.columns
    .str.strip()
    .str.lower()
    .str.replace(' ', '_')
    .str.replace('/', '_')
```

```
    .str.replace('_+', '_', regex=True)
)
df.columns
```

[23]:
```
Index(['country_region', 'confirmed', 'deaths', 'recovered', 'active',
       'new_cases', 'new_deaths', 'new_recovered', 'deaths_100_cases',
       'recovered_100_cases', 'deaths_100_recovered', 'confirmed_last_week',
       '1_week_change', '1_week_%_increase', 'who_region'],
      dtype='object')
```

Check Duplicates

[24]:
```
df.duplicated().sum()
```

[24]:
```
np.int64(0)
```

[26]:
```
df['country_region'].nunique()
```

[26]:
```
187
```

[27]:
```
df['country_region'] = df['country_region'].str.strip()
```

Recalculate or Validate Key Columns Key Col here is active which should justify logic, active = confirmed - deaths - recovered

[29]:
```
df['check_active'] = df['confirmed'] - df['deaths'] - df['recovered']
```

[31]:
```
(df['active'] != df['check_active']).sum()
```

[31]:
```
np.int64(0)
```

[34]:
```
df.drop(columns='check_active', inplace=True)
```

Clean Text Columns

[36]:
```
#strip / remove white spaces from text columns
df['country_region'] = df['country_region'].str.strip()
df['who_region'] = df['who_region'].str.strip()
```

3 Step 3: Exploratory Data Analysis (EDA)

3.1 A. General Global Insights

[96]:
```
# What is the total number of confirmed cases, deaths, and recoveries worldwide?
# What is the global death rate and recovery rate?
# Which WHO region has the highest number of confirmed cases?
# Calculate Global Death & Recovery Rate
```

```
total_confirmed = df['confirmed'].sum()
print(f" Total Confirmed Cases: {total_confirmed:,}")
total_deaths = df['deaths'].sum()
print(f" Total Deaths: {total_deaths:,}")
total_recovered = df['recovered'].sum()
print(f" Total Recovered: {total_recovered:,}")
death_rate = (total_deaths / total_confirmed) * 100
print(f" Global Death Rate: {death_rate:.2f}%")
recovery_rate = (total_recovered / total_confirmed) * 100
print(f" Global Recovery Rate: {recovery_rate:.2f}%")
```

```
Total Confirmed Cases: 16,480,485
Total Deaths: 654,036
Total Recovered: 9,468,087
Global Death Rate: 3.97%
Global Recovery Rate: 57.45%
```

[97]:
```
#Which WHO region has the highest number of confirmed cases?
region_cases = df.groupby('who_region')['confirmed'].sum()
top_region = region_cases.idxmax()
top_cases = region_cases.max()
print(f" WHO Region with highest confirmed cases: {top_region} ({top_cases:,}␣
 ↪cases)")
```

```
WHO Region with highest confirmed cases: Americas (8,839,286 cases)
```

3.2 B. Country-Level Analysis

[120]:
```
# Top 10 countries by confirmed cases
country_cases = df.groupby('country_region')['confirmed'].max().
 ↪sort_values(ascending=False).head(10)
print(f" Top 10 countries by confirmed cases: {country_cases:}")

# Top 10 Countries by deaths
country_deaths = df.groupby('country_region')['deaths'].max().
 ↪sort_values(ascending=False).head(10)
print(f" Top 10 Countries by deaths: {country_deaths:}")

# Top 10 countries with highest death rate (deaths per 100 cases)
country_death_rate = df.groupby('country_region')['deaths_100_cases'].max().
 ↪sort_values(ascending=False).head(10)
print(f" Top 10 countries with highest death rate: {country_death_rate:}")

# Top 10 countries by recovery rate
country_recovery = (df['recovered'] / df['confirmed']) * 100
country_recovery.head(10)
```

```
 Top 10 countries by confirmed cases: country_region
US                4290259
Brazil            2442375
India             1480073
Russia             816680
South Africa       452529
Mexico             395489
Peru               389717
Chile              347923
United Kingdom     301708
Iran               293606
Name: confirmed, dtype: int64
 Top 10 Countries by deaths: country_region
US                148011
Brazil             87618
United Kingdom     45844
Mexico             44022
Italy              35112
India              33408
France             30212
Spain              28432
Peru               18418
Iran               15912
Name: deaths, dtype: int64
 Top 10 countries with highest death rate: country_region
Yemen             28.56
United Kingdom    15.19
Belgium           14.79
Italy             14.26
France            13.71
Hungary           13.40
Netherlands       11.53
Mexico            11.13
Spain             10.44
Western Sahara    10.00
Name: deaths_100_cases, dtype: float64
```

[120]:
```
0    69.486805
1    56.250000
2    67.339935
3    88.533627
4    25.473684
5    75.581395
6    43.350098
7    71.315860
8    60.844279
9    88.753770
```

```
dtype: float64
```

4 C. Growth & Trends

[126]:
```
# Which countries had the largest 1-week increase in confirmed cases?
largest_confirmed_cases = df.sort_values(by='1_week_change',␣
 ↪ascending=False)[['country_region', '1_week_change']].head(10)
print(f" Top 10 countries with largest 1-week increase in confirmed cases:␣
 ↪{largest_confirmed_cases :}")

# Which countries had the highest 1-week % increase in cases?
highest_week_percent = df.sort_values(by='1_week_%_increase',␣
 ↪ascending=False)[['country_region', '1_week_%_increase']].head(10)
print(f" Top 10 countries with highest 1-week % increase in cases:␣
 ↪{highest_week_percent :}")

# Which countries are recovering fastest this week (new recovered > new cases)?
country_recovery = df[df['new_recovered'] > df['new_cases']][['country_region',␣
 ↪'new_recovered', 'new_cases']].head(10)
print(f" countries with fastest recovery this week: {country_recovery :}")
```

```
 Top 10 countries with largest 1-week increase in confirmed cases:
country_region  1_week_change
173             US         455582
79           India         324735
23          Brazil         323729
154   South Africa          78901
37        Colombia          53096
111         Mexico          46093
138         Russia          40468
6        Argentina          36642
132           Peru          32036
13      Bangladesh          18772
 Top 10 countries with highest 1-week % increase in cases:        country_region
1_week_%_increase
130  Papua New Guinea             226.32
63             Gambia             191.07
11            Bahamas             119.54
186          Zimbabwe              57.85
99              Libya              42.78
58           Ethiopia              42.52
22           Botswana              41.57
97            Lesotho              40.67
160          Suriname              37.44
41         Costa Rica              37.34
 countries with fastest recovery this week:          country_region
```

```
new_recovered  new_cases
2                 Algeria            749        616
5   Antigua and Barbuda                5          4
7                 Armenia            187         73
10             Azerbaijan            558        396
12                Bahrain            421        351
23                 Brazil          33728      23284
25               Bulgaria            230        194
27                  Burma              2          0
28                Burundi             22         17
29             Cabo Verde            103         21
```

[128]:
```
# Any countries with 0 deaths - are they small or underreporting?
zero_deaths = df[df['deaths'] == 0][['country_region','deaths', 'confirmed',␣
 ↪'recovered', 'active']]
zero_deaths
# Out of 187 countries, 17 reported zero COVID-19 deaths. Most are small␣
 ↪nations with very few confirmed cases. While this could reflect successful
# containment strategies, countries like Timor-Leste and Papua New Guinea show␣
 ↪signs of delayed or incomplete reporting, with missing recoveries or
# unusually high active cases. These findings highlight the importance of␣
 ↪interpreting health data in the context of population, infrastructure, and
# transparency.
```

[128]:

	country_region	deaths	confirmed	recovered	active
19	Bhutan	0	99	86	13
30	Cambodia	0	226	147	79
49	Dominica	0	18	18	0
55	Eritrea	0	265	191	74
59	Fiji	0	27	18	9
68	Greenland	0	14	13	1
69	Grenada	0	23	23	0
75	Holy See	0	12	12	0
94	Laos	0	20	19	1
114	Mongolia	0	289	222	67
130	Papua New Guinea	0	62	11	51
140	Saint Kitts and Nevis	0	17	15	2
141	Saint Lucia	0	24	22	2
142	Saint Vincent and the Grenadines	0	52	39	13
148	Seychelles	0	114	39	75
168	Timor-Leste	0	24	0	24
181	Vietnam	0	431	365	66

[ ]: