Exploratory Data Analysis on Covid Data

1.Importing data using pandas

```
import pandas as pd
data = pd.read_csv("/content/covid-data.csv")
```

2.High level data understanding

a.Number of rows and columns

```
print(data.shape)
```

```
(20283, 49)
```

b.Datatypes,Information,Describe

```
print(data.dtypes)
```

```
iso_code                                 object
continent                                object
location                                 object
date                                     object
total_cases                             float64
new_cases                               float64
new_cases_smoothed                      float64
total_deaths                            float64
new_deaths                              float64
new_deaths_smoothed                     float64
total_cases_per_million                 float64
new_cases_per_million                   float64
new_cases_smoothed_per_million          float64
total_deaths_per_million                float64
new_deaths_per_million                  float64
new_deaths_smoothed_per_million         float64
reproduction_rate                       float64
icu_patients                            float64
icu_patients_per_million                float64
hosp_patients                           float64
hosp_patients_per_million               float64
weekly_icu_admissions                   float64
weekly_icu_admissions_per_million       float64
weekly_hosp_admissions                  float64
weekly_hosp_admissions_per_million      float64
total_tests                             float64
new_tests                               float64
total_tests_per_thousand                float64
new_tests_per_thousand                  float64
new_tests_smoothed                      float64
new_tests_smoothed_per_thousand         float64
```

```
tests_per_case                       float64
positive_rate                        float64
stringency_index                     float64
population                             int64
population_density                   float64
median_age                           float64
aged_65_older                        float64
aged_70_older                        float64
gdp_per_capita                       float64
extreme_poverty                      float64
cardiovasc_death_rate                float64
diabetes_prevalence                  float64
female_smokers                       float64
male_smokers                         float64
handwashing_facilities               float64
hospital_beds_per_thousand           float64
life_expectancy                      float64
human_development_index              float64
dtype: object
```

```
print(data.info)
```

```
<bound method DataFrame.info of       iso_code continent     location      dat
0          AFG      Asia  Afghanistan  31-12-2019         NaN         0.0
1          AFG      Asia  Afghanistan  01-01-2020         NaN         0.0
2          AFG      Asia  Afghanistan  02-01-2020         NaN         0.0
3          AFG      Asia  Afghanistan  03-01-2020         NaN         0.0
4          AFG      Asia  Afghanistan  04-01-2020         NaN         0.0
...        ...       ...          ...         ...         ...         ...
20278      GHA    Africa        Ghana  11-10-2020     47005.0        18.0
20279      GHA    Africa        Ghana  12-10-2020     47005.0         0.0
20280      GHA    Africa        Ghana  13-10-2020     47030.0        25.0
20281      GHA    Africa        Ghana  14-10-2020     47126.0        96.0
20282      GHA    Africa        Ghana  15-10-2020     47126.0         0.0

       new_cases_smoothed  total_deaths  new_deaths  new_deaths_smoothed  ...  \
0                     NaN           NaN         0.0                  NaN  ...
1                     NaN           NaN         0.0                  NaN  ...
2                     NaN           NaN         0.0                  NaN  ...
3                     NaN           NaN         0.0                  NaN  ...
4                     NaN           NaN         0.0                  NaN  ...
...                   ...           ...         ...                  ...  ...
20278              28.857         306.0         0.0                0.429  ...
20279              25.143         306.0         0.0                0.429  ...
20280              28.714         308.0         2.0                0.714  ...
20281              42.429         310.0         2.0                1.000  ...
20282              42.429         310.0         0.0                1.000  ...

       gdp_per_capita  extreme_poverty  cardiovasc_death_rate  \
0            1803.987              NaN                597.029
1            1803.987              NaN                597.029
2            1803.987              NaN                597.029
3            1803.987              NaN                597.029
4            1803.987              NaN                597.029
...               ...              ...                    ...
20278        4227.630             12.0                298.245
20279        4227.630             12.0                298.245
20280        4227.630             12.0                298.245
20281        4227.630             12.0                298.245
```

```
20282        4227.630            12.0                    NaN

       diabetes_prevalence  female_smokers  male_smokers  \
0                     9.59             NaN           NaN
1                     9.59             NaN           NaN
2                     9.59             NaN           NaN
3                     9.59             NaN           NaN
4                     9.59             NaN           NaN
...                    ...             ...           ...
20278                 4.97             0.3           7.7
20279                 4.97             0.3           7.7
20280                 4.97             0.3           7.7
20281                 4.97             0.3           7.7
20282                  NaN             NaN           NaN

       handwashing_facilities  hospital_beds_per_thousand  life_expectancy  \
0                      37.746                         0.5            64.83
1                      37.746                         0.5            64.83
2                      37.746                         0.5            64.83
```

```
print(data.describe)
```

```
20281            42.429        310.0        2.0              1.000  ...
20282            42.429        310.0        0.0              1.000  ...

       gdp_per_capita  extreme_poverty  cardiovasc_death_rate  \
0            1803.987              NaN                597.029
1            1803.987              NaN                597.029
2            1803.987              NaN                597.029
3            1803.987              NaN                597.029
4            1803.987              NaN                597.029
...               ...              ...                    ...
20278        4227.630             12.0                298.245
20279        4227.630             12.0                298.245
20280        4227.630             12.0                298.245
```

```
20278                 41.047                           0.9         64.07
20279                 41.047                           0.9         64.07
20280                 41.047                           0.9         64.07
20281                 41.047                           0.9         64.07
20282                    NaN                           NaN           NaN

       human_development_index
0                        0.498
1                        0.498
2                        0.498
3                        0.498
4                        0.498
...                        ...
20278                    0.592
20279                    0.592
20280                    0.592
20281                    0.592
20282                      NaN

[20283 rows x 49 columns]>
```

3.Low level data understandng

a.Count unique values in each columns

```
print(data['new_deaths'].nunique)
```

```
<bound method IndexOpsMixin.nunique of 0        0.0
1        0.0
2        0.0
3        0.0
4        0.0
        ...
20278    0.0
20279    0.0
20280    2.0
20281    2.0
20282    0.0
Name: new_deaths, Length: 20283, dtype: float64>
```

```
print(data.nunique)
```

```
<bound method DataFrame.nunique of       iso_code continent     location
0          AFG      Asia  Afghanistan  31-12-2019         NaN       0.0
1          AFG      Asia  Afghanistan  01-01-2020         NaN       0.0
2          AFG      Asia  Afghanistan  02-01-2020         NaN       0.0
3          AFG      Asia  Afghanistan  03-01-2020         NaN       0.0
4          AFG      Asia  Afghanistan  04-01-2020         NaN       0.0
...        ...       ...          ...         ...         ...       ...
20278      GHA    Africa        Ghana  11-10-2020     47005.0      18.0
20279      GHA    Africa        Ghana  12-10-2020     47005.0       0.0
20280      GHA    Africa        Ghana  13-10-2020     47030.0      25.0
20281      GHA    Africa        Ghana  14-10-2020     47126.0      96.0
20282      GHA    Africa        Ghana  15-10-2020     47126.0       0.0
```

```
      new_cases_smoothed  total_deaths  new_deaths  new_deaths_smoothed  ...  \
0                    NaN           NaN         0.0                  NaN  ...
1                    NaN           NaN         0.0                  NaN  ...
2                    NaN           NaN         0.0                  NaN  ...
3                    NaN           NaN         0.0                  NaN  ...
4                    NaN           NaN         0.0                  NaN  ...
...                  ...           ...         ...                  ...  ...
20278             28.857         306.0         0.0                0.429  ...
20279             25.143         306.0         0.0                0.429  ...
20280             28.714         308.0         2.0                0.714  ...
20281             42.429         310.0         2.0                1.000  ...
20282             42.429         310.0         0.0                1.000  ...

      gdp_per_capita  extreme_poverty  cardiovasc_death_rate  \
0           1803.987              NaN                597.029
1           1803.987              NaN                597.029
2           1803.987              NaN                597.029
3           1803.987              NaN                597.029
4           1803.987              NaN                597.029
...              ...              ...                    ...
20278       4227.630             12.0                298.245
20279       4227.630             12.0                298.245
20280       4227.630             12.0                298.245
20281       4227.630             12.0                298.245
20282       4227.630             12.0                    NaN

      diabetes_prevalence  female_smokers  male_smokers  \
0                    9.59             NaN           NaN
1                    9.59             NaN           NaN
2                    9.59             NaN           NaN
3                    9.59             NaN           NaN
4                    9.59             NaN           NaN
...                   ...             ...           ...
20278                4.97             0.3           7.7
20279                4.97             0.3           7.7
20280                4.97             0.3           7.7
20281                4.97             0.3           7.7
20282                 NaN             NaN           NaN

      handwashing_facilities  hospital_beds_per_thousand  life_expectancy  \
0                     37.746                         0.5            64.83
1                     37.746                         0.5            64.83
2                     37.746                         0.5            64.83
```

b.Find which continent have maximum frequency of values.

```
Maximum_Frequency=data['continent'].value_counts()
Maximum_Frequency_continent=Maximum_Frequency.idxmax()
print(Maximum_Frequency_continent,Maximum_Frequency.max())
```

```
Africa 5468
```

c.max & mean value in total_cases

```
print("Maximum Value",data['total_cases'].max)
print("Mean",data['total_cases'].mean)
```

```
Maximum Value <bound method Series.max of 0            NaN
1            NaN
2            NaN
3            NaN
4            NaN
          ...
20278    47005.0
20279    47005.0
20280    47030.0
20281    47126.0
20282    47126.0
Name: total_cases, Length: 20283, dtype: float64>
Mean <bound method Series.mean of 0            NaN
1            NaN
2            NaN
3            NaN
4            NaN
          ...
20278    47005.0
20279    47005.0
20280    47030.0
20281    47126.0
20282    47126.0
Name: total_cases, Length: 20283, dtype: float64>
```

d.find interquartiles

```
print(data['total_deaths'].describe(percentiles=[.25,.50,.75]))
```

```
count     15625.000000
mean       2936.541056
std       12890.630598
min           1.000000
25%           9.000000
50%          83.000000
75%         712.000000
max      166014.000000
Name: total_deaths, dtype: float64
```

e.Find which continent have highest human_value_development_index.

```
df=data.groupby('continent')['human_development_index'].max()
continent=df.idxmax()
df1=df.max()
print(continent,df1)
```

```
Oceania 0.939
```

f.Find Which continent have minimum gdp_per_captia

```
df2=data.groupby('continent')['gdp_per_capita'].min()
continent1=df2.idxmin()
mini=df2.min()
print(continent1,mini)
```

```
Africa 661.24
```

4.Filtering

```
tokeep=['continent','location','date','total_cases','total_deaths','gdp_per_capita','huma
df=data[tokeep]
print(df)
```

```
      continent     location        date  total_cases  total_deaths  \
0          Asia  Afghanistan  31-12-2019          NaN           NaN
1          Asia  Afghanistan  01-01-2020          NaN           NaN
2          Asia  Afghanistan  02-01-2020          NaN           NaN
3          Asia  Afghanistan  03-01-2020          NaN           NaN
4          Asia  Afghanistan  04-01-2020          NaN           NaN
...         ...          ...         ...          ...           ...
20278    Africa        Ghana  11-10-2020      47005.0         306.0
20279    Africa        Ghana  12-10-2020      47005.0         306.0
20280    Africa        Ghana  13-10-2020      47030.0         308.0
20281    Africa        Ghana  14-10-2020      47126.0         310.0
20282    Africa        Ghana  15-10-2020      47126.0         310.0

       gdp_per_capita  human_development_index
0            1803.987                    0.498
1            1803.987                    0.498
2            1803.987                    0.498
3            1803.987                    0.498
4            1803.987                    0.498
...               ...                      ...
20278        4227.630                    0.592
20279        4227.630                    0.592
20280        4227.630                    0.592
20281        4227.630                    0.592
20282        4227.630                      NaN

[20283 rows x 7 columns]
```

5.Data Cleaning.

```
df_cleaned=df.drop_duplicates()
df_3=df_cleaned.isna().sum()
df_cleaned = df_cleaned.dropna(subset=['continent'])
```

```
df_cleaned=df_cleaned.fillna(0)
print(df_cleaned)
```

```
      continent     location        date  total_cases  total_deaths  \
0          Asia  Afghanistan  31-12-2019          0.0           0.0
1          Asia  Afghanistan  01-01-2020          0.0           0.0
```

```
2          Asia  Afghanistan  02-01-2020        0.0          0.0
3          Asia  Afghanistan  03-01-2020        0.0          0.0
4          Asia  Afghanistan  04-01-2020        0.0          0.0
...         ...          ...         ...        ...          ...
20278    Africa        Ghana  11-10-2020    47005.0        306.0
20279    Africa        Ghana  12-10-2020    47005.0        306.0
20280    Africa        Ghana  13-10-2020    47030.0        308.0
20281    Africa        Ghana  14-10-2020    47126.0        310.0
20282    Africa        Ghana  15-10-2020    47126.0        310.0

       gdp_per_capita  human_development_index
0            1803.987                    0.498
1            1803.987                    0.498
2            1803.987                    0.498
3            1803.987                    0.498
4            1803.987                    0.498
...               ...                      ...
20278        4227.630                    0.592
20279        4227.630                    0.592
20280        4227.630                    0.592
20281        4227.630                    0.592
20282        4227.630                    0.000

[20283 rows x 7 columns]
```

6.a.date time format

```
df_cleaned['date'] = pd.to_datetime(df_cleaned['date'], errors='coerce')
print(df_cleaned['date'])
```

```
0       2019-12-31
1       2020-01-01
2       2020-01-02
3       2020-01-03
4       2020-01-04
           ...
20278   2020-10-11
20279   2020-10-12
20280   2020-10-13
20281   2020-10-14
20282   2020-10-15
Name: date, Length: 20283, dtype: datetime64[ns]
<ipython-input-19-3aa3f2ac1bfa>:1: UserWarning: Parsing dates in %d-%m-%Y format wher
  df_cleaned['date'] = pd.to_datetime(df_cleaned['date'], errors='coerce')
```

b.Extract month from date column

```
df_cleaned['month']=df_cleaned['date'].dt.month
print("New Extracted Month Column",df_cleaned['month'])
```

```
New Extracted Month Column 0        12
1         1
2         1
```

```
3        1
4        1
        ..
20278    10
20279    10
20280    10
20281    10
20282    10
Name: month, Length: 20283, dtype: int32
```

7.a Data Aggregation - Find maximum values in every columns grouped by continent

```
df_groupby=df_cleaned.groupby('continent').max()
print(df_groupby.reset_index())
```

```
       continent          location       date  total_cases  total_deaths  \
0         Africa             Ghana 2020-11-17     111009.0        6465.0
1           Asia           Georgia 2020-11-17     434472.0        6215.0
2         Europe           Germany 2020-11-17    1991233.0       45054.0
3  North America       El Salvador 2020-11-17     302192.0       11027.0
4        Oceania  French Polynesia 2020-11-17      27750.0         907.0
5  South America  Falkland Islands 2020-11-17    5876464.0      166014.0

   gdp_per_capita  human_development_index  month
0       22604.873                    0.754     12
1       71809.251                    0.853     12
2       46682.515                    0.936     12
3       50669.315                    0.926     12
4       44648.710                    0.939     12
5       22767.037                    0.843     12
```

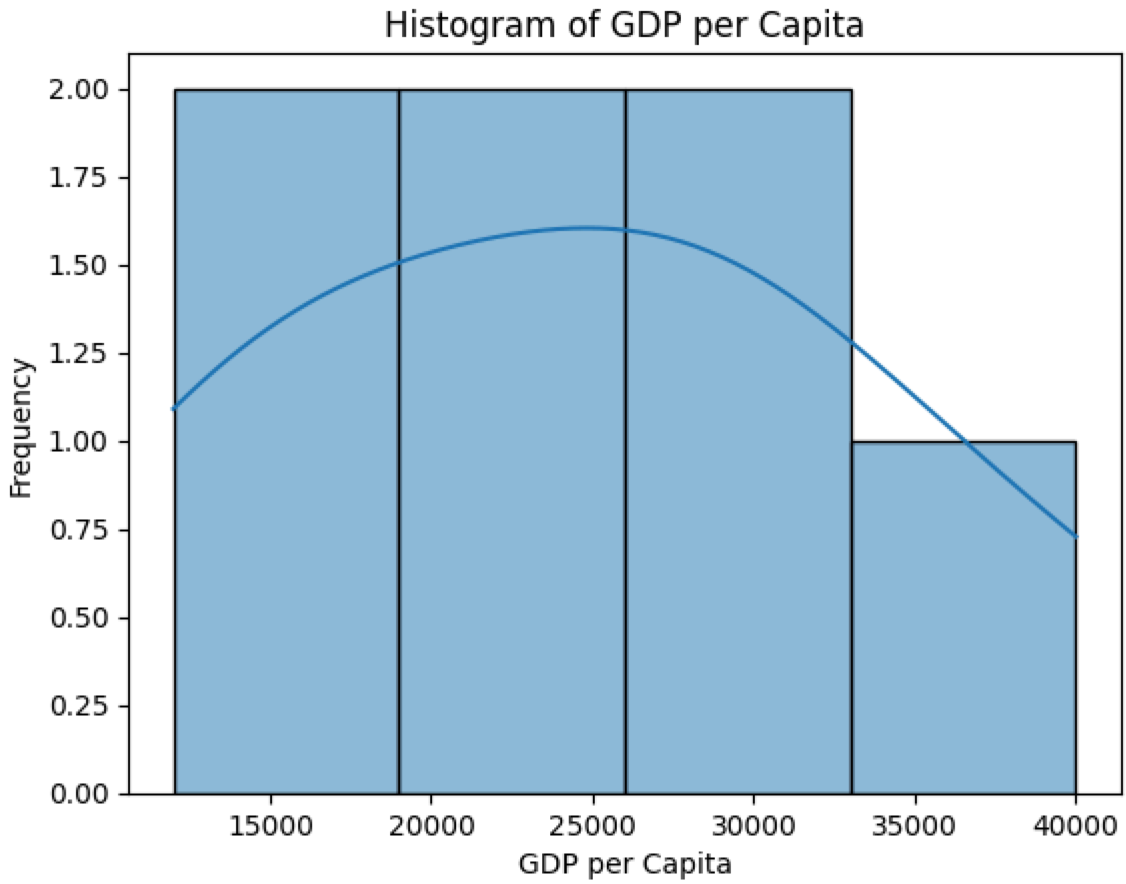
8.Feature Engineering

```
df_groupby['totaldeaths_totalcases']=df_groupby['total_deaths']/df_groupby['total_cases']
print(df_groupby['totaldeaths_totalcases'])
```

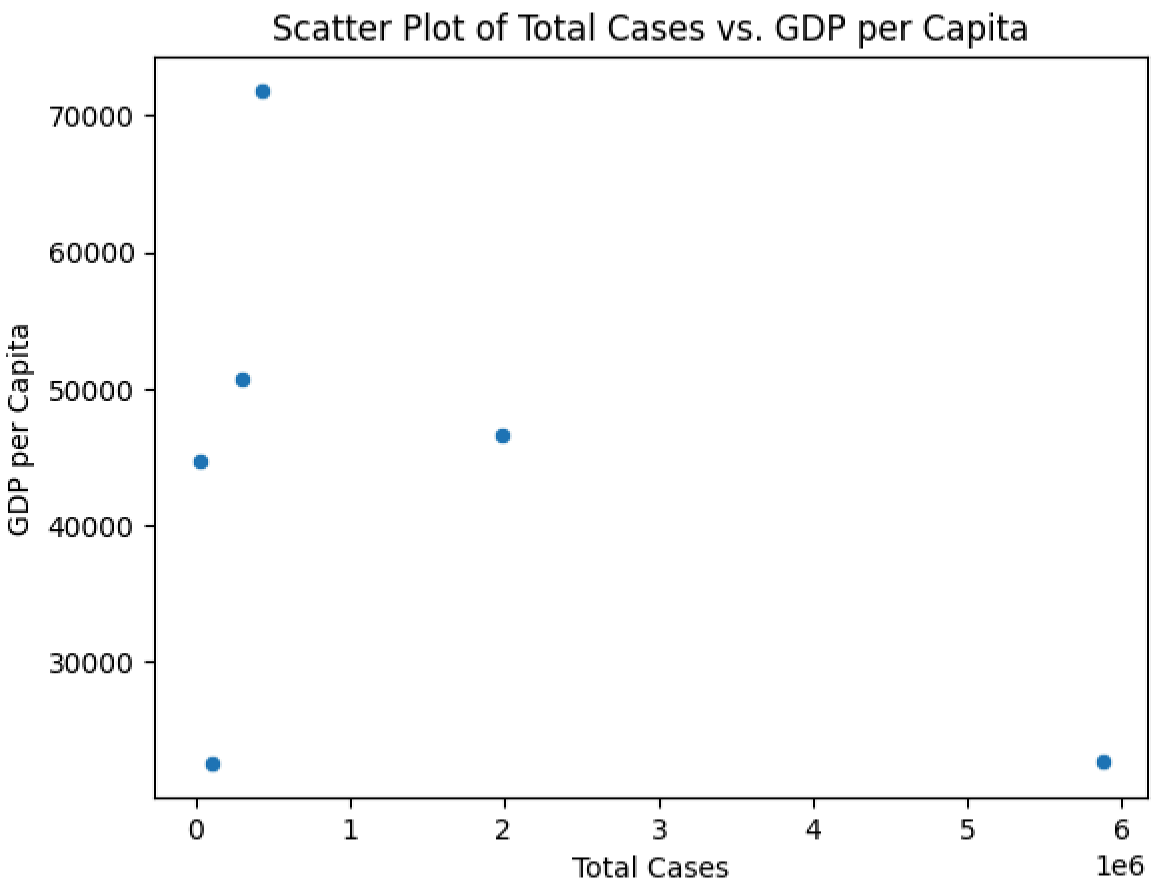
```
continent
Africa           0.058239
Asia             0.014305
Europe           0.022626
North America    0.036490
Oceania          0.032685
South America    0.028251
Name: totaldeaths_totalcases, dtype: float64
```

9.Data Visualization

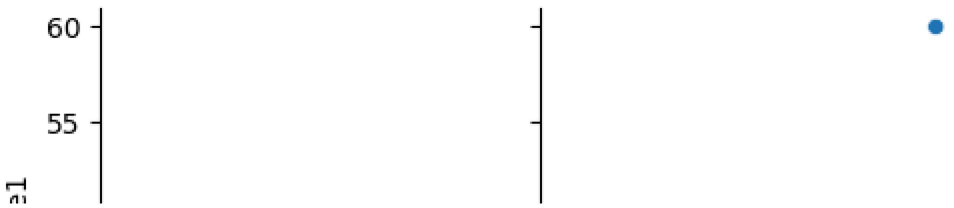
```
import seaborn as sns
import matplotlib.pyplot as plt
df = pd.DataFrame({
    'gdp_per_capita': [30000, 15000, 20000, 40000, 25000, 30000, 12000]
})
sns.histplot(df['gdp_per_capita'], kde=True)  # `distplot` is deprecated; use `histplot`
plt.title('Histogram of GDP per Capita')
plt.xlabel('GDP per Capita')
plt.ylabel('Frequency')
plt.show()
```



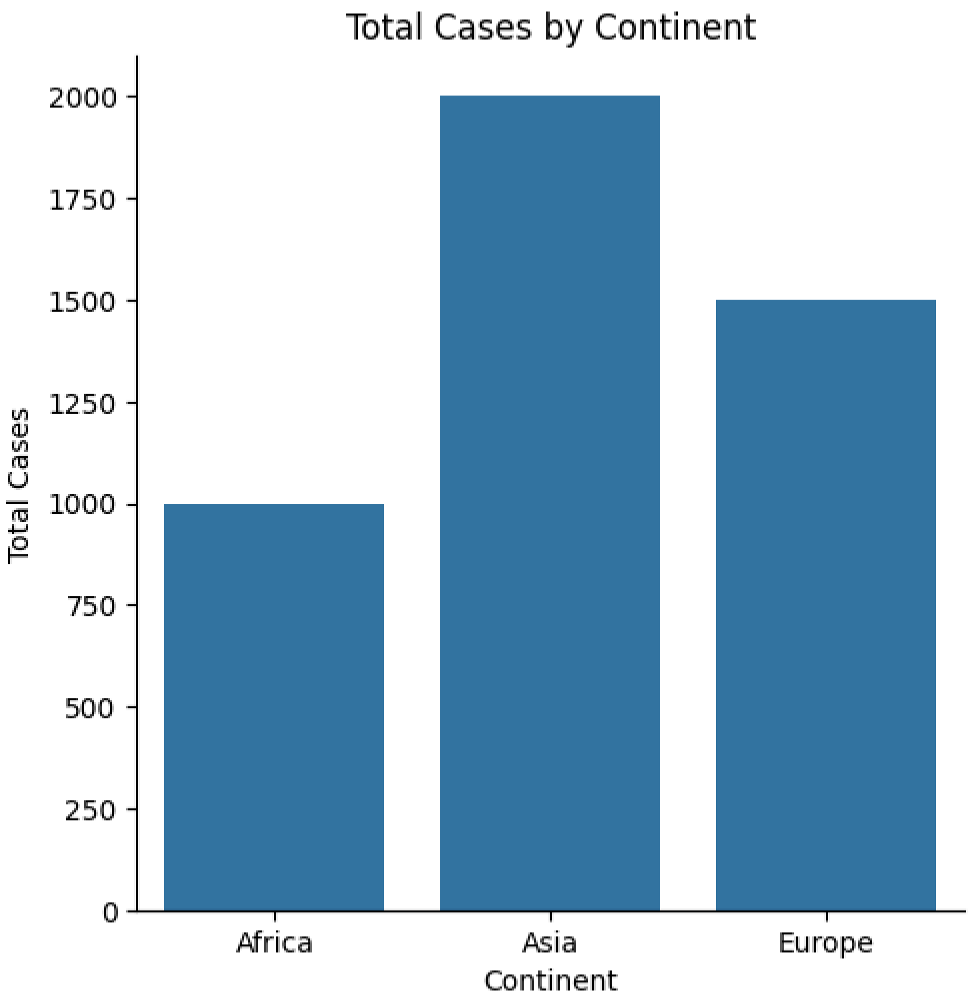

```
import seaborn as sns
import matplotlib.pyplot as ply
sns.scatterplot(x=df_groupby['total_cases'], y=df_groupby['gdp_per_capita'], data=df_grou
plt.title('Scatter Plot of Total Cases vs. GDP per Capita')
plt.xlabel('Total Cases')
plt.ylabel('GDP per Capita')
plt.show()
```




```
df_groupby = pd.DataFrame({
    'continent': ['Africa', 'Asia', 'Europe'],
    'value1': [60, 50, 40],
    'value2': [55, 45, 35]
})
sns.pairplot(df_groupby, hue='continent')
plt.show()
```




```
df = pd.DataFrame({
    'continent': ['Africa', 'Asia', 'Europe'],
    'total_cases': [1000, 2000, 1500]
})
sns.catplot(x='continent', y='total_cases', data=df, kind='bar')
plt.title('Total Cases by Continent')
plt.xlabel('Continent')
plt.ylabel('Total Cases')
plt.show()
```




```
df_groupby.to_csv('df_groupby.csv', index=False)
```