world-data

November 24, 2024

0.1 Importing Libraries

```
[2]: import pandas as pd
     import numpy as np
     import seaborn as sns
     import matplotlib.pyplot as plt
```

0.2 Load and Check the DataSet

```
[4]: df = pd.read_csv(r"D:\ANURAG KUMAR\Project\Data Analysis\Projeact 4\WorldData.
      ↪csv")
     df.head()
```

```
[4]:   Code         Name      Continent                     Region  SurfaceArea  \
     0  ABW        Aruba  North America                  Caribbean        193.0
     1  AFG  Afghanistan           Asia  Southern and Central Asia     652090.0
     2  AGO       Angola         Africa             Central Africa    1246700.0
     3  AIA     Anguilla  North America                  Caribbean         96.0
     4  ALB      Albania         Europe            Southern Europe      28748.0

        IndepYear  Population  LifeExpectancy     GNP  GNPOld  \
     0        NaN      103000            78.4   828.0   793.0
     1     1919.0    22720000            45.9  5976.0     NaN
     2     1975.0    12878000            38.3  6648.0  7984.0
     3        NaN        8000            76.1    63.2     NaN
     4     1912.0     3401200            71.6  3205.0  2500.0

                    LocalName                                GovernmentForm  \
     0                  Aruba  Nonmetropolitan Territory of The Netherlands
     1  Afganistan/Afqanestan                               Islamic Emirate
     2                 Angola                                      Republic
     3               Anguilla                    Dependent Territory of the UK
     4              Shqipëria                                      Republic

                    HeadOfState  Capital Code2
     0                  Beatrix    129.0    AW
     1            Mohammad Omar      1.0    AF
     2  José Eduardo dos Santos     56.0    AO
```

```
3             Elisabeth II     62.0    AI
4           Rexhep Mejdani     34.0    AL
```

0.3 Get all the Information About the DataSet

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 239 entries, 0 to 238
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Code            239 non-null    object
 1   Name            239 non-null    object
 2   Continent       239 non-null    object
 3   Region          239 non-null    object
 4   SurfaceArea     239 non-null    float64
 5   IndepYear       192 non-null    float64
 6   Population      239 non-null    int64
 7   LifeExpectancy  222 non-null    float64
 8   GNP             239 non-null    float64
 9   GNPOld          178 non-null    float64
 10  LocalName       239 non-null    object
 11  GovernmentForm  239 non-null    object
 12  HeadOfState     236 non-null    object
 13  Capital         232 non-null    float64
 14  Code2           238 non-null    object
dtypes: float64(6), int64(1), object(8)
memory usage: 28.1+ KB
```

0.4 Get Statistical Data on DataSet

```
[8]: df.describe()
```

```
[8]:        SurfaceArea    IndepYear    Population  LifeExpectancy           GNP  \
     count  2.390000e+02   192.000000  2.390000e+02      222.000000  2.390000e+02
     mean   6.232481e+05  1847.260417  2.543410e+07       66.486036  1.228239e+05
     std    1.924140e+06   420.831370  1.093398e+08       11.519267  6.379976e+05
     min    4.000000e-01 -1523.000000  0.000000e+00       37.200000  0.000000e+00
     25%    2.275000e+03  1906.750000  2.380000e+05       60.300000  6.400000e+02
     50%    7.174000e+04  1960.000000  3.869000e+06       70.150000  4.787000e+03
     75%    3.987545e+05  1974.000000  1.493550e+07       75.500000  2.994450e+04
     max    1.707540e+07  1994.000000  1.277558e+09       83.500000  8.510700e+06

                 GNPOld      Capital
     count  1.780000e+02   232.000000
     mean   1.655343e+05  2071.306034
```

```
std    7.204689e+05  1184.095609
min    1.570000e+02     1.000000
25%    2.187000e+03   915.750000
50%    8.421000e+03  2449.500000
75%    7.114550e+04  3065.250000
max    8.110900e+06  4074.000000
```

0.5 Check For Missing Values

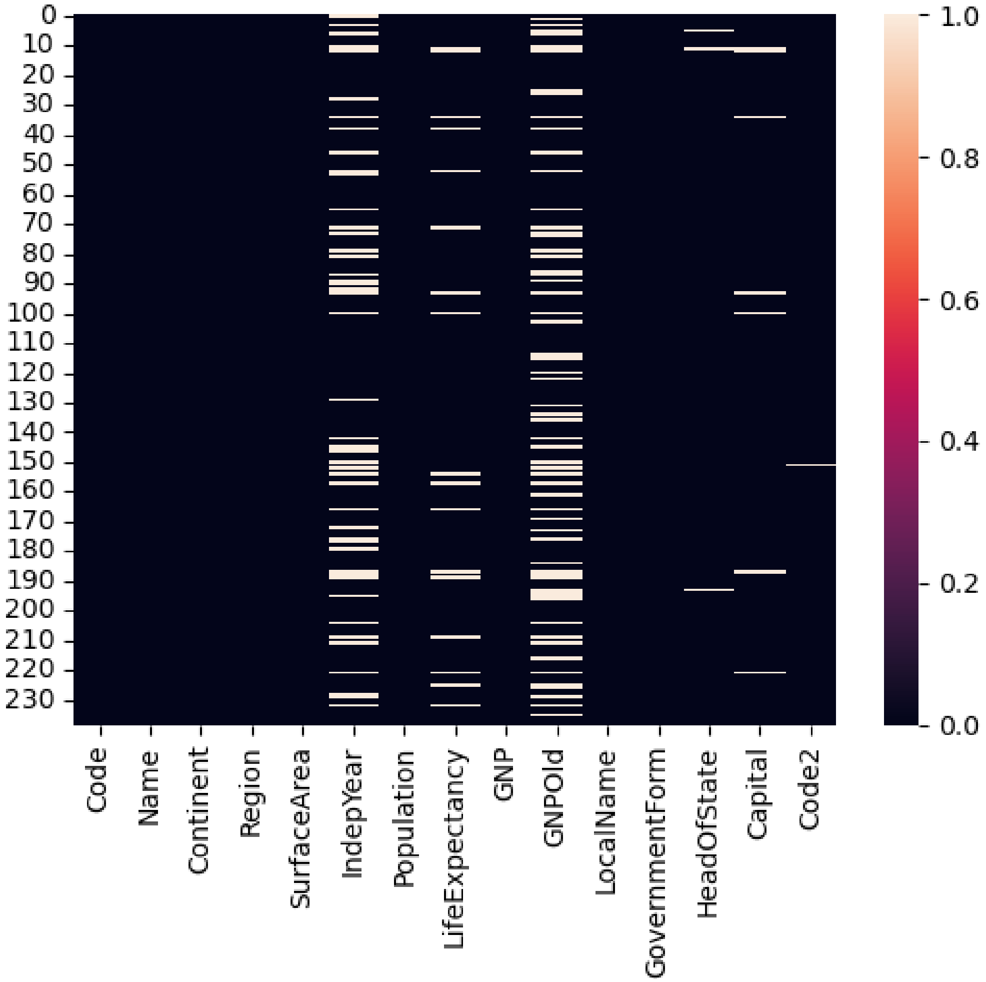
[10]:
```
df.isnull().sum()
# OR
df.isna().sum()
# Both Gives Same result
```

[10]:
```
Code               0
Name               0
Continent          0
Region             0
SurfaceArea        0
IndepYear         47
Population         0
LifeExpectancy    17
GNP                0
GNPOld            61
LocalName          0
GovernmentForm     0
HeadOfState        3
Capital            7
Code2              1
dtype: int64
```

0.6 Visualizing Missing values

[12]:
```
sns.heatmap(df.isnull())
```

[12]:
```
<Axes: >
```



0.7 Remove Unwanted Columns

[14]:
```
df.drop(columns='GNP', inplace=True)
df.drop(columns='GNPOld', inplace=True)
df.head()
```

[14]:
```
  Code         Name      Continent                     Region  SurfaceArea  \
0  ABW        Aruba  North America                  Caribbean        193.0
1  AFG  Afghanistan           Asia  Southern and Central Asia     652090.0
2  AGO       Angola         Africa             Central Africa    1246700.0
3  AIA     Anguilla  North America                  Caribbean         96.0
4  ALB      Albania         Europe            Southern Europe      28748.0

   IndepYear  Population  LifeExpectancy             LocalName  \
0        NaN      103000            78.4                 Aruba
1     1919.0    22720000            45.9  Afganistan/Afqanestan
```

```
2     1975.0    12878000            38.3                  Angola
3        NaN        8000            76.1                Anguilla
4     1912.0     3401200            71.6               Shqipëria

                                GovernmentForm              HeadOfState  \
0  Nonmetropolitan Territory of The Netherlands                  Beatrix
1                              Islamic Emirate             Mohammad Omar
2                                     Republic  José Eduardo dos Santos
3                Dependent Territory of the UK             Elisabeth II
4                                     Republic           Rexhep Mejdani

   Capital Code2
0    129.0    AW
1      1.0    AF
2     56.0    AO
3     62.0    AI
4     34.0    AL
```

0.8 Renaming a Column

[17]:
```
df.rename(columns = {'IndepYear' : 'IndependentYear'}, inplace = True)
```

0.9 Extract the names of all the columns in the Dataset

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

[20]:
```
Index(['Code', 'Name', 'Continent', 'Region', 'SurfaceArea', 'IndependentYear',
       'Population', 'LifeExpectancy', 'LocalName', 'GovernmentForm',
       'HeadOfState', 'Capital', 'Code2'],
      dtype='object')
```

0.10 Check Datatypes of all the Columns

[23]:
```
df.dtypes
```

[23]:
```
Code                object
Name                object
Continent           object
Region              object
SurfaceArea        float64
IndependentYear    float64
Population           int64
LifeExpectancy     float64
LocalName           object
GovernmentForm      object
HeadOfState         object
```

```
Capital            float64
Code2               object
dtype: object
```

0.11 Handling Missing or NULL Values

[26]:
```
df['IndependentYear'] = df['IndependentYear'].fillna(0) #Replacing With 0

df['LifeExpectancy'] = df['LifeExpectancy'].fillna(df['LifeExpectancy'].mean())␣
 ↪ # Replace with mean
df['Capital'] = df['Capital'].fillna(df['Capital'].mean())  # Replace with mean

df['HeadOfState'] = df['HeadOfState'].fillna('Unknown')   # Replace with␣
 ↪"Unknown"
df['Code2'] = df['Code2'].fillna('N/A')   # Replace with "N/A"
```

[28]:
```
df.isnull().sum()
```

[28]:
```
Code               0
Name               0
Continent          0
Region             0
SurfaceArea        0
IndependentYear    0
Population         0
LifeExpectancy     0
LocalName          0
GovernmentForm     0
HeadOfState        0
Capital            0
Code2              0
dtype: int64
```

0.12 Count Number of Unique Values in Each Column

[31]:
```
df.nunique()
```

[31]:
```
Code               239
Name               239
Continent            7
Region              25
SurfaceArea        238
IndependentYear     89
Population         226
LifeExpectancy     161
LocalName          239
```

```
GovernmentForm      35
HeadOfState        179
Capital            233
Code2              239
dtype: int64
```

0.13 Name All the Unique Values in ‘Continent’ Column

[34]:

```
df['Continent'].value_counts()
      # OR
df.Continent.value_counts()    # Both gives same result
```

[34]:

```
Continent
Africa           58
Asia             51
Europe           46
North America    37
Oceania          28
South America    14
Antarctica        5
Name: count, dtype: int64
```

0.14 Get all the values of ‘Antarctica’ Continent

[37]:

```
# Using Filter Method
df[df['Continent'] == 'Antarctica']

# Using groupby() method
df.groupby('Continent').get_group('Antarctica')
```

[37]:

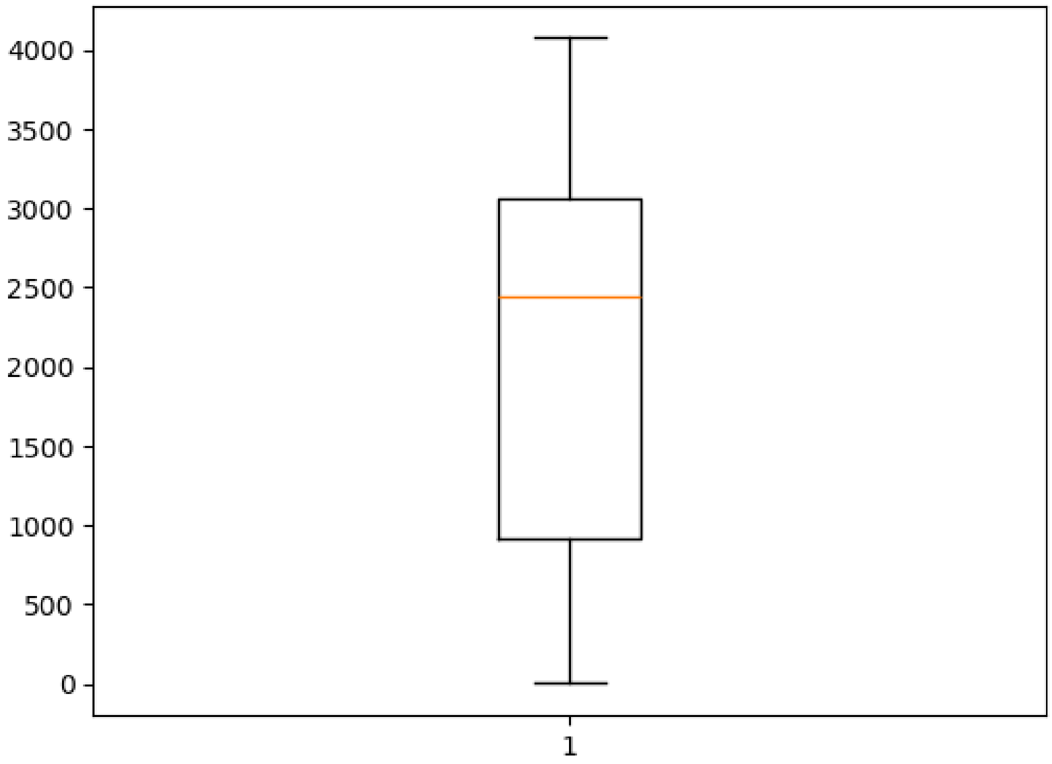
```
    Code                                          Name   Continent  \
11   ATA                                    Antarctica  Antarctica
12   ATF                   French Southern territories  Antarctica
34   BVT                                 Bouvet Island  Antarctica
93   HMD             Heard Island and McDonald Islands  Antarctica
187  SGS  South Georgia and the South Sandwich Islands  Antarctica

         Region  SurfaceArea  IndependentYear  Population  LifeExpectancy  \
11   Antarctica   13120000.0              0.0           0       66.486036
12   Antarctica       7780.0              0.0           0       66.486036
34   Antarctica         59.0              0.0           0       66.486036
93   Antarctica        359.0              0.0           0       66.486036
187  Antarctica       3903.0              0.0           0       66.486036

                                        LocalName  \
11                                              -
```

```
12                      Terres australes françaises
34                                        Bouvetøya
93                       Heard and McDonald Islands
187  South Georgia and the South Sandwich Islands

                          GovernmentForm     HeadOfState      Capital Code2
11                      Co-administrated         Unknown  2071.306034    AQ
12   Nonmetropolitan Territory of France  Jacques Chirac  2071.306034    TF
34         Dependent Territory of Norway        Harald V  2071.306034    BV
93                Territory of Australia    Elisabeth II  2071.306034    HM
187        Dependent Territory of the UK    Elisabeth II  2071.306034    GS
```

0.15 Visualize and analyse the "Capital" Column

[40]:
```
y = list(df.Capital)
 plt.boxplot(y)
 plt.show
```

[40]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



[42]:
```
df['Capital'].max()   # Maximum Value
```

```
[42]: 4074.0
```

```
[44]: df['Capital'].min()  # Minimum Value
```

```
[44]: 1.0
```

```
[46]: df['Capital'].mean()  # Mean Value
```

```
[46]: 2071.3060344827586
```

```
[50]: df['Capital'].sum()  # Sum of all values
```

```
[50]: 495042.1422413793
```

0.16 Find Out the Maximum Capital Of Each Region According to their Continents

```
[53]: df.groupby(['Continent','Region']).max('Region')
```

```
[53]:                                         SurfaceArea  IndependentYear  \
      Continent     Region
      Africa        Central Africa              2344858.0           1975.0
                    Eastern Africa              1104300.0           1993.0
                    Northern Africa             2505813.0           1962.0
                    Southern Africa             1221037.0           1990.0
                    Western Africa              1267000.0           1975.0
      Antarctica    Antarctica                 13120000.0              0.0
      Asia          Eastern Asia                9572900.0           1948.0
                    Middle East                 2149690.0           1991.0
                    Southeast Asia              1904569.0           1984.0
                    Southern and Central Asia   3287263.0           1991.0
      Europe        Baltic Countries              65301.0           1991.0
                    British Islands              242900.0           1921.0
                    Eastern Europe             17075400.0           1993.0
                    Nordic Countries             449964.0           1944.0
                    Southern Europe              505992.0           1992.0
                    Western Europe               551500.0           1955.0
      North America Caribbean                    110861.0           1983.0
                    Central America             1958201.0           1981.0
                    North America               9970610.0           1867.0
      Oceania       Australia and New Zealand   7741220.0           1907.0
                    Melanesia                    462840.0           1980.0
                    Micronesia                      726.0           1994.0
                    Micronesia/Caribbean             16.0              0.0
                    Polynesia                      4000.0           1978.0
      South America South America               8547403.0           1975.0
```

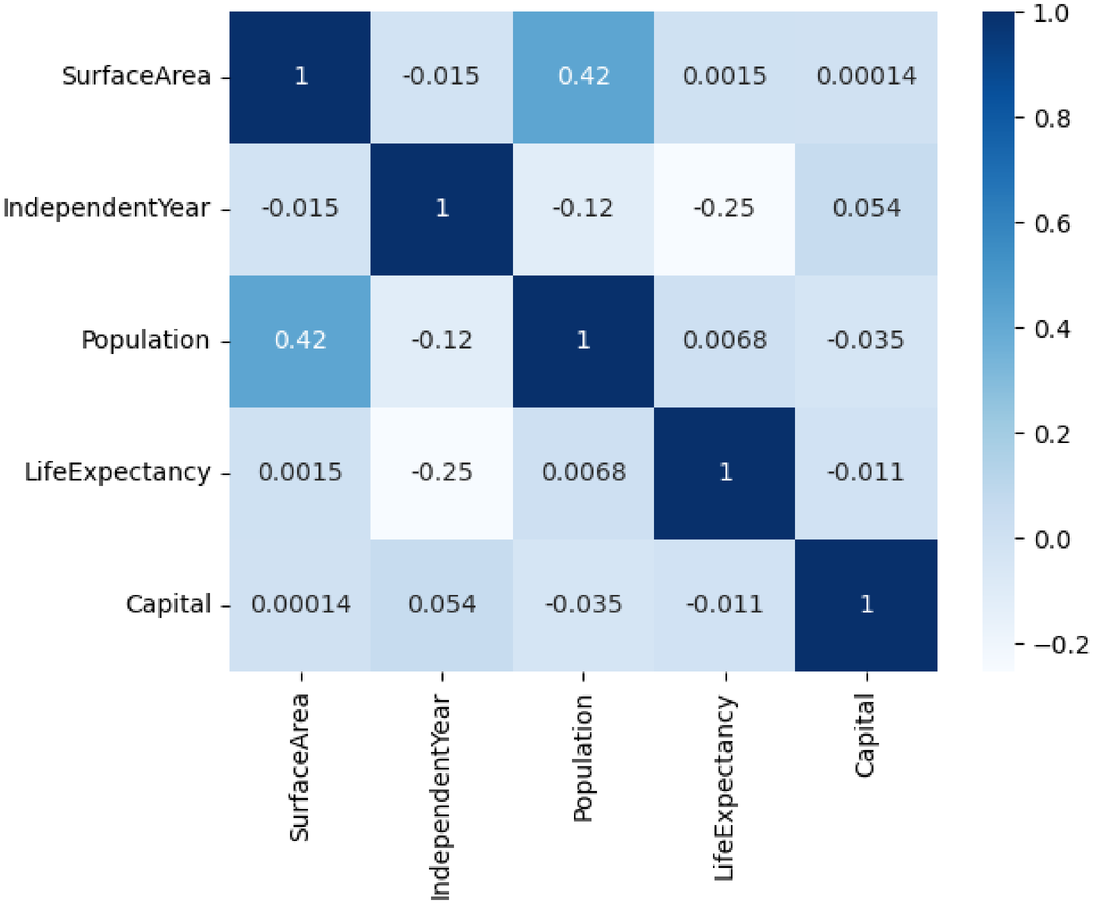
```
                                      Population  LifeExpectancy  \
Continent     Region
Africa        Central Africa            51654000       65.300000
              Eastern Africa            62565000       72.700000
              Northern Africa           68470000       75.500000
              Southern Africa           40377000       51.100000
              Western Africa           111506000       76.800000
Antarctica    Antarctica                       0       66.486036
Asia          Eastern Asia            1277558000       81.600000
              Middle East               66591000       78.600000
              Southeast Asia           212107000       80.100000
              Southern and Central Asia  1013662000       71.800000
Europe        Baltic Countries           3698500       69.500000
              British Islands           59623400       77.700000
              Eastern Europe           146934000       74.500000
              Nordic Countries           8861400       79.600000
              Southern Europe           57680000       83.500000
              Western Europe            82164700       79.600000
North America Caribbean                 11201000       78.900000
              Central America           98881000       75.800000
              North America            278357000       79.400000
Oceania       Australia and New Zealand   18886000       79.800000
              Melanesia                  4807000       72.800000
              Micronesia                  168000       77.800000
              Micronesia/Caribbean             0       66.486036
              Polynesia                   235000       75.100000
South America South America            170115000       76.100000

                                            Capital
Continent     Region
Africa        Central Africa            3337.000000
              Eastern Africa            4068.000000
              Northern Africa           3349.000000
              Southern Africa           3244.000000
              Western Africa            3332.000000
Antarctica    Antarctica                2071.306034
Asia          Eastern Asia              3263.000000
              Middle East               4074.000000
              Southeast Asia            3770.000000
              Southern and Central Asia  3503.000000
Europe        Baltic Countries          3791.000000
              British Islands           1447.000000
              Eastern Europe            3580.000000
              Nordic Countries          3315.000000
              Southern Europe           3538.000000
              Western Europe            3248.000000
North America Caribbean                 4067.000000
```

```
              Central America           2882.000000
              North America             3813.000000
Oceania       Australia and New Zealand  3499.000000
              Melanesia                 3537.000000
              Micronesia                2913.000000
              Micronesia/Caribbean      2071.306034
              Polynesia                 3536.000000
South America South America             3539.000000
```

0.17 Find out the correlation between different variables in the given dataset

```
[56]: numeric_df = df.select_dtypes(include=['number'])
      sns.heatmap(numeric_df.corr(), cbar=True,annot=True,cmap='Blues')
```

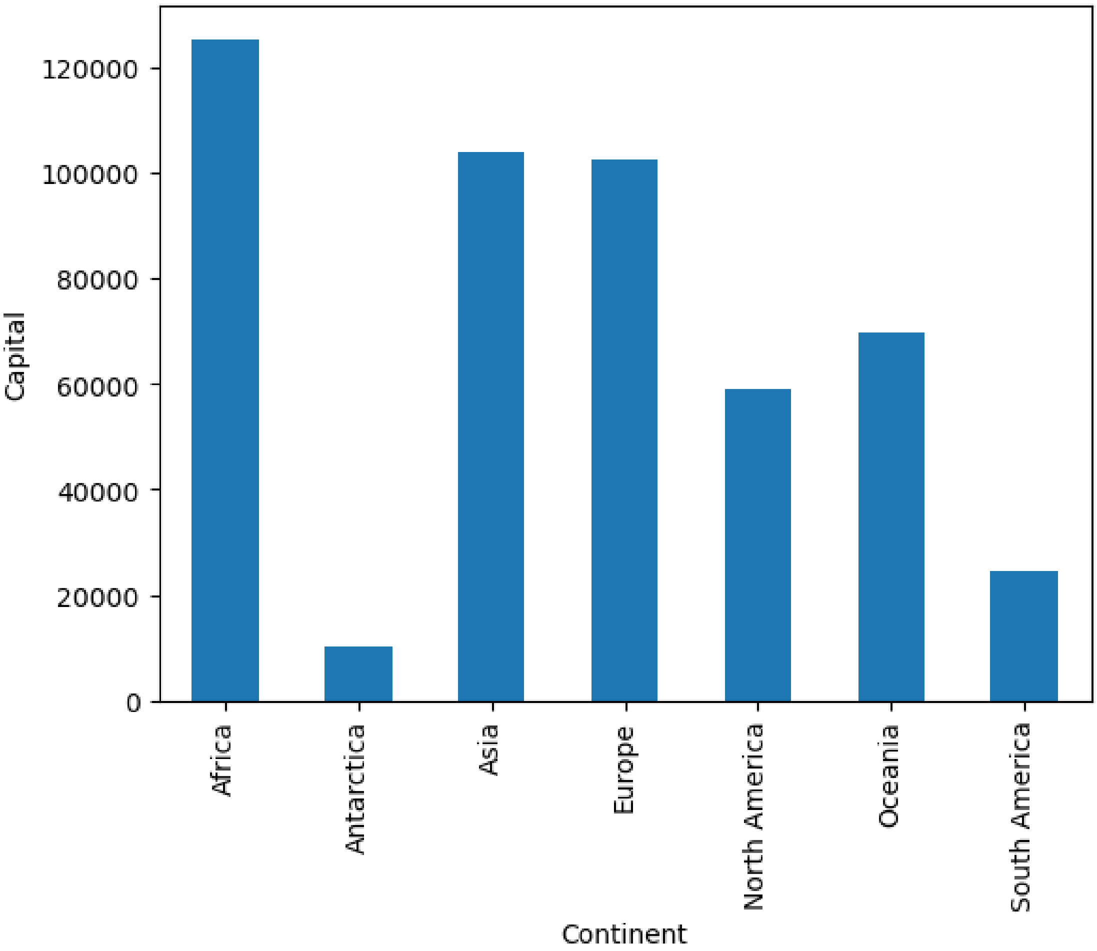
```
[56]: <Axes: >
```



0.18 Analyse how the capital of a region affects the Continent

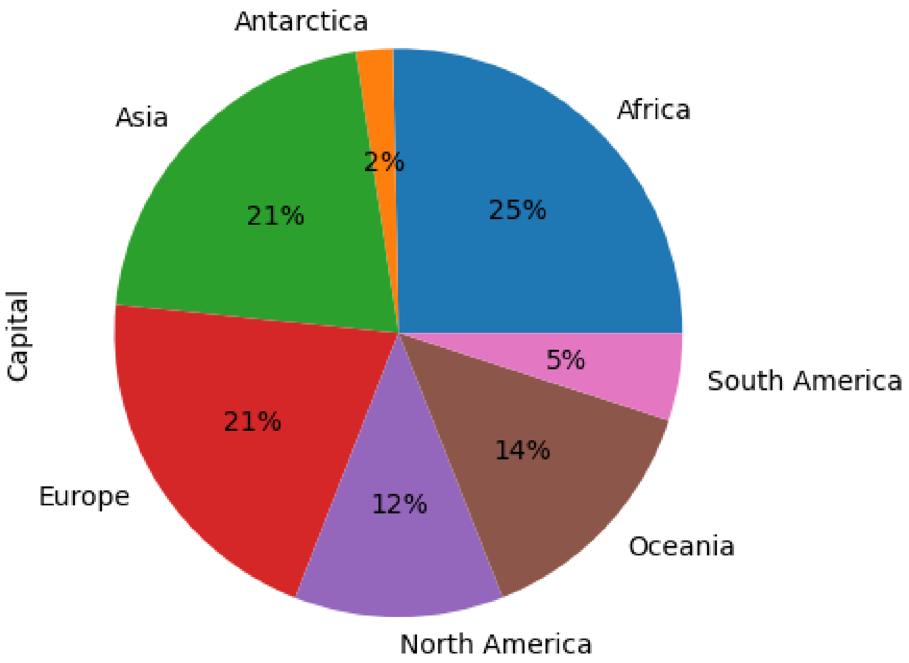
```
[60]: df.groupby("Continent")['Capital'].sum().plot.bar()
      plt.xlabel('Continent')
      plt.ylabel('Capital')
```

```
[60]: Text(0, 0.5, 'Capital')
```



```
[62]: df.groupby("Continent")['Capital'].sum().plot.pie(autopct="%1.0f%%")
```

```
[62]: <Axes: ylabel='Capital'>
```


Antarctica
Africa
Asia
2%
25%
21%
Capital
5%
South America
21%
14%
Europe
12%
Oceania
North America