# zerotopandas-course-project

December 10, 2020

## 1 COVID DATA ANALYSIS

### 1.1 Introduction

As we know that the whole world is going to the pandemic so there are many different datas for each country i.e. new_cases,deaths,recovery and many more. So, we will analysis the data for country where the population is very high and day-by-day the cases are increasing.The course (Data Analysis with Python) by jovian.ml has made learn a lot of things in this course like data cleaning,load a file (CSV,Json etc) and most important is visualization and analysis of the datasets.

In this block we will first load the csv file into the jupyter notebook. The file name(owid-covid-data-csv) this file contains the data all over world and we will analysis for india.

[5]:
```
project_name = "covid-data-analysis-project"
```

[6]:
```
!pip install jovian --upgrade -q
```

We will load the file with the help of pandas and there are many more ways that you can load the file like using from urllib.urlrequest and importing urlretireve if you are having the link or you can use opendatasets which is created by jovian.ml itself.

[7]:
```
import pandas as pd
```

[8]:
```
survy_raw_df = pd.read_csv('owid-covid-data.csv')
```

### 1.2 Explore the data.

[9]:
```
survy_raw_df
```

[9]:

| | iso_code | continent | location | date | total_cases | new_cases | \ |
|---|---|---|---|---|---|---|---|
| 0 | AFG | Asia | Afghanistan | 2019-12-31 | 0.0 | 0.0 | |
| 1 | AFG | Asia | Afghanistan | 2020-01-01 | 0.0 | 0.0 | |
| 2 | AFG | Asia | Afghanistan | 2020-01-02 | 0.0 | 0.0 | |
| 3 | AFG | Asia | Afghanistan | 2020-01-03 | 0.0 | 0.0 | |
| 4 | AFG | Asia | Afghanistan | 2020-01-04 | 0.0 | 0.0 | |
| … | … | … | … | … | … | … | |
| 45634 | NaN | NaN | International | 2020-09-19 | 696.0 | NaN | |
| 45635 | NaN | NaN | International | 2020-09-20 | 696.0 | NaN | |
| 45636 | NaN | NaN | International | 2020-09-21 | 696.0 | NaN | |

```
45637      NaN       NaN  International  2020-09-22        696.0        NaN
45638      NaN       NaN  International  2020-09-23        696.0        NaN

       new_cases_smoothed  total_deaths  new_deaths  new_deaths_smoothed  …  \
0                     NaN           0.0         0.0                  NaN  …
1                     NaN           0.0         0.0                  NaN  …
2                     NaN           0.0         0.0                  NaN  …
3                     NaN           0.0         0.0                  NaN  …
4                     NaN           0.0         0.0                  NaN  …
…                     …             …           …                    …  …
45634                 NaN           7.0         NaN                  NaN  …
45635                 NaN           7.0         NaN                  NaN  …
45636                 NaN           7.0         NaN                  NaN  …
45637                 NaN           7.0         NaN                  NaN  …
45638                 NaN           7.0         NaN                  NaN  …

       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
…               …                  …                    …
45634             NaN              NaN                    NaN
45635             NaN              NaN                    NaN
45636             NaN              NaN                    NaN
45637             NaN              NaN                    NaN
45638             NaN              NaN                    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
…                     …                …             …
45634                  NaN             NaN           NaN
45635                  NaN             NaN           NaN
45636                  NaN             NaN           NaN
45637                  NaN             NaN           NaN
45638                  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
3                      37.746                         0.5            64.83
```

```
4                      37.746                        0.5            64.83
…                         …                          …               …
45634                    NaN                         NaN              NaN
45635                    NaN                         NaN              NaN
45636                    NaN                         NaN              NaN
45637                    NaN                         NaN              NaN
45638                    NaN                         NaN              NaN

       human_development_index
0                        0.498
1                        0.498
2                        0.498
3                        0.498
4                        0.498
…                          …
45634                    NaN
45635                    NaN
45636                    NaN
45637                    NaN
45638                    NaN

[45639 rows x 41 columns]
```

The dataset contains 45639 rows and 41 columns (in which some of the columns are optional).We need to reduce or optimize are the data to a bit.

Lets see the columns of the dataset

[10]:
```
survy_raw_df.columns
```

[10]:
```
Index(['iso_code', 'continent', 'location', 'date', 'total_cases', 'new_cases',
       'new_cases_smoothed', 'total_deaths', 'new_deaths',
       'new_deaths_smoothed', 'total_cases_per_million',
       'new_cases_per_million', 'new_cases_smoothed_per_million',
       'total_deaths_per_million', 'new_deaths_per_million',
       'new_deaths_smoothed_per_million', 'new_tests', 'total_tests',
       'total_tests_per_thousand', 'new_tests_per_thousand',
       'new_tests_smoothed', 'new_tests_smoothed_per_thousand',
       'tests_per_case', 'positive_rate', 'tests_units', 'stringency_index',
       'population', 'population_density', 'median_age', 'aged_65_older',
       'aged_70_older', 'gdp_per_capita', 'extreme_poverty',
       'cardiovasc_death_rate', 'diabetes_prevalence', 'female_smokers',
       'male_smokers', 'handwashing_facilities', 'hospital_beds_per_thousand',
       'life_expectancy', 'human_development_index'],
      dtype='object')
```

The dataset contains many rows and we need to sort them as our requirement for analysis.

### 1.3 Save the work

[11]:
```
import jovian
```

[13]:
```
jovian.commit(project=project_name)
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Attempting to save notebook..
[jovian] Updating notebook "jain-hrsh01/covid-data-analysis-project" on
https://jovian.ml/
[jovian] Uploading notebook..
[jovian] Capturing environment..
[jovian] Committed successfully! https://jovian.ml/jain-hrsh01/covid-data-
analysis-project
```

[13]:
```
'https://jovian.ml/jain-hrsh01/covid-data-analysis-project'
```

### 1.4 Data Preparation and Cleaning

While the survey data set contains many things we need to sort things out and keep our analysis for limited columns.

lets create a subset of the data for our analysis

[14]:
```
selected_columns = ['iso_code',
                    'continent',
                    'location',
                    'date',
                    'total_cases',
                    'total_deaths',
                    'new_cases',
                    'new_deaths',
                    'new_tests',
                    'total_tests',
                    'population',
                    'gdp_per_capita',
                    'cardiovasc_death_rate',
                    'diabetes_prevalence',
                    'male_smokers',
                   ␣
 ↪'female_smokers','hospital_beds_per_thousand','life_expectancy','positive_rate']
```

Let's extract a copy of the data from these columns into a new data frame covid_df, which we can continue to modify further without affecting the original data frame.

[15]:
```
len(selected_columns)
```

[15]:
```
19
```

[16]:
```
covid_df = survy_raw_df[selected_columns].copy()
```

[17]:
```
covid_df
```

[17]:

| | iso_code | continent | location | date | total_cases | \ |
|---|---|---|---|---|---|---|
| 0 | AFG | Asia | Afghanistan | 2019-12-31 | 0.0 | |
| 1 | AFG | Asia | Afghanistan | 2020-01-01 | 0.0 | |
| 2 | AFG | Asia | Afghanistan | 2020-01-02 | 0.0 | |
| 3 | AFG | Asia | Afghanistan | 2020-01-03 | 0.0 | |
| 4 | AFG | Asia | Afghanistan | 2020-01-04 | 0.0 | |
| … | … | … | … | … | … | |
| 45634 | NaN | NaN | International | 2020-09-19 | 696.0 | |
| 45635 | NaN | NaN | International | 2020-09-20 | 696.0 | |
| 45636 | NaN | NaN | International | 2020-09-21 | 696.0 | |
| 45637 | NaN | NaN | International | 2020-09-22 | 696.0 | |
| 45638 | NaN | NaN | International | 2020-09-23 | 696.0 | |

| | total_deaths | new_cases | new_deaths | new_tests | total_tests | \ |
|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | NaN | NaN | |
| 1 | 0.0 | 0.0 | 0.0 | NaN | NaN | |
| 2 | 0.0 | 0.0 | 0.0 | NaN | NaN | |
| 3 | 0.0 | 0.0 | 0.0 | NaN | NaN | |
| 4 | 0.0 | 0.0 | 0.0 | NaN | NaN | |
| … | … | … | … | … | … | |
| 45634 | 7.0 | NaN | NaN | NaN | NaN | |
| 45635 | 7.0 | NaN | NaN | NaN | NaN | |
| 45636 | 7.0 | NaN | NaN | NaN | NaN | |
| 45637 | 7.0 | NaN | NaN | NaN | NaN | |
| 45638 | 7.0 | NaN | NaN | NaN | NaN | |

| | population | gdp_per_capita | cardiovasc_death_rate | diabetes_prevalence | \ |
|---|---|---|---|---|---|
| 0 | 38928341.0 | 1803.987 | 597.029 | 9.59 | |
| 1 | 38928341.0 | 1803.987 | 597.029 | 9.59 | |
| 2 | 38928341.0 | 1803.987 | 597.029 | 9.59 | |
| 3 | 38928341.0 | 1803.987 | 597.029 | 9.59 | |
| 4 | 38928341.0 | 1803.987 | 597.029 | 9.59 | |
| … | … | … | … | … | |
| 45634 | NaN | NaN | NaN | NaN | |
| 45635 | NaN | NaN | NaN | NaN | |
| 45636 | NaN | NaN | NaN | NaN | |
| 45637 | NaN | NaN | NaN | NaN | |
| 45638 | NaN | NaN | NaN | NaN | |

| | male_smokers | female_smokers | hospital_beds_per_thousand | \ |
|---|---|---|---|---|
| 0 | NaN | NaN | 0.5 | |
| 1 | NaN | NaN | 0.5 | |
| 2 | NaN | NaN | 0.5 | |

```
3               NaN             NaN                          0.5
4               NaN             NaN                          0.5
…               …               …                            …
45634           NaN             NaN                          NaN
45635           NaN             NaN                          NaN
45636           NaN             NaN                          NaN
45637           NaN             NaN                          NaN
45638           NaN             NaN                          NaN

       life_expectancy  positive_rate
0                64.83            NaN
1                64.83            NaN
2                64.83            NaN
3                64.83            NaN
4                64.83            NaN
…                …                …
45634              NaN            NaN
45635              NaN            NaN
45636              NaN            NaN
45637              NaN            NaN
45638              NaN            NaN

[45639 rows x 19 columns]
```

Lets see some of the basic information about the dataframe.

[18]:
```
covid_df.shape
```

[18]:
```
(45639, 19)
```

The function shape is used to return the number of rows and columns there are in the dataframe and the output is basically the tuple.

[19]:
```
covid_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45639 entries, 0 to 45638
Data columns (total 19 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   iso_code                    45371 non-null  object
 1   continent                   45103 non-null  object
 2   location                    45639 non-null  object
 3   date                        45639 non-null  object
 4   total_cases                 45025 non-null  float64
 5   total_deaths                45025 non-null  float64
 6   new_cases                   44821 non-null  float64
 7   new_deaths                  44821 non-null  float64
```

```
 8   new_tests                   16212 non-null  float64
 9   total_tests                 16608 non-null  float64
 10  population                  45371 non-null  float64
 11  gdp_per_capita              40184 non-null  float64
 12  cardiovasc_death_rate       40714 non-null  float64
 13  diabetes_prevalence         42148 non-null  float64
 14  male_smokers                31522 non-null  float64
 15  female_smokers              31925 non-null  float64
 16  hospital_beds_per_thousand  36799 non-null  float64
 17  life_expectancy             44801 non-null  float64
 18  positive_rate               17111 non-null  float64
dtypes: float64(15), object(4)
memory usage: 6.6+ MB
```

It seems like the dataframe india.df contains many of the floating dtatype and also some of the empty cells NaN.

[20]:
```
covid_df.describe()
```

[20]:

| | total_cases | total_deaths | new_cases | new_deaths | new_tests |
|---|---|---|---|---|---|
| count | 4.502500e+04 | 45025.000000 | 44821.000000 | 44821.000000 | 1.621200e+04 |
| mean | 9.751751e+04 | 3983.616635 | 1412.666964 | 43.366681 | 2.239236e+04 |
| std | 1.009840e+06 | 37103.278332 | 12854.763120 | 366.241245 | 9.170192e+04 |
| min | 0.000000e+00 | 0.000000 | -8261.000000 | -1918.000000 | -3.743000e+03 |
| 25% | 5.700000e+01 | 1.000000 | 0.000000 | 0.000000 | 9.580000e+02 |
| 50% | 9.840000e+02 | 19.000000 | 10.000000 | 0.000000 | 3.191000e+03 |
| 75% | 1.043800e+04 | 225.000000 | 160.000000 | 3.000000 | 1.173925e+04 |
| max | 3.165857e+07 | 971869.000000 | 315329.000000 | 10491.000000 | 1.240829e+06 |

| | total_tests | population | gdp_per_capita | cardiovasc_death_rate |
|---|---|---|---|---|
| count | 1.660800e+04 | 4.537100e+04 | 40184.000000 | 40714.000000 |
| mean | 1.437130e+06 | 8.883081e+07 | 20952.291024 | 251.433578 |
| std | 6.443918e+06 | 6.145122e+08 | 20454.822744 | 117.585866 |
| min | 1.000000e+00 | 8.090000e+02 | 661.240000 | 79.370000 |
| 25% | 4.360500e+04 | 1.399491e+06 | 5338.454000 | 155.898000 |
| 50% | 1.843415e+05 | 8.654618e+06 | 14103.452000 | 238.339000 |
| 75% | 7.178142e+05 | 3.107294e+07 | 32415.132000 | 318.949000 |
| max | 1.061536e+08 | 7.794799e+09 | 116935.600000 | 724.417000 |

| | diabetes_prevalence | male_smokers | female_smokers |
|---|---|---|---|
| count | 42148.000000 | 31522.000000 | 31925.000000 |
| mean | 8.048859 | 32.635178 | 10.825012 |
| std | 4.144628 | 13.412935 | 10.480915 |
| min | 0.990000 | 7.700000 | 0.100000 |
| 25% | 5.310000 | 21.400000 | 1.900000 |
| 50% | 7.110000 | 31.400000 | 6.400000 |
| 75% | 10.180000 | 40.900000 | 19.600000 |
| max | 23.360000 | 78.100000 | 44.000000 |

```
       hospital_beds_per_thousand  life_expectancy  positive_rate
count                36799.000000     44801.000000   17111.000000
mean                     3.113499        74.042755       0.063543
std                      2.529090         7.367896       0.089373
min                      0.100000        53.280000       0.000000
25%                      1.300000        69.910000       0.008000
50%                      2.500000        75.490000       0.028000
75%                      4.200000        79.930000       0.081000
max                     13.800000        86.750000       0.651000
```

The describe() function lets us now the overall view of the dataframe and in this there were the negative values in some of the columns and we need to modify some of the data.

1. Replace the data with 0
2. Discard that column entirely
3. By mistakely the values where entered negative.
4. Take the average.

[21]:
```
covid = covid_df[covid_df['new_cases'] > 0]
covid = covid_df[covid_df['new_deaths'] > 0]
covid = covid_df[covid_df['new_tests'] > 0]
covid
```

[21]:
```
      iso_code      continent   location        date  total_cases  \
1490       ARG  South America  Argentina  2020-02-11          NaN
1499       ARG  South America  Argentina  2020-02-20          NaN
1504       ARG  South America  Argentina  2020-02-25          NaN
1506       ARG  South America  Argentina  2020-02-27          NaN
1507       ARG  South America  Argentina  2020-02-28          NaN
…          …              …          …           …           …
45095      ZWE         Africa   Zimbabwe  2020-09-16       7576.0
45096      ZWE         Africa   Zimbabwe  2020-09-17       7598.0
45097      ZWE         Africa   Zimbabwe  2020-09-18       7633.0
45098      ZWE         Africa   Zimbabwe  2020-09-19       7647.0
45099      ZWE         Africa   Zimbabwe  2020-09-20       7672.0

       total_deaths  new_cases  new_deaths  new_tests  total_tests  \
1490            NaN        NaN         NaN        2.0          2.0
1499            NaN        NaN         NaN        2.0          4.0
1504            NaN        NaN         NaN        1.0          5.0
1506            NaN        NaN         NaN        5.0         10.0
1507            NaN        NaN         NaN        9.0         19.0
…               …          …           …          …           …
45095         224.0       45.0         0.0      647.0     111421.0
45096         224.0       22.0         0.0     1011.0     112432.0
45097         224.0       35.0         0.0      755.0     113187.0
45098         224.0       14.0         0.0      539.0     113726.0
```

```
45099         225.0       25.0          1.0      628.0     114354.0

       population  gdp_per_capita  cardiovasc_death_rate  diabetes_prevalence  \
1490   45195777.0       18933.907                191.032                 5.50
1499   45195777.0       18933.907                191.032                 5.50
1504   45195777.0       18933.907                191.032                 5.50
1506   45195777.0       18933.907                191.032                 5.50
1507   45195777.0       18933.907                191.032                 5.50
…           …              …                      …                      …
45095  14862927.0        1899.775                307.846                 1.82
45096  14862927.0        1899.775                307.846                 1.82
45097  14862927.0        1899.775                307.846                 1.82
45098  14862927.0        1899.775                307.846                 1.82
45099  14862927.0        1899.775                307.846                 1.82

       male_smokers  female_smokers  hospital_beds_per_thousand  \
1490           27.7            16.2                         5.0
1499           27.7            16.2                         5.0
1504           27.7            16.2                         5.0
1506           27.7            16.2                         5.0
1507           27.7            16.2                         5.0
…              …               …                            …
45095          30.7             1.6                         1.7
45096          30.7             1.6                         1.7
45097          30.7             1.6                         1.7
45098          30.7             1.6                         1.7
45099          30.7             1.6                         1.7

       life_expectancy  positive_rate
1490             76.67            NaN
1499             76.67            NaN
1504             76.67            NaN
1506             76.67            NaN
1507             76.67            NaN
…                …              …
45095            61.49          0.038
45096            61.49          0.034
45097            61.49          0.038
45098            61.49          0.041
45099            61.49          0.038

[16207 rows x 19 columns]
```

In this block of cell we have discard the rows whose input are negative so that the min of overall data is cleaned and it comes to the positive value as the value can never be negative.

[22]:
```
import jovian
```

[23]:
```
jovian.commit(file = 'owid-covid-data.csv')
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Attempting to save notebook..
[jovian] Updating notebook "jain-hrsh01/covid-data-analysis-project" on
https://jovian.ml/
[jovian] Uploading notebook..
[jovian] Capturing environment..
[jovian] Committed successfully! https://jovian.ml/jain-hrsh01/covid-data-
analysis-project
```

[23]:
```
'https://jovian.ml/jain-hrsh01/covid-data-analysis-project'
```

### 1.5 Exploratory Analysis and Visualization

As the subset of the data is been taken and cleaned we will go further for the visualization of the data i.e. cases, deaths, tests through many different graphs and figure.

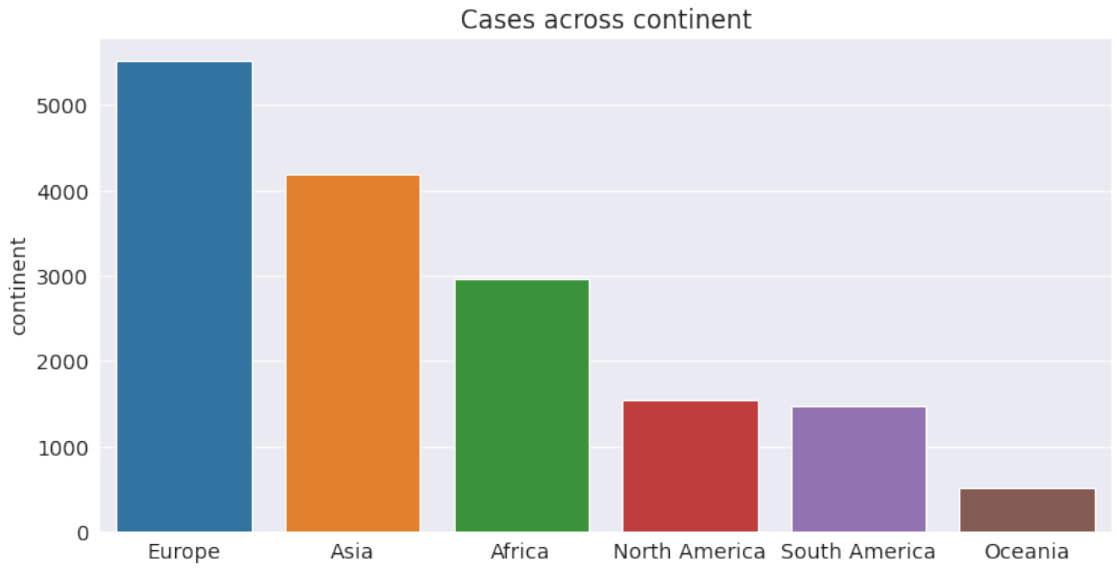
[24]:
```
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
%matplotlib inline

sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

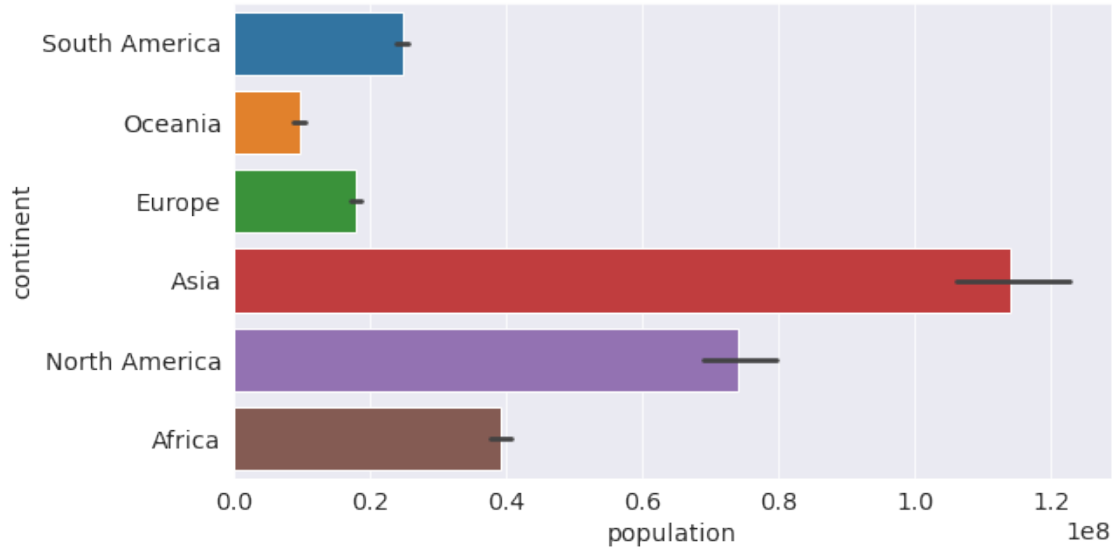
[25]:
```
top_continent = covid.continent.value_counts().head(10)
top_continent
```

[25]:
```
Europe           5520
Asia             4197
Africa           2965
North America    1549
South America    1465
Oceania           511
Name: continent, dtype: int64
```

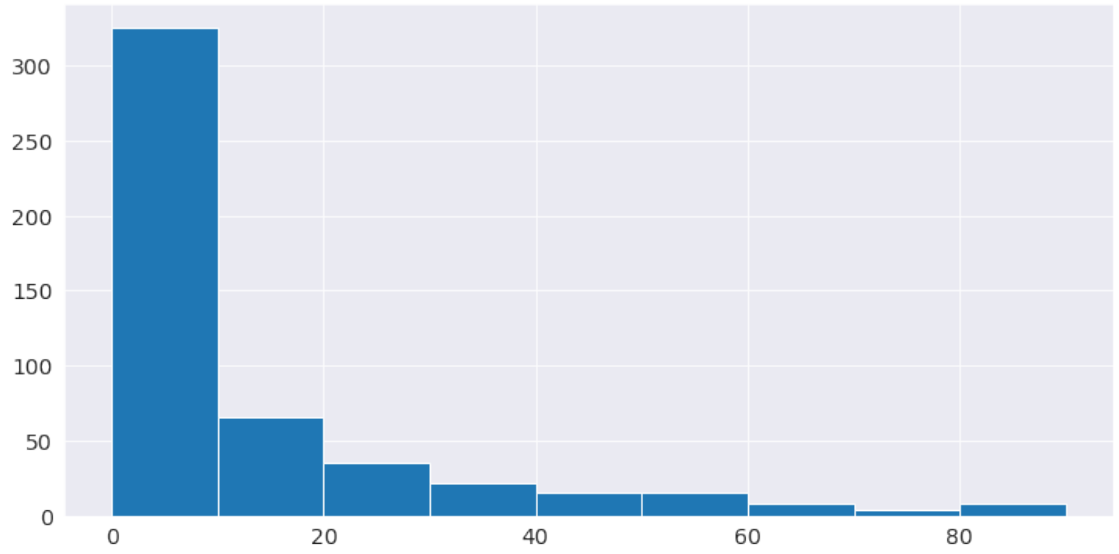
[26]:
```
plt.figure(figsize=(12,6))
plt.title('Cases across continent')
sns.barplot(top_continent.index,top_continent);
```



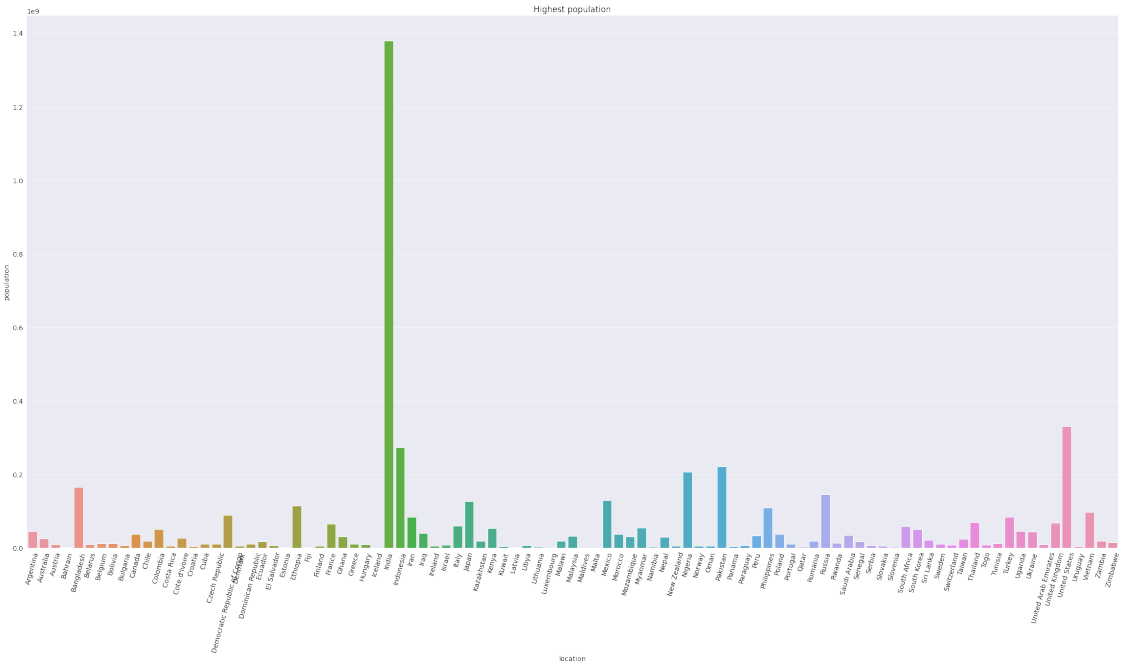

[27]:
```
sns.barplot(covid.population,covid.continent);
```




[28]:
```
plt.figure(figsize=(12,6))
gender_counts = covid.positive_rate.value_counts()
gender_counts
plt.hist(gender_counts,bins = np.arange(0,100,10));
```




[29]:
```
plt.figure(figsize=(40,20))
plt.title('Highest population')
plt.xticks(rotation=75)
sns.barplot(covid.location,covid.population);
```




[30]:
```
plt.figure(figsize=(12, 6))
plt.title('Sepal Dimensions')
```

```
sns.scatterplot(covid.cardiovasc_death_rate,
                covid.diabetes_prevalence,
                hue=covid.location,
                s=100);
```


Sepal Dimensions
diabetes_prevalence
cardiovasc_death_rate
location
Argentina
Australia
Austria
Bahrain
Bangladesh
Belarus
Belgium
Bolivia
Bulgaria
Canada
Chile
Colombia
Costa Rica
Cote d'Ivoire
Croatia
Cuba
Czech Republic
Democratic Republic of Congo
Denmark
Dominican Republic
Ecuador
El Salvador
Estonia
Ethiopia
Fiji
Finland
France
Ghana
Greece
Hungary
Iceland
India
Indonesia
Iran
Iraq
Ireland
Israel
Italy
Japan
Kazakhstan
Kenya
Kuwait
Latvia
Libya
Lithuania
Luxembourg
Malawi
Malaysia
Maldives
Malta
Mexico
Morocco
Mozambique
Myanmar
Namibia
Nepal
New Zealand
Nigeria
Norway
Oman
Pakistan
Panama
Paraguay
Peru
Philippines
Poland
Portugal
Qatar
Romania
Russia
Rwanda
Saudi Arabia
Senegal
Serbia
Slovakia
Slovenia
South Africa
South Korea
Sri Lanka
Sweden
Switzerland
Taiwan
Thailand
Togo
Tunisia
Turkey
Uganda
Ukraine
United Arab Emirates
United Kingdom
United States
Uruguay
Vietnam
Zambia
Zimbabwe

[31]:
```
import jovian
```

[32]:
```
jovian.commit()
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Attempting to save notebook..
[jovian] Updating notebook "jain-hrsh01/covid-data-analysis-project" on
https://jovian.ml/
[jovian] Uploading notebook..
[jovian] Capturing environment..
[jovian] Committed successfully! https://jovian.ml/jain-hrsh01/covid-data-
analysis-project
```

[32]:
```
'https://jovian.ml/jain-hrsh01/covid-data-analysis-project'
```

### 1.6 Asking and Answering Questions

TODO

Q. What is the total number of reported cases and deaths due to Covid-19 in Canada?

[33]:
```
cov = covid[covid['location']=='Canada'][['total_cases','total_deaths']]
cov
```

[33]:

| | total_cases | total_deaths |
|---|---|---|
| 7629 | 690.0 | 9.0 |
| 7630 | 846.0 | 10.0 |
| 7631 | 971.0 | 12.0 |
| 7632 | 1302.0 | 18.0 |
| 7633 | 1430.0 | 20.0 |
| … | … | … |
| 7811 | 139747.0 | 9193.0 |
| 7812 | 140867.0 | 9200.0 |
| 7813 | 141911.0 | 9205.0 |
| 7816 | 145415.0 | 9228.0 |
| 7817 | 146663.0 | 9234.0 |

[173 rows x 2 columns]

[34]:
```
total_cases = cov.sum()
total_cases
```

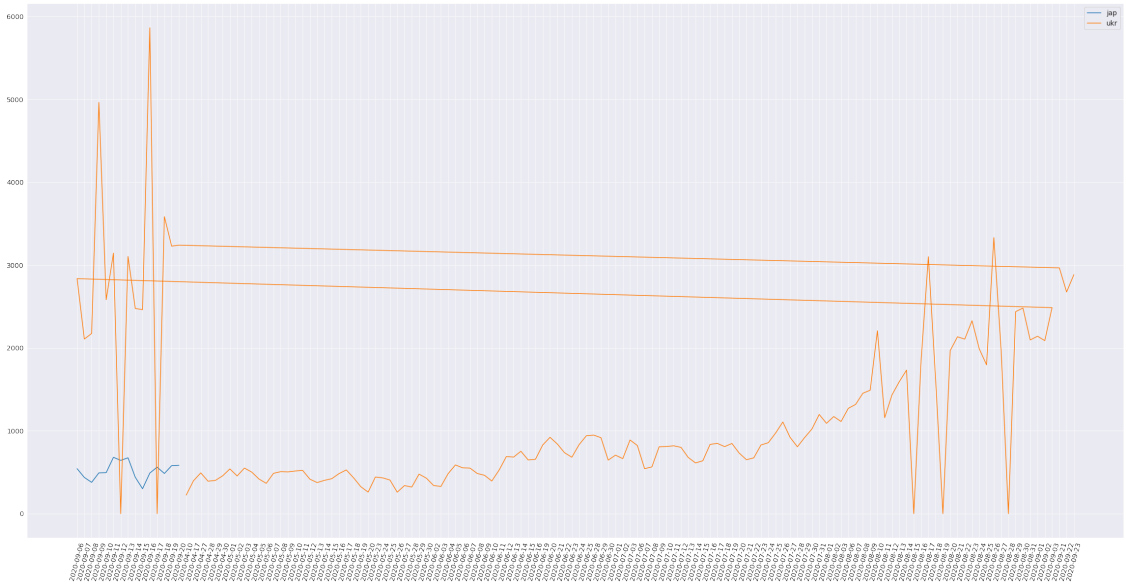
[34]:
```
total_cases     14807199.0
total_deaths     1093359.0
dtype: float64
```

Till now the cases and deaths of Canada is 14807199 and 1093359

Q. Analyis by plotting the new_cases of the location in japan and Ukraine?

[35]:
```
jap = covid[covid['location']=='Japan'][['date','new_cases']]
ukr = covid[covid['location']=='Ukraine'][['date','new_cases']]
jap = jap.set_index('date')
ukr = ukr.set_index('date')
```

[44]:
```
plt.figure(figsize=(40,20))
plt.plot(jap);
plt.plot(ukr);
plt.xticks(rotation=75);
plt.legend(['jap','ukr']);
```




From the above graph we can analysis that ukraine had more new cases in day to day basis rather than japan and also for the particular date we can see a higher peak in the graph.

Q.What is overall number of tests conducted in “India”?A total of 9325460 tests were conducted before daily test numbers were being reported?

[46]:
```
ind = covid[covid['location']=="India"]['new_tests']
ind
```

[46]:
```
19138      191.0
19139     1060.0
19140     1325.0
19141     1298.0
19142     3708.0
```

```
          ...
19321    1006615.0
19322     881911.0
19323    1206806.0
19324     731534.0
19325     933185.0
Name: new_tests, Length: 176, dtype: float64
```

[49]:
```
initial_tests = 9325460
total_tests = initial_tests + ind.sum()
print('The total number of test in India conducted till date is',total_tests)
```

```
The total number of test in India conducted till date is 74516351.0
```

Q.Analysis the number of new_cases,new_deaths,total_cases,total_deaths of the month July of Libya and also print the mean of the above data mentioned according to the weekday 'thursday'?

First to analysis the data we have to see the data column and also need to add the month column in the covid dataframe so to do that we will execute some command.

[52]:
```
covid['date'] = pd.to_datetime(covid.date);
covid['year'] = pd.DatetimeIndex(covid.date).year;
covid['month'] = pd.DatetimeIndex(covid.date).month;
covid['day'] = pd.DatetimeIndex(covid.date).day;
covid['weekday'] = pd.DatetimeIndex(covid.date).weekday;
```

```
<ipython-input-52-08bff3be7d29>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  covid['date'] = pd.to_datetime(covid.date);
<ipython-input-52-08bff3be7d29>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  covid['year'] = pd.DatetimeIndex(covid.date).year;
<ipython-input-52-08bff3be7d29>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  covid['month'] = pd.DatetimeIndex(covid.date).month;
<ipython-input-52-08bff3be7d29>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

```
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  covid['day'] = pd.DatetimeIndex(covid.date).day;
<ipython-input-52-08bff3be7d29>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  covid['weekday'] = pd.DatetimeIndex(covid.date).weekday;
```

After the follwoing execution of the command we see that 4 new columns are added in the datafame of the covid.

[51]:
```
covid
```

[51]:
```
      iso_code      continent   location       date  total_cases  \
1490       ARG  South America  Argentina 2020-02-11          NaN
1499       ARG  South America  Argentina 2020-02-20          NaN
1504       ARG  South America  Argentina 2020-02-25          NaN
1506       ARG  South America  Argentina 2020-02-27          NaN
1507       ARG  South America  Argentina 2020-02-28          NaN
…          …             …         …          …           …
45095      ZWE         Africa   Zimbabwe 2020-09-16       7576.0
45096      ZWE         Africa   Zimbabwe 2020-09-17       7598.0
45097      ZWE         Africa   Zimbabwe 2020-09-18       7633.0
45098      ZWE         Africa   Zimbabwe 2020-09-19       7647.0
45099      ZWE         Africa   Zimbabwe 2020-09-20       7672.0

       total_deaths  new_cases  new_deaths  new_tests  total_tests  …  \
1490            NaN        NaN         NaN        2.0          2.0  …
1499            NaN        NaN         NaN        2.0          4.0  …
1504            NaN        NaN         NaN        1.0          5.0  …
1506            NaN        NaN         NaN        5.0         10.0  …
1507            NaN        NaN         NaN        9.0         19.0  …
…               …          …           …          …           …  …
45095         224.0       45.0         0.0      647.0     111421.0  …
45096         224.0       22.0         0.0     1011.0     112432.0  …
45097         224.0       35.0         0.0      755.0     113187.0  …
45098         224.0       14.0         0.0      539.0     113726.0  …
45099         225.0       25.0         1.0      628.0     114354.0  …

       diabetes_prevalence  male_smokers  female_smokers  \
1490                  5.50          27.7            16.2
1499                  5.50          27.7            16.2
1504                  5.50          27.7            16.2
```

```
1506                  5.50          27.7            16.2
1507                  5.50          27.7            16.2
…                     …             …               …
45095                 1.82          30.7             1.6
45096                 1.82          30.7             1.6
45097                 1.82          30.7             1.6
45098                 1.82          30.7             1.6
45099                 1.82          30.7             1.6

       hospital_beds_per_thousand  life_expectancy  positive_rate  year  \
1490                          5.0            76.67            NaN  2020
1499                          5.0            76.67            NaN  2020
1504                          5.0            76.67            NaN  2020
1506                          5.0            76.67            NaN  2020
1507                          5.0            76.67            NaN  2020
…                             …              …                …    …
45095                         1.7            61.49          0.038  2020
45096                         1.7            61.49          0.034  2020
45097                         1.7            61.49          0.038  2020
45098                         1.7            61.49          0.041  2020
45099                         1.7            61.49          0.038  2020

       month  day  weekday
1490       2   11        1
1499       2   20        3
1504       2   25        1
1506       2   27        3
1507       2   28        4
…      …   …   …
45095      9   16        2
45096      9   17        3
45097      9   18        4
45098      9   19        5
45099      9   20        6

[16207 rows x 23 columns]
```

So, now for the given question we can analysis in the dataframe.

[58]:
```
li = covid[covid['location']=='Libya']
l = covid[covid['month']==7][['new_cases','new_deaths','total_cases','total_deaths']].sum()
l
```

[58]:
```
new_cases          5313085.0
new_deaths          119976.0
```

```
total_cases     313907604.0
total_deaths     13383689.0
dtype: float64
```

We also need to print the mean of the above.

[61]:
```
li[li.weekday == 4][['new_cases','new_deaths','total_cases','total_deaths']].
↪mean()
```

[61]:
```
new_cases         171.538462
new_deaths          2.730769
total_cases      4132.307692
total_deaths       73.807692
dtype: float64
```

Q.Print the total number of cases and deaths of all the country in monthwise for?

[64]:
```
covid_country = covid.groupby('month')[['total_cases','new_cases']].sum()
covid_country
```

[64]:

| month | total_cases | new_cases |
|---|---|---|
| 1 | 60.0 | 11.0 |
| 2 | 11935.0 | 2999.0 |
| 3 | 1937340.0 | 284306.0 |
| 4 | 38041462.0 | 1696348.0 |
| 5 | 107069097.0 | 2175430.0 |
| 6 | 181848811.0 | 3041195.0 |
| 7 | 313907604.0 | 5313085.0 |
| 8 | 494248820.0 | 6076404.0 |
| 9 | 421703350.0 | 4375590.0 |

[97]:
```
import jovian
```

[ ]:
```
jovian.commit()
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Attempting to save notebook..
```

### 1.7 Inferences and Conclusion

TODO

[ ]:

[ ]:

[ ]:

[11]:
```
import jovian
```

[12]:
```
jovian.commit()
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Attempting to save notebook..
[jovian] Updating notebook "aakashns/zerotopandas-course-project-starter" on
https://jovian.ml/
[jovian] Uploading notebook..
[jovian] Capturing environment..
[jovian] Committed successfully! https://jovian.ml/aakashns/zerotopandas-course-
project-starter
```

[12]:
```
'https://jovian.ml/aakashns/zerotopandas-course-project-starter'
```

### 1.8 References and Future Work

TODO

[13]:
```
import jovian
```

[ ]:
```
jovian.commit()
```

```
<IPython.core.display.Javascript object>
```

```
[jovian] Attempting to save notebook..
```

[ ]:
```

```