Exploratory Data Analysis (World population Analysis) IN PYTHON

Install all the libraries Pandas, matplotlib,seaborn

In [24]:
```
!pip install matplotlib
```

```
Requirement already satisfied: matplotlib in c:\users\joybose\anaconda2\lib\site-package
s (3.4.3)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\joybose\anaconda2\lib\site-p
ackages (from matplotlib) (3.0.4)
Requirement already satisfied: cycler>=0.10 in c:\users\joybose\anaconda2\lib\site-packa
ges (from matplotlib) (0.10.0)
Requirement already satisfied: pillow>=6.2.0 in c:\users\joybose\anaconda2\lib\site-pack
ages (from matplotlib) (8.4.0)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\joybose\anaconda2\lib\si
te-packages (from matplotlib) (2.8.2)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\joybose\anaconda2\lib\site-
packages (from matplotlib) (1.3.1)
Requirement already satisfied: numpy>=1.16 in c:\users\joybose\anaconda2\lib\site-packag
es (from matplotlib) (1.20.3)
Requirement already satisfied: six in c:\users\joybose\anaconda2\lib\site-packages (from
cycler>=0.10->matplotlib) (1.16.0)
```

In [46]:
```
from pylab import *
```

In [25]:
```
import pandas as pd
import seaborn as sns
import matplotlib as plt
```

In [26]:
```
import matplotlib as plt
```

Read the data from the files and get the values into a pandas dataframe

In [3]:
```
df = pd.read_csv(r"C:/Users/joybose/Downloads\world_population.csv")

df
```

Out[3]:

	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	2015 Population	2010 Population	Pop
0	36	AFG	Afghanistan	Kabul	Asia	41128771.0	38972230.0	33753499.0	28189672.0	195

	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	2015 Population	2010 Population	Pop
1	138	ALB	Albania	Tirana	Europe	2842321.0	2866849.0	2882481.0	2913399.0	31
2	34	DZA	Algeria	Algiers	Africa	44903225.0	43451666.0	39543154.0	35856344.0	307
3	213	ASM	American Samoa	Pago Pago	Oceania	44273.0	46189.0	51368.0	54849.0	
4	203	AND	Andorra	Andorra la Vella	Europe	79824.0	77700.0	71746.0	71519.0	
...	...	...	...	...	...	...	...	...	...	
229	226	WLF	Wallis and Futuna	Mata-Utu	Oceania	11572.0	11655.0	12182.0	13142.0	
230	172	ESH	Western Sahara	El Aaiún	Africa	575986.0	556048.0	491824.0	413296.0	2
231	46	YEM	Yemen	Sanaa	Asia	33696614.0	32284046.0	28516545.0	24743946.0	186
232	63	ZMB	Zambia	Lusaka	Africa	20017675.0	18927715.0	NaN	13792086.0	98
233	74	ZWE	Zimbabwe	Harare	Africa	16320537.0	15669666.0	14154937.0	12839771.0	118

234 rows × 17 columns

Format the data to remove the scientific notation

In [4]:
```
pd.set_option('display.float_format',lambda x: '%.2f' %x)
```

Take a look at the information in the dataframe

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 234 entries, 0 to 233
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   Rank                        234 non-null    int64  
 1   CCA3                        234 non-null    object 
 2   Country                     234 non-null    object 
 3   Capital                     234 non-null    object 
 4   Continent                   234 non-null    object 
 5   2022 Population             230 non-null    float64
 6   2020 Population             233 non-null    float64
 7   2015 Population             230 non-null    float64
 8   2010 Population             227 non-null    float64
```

```
 9   2000 Population              227 non-null    float64
 10  1990 Population              229 non-null    float64
 11  1980 Population              229 non-null    float64
 12  1970 Population              230 non-null    float64
 13  Area (km²)                   232 non-null    float64
 14  Density (per km²)            230 non-null    float64
 15  Growth Rate                  232 non-null    float64
 16  World Population Percentage  234 non-null    float64
dtypes: float64(12), int64(1), object(4)
memory usage: 27.5+ KB
```

Describe the contents inside the dataframe parameters like count,mean,minimum,maximum

In [6]:
```
df.describe()
```

Out[6]:

	Rank	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	199(Populatio
count	234.00	230.00	233.00	230.00	227.00	227.00	229.0(
mean	117.50	34632250.88	33600710.95	32066004.16	30270164.48	26840495.26	19330463.9
std	67.69	137889172.44	135873196.61	131507146.34	126074183.54	113352454.57	81309624.9(
min	1.00	510.00	520.00	564.00	596.00	651.00	700.0(
25%	59.25	419738.50	406471.00	394295.00	382726.50	329470.00	261928.0(
50%	117.50	5762857.00	5456681.00	5244415.00	4889741.00	4491202.00	3785847.0(
75%	175.75	22653719.00	21522626.00	19730853.75	16825852.50	15625467.00	11882762.0(
max	234.00	1425887337.00	1424929781.00	1393715448.00	1348191368.00	1264099069.00	1153704252.0(

Count the amount of null values in each column

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
Rank                        0
CCA3                        0
Country                     0
Capital                     0
Continent                   0
2022 Population             4
2020 Population             1
2015 Population             4
2010 Population             7
2000 Population             7
1990 Population             5
```

```
1980 Population               5
1970 Population               4
Area (km²)                    2
Density (per km²)             4
Growth Rate                   2
World Population Percentage   0
dtype: int64
```

Identify the amount of unique values in each column

In [13]:
```
df.nunique()
```

Out[13]:
```
Rank                          234
CCA3                          234
Country                       234
Capital                       234
Continent                       6
2022 Population               230
2020 Population               233
2015 Population               230
2010 Population               227
2000 Population               227
1990 Population               229
1980 Population               229
1970 Population               230
Area (km²)                    231
Density (per km²)             230
Growth Rate                   178
World Population Percentage    70
dtype: int64
```

Sort the values from largest to smallest based on a column and view only the top 10

In [16]:
```
df.sort_values(by="2022 Population",ascending=False).head(10)
```

Out[16]:

	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	2015 Population	
41	1	CHN	China	Beijing	Asia	1425887337.00	1424929781.00	1393715448.00	134
92	2	IND	India	New Delhi	Asia	1417173173.00	1396387127.00	1322866505.00	124
221	3	USA	United States	Washington, D.C.	North America	338289857.00	335942003.00	324607776.00	31
93	4	IDN	Indonesia	Jakarta	Asia	275501339.00	271857970.00	259091970.00	24
156	5	PAK	Pakistan	Islamabad	Asia	235824862.00	227196741.00	210969298.00	19
149	6	NGA	Nigeria	Abuja	Africa	218541212.00	208327405.00	183995785.00	16

	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	2015 Population	
27	7	BRA	Brazil	Brasilia	South America	215313498.00	213196304.00	205188205.00	19
16	8	BGD	Bangladesh	Dhaka	Asia	171186372.00	167420951.00	157830000.00	14
171	9	RUS	Russia	Moscow	Europe	144713314.00	145617329.00	144668389.00	14
131	10	MEX	Mexico	Mexico City	North America	127504125.00	125998302.00	120149897.00	11

View the columns with only data types as the number

In [51]:
```
df.select_dtypes(include='number')
```

Out[51]:

	Rank	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	1990 Population	1980 Population
0	36	41128771.00	38972230.00	33753499.00	28189672.00	19542982.00	10694796.00	12486631.00
1	138	2842321.00	2866849.00	2882481.00	2913399.00	3182021.00	3295066.00	2941651.00
2	34	44903225.00	43451666.00	39543154.00	35856344.00	30774621.00	25518074.00	18739378.00
3	213	44273.00	46189.00	51368.00	54849.00	58230.00	47818.00	32886.00
4	203	79824.00	77700.00	71746.00	71519.00	66097.00	53569.00	35611.00
...	...	...	...	...	...	...	...	...
229	226	11572.00	11655.00	12182.00	13142.00	14723.00	13454.00	11315.00
230	172	575986.00	556048.00	491824.00	413296.00	270375.00	178529.00	116775.00
231	46	33696614.00	32284046.00	28516545.00	24743946.00	18628700.00	13375121.00	9204938.00
232	63	20017675.00	18927715.00	NaN	13792086.00	9891136.00	7686401.00	5720438.00
233	74	16320537.00	15669666.00	14154937.00	12839771.00	11834676.00	10113893.00	7049926.00

234 rows × 13 columns

View the columns with only data types as Object

In [52]:
```
df.select_dtypes(include='object')
```

Out[52]:

	CCA3	Country	Capital	Continent
0	AFG	Afghanistan	Kabul	Asia
1	ALB	Albania	Tirana	Europe
2	DZA	Algeria	Algiers	Africa
3	ASM	American Samoa	Pago Pago	Oceania
4	AND	Andorra	Andorra la Vella	Europe
...	...	...	...	...
229	WLF	Wallis and Futuna	Mata-Utu	Oceania
230	ESH	Western Sahara	El Aaiún	Africa
231	YEM	Yemen	Sanaa	Asia
232	ZMB	Zambia	Lusaka	Africa
233	ZWE	Zimbabwe	Harare	Africa

234 rows × 4 columns

Find the correlation between the column values

In [17]:
```
df.corr()
```

Out[17]:

	Rank	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	1990 Population	1980 Population
Rank	1.00	-0.36	-0.36	-0.35	-0.35	-0.34	-0.33	-0.33
2022 Population	-0.36	1.00	1.00	1.00	1.00	0.99	0.99	0.99
2020 Population	-0.36	1.00	1.00	1.00	1.00	1.00	0.99	0.99
2015 Population	-0.35	1.00	1.00	1.00	1.00	1.00	0.99	0.99
2010 Population	-0.35	1.00	1.00	1.00	1.00	1.00	1.00	0.99
2000 Population	-0.34	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1990 Population	-0.33	0.99	0.99	0.99	1.00	1.00	1.00	1.00
1980 Population	-0.33	0.99	0.99	0.99	0.99	1.00	1.00	1.00
1970 Population	-0.34	0.97	0.98	0.98	0.98	0.99	1.00	1.00

	Rank	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	1990 Population	1980 Population
Area (km²)	-0.38	0.45	0.45	0.46	0.46	0.47	0.52	0.53
Density (per km²)	0.13	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
Growth Rate	-0.22	-0.02	-0.03	-0.03	-0.04	-0.05	-0.07	-0.08
World Population Percentage	-0.36	1.00	1.00	1.00	1.00	0.99	0.99	0.99

Plot the correlation graph into a HeatMap

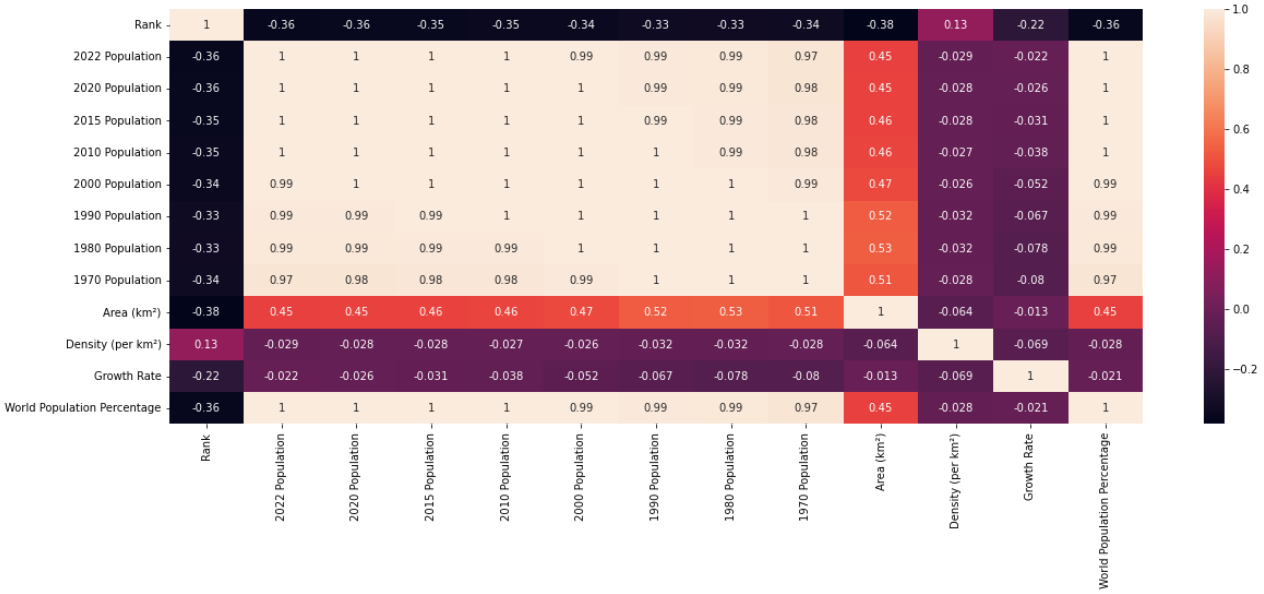
In [27]:
```
sns.heatmap(df.corr(), annot=True)

plt.rcParams['figure.figsize']=(20,7)

plt.show()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_7592/3950113531.py in <module>
      3 plt.rcParams['figure.figsize']=(20,7)
      4
----> 5 plt.show()

AttributeError: module 'matplotlib' has no attribute 'show'
```



Analyze by grouping the columns into specific category

In [33]:
```
df.groupby('Continent').mean().sort_values(by='2022 Population', ascending=False)
```

Out[33]:

Continent	Rank	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	1990 Population	Popul
Asia	77.56	96327387.31	94955134.37	89165003.64	89087770.00	80580835.11	48639995.33	402783
South America	97.57	31201186.29	30823574.50	29509599.71	26789395.54	25015888.69	21224743.93	172706
Africa	92.16	25455879.68	23871435.26	21419703.57	18898197.31	14598365.95	11376964.52	85860
Europe	124.50	15055371.82	14915843.92	15027454.12	14712278.68	14817685.71	14785203.94	142000
North America	160.93	15007403.40	14855914.82	14259596.25	13568016.28	12151739.60	10531660.62	92073
Oceania	188.52	2046386.32	1910148.96	1756664.48	1613163.65	1357512.09	1162774.87	9965

Deeper look at a specific value in a column

In [31]:
```
df[df['Continent'].str.contains('Oceania')]
```

Out[31]:

	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	2015 Population	2010 Population
3	213	ASM	American Samoa	Pago Pago	Oceania	44273.00	46189.00	51368.00	54849.00
11	55	AUS	Australia	Canberra	Oceania	26177413.00	25670051.00	23820236.00	22019168.00
44	223	COK	Cook Islands	Avarua	Oceania	17011.00	17029.00	17695.00	17212.00
66	162	FJI	Fiji	Suva	Oceania	929766.00	920422.00	917200.00	905169.00
70	183	PYF	French Polynesia	Papeete	Oceania	306279.00	301920.00	291787.00	283788.00
81	191	GUM	Guam	Hagåtña	Oceania	171774.00	169231.00	167978.00	164905.00
107	192	KIR	Kiribati	Tarawa	Oceania	131232.00	126463.00	116707.00	107995.00
126	215	MHL	Marshall Islands	Majuro	Oceania	41569.00	43413.00	49410.00	53416.00
132	194	FSM	Micronesia	Palikir	Oceania	114164.00	112106.00	109462.00	107588.00
142	225	NRU	Nauru	Yaren	Oceania	12668.00	12315.00	11185.00	10241.00
145	185	NCL	New Caledonia	Nouméa	Oceania	289950.00	286403.00	283032.00	261426.00
146	123	NZL	New	Wellington	Oceania	5185288.00	5061133.00	4590590.00	4346338.00

	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	2015 Population	2010 Population
			Zealand						
150	232	NIU	Niue	Alofi	Oceania	1934.00	1942.00	1847.00	1812.00
153	210	NFK	Northern Mariana Islands	Saipan	Oceania	49551.00	49587.00	51514.00	54087.00
157	222	PLW	Palau	Ngerulmud	Oceania	NaN	17972.00	17794.00	18540.00
160	93	PNG	Papua New Guinea	Port Moresby	Oceania	10142619.00	9749640.00	8682174.00	7583269.00
179	188	WSM	Samoa	Apia	Oceania	222382.00	214929.00	203571.00	194672.00
191	166	SLB	Solomon Islands	Honiara	Oceania	724273.00	691191.00	612660.00	540394.00
209	233	TKL	Tokelau	Nukunonu	Oceania	1871.00	1827.00	1454.00	1367.00
210	197	TON	Tonga	Nuku'alofa	Oceania	106858.00	105254.00	106122.00	107383.00
216	227	TUV	Tuvalu	Funafuti	Oceania	11312.00	11069.00	10877.00	10550.00
225	181	VUT	Vanuatu	Port-Vila	Oceania	326740.00	311685.00	276438.00	245453.00
229	226	WLF	Wallis and Futuna	Mata-Utu	Oceania	11572.00	11655.00	12182.00	13142.00

In [39]:
```
df2=df.groupby('Continent')[df.columns[5:13]].mean().sort_values(by='2022 Population',

df2
```

Out[39]:

Continent	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	1990 Population	1980 Population	
Asia	96327387.31	94955134.37	89165003.64	89087770.00	80580835.11	48639995.33	40278333.33	4
South America	31201186.29	30823574.50	29509599.71	26789395.54	25015888.69	21224743.93	17270643.29	1
Africa	25455879.68	23871435.26	21419703.57	18898197.31	14598365.95	11376964.52	8586031.98	
Europe	15055371.82	14915843.92	15027454.12	14712278.68	14817685.71	14785203.94	14200004.52	1
North America	15007403.40	14855914.82	14259596.25	13568016.28	12151739.60	10531660.62	9207334.03	
Oceania	2046386.32	1910148.96	1756664.48	1613163.65	1357512.09	1162774.87	996532.17	

Use transpose

In [40]:
```
df3=df2.transpose()

df3
```

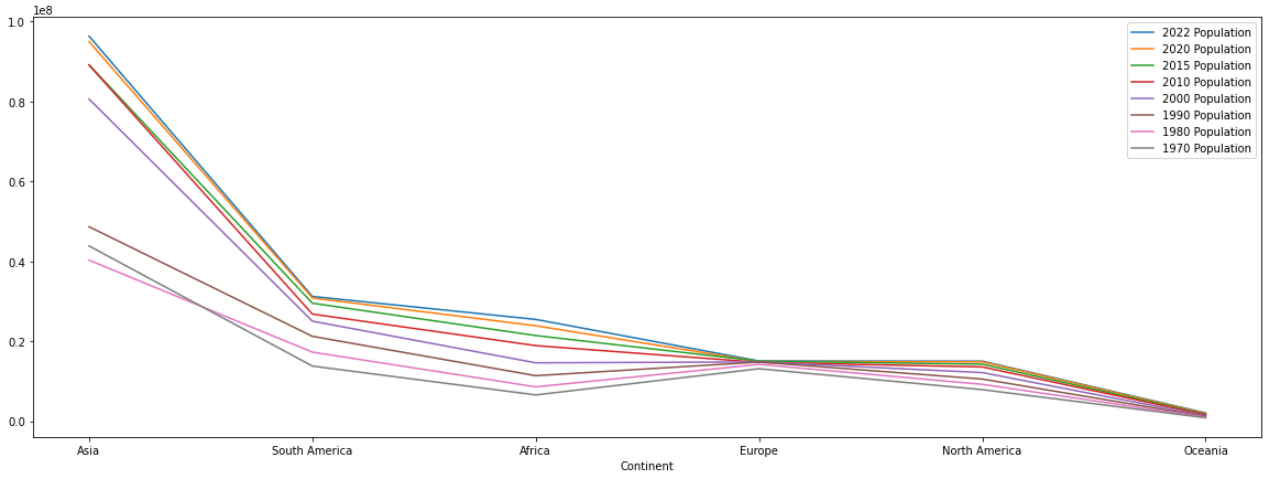
Out[40]:

Continent	Asia	South America	Africa	Europe	North America	Oceania
2022 Population	96327387.31	31201186.29	25455879.68	15055371.82	15007403.40	2046386.32
2020 Population	94955134.37	30823574.50	23871435.26	14915843.92	14855914.82	1910148.96
2015 Population	89165003.64	29509599.71	21419703.57	15027454.12	14259596.25	1756664.48
2010 Population	89087770.00	26789395.54	18898197.31	14712278.68	13568016.28	1613163.65
2000 Population	80580835.11	25015888.69	14598365.95	14817685.71	12151739.60	1357512.09
1990 Population	48639995.33	21224743.93	11376964.52	14785203.94	10531660.62	1162774.87
1980 Population	40278333.33	17270643.29	8586031.98	14200004.52	9207334.03	996532.17
1970 Population	43839877.83	13781939.71	6567175.27	13118479.82	7885865.15	846968.26

Plot the initial dataframe

In [41]:
```
df2.plot()
```

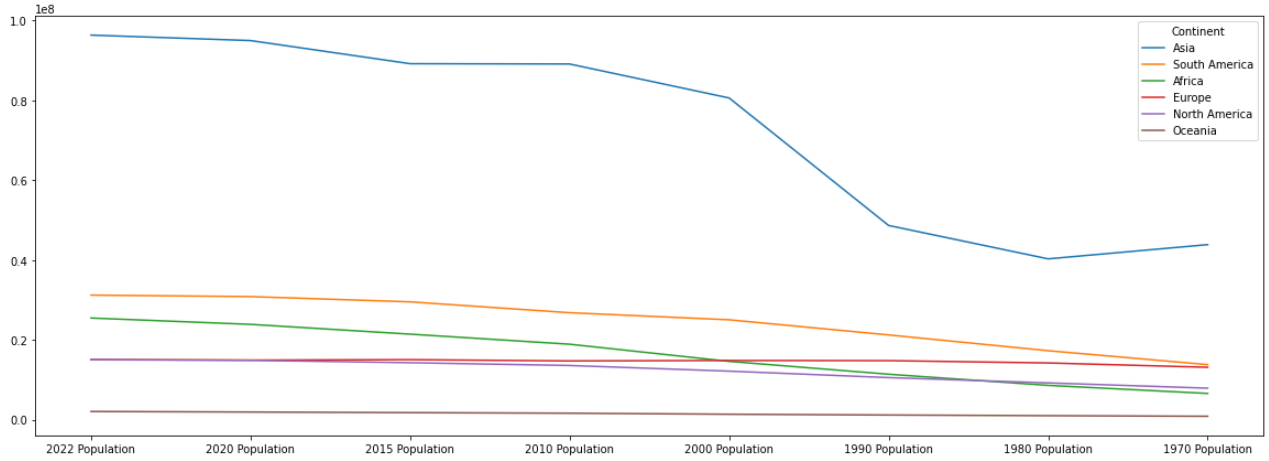
Out[41]:
```
<AxesSubplot:xlabel='Continent'>
```



Plot the transpose dataframe

In [42]:
```
df3.plot()
```

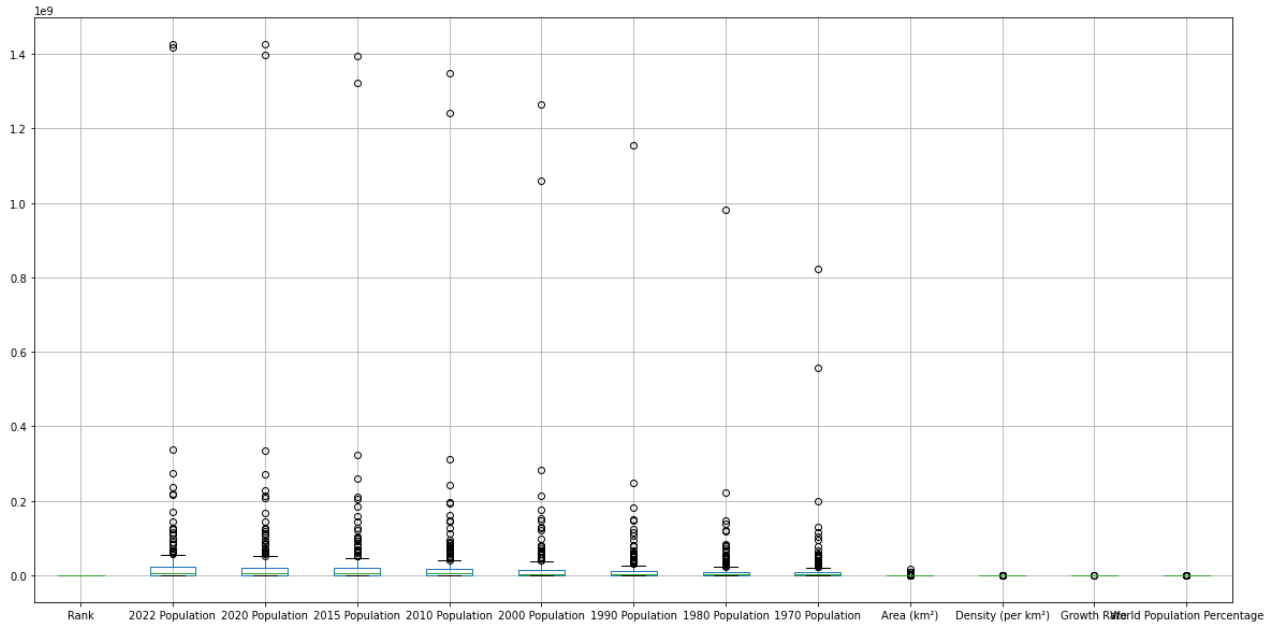
Out[42]:
```
<AxesSubplot:>
```



Plot a BoxPlot to know about the Outliers

In [48]:
```
df.boxplot(figsize=(20,10))
```

Out[48]: <AxesSubplot:>



In [ ]: