EDA in Pandas

In [3]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:
```
df=pd.read_csv(r"D:\COURSES\YOUTUBE\ALEX THE ANALYST\PYTHON\world_population2.c
```

In [5]:
```
df
```

Out[5]:

	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	2015 Population	2 Popula
0	36	AFG	Afghanistan	Kabul	Asia	41128771.0	38972230.0	33753499.0	2818967
1	138	ALB	Albania	Tirana	Europe	2842321.0	2866849.0	2882481.0	291339
2	34	DZA	Algeria	Algiers	Africa	44903225.0	43451666.0	39543154.0	3585634
3	213	ASM	American Samoa	Pago Pago	Oceania	44273.0	46189.0	51368.0	5484
4	203	AND	Andorra	Andorra la Vella	Europe	79824.0	77700.0	71746.0	7151
...	...	...	...	...	...	...	...	...	
229	226	WLF	Wallis and Futuna	Mata-Utu	Oceania	11572.0	11655.0	12182.0	1314
230	172	ESH	Western	El	Africa	575986.0	556048.0	491824.0	41329

In [49]:
```
pd.set_option('display.float_format', lambda x: '%.2f' % x)
```

In [50]:
```
df
```

Out[50]:

	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	2015 Population	Popu
0	36	AFG	Afghanistan	Kabul	Asia	41128771.00	38972230.00	33753499.00	281896
1	138	ALB	Albania	Tirana	Europe	2842321.00	2866849.00	2882481.00	29133
2	34	DZA	Algeria	Algiers	Africa	44903225.00	43451666.00	39543154.00	358563
3	213	ASM	American Samoa	Pago Pago	Oceania	44273.00	46189.00	51368.00	548
4	203	AND	Andorra	Andorra la Vella	Europe	79824.00	77700.00	71746.00	715
...	...	...	...	...	...	...	...	...	
229	226	WLF	Wallis and Futuna	Mata-Utu	Oceania	11572.00	11655.00	12182.00	131
230	172	ESH	Western Sahara	El Aaiún	Africa	575986.00	556048.00	491824.00	4132
231	46	YEM	Yemen	Sanaa	Asia	33696614.00	32284046.00	28516545.00	247439
232	63	ZMB	Zambia	Lusaka	Africa	20017675.00	18927715.00	NaN	137920
233	74	ZWE	Zimbabwe	Harare	Africa	16320537.00	15669666.00	14154937.00	128397

234 rows × 17 columns

In [51]:
```
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: 31.2+ KB
```

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

Out[8]:

	Rank	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	P
count	234.00	230.00	233.00	230.00	227.00	227.00	
mean	117.50	34632250.88	33600710.95	32066004.16	30270164.48	26840495.26	193
std	67.69	137889172.44	135873196.61	131507146.34	126074183.54	113352454.57	813
min	1.00	510.00	520.00	564.00	596.00	651.00	
25%	59.25	419738.50	406471.00	394295.00	382726.50	329470.00	2
50%	117.50	5762857.00	5456681.00	5244415.00	4889741.00	4491202.00	37
75%	175.75	22653719.00	21522626.00	19730853.75	16825852.50	15625467.00	118
max	234.00	1425887337.00	1424929781.00	1393715448.00	1348191368.00	1264099069.00	11537

In [52]:
```
df.isnull().sum()
```

Out[52]:
```
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
```

In [53]:
```
df.nunique()
```

Out[53]:
```
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
```

In [54]:
```
df.sort_values(by='World Population Percentage', ascending=False).head(10)
```

Out[54]:

	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	20 Populati
41	1	CHN	China	Beijing	Asia	1425887337.00	1424929781.00	1393715448.
92	2	IND	India	New Delhi	Asia	1417173173.00	1396387127.00	1322866505.
221	3	USA	United States	Washington, D.C.	North America	338289857.00	335942003.00	324607776.
93	4	IDN	Indonesia	Jakarta	Asia	275501339.00	271857970.00	259091970.
156	5	PAK	Pakistan	Islamabad	Asia	235824862.00	227196741.00	210969298.
149	6	NGA	Nigeria	Abuja	Africa	218541212.00	208327405.00	183995785.
27	7	BRA	Brazil	Brasilia	South America	215313498.00	213196304.00	205188205.
16	8	BGD	Bangladesh	Dhaka	Asia	171186372.00	167420951.00	157830000.
171	9	RUS	Russia	Moscow	Europe	144713314.00	145617329.00	144668389.
131	10	MEX	Mexico	Mexico City	North America	127504125.00	125998302.00	120149897.

In [12]:
```
numeric_df=df.select_dtypes(include=['float64','int64'])
```

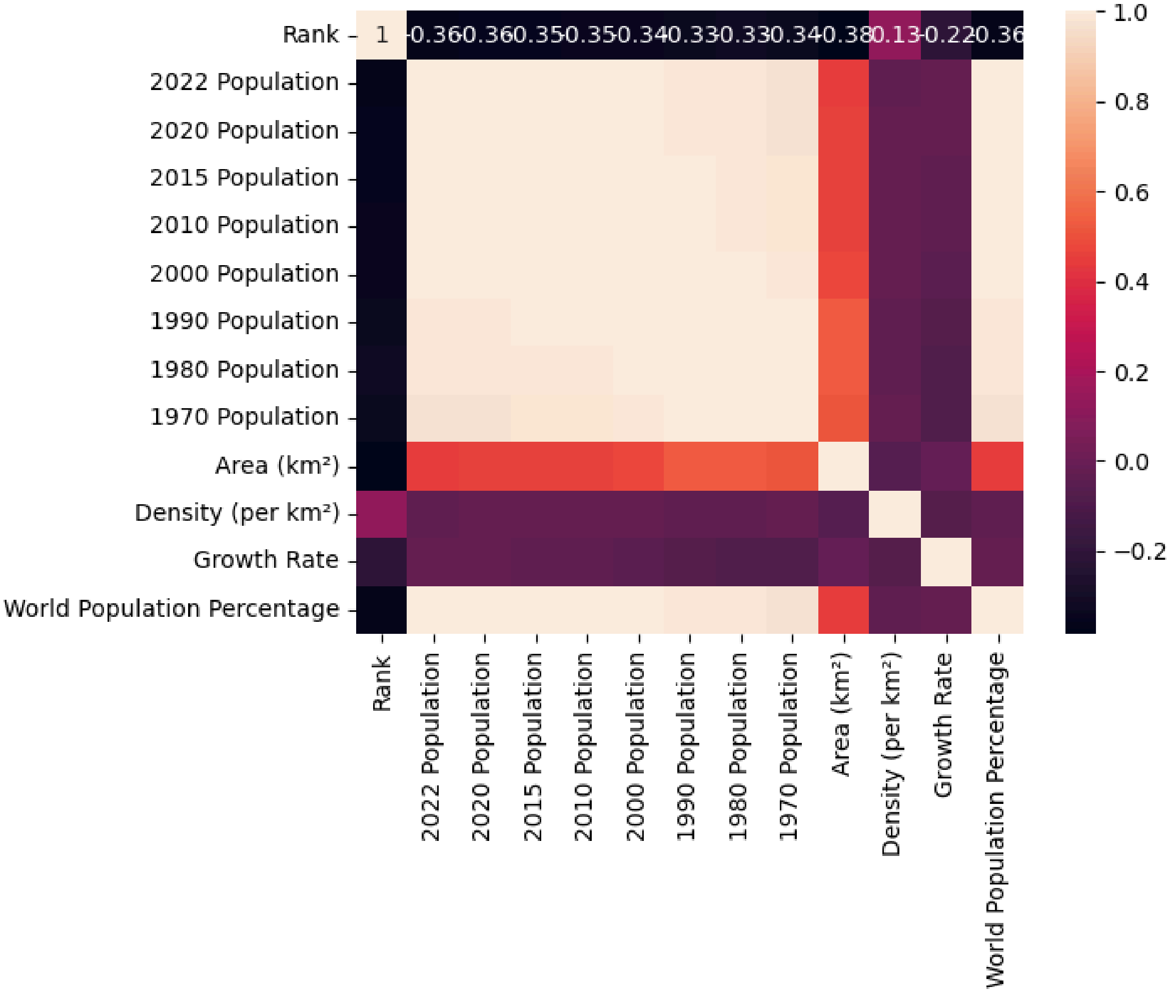
In [13]:

```
numeric_df.corr()
```

Out[13]:

	Rank	2022 Population	2020 Population	2015 Population	2010 Population	2000 Population	1990 Population	Popula
Rank	1.00	-0.36	-0.36	-0.35	-0.35	-0.34	-0.33	-
2022 Population	-0.36	1.00	1.00	1.00	1.00	0.99	0.99	
2020 Population	-0.36	1.00	1.00	1.00	1.00	1.00	0.99	
2015 Population	-0.35	1.00	1.00	1.00	1.00	1.00	0.99	
2010 Population	-0.35	1.00	1.00	1.00	1.00	1.00	1.00	
2000 Population	-0.34	0.99	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	
1980 Population	-0.33	0.99	0.99	0.99	0.99	1.00	1.00	
1970 Population	-0.34	0.97	0.98	0.98	0.98	0.99	1.00	
Area (km²)	-0.38	0.45	0.45	0.46	0.46	0.47	0.52	
Density (per km²)	0.13	-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	-
World Population Percentage	-0.36	1.00	1.00	1.00	1.00	0.99	0.99	

In [14]:
```
sns.heatmap(numeric_df.corr(), annot=True)
plt.rcParams['figure.figsize']=(20,10)
plt.show()
```



In [55]:
```
df
```

Out[55]:

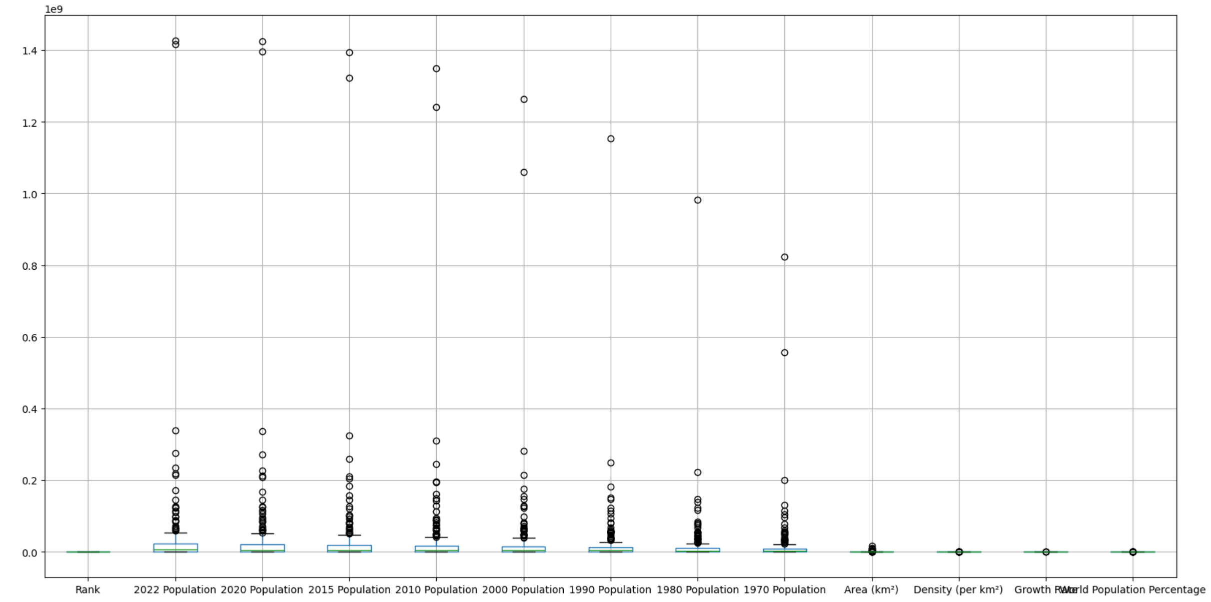
	Rank	CCA3	Country	Capital	Continent	2022 Population	2020 Population	2015 Population	Popu
0	36	AFG	Afghanistan	Kabul	Asia	41128771.00	38972230.00	33753499.00	281896
1	138	ALB	Albania	Tirana	Europe	2842321.00	2866849.00	2882481.00	29133
2	34	DZA	Algeria	Algiers	Africa	44903225.00	43451666.00	39543154.00	358563
3	213	ASM	American Samoa	Pago Pago	Oceania	44273.00	46189.00	51368.00	548
4	203	AND	Andorra	Andorra la Vella	Europe	79824.00	77700.00	71746.00	715
...	...	...	...	...	...	...	...	...	
229	226	WLF	Wallis and Futuna	Mata-Utu	Oceania	11572.00	11655.00	12182.00	131
230	172	ESH	Western Sahara	El Aaiún	Africa	575986.00	556048.00	491824.00	4132
231	46	YEM	Yemen	Sanaa	Asia	33696614.00	32284046.00	28516545.00	247439
232	63	ZMB	Zambia	Lusaka	Africa	20017675.00	18927715.00	NaN	137920
233	74	ZWE	Zimbabwe	Harare	Africa	16320537.00	15669666.00	14154937.00	128397

234 rows × 17 columns

In [1]:
```
#for groupby queries and visualization watch the video
```

In [6]:
```
df.boxplot(figsize=(20,10))
```

Out[6]:
```
<Axes: >
```



In [7]:

```
#df.select_dtypes(include='number/object/float')
```

In [ ]: