EXPLORATORY DATA ANALYSIS AND VISUALIZATION

Information about the dataset:

- How many rows and attributes?

DATASET-1(GDP):

Rows:2248 entries
Cols:3

#	Column	Non-Null Count	Dtype
1	LOCATION	2248 non-null	object
2	TIME	2248 non-null	int64
3	Value	2248 non-null	float64

DATASET-2(Unemployment):

Rows: 931
Columns: 3

#	Column	Non-Null Count	Dtype
1	LOCATION	931 non-null	object
2	TIME	931 non-null	int64
3	Value	931 non-null	float64

dtypes: float64(1), int64(1), object(1)

- How many missing data?

No missing data.

- Any inconsistent, incomplete, duplicate, or incorrect data?

In the Dataset consisting of GDP rates, we have dropped a few rows consisting of a specific way to record GDP values as it did not prove to be helpful while analyzing the data.
All the records with the SUBJECT field equal to 'VOLIDX' were dropped.

For GDP:

The GDP vs Time dataset had rates instead of absolute values, the line graph wasn't showing a specific trend. Thus the cumulative GDP rates were considered for the line graph.

- Are the variables correlated to each other?

In dataset one, we considered the cumulative values of the GDP rates and found the Pearson correlation between year and GDP rates. The attributes were highly correlated.
In dataset two, we found the Pearson correlation between year and Unemployment rates. The attributes were correlated to a few countries. (more under insights)

- Are any of the preprocessing techniques needed: dimensionality reduction, range transformation, standardization, etc.?
No transformation, standardization, or dimensionality reduction required.

- Does PCA help visualize the data? Do we get any insights from histograms/ bar charts/ line plots, etc.?
We analyzed two datasets, GDP vs Time, and Unemployment rate vs Time

Line plots between cumulative GDP rates and year imply that the GDP rates increase with time, as seen from the highly positive correlation coefficient values, and that there is a jump in the rate of increase of the GDP around the early 2000s, We see that the rate of increase for GDP is higher for developing countries compared to developing countries

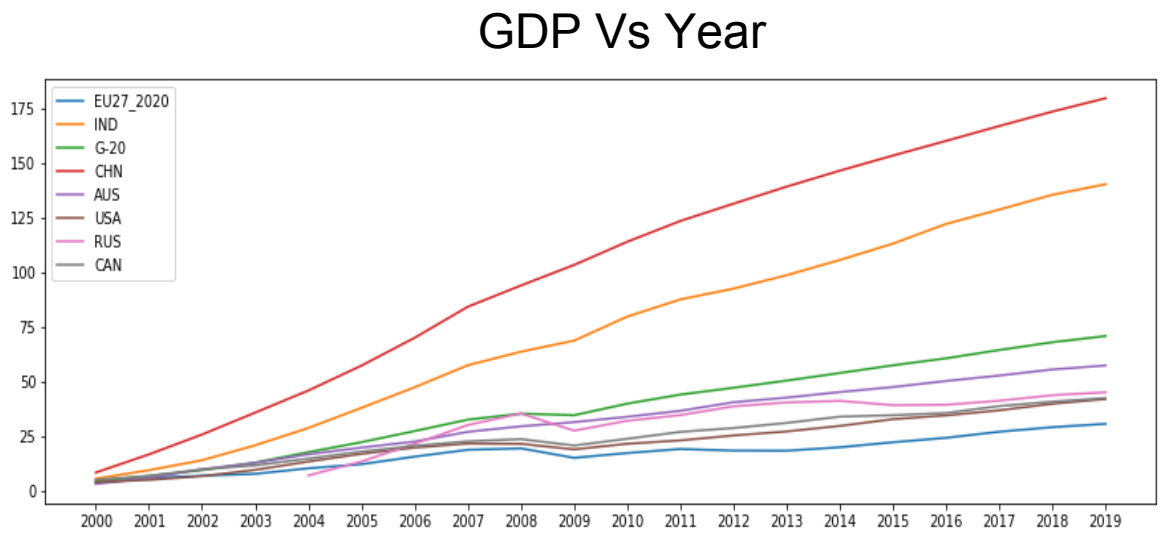
Line plots between employment rates and year show that there is no strong relationship for most of the countries (from the correlation coefficients) but for Russia and Japan the correlation showed a strong negative correlation between time and unemployment rates, these are very strong economic indicators, some countries in showed an increase in unemployment pre-pandemic.

Visualization:

Observations:

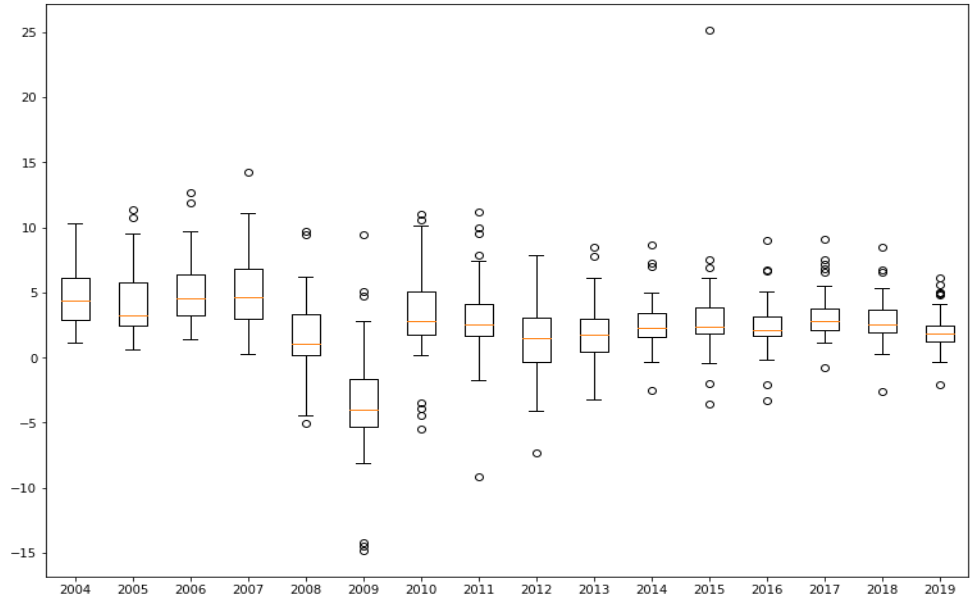
The dataset had rates instead of absolute values, the line graph wasn't showing a specific trend. Thus the cumulative GDP rates were considered for the line graph. The below line graph compares the GDP rates of 8 countries.

Line plots between cumulative GDP rates and year imply that the GDP rates increase with time, as seen from the highly positive correlation coefficient values, and that there is a jump in the rate of increase of the GDP around the early 2000s, We see that the rate of increase for GDP is higher for developing countries compared to developing countries



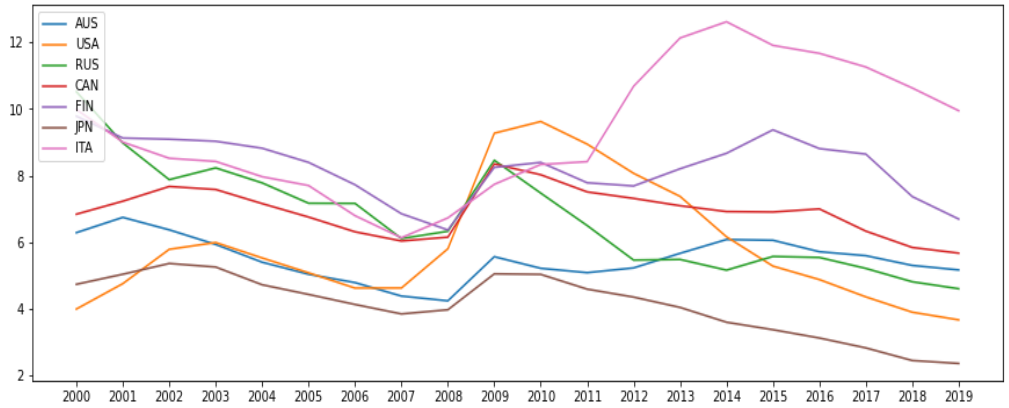
Box Plot for GDP for various years, showing outliers:

The outliers here are values that go beyond the +IQR and the -IQR in the box plot.



Unemployment Rates Vs Year

The unemployment rates for 7 countries were plotted and compared.



Line plots between employment rates and year show that there is no strong relationship for most of the countries (from the correlation coefficients) but for Russia and Japan the correlation showed a strong negative correlation between time and unemployment rates, these are very strong economic indicators, some countries in showed an increase in unemployment pre-pandemic.