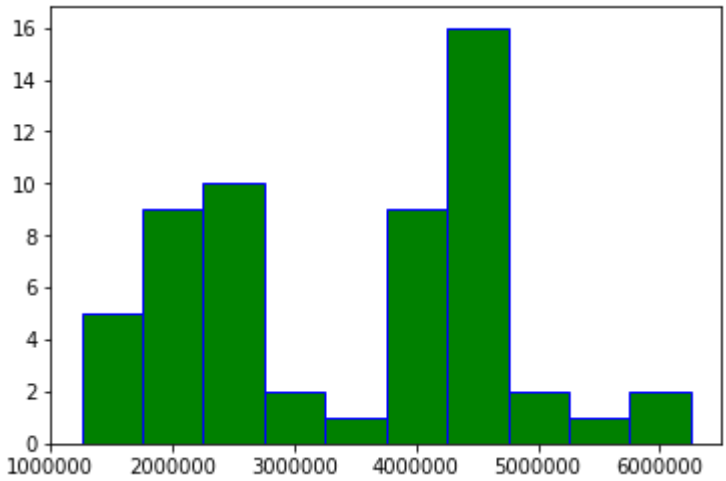
HISTOGRAMS

- We should use histogram when we need the count of the variable in a plot. eg: Number of particular games sold in a store.

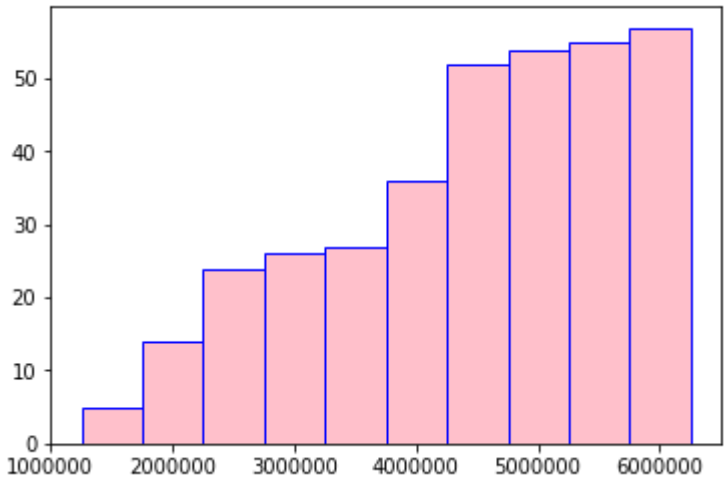
Out[9]:

	Year	Badlands	GrandCanyon	BryceCanyon
0	1961	833300	1253000	264800
1	1962	1044800	1447400	251000



The above figure gives the frequency of data in each bin

```
n =  [ 5. 14. 24. 26. 27. 36. 52. 54. 55. 57.]
--------------------------
bins  [1253000.  1753123.8 2253247.6 2753371.4 3253495.2 3753619.  425
3742.8
 4753866.6 5253990.4 5754114.2 6254238. ]
--------------------------
patches  <a list of 10 Patch objects>
```

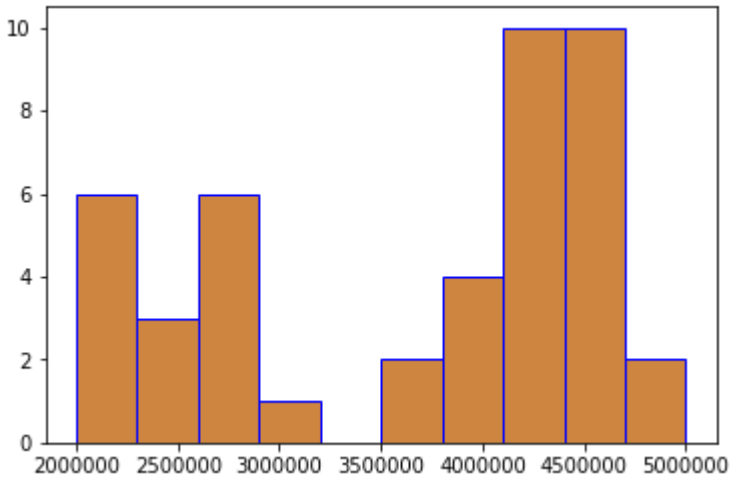


In above figure, the figure is using cumulative bins, each bin adds up the frequecy of the previous bins

- The last bin gives the total frequency of the data i.e 57

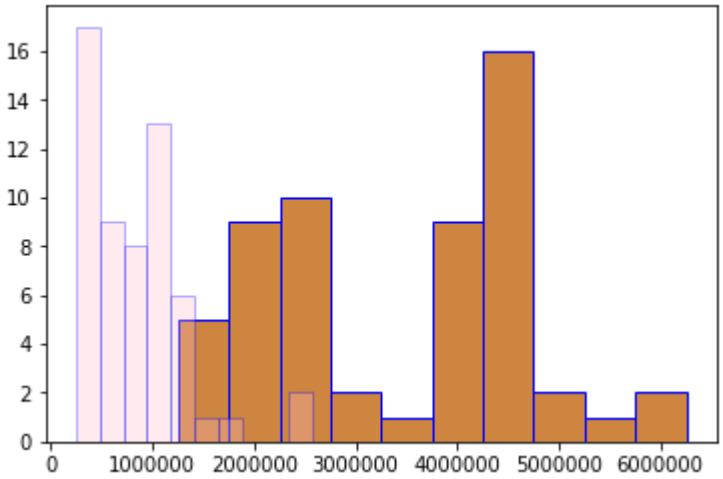
Now what if we want to lookup the data at a specific range?

- Range helps us in understanding value distribution between specified values.



Multiple Histograms

- They are useful in understanding the distribution between two entites variables.



- From the above figure we can see that GrandCanyon has comparably more visitors than BryceCanyon.

Line plot

- Time series data is usually analyzed by line plots
- It is used to represent the trend in the data continuously
- Used in `'Weather Forecasting'`, `'Stock-market'`

Out[40]:

	Date	AAPL	ADBE	CVX	GOOG	IBM	MDLZ	MSFT	NFL
0	3-Jan-07	11.107141	38.869999	50.777351	251.001007	79.242500	17.519524	24.118483	3.25857
1	1-Feb-07	10.962033	39.250000	48.082939	224.949951	74.503204	16.019426	22.092464	3.21857
2	1-Mar-07	12.037377	41.700001	51.900383	229.309311	75.561348	16.009354	21.857189	3.31285

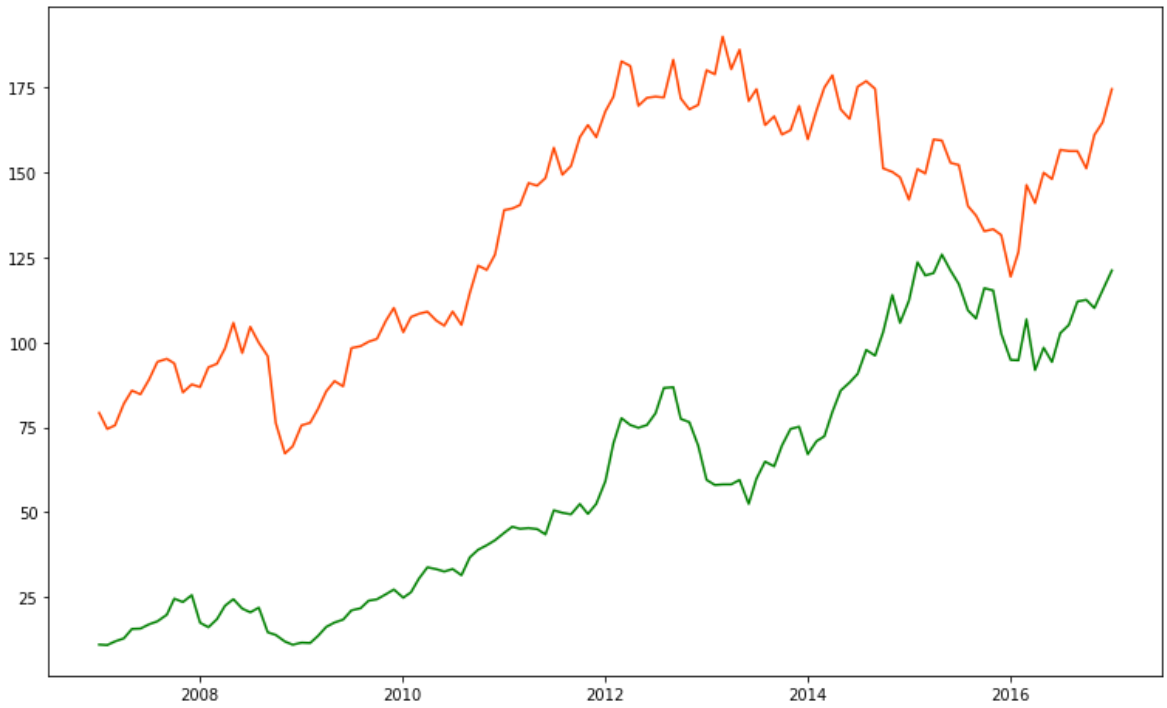
- It is recommended to convert date to pandas DateTime format

Out[42]:

	Date	AAPL	ADBE	CVX	GOOG	IBM	MDLZ	MSFT	NFI
0	2007-01-03	11.107141	38.869999	50.777351	251.001007	79.242500	17.519524	24.118483	3.2585
1	2007-02-01	10.962033	39.250000	48.082939	224.949951	74.503204	16.019426	22.092464	3.2185
2	2007-03-01	12.037377	41.700001	51.900383	229.309311	75.561348	16.009354	21.857189	3.3128

Out[51]:

```
[<matplotlib.lines.Line2D at 0x7f12b9ad1320>]
```



- from the above figure we can say that there is an upward trend

BOXPLOT

- Boxplot is used to have a statistical understanding of the data.
- It is used to see the outliers

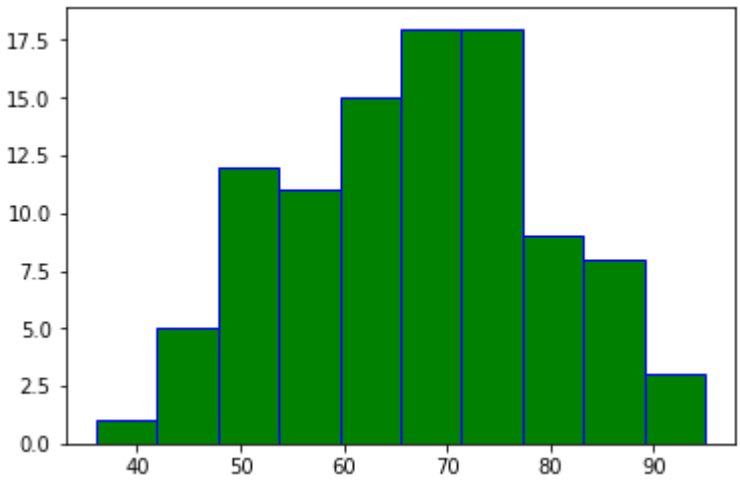
Out[6]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	male	group C	some high school	free/reduced	none	69	61	58
1	female	group C	some college	free/reduced	completed	47	65	69

- Now we can use boxplot to have a detailed overview of the marks distrbution in each subjects.
- We can use `'Barplot'` also but it would only represent frequency distribution but it wont be help full much.

For example I have plottted this for your reference

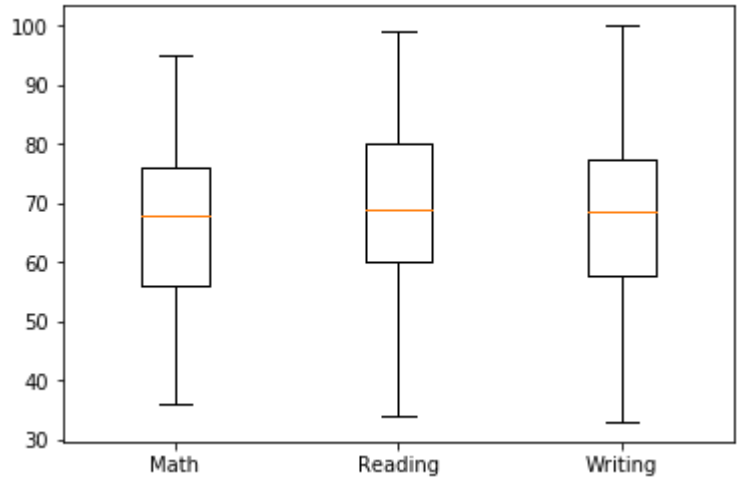
- From this plot we are still confusing regarding the math scores. SO to have better understanding lets do something different.



Doing same thing using `'Boxplot'`

Out[21]:

	math score	reading score	writing score
0	69	61	58
1	47	65	69



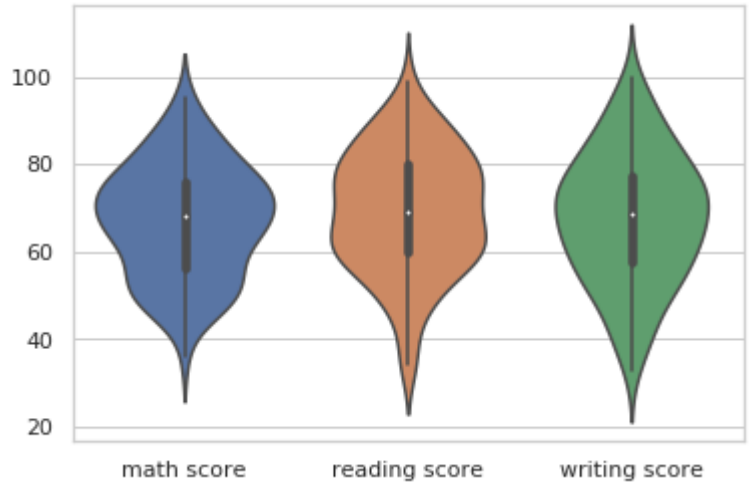
The above figure clearly represents the following:

- Range of marks in each subject
- 1,2,3 Quartiles percentage
- Median marks of each subject

Violin Plot

A violin plot plays a similar role as a boxplot. It shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared. Unlike a box plot, in which all of the plot components correspond to actual datapoints, the violin plot features a kernel density estimation of the underlying distribution.

This can be an effective and attractive way to show multiple distributions of data at once, but keep in mind that the estimation procedure is influenced by the sample size, and violins for relatively small samples might look misleadingly smooth.

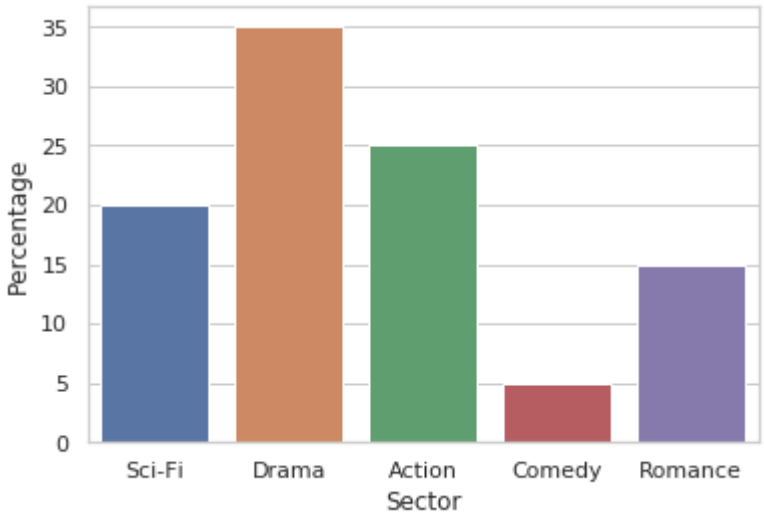


Barplot

Bar Plot shows the distribution of data over several groups. It is commonly confused with a histogram which only takes numerical data for plotting. It is used when to compare between several groups.

Out[106]:

	Sector	Percentage
0	Sci-Fi	20
1	Drama	35
2	Action	25
3	Comedy	5
4	Romance	15

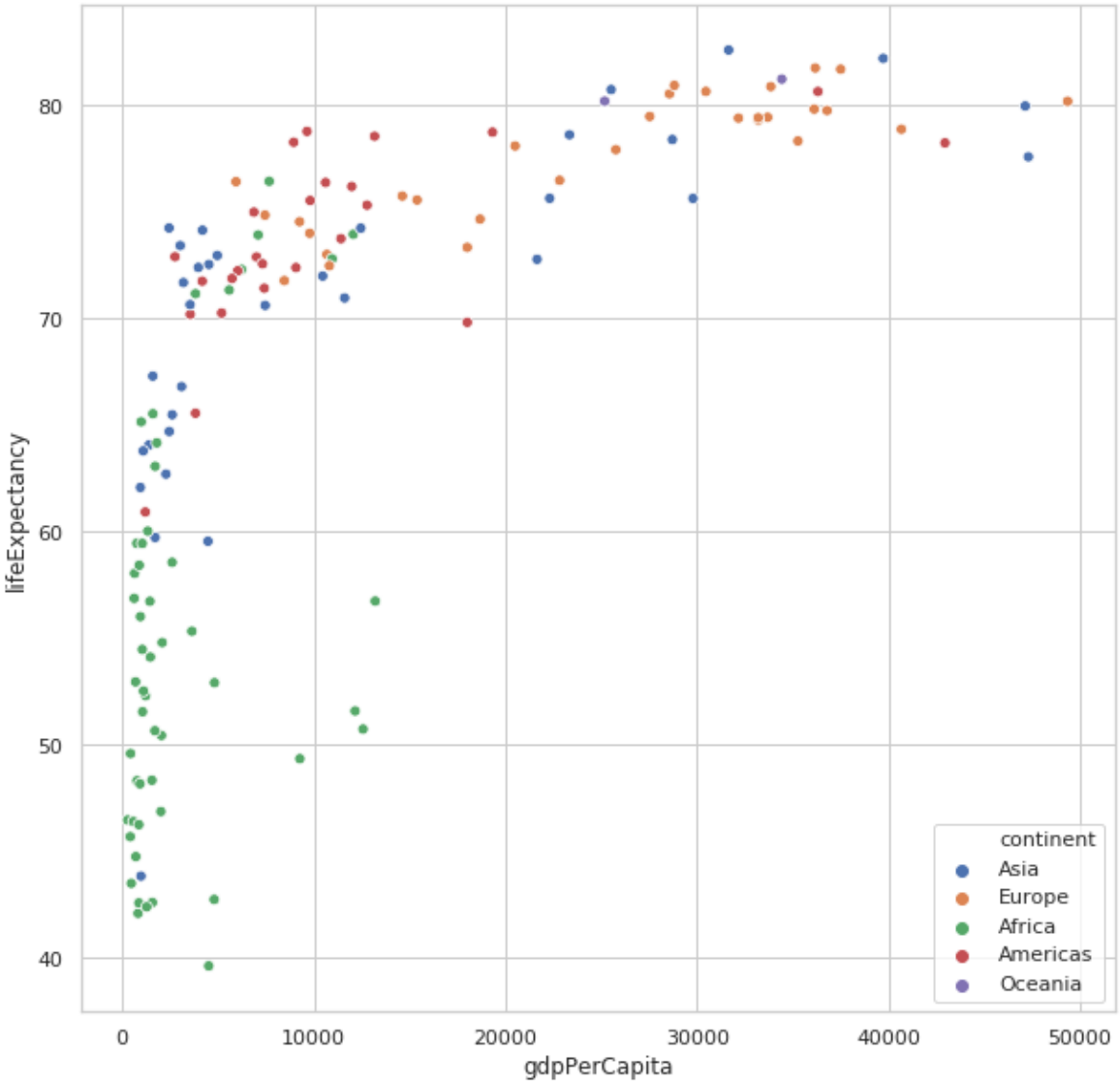


Scatter plot

Scatter plot helps in visualizing 2 numeric variables. It helps in identifying the relationship of the data with each variable i.e correlation or trend patterns. It also helps in detecting outliers in the plot.

Out[124]:

	country	continent	year	lifeExpectancy	population	gdpPerCapita
11	Afghanistan	Asia	2007	43.828	31889923	974.580338
23	Albania	Europe	2007	76.423	3600523	5937.029526


lifeExpectancy
gdpPerCapita
80
70
60
50
40
0
10000
20000
30000
40000
50000
continent
Asia
Europe
Africa
Americas
Oceania

Show code