DATA SCIENCE-UNIVARIATE ANALYSIS-

INTERQUARTILE RANGE (IQR)

IQR:

- The IQR is a robust measure of spreading, meaning it's less affected by outliers compared to measures like standard deviation or variance.
- The IQR can be used to identify potential outliers, which are data points that fall significantly outside the typical range.

HOW IQR IS CALCULATED:

- Arrange the dataset from smallest to largest.
- Find the median Q2, The middle value that divides the data into two halves.
- Find the first quartile Q1, The median of the lower half of the data.
- Find the first quartile Q3, The median of the upper half of the data.
- IQR=Q3-Q1

TYPES OF OUTLIERS:

LESSER OUTLIER:

Less than outlier range= Q1-1.5*IQR (In the dataset no value should be less than this range otherwise it is considered as outlier)

GREATER OUTLIER:

Greater than outlier range= Q3+1.5*IQR (In the dataset no value should be greater than this range otherwise it is considered as outlier)

WHY 1.5?

- The 1.5 in the 1.5IQR rule for outlier detection is used because it provides a good balance between sensitivity and stringency, identifying outliers that deviate significantly from the expected range without being overly stringent or permissive.
- The 1.5IQR rule defines outliers as datapoints that fall below the first quartile (Q1) minus 1.5 times the interquartile range (IQR) or above the third quartile (Q3) plus 1.5 times the IQR.
- 1.5 : Supports Gaussian Distribution, It is convenience to use, Turkey's choice.

Example:

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
Mean	108.0	67.303395	66.333163	66.370186	72.100558	62.278186	288655.405405
Median	108.0	67.0	65.0	66.0	71.0	62.0	265000.0
Mode	1	62.0	63.0	65.0	60.0	56.7	300000.0
Q1:25%	54.5	60.6	60.9	61.0	60.0	57.945	240000.0
Q2:50%	108.0	67.0	65.0	66.0	71.0	62.0	265000.0
Q3:75%	161.5	75.7	73.0	72.0	83.5	66.255	300000.0
99%	212.86	87.0	91.86	83.86	97.0	76.1142	NaN
Q4:100%	215.0	89.4	97.7	91.0	98.0	77.89	940000.0
IQR	107.0	15.1	12.1	11.0	23.5	8.31	60000.0
1.5rule	160.5	22.65	18.15	16.5	35.25	12.465	90000.0
min	1	40.89	37.0	50.0	50.0	51.21	200000.0
max	215	89.4	97.7	91.0	98.0	77.89	940000.0
Lesser	-106.0	37.95	42.75	44.5	24.75	45.48	150000.0
Greater	322.0	98.35	91.15	88.5	118.75	78.72	390000.0

INFERENCE:

The above given table is calculated values from a dataset. First Mean, Median, Mode of the values are calculated for the values in the dataset. Then Q1,Q2,Q3 and Q4 are calculated. Then IQR values are calculated.

HSC_P COLUMN IQR CALCULATION EXPLAINED:

IQR =Q3-Q1 =73-60.9 =12.1

1.5*IQR = 1.5*12.1 = 18.15

Lesser= Q1-1.5*IQR=60.9-18.15 = 42.75

Greater = Q3+1.5*IQR = 73+18.15=91.15

Minimum value in the HSC Pass mark column = 37.0

Maximum value in the HSC Pass mark column = 97.7

Hence the minimum value of the column is less than Lesser outlier value **(37.0<42.75)** and also the maximum value of the column is higher than Greater outlier value **(97.7>91.15)** . Thus both lesser and greater outliers are present in the HSC pass mark column. Likewise for all the columns are calculated and outliers are found and replaced.