OUTLIER

Reason of multiplying 1.5 with IQR:

- Multiplying 1.5 with the interquartile range (IQR) is a common technique used in statistics to identify outliers using the concept of the "Tukey method" or "Tukey's fences."
- The interquartile range is a measure of statistical dispersion, representing the range between the first quartile (25th percentile) and the third quartile (75th percentile) of a dataset.

Tukey's Fences: Tukey's fences are thresholds used to define the bounds beyond which data points are considered outliers. These fences are calculated as follows:

```
Lower Bound: (Q1 - 1.5 * IQR)
Upper Bound: (Q3 + 1.5 * IQR)
```

- The choice of 1.5 as the multiplier is somewhat arbitrary but is widely accepted as a balance between detecting genuine outliers and not excessively labelling data points as outliers.
- This value provides a reasonable compromise between sensitivity to potential outliers and the risk of falsely identifying normal data points as outliers.
- John Tukey, a prominent statistician, introduced this method as part of exploratory data analysis. While other multipliers can be used, 1.5 has become a standard in many statistical analyses due to its effectiveness in identifying potential outliers without overly inflating the outlier count.

Example:

a. The interquartile range. Compare the two interquartile ranges.

b. Any outliers in either set.

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

Solution:

IQR = Q3-Q1 = 82.5 - 56 =26.5

IQR=26.5

LESSER RANGE OUTLIER:

Q1-1.5(IQR) = 56 – 1.5(26.5)

= 56 – 39.75

= 16.25

Q1-1.5(IQR) = 16.25

GREATER RANGE OF OUTLIER:

Q3 + 1.5(IQR) = 82.5 + 1.5(26.5)

= 82.5 + 39.75

= 122.75

Q3 + 1.5(IQR)= 122.75

CONCLUSION:

Day and night classes lower than 16.25 are lesser range outliers and classes higher than 122.25 are higher range outliers.

IQR Calculated for Placement Sample dataset:

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
Mean	108	67.3034	66.3332	66.3702	72.1006	62.2782	288655
Median	108	67	65	66	71	62	265000
Mode	1	62	63	65	60	56.7	300000
Q1:25%	54.5	60.6	60.9	61	60	57.945	240000
Q2:50%	108	67	65	66	71	62	265000
Q3:75%	161.5	75.7	73	72	83.5	66.255	300000
Q4:100%	215	89.4	97.7	91	98	77.89	940000
IQR	107	15.1	12.1	11	23.5	8.31	60000
1.5rule	160.5	22.65	18.15	16.5	35.25	12.465	90000
Lesser	-106	37.95	42.75	44.5	24.75	45.48	150000
Greater	322	98.35	91.15	88.5	118.75	78.72	390000
Min	1	40.89	37	50	50	51.21	200000
Max	215	89.4	97.7	91	98	77.89	940000

Baby step: Finding outliers manually

Min and Max:

While comparing the min and max values to lesser and greater outliers then hsc_p min column has lesser outlier; hsc_p max, degree_p max and salary have greater outlier.