IQR

INTER QUARTILE RANGE

What is the purpose of IQR?

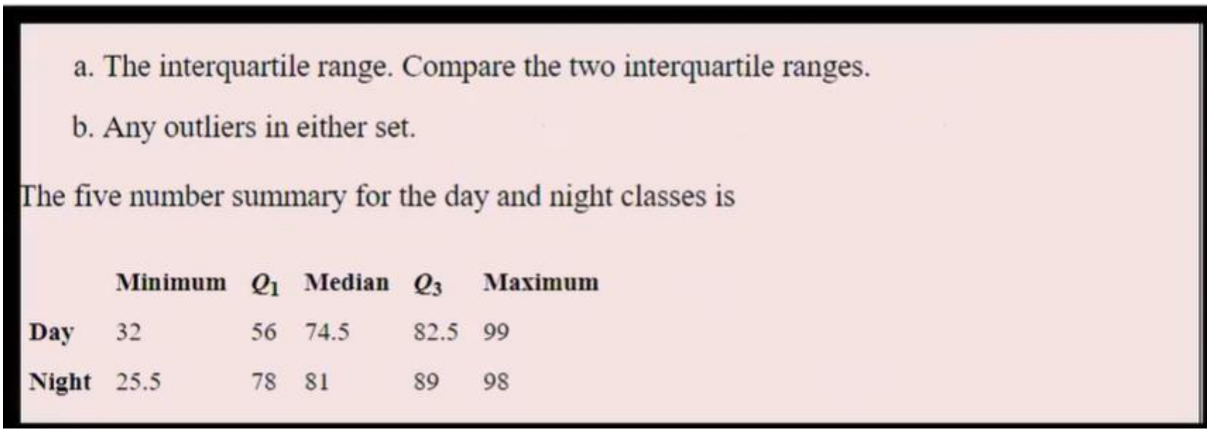
- ❖ The IQR can help to determine potential outliers. A value is Suspected to be a potential outlier if it is less than (1.5)(IQR) below the first quartiler or more than (1.5)(IQR) above the third quartile. Potential outliers always require investigation.

Reason for 1.5*IQR:

- ❖ The choice of 1.5 as the multiplier is a convention that has been found to work well in practice for many datasets, especially those with approximately normal distributions. It provides a balance between sensitivity to outliers and avoiding false positives.

- In essence, the 1.5*IQR rule provides a simple yet effective way to identify data points that deviate significantly from the expected range based on the quartiles of the dataset.
- To know the Outlier range present in the dataset. IQR=Q3-Q1.
  - Lesser Outlier Less than Outlier range =Q1-1.5*IQR
  - Greater Outlier greater than outlier range=Q3+1.5*IQR

INTER QUARTILE RANGE(IQR)



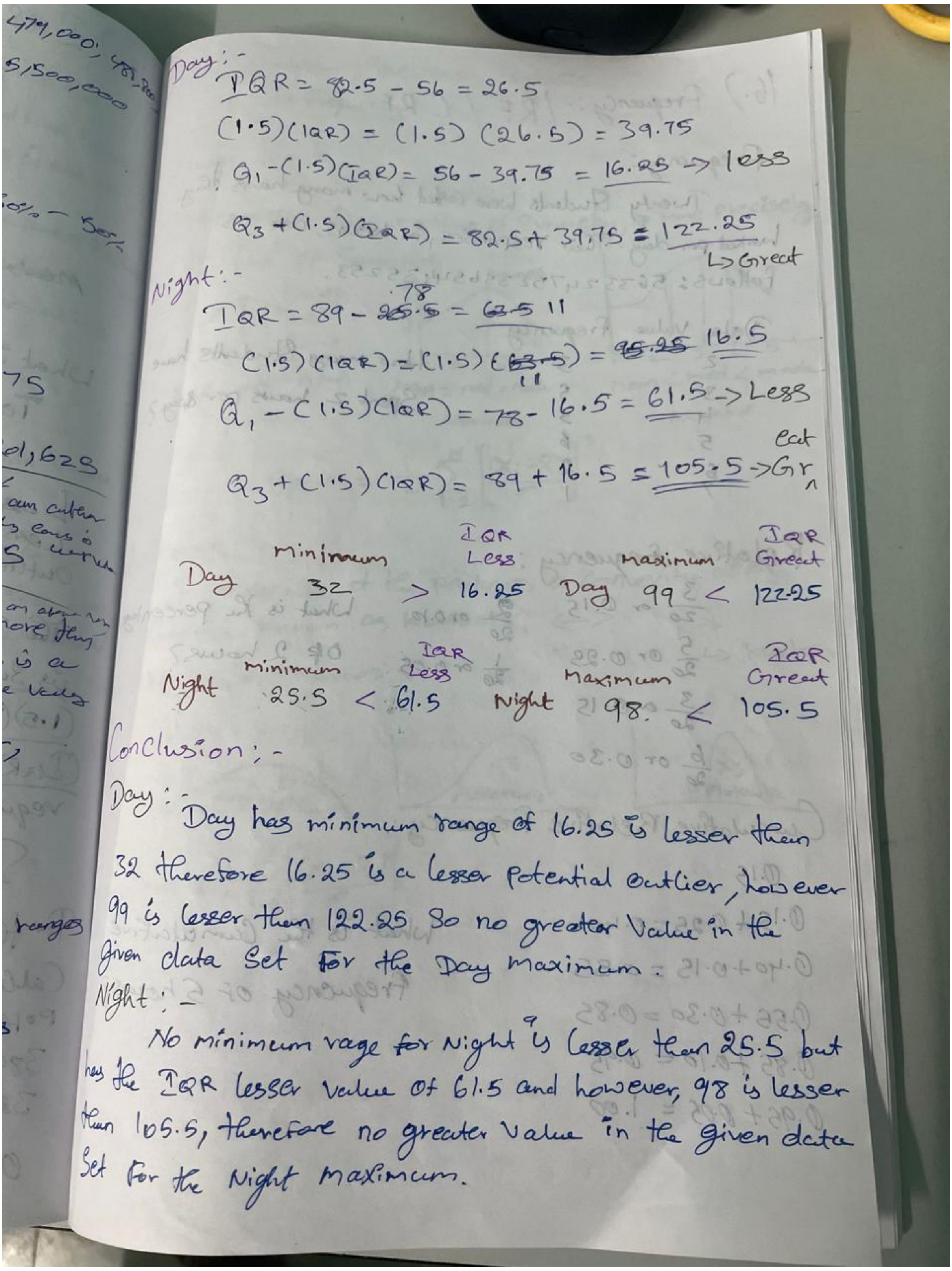
a. The interquartile range. Compare the two interquartile ranges.

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

ANSWER:



Day :-

$IQR = 82.5 - 56 = 26.5$

$(1.5)(IQR) = (1.5)(26.5) = 39.75$

$Q_1 - (1.5)(IQR) = 56 - 39.75 = \underline{16.25}$ -> less

$Q_3 + (1.5)(IQR) = 82.5 + 39.75 = \underline{122.25}$ -> Great

Night :-

$IQR = 89 - \overset{.78}{\cancel{25.5}} = \underline{\cancel{63.5}\ 11}$

$(1.5)(IQR) = (1.5)(\underset{11}{\cancel{63.5}}) = \cancel{95.25}\ \underline{16.5}$

$Q_1 - (1.5)(IQR) = 78 - 16.5 = \underline{61.5}$ -> Less

$Q_3 + (1.5)(IQR) = 89 + 16.5 = \underline{105.5}$ -> Great

	minimum		IQR Less		maximum		IQR Great
Day	32	>	16.25	Day	99	<	122.25
Night	25.5	<	61.5	Night	98.	<	105.5

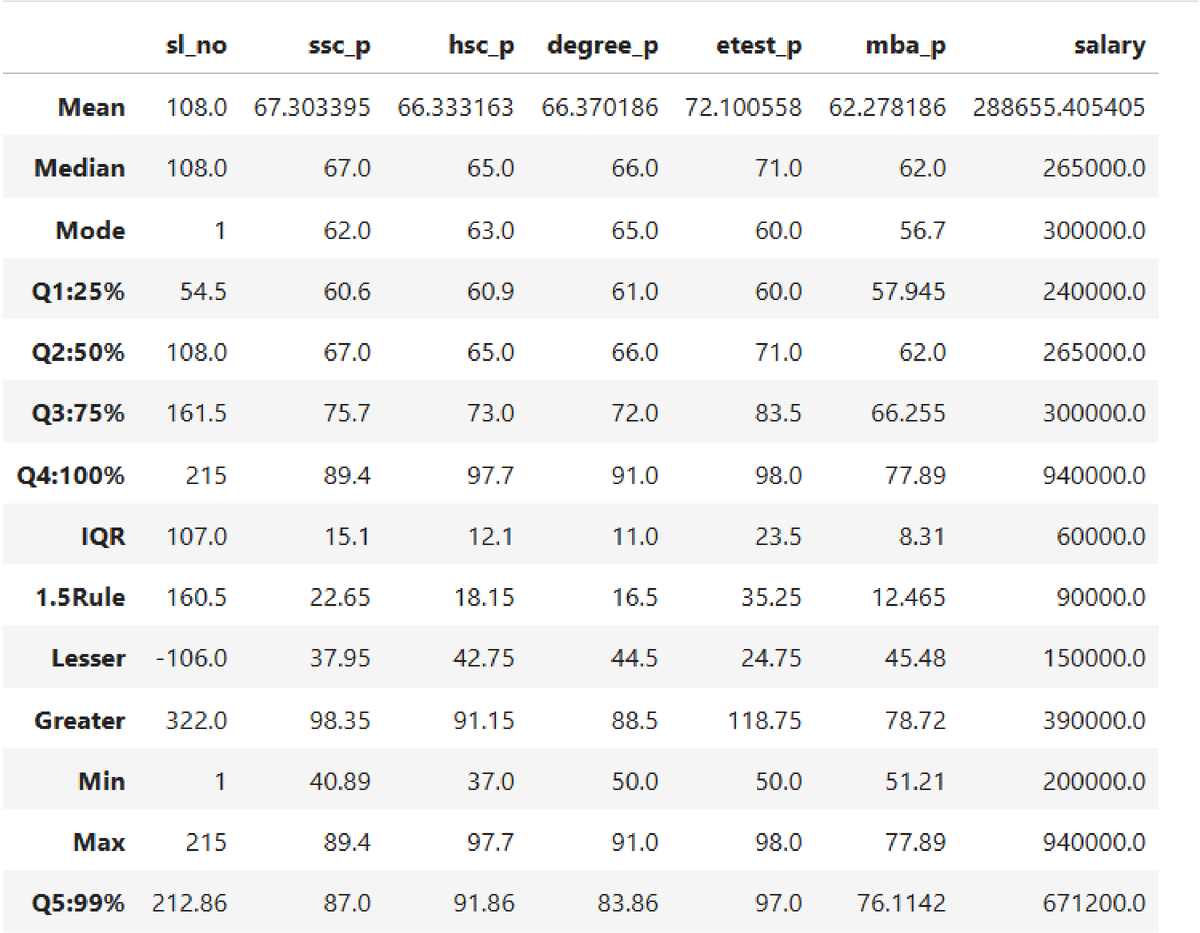
Conclusion :-

Day :- Day has minimum range of 16.25 is lesser than 32 therefore 16.25 is a lesser potential outlier, however 99 is lesser than 122.25 So no greater Value in the given data Set for the Day maximum.

Night :-

No minimum rage for Night is lesser than 25.5 but has the IQR lesser value of 61.5 and however, 98 is lesser than 105.5, therefore no greater Value in the given data Set for the Night maximum.

IQR



	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
Q4:100%	215	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
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
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
Q5:99%	212.86	87.0	91.86	83.86	97.0	76.1142	671200.0

IQR (Interquartile Range):

1. The IQR represents the middle 50% of the data, measuring the spread between the 75th percentile (Q3) and the 25th percentile (Q1).

2. For the variables in the dataset, IQR values vary, with the highest spread observed in the salary column (600000) and the lowest in the MBA percentage (8.31).

3. A higher IQR suggests greater variability in data, whereas a lower IQR means the data points are more tightly clustered.

4. The significant IQR for salary indicates diverse salary distributions, likely affected by different education levels and other parameters.

1.5 Rule:

1. The 1.5 Rule is a common method for detecting potential outliers by

extending 1.5 times the IQR above Q3 and below Q1.

2. This rule helps to identify values that may be unusually high or low compared to the normal distribution.
3. For salary, the calculated value for the 1.5 Rule is 90000, meaning salaries significantly beyond this range could be outliers.
4. The highest value for the 1.5 Rule is found in salary, indicating a broader spread in earnings compared to academic percentages.

Lesser (Lower Bound Outliers):

1. Lesser values are the lower threshold, calculated as **Q1 - 1.5 * IQR**,

meaning any value below this is an outlier.

2. The salary's lower bound is **150000**, suggesting salaries below this are significantly different from the normal distribution.

3. Negative values in some variables (like -106 for sl_no) suggest a data entry or calculation issue.

4. Academic percentages have lesser thresholds ranging from 37.95 to 45.48, meaning very low scores in these ranges could be unusual.

Greater (Upper Bound Outliers):

1. Greater values represent the higher threshold, calculated as **Q3 +**

1.5 * IQR, meaning any value above this is an outlier.

2. The highest greater bound is observed in salary (390000), indicating that salaries above this might be considered anomalies.
3. In educational scores, the highest outlier threshold is for hsc_p (91.15), meaning scores above this are rare.
4. These values help in detecting anomalies in datasets, ensuring proper data cleaning and preprocessing.

Min (Minimum Values in Data):

1. The minimum value represents the smallest recorded data point for each variable.
2. The lowest recorded salary is **200000**, which could be the starting salary for freshers.
3. The minimum scores for various academic categories show variations, with the lowest being 37.0 for hsc_p and 40.89 for ssc_p.
4. These values are essential in understanding the lowest boundaries of observed data.

Max (Maximum Values in Data):

1. The maximum value represents the highest recorded data point for each variable.
2. The highest salary recorded in the dataset is **940000**, which is a significant range from the minimum.
3. Among academic scores, the highest recorded value is 97.7 (hsc_p), indicating a few students achieved nearly perfect scores.
4. The max values help in understanding the upper limits of data and possible high-performing individuals.

CONCLUSION:

- The dataset exhibits a high variation in salary, as indicated by the large IQR and max values. The 1.5 Rule helps identify potential outliers, with a clear distinction in salary brackets. The lower and upper outlier thresholds (Lesser & Greater) show a few extreme values, particularly in salary and academic percentages. The minimum and maximum values confirm the diversity in academic performance and salary distribution. Overall, the dataset suggests a wide range of salaries influenced by multiple factors, requiring further analysis to determine key driving elements.