1.TABLE EXPLANATION

	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

Mean:

The students in the 10th and 12th standards had average marks, but they scored well in the entrance test. Their MBA marks were also average. They received salaries ranging from 288,655 to 405,405.

Mode:

Most students in the 10th standard scored 62.0 marks, while in the 12th standard, most scored 63.0 marks. For undergraduate studies, the majority of students scored 65.0 marks. In the entrance test, most students scored 60.0 marks, and in the MBA, the majority scored 56.7 marks. Most people received a salary of 3,000,000.

Difference between Mean,Median,Mode:

The mean is the average of all data points, including outliers. The median does not consider outliers and represents the middle value when the data is sorted. The mode reflects the most frequently occurring values in the dataset. Each of these measures provides insight into the central tendency of the data.

2.Percentile Report

	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

- From Q1 :25% to Q2 50%: There is a 25% different,that is a percentile.
  - ssc_p marks increased by 7 marks.
  - hsc_p and degree_p marks both increased by 5 marks.
  - etest_p marks increased by 11 marks.
  - mba_p marks increased by 5 marks.
  - The salary difference is ₹2,50,000.

- From Q2 50% to Q3 75%: There is a 25% different.
  - ssc_p marks increased by 8 marks.
  - hsc_p marks increased by 18 marks.
  - degree_p marks increased by 6 marks.
  - etest_p marks increased by 12 marks.
  - mba_p marks increased by 4 marks.

- The salary difference is only ₹35,000.

From Q3 :75% to 99%: There is a 25% different.

- ssc_p marks increased by 12 marks.
- hsc_p marks increased by 19 marks.
- degree_p marks increased by 11 marks.
- etest_p marks increased by 14 marks.
- mba_p marks increased by 10 marks.

From 99% to Q4:100%: There is only a 1% different.

- ssc_p marks increased by 2 marks.
- hsc_p marks by 5 marks.
- degree_p marks by 8 marks.
- etest_p marks increased by 11 marks.
- mba_p marks increased by just 1 mark.
- The salary difference from Q3 (75%) to Q4 (100%) is ₹6,40,000.

3.Percentage Report

S.NO:	*SUBJECTS*	*MARKS*
1.	*TAMIL*	*89/100*
2.	*ENGLISH*	*59/100*
3.	*MATHS*	*60/100*
4.	*SCIENCE*	*49/100*
5.	*SOCIAL*	*86/100*
	TOTAL	*343/500*

$$\text{Percentage} = \left(\frac{\text{Total Marks Obtained}}{\text{Total Marks}}\right) \times 100$$

In your case:

$$\text{Percentage} = \left(\frac{343}{500}\right) \times 100 = 68.6\%$$

So, your percentage is 68.6%.

Difference between percentage and percentile

Percentage

- It represents a value out of 100.
- For example, if you score 80 out of 100 marks in a subject, your percentage is

  80%.

- Formula:

$$\text{Percentage} = \left(\frac{\text{Obtained Marks}}{\text{Total Marks}}\right) \times 100$$

Percentile

- Percentile indicates the position or rank of a score within a group.
- For example, if you are in the 90th percentile, it means you scored better than 90% of the people.
- Percentile is often used in competitive exams to show how your performance compares with others.

4.Inter Quartile Range(IQR)

Question:1

Why we are using 1.5 rule in (IQR) ?

The 1.5 rule is an important concept in statistics. It is used in the interquartile range (IQR) method to identify outliers. By applying this rule, we can effectively detect both lower and upper outliers in a dataset.

Question:2

- The Inter Quartile Range. Compare the two inter quartile ranges.
- Any outliers in either set.

The five number summary for the day and night classes is

	Initial Value	Q1:25%	Median Q2:50%	Q3:75%	Max Q4:100%
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

Answer:

(i)

	Initial Value	Q1:25%	Median Q2:50%	Q3:75%	Max Q4:100%
Day	32	56	74.5	82.5	99

IQR=Q3 - Q1 $\quad$ IQR = 82.5 – 56 = 26.5

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

Q1 – 1.5 (IQR) 56 – 39.75 = 16.25

Q3 + 1.5 (IQR) 82.5 + 39.75 = 122.25

- The initial value is 16.25. If there is a lesser outlier, the value should be below that; however, since the value here is 32, it is not considered an outlier.
- The maximum value is 122.25. If there is a greater outlier, the value should be above that; however, since the value here is 99, it is not considered an outlier.

Answer:

(ii)

	Initial Value	Q1:25%	Median Q2:50%	Q3:75%	Max Q4:100%
Night	25.5	78	81	89	98

IQR=Q3 - Q1 IQR = 89 – 78 = 11

(1.5)(IQR) 1.5 (11) = 16.5

Q1 – 1.5 (IQR) 78 – 16.5 = 61.5

Q3 + 1.5 (IQR) 89 + 16.5 = 105.5

- The initial value is 61.5. If there is a lesser outlier, the value should be below that; however, since the value here is 25.5, it is considered a lesser outlier.

- The maximum value is 105.5. If there is a greater outlier, the value should be above that; however, since the value here is 99, it is not considered an outlier.

5.TABLE EXPLANATION

	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
Lesser Outlier	-106.0	37.95	42.75	44.5	24.75	45.48	150000.0
Greater Outlier	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

ssc_p column:

- **In the dataset, the min value in the ssc_p column is 40.89, and the calculated lower outlier value is 37.95. Normally, if the min value is less than the lower outlier value, it is considered an outlier. However, in this case, the min value is greater than the lower outlier value, so there is no lesser outlier.**

- In the dataset, the max value in the ssc_p column is 89.4, and the

- calculated greater outlier value is 98.35. Normally, if the max value is greater than the greater outlier value, it is considered an outlier. However, in this case, the max value is less than the greater outlier value, so there is no greater outlier.

hsc_p column:

- In the dataset, the minimum value in the hsc_p column is 37.0, and the calculated lower outlier value is 42.75. Normally, if the minimum value is less than the lower outlier value, it is considered an outlier. However In this case, minimum value is lesser than the lower outlier value, it is considered a lower outlier.

- In the dataset, the maximum value in the hsc_p column is 97.7, and the calculated greater outlier value is 91.15. Normally, if the maximum value is greater than the greater outlier value, it is considered an outlier. In this case, since the maximum value is greater than the greater outlier value, it is considered a greater outlier."

degree_p column:

- In the dataset, the min value in the degree_p column is 50.0, and the calculated lower outlier value is 44.5. Normally, if the min value is less than the lower outlier value, it is considered an outlier. However, in this case, the min value is greater than the lower outlier value, so there is no lesser outlier.

- In the dataset, the max value in the degree_p column is 91.0, and the calculated greater outlier value is 88.5. Normally, if the max value is greater than the greater outlier value, it is considered an outlier. However, in this case, the max value is less than the greater outlier value, so there is no greater outlier.

etest_p column:

- In the dataset, the min value in the etest_p column is 50.0, and the calculated lower outlier value is 24.75. Normally, if the min value is less than the lower outlier value, it is considered an outlier. However, in this case, the min value is greater than the lower outlier value, so there is no lesser outlier.

- In the dataset, the max value in the etest_p column is 98.0, and the calculated greater outlier value is 118.75. Normally, if the max value is greater than the greater outlier value, it is considered an outlier. However, in this case, the max value is less than the greater outlier value, so there is no greater outlier.

mba_p column:

- In the dataset, the min value in the mba_p column is 51.21, and the calculated lower outlier value is 45.48. Normally, if the min value is less than the lower outlier value, it is considered an outlier. However, in this case, the min value is greater than the lower outlier value, so there is no lesser outlier.

- In the dataset, the max value in the mba_p column is 77.89, and the calculated greater outlier value is 78.72. Normally, if the max value is greater than the greater outlier value, it is considered an outlier. However, in this case, the max value is less than the greater outlier value, so there is no greater outlier.

salary column:

- In the dataset, the min value in the salary column is 200000, and the calculated lower outlier value is 150000. Normally, if the min value is less than the lower outlier value, it is considered an outlier. However, in this case, the min value is greater than the lower outlier value, so there is no lesser outlier.

    - In the dataset, the maximum value in the salary column is 940000, and the calculated greater outlier value is 390000.

Normally, if the maximum value is greater than the greater outlier value, it is considered an outlier. In this case, since the maximum value is greater than the greater outlier value, it is considered a greater outlier.

6.TABLE EXPLANATION

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
Mean	108.0	67.303395	66.334744	66.358558	72.100558	62.278186	277648.648649
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	91.129	83.86	97.0	76.1142	NaN
Q4:100%	215.0	89.4	91.15	88.5	98.0	77.89	390000.0
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 Outlier	-106.0	37.95	42.75	44.5	24.75	45.48	150000.0
Greater Outlier	322.0	98.35	91.15	88.5	118.75	78.72	390000.0
Min	1	40.89	42.75	50.0	50.0	51.21	200000.0
Max	215	89.4	91.15	88.5	98.0	77.89	390000.0
kurtosis	-1.2	-0.60751	0.086901	-0.09749	-1.08858	-0.470723	-0.239837
skew	0.0	-0.132649	0.162611	0.204164	0.282308	0.313576	0.8067

Kurtosis:

The values for ssc_p (-0.60751), hsc_p (0.086901), degree_p (-0.09749), etest_p (-1.08858), mba_p (-0.470723), and salary (-0.239837) are all less than 3 (<3), Therefore we decided to apply kurtosis analysis.

Skewness:

he ssc_p value is -0.132649, indicating a negative value. This column predominantly contains values close to the mean.

For the other columns:

- hsc_p (0.162611)
- degree_p (0.204164)
- etest_p (0.282308)
- mba_p (313576)
- salary (8067)

All these values are positive and represent the mode values in their respective columns.

7. PROBABILITY DENSITY FUNCTION

EXPLANATION OF THE CODE

```python
#find out probability density function value
def get_pdf_probability(dataset,startrange,endrange):
    from matplotlib import pyplot
    from scipy.stats import norm
    import seaborn as sns
    ax = sns.distplot(dataset,kde=True,kde_kws={'color':'blue'},color='Green')
    pyplot.axvline(startrange,color='Red')
    pyplot.axvline(endrange,color='Red')
    # generate a sample
    sample = dataset
    # calculate parameters
    sample_mean =sample.mean()
    sample_std = sample.std()
    print('Mean=%.3f, Standard Deviation=%.3f' % (sample_mean, sample_std))
    # define the distribution
    dist = norm(sample_mean, sample_std)

    # sample probabilities for a range of outcomes
    values = [value for value in range(startrange, endrange)]
    probabilities = [dist.pdf(value) for value in values]
    prob=sum(probabilities)
    print("The area between range({},{}):{}".format(startrange,endrange,sum(probabilities)))
    return prob
```

- Matplotlib.pyplot – For plotting the graph.
- Scipy.stats.norm – Provides tools to work with normal distributions
- Seaborn – For data visualizations

- sns.distplot – The Histogram and the Kernel Density Estimate (KDE) curve of the dataset.
- Kde=True – Enable the KDE curve
- Kde_kws ={'colour':'blue'}: - Specifies the KDE curve colour as blue.
- Color = 'Green': Histogram bars color to green
- Pyplot.axvline: Draws vertical red lines at starting range and end range on the plot, for which we are calculating the probability .

- Sample_mean : The average mean of the dataset.
- Sample_std : The standard deviation of the dataset.

These values are printed for reference.

- dist: - creates a normal distribution object using (scipy.stats.norm) with the dataset's mean and standard deviation.

- values = [value for value in range(startrange, endrange)] – Create a list of integer values from startrange to endrange.
  If startrange = 5 and endrange = 10, values will be [5,6,7,8,9,]

- probabilities = [dist.pdf(value) for value in values] – This line calculates the probability density function (PDF) value for each value in the values list. For each integer value in values, it computes dist.pdf(values) and stores the result in the probabilities list.
- prop = sum(probabilities) – This gives a rough estimate of the probability area, assuming that the (PDF) values are similar to point heights in a histogram.

8. STANDARD NORMAL DISTRIBUTION

EXPLANATION OF THE CODE

```python
#standard normal distribution
def stdNBgraph(dataset):
    # Coverted to standard Normal Distribution
    import seaborn as sns
    mean=dataset.mean()
    std=dataset.std()

    values=[i for i in dataset]

    z_score=[((j-mean)/std) for j in values]

    sns.distplot(z_score,kde=True)

    sum(z_score)/len(z_score)
    #z_score.std()
```

- def stdNBgraph(dataset): - This is a function named stdNBgraph that takes a list or series of numbers .

- import seaborn as sns – We import the seaborn library, which helps create graphs.

- Mean = dataset.mean()

- Std = dataset.std()

  Mean : calculates the average of the dataset. Std: Calculates the standard deviation (spread out the numbers).

- Values = [i for i in dataset] – Creates a list values containing all numbers from the dataset.
- Z_score = [((j – mean) / std) for j in values] – Calculates the z-score for each number. It converts the dataset to a standard normal form with mean 0 and standard deviation 1.
- Sns.distplot (z_score, kde = True) – Plots a graph (Histogram) of z_score using seaborn. It shows the frequency and a smooth curve.

- Sum(z_score) / len(z_score) – **Calculates the mean of the z-scores, which should be close to 0.**