Measure of Central Tendency:

```
]: descriptive
```

]:

	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

Data Types:
Secondary Education Percentage (ssc_p): Float
Higher Secondary Education Percentage (hsc_p): Float
Degree Percentage (degree_p): Float
Employability Test Percentage (etest_p): Float
MBA Percentage (mba_p): Float
Salary: Float

Statistical Measures:
Mean:
ssc_p: 67.303395
hsc_p: 66.333163
degree_p: 66.370186
etest_p: 72.100558
mba_p: 62.278186
salary: 288655.405405

Mean / Arithmetic mean :
Average which gives (OVERALL CENTRAL VALUE).removing outlier
none null value must be not there

Median (Mid Point of the data)

ssc_p: 67.0
hsc_p: 67.0
degree_p: 65.0
etest_p: 66.0
mba_p: 71.0
salary: 265000.0

They represent the middle values of each parameter when sorted in ascending order. Therefore, there's no need for additional calculations to find the overall median. Each parameter's median stands on its own

MODE:

The mode is useful for identifying central tendencies in a dataset, especially when dealing with categorical or discrete data.(MOST REPEATED DATAPOINT)
EX:[2,22,24,22,54,22]-mode=3

Mode:
ssc_p: 62.0
hsc_p: 63.0
degree_p: 65.0
etest_p: 60.0
mba_p: 56.7
salary: 300000.0

]: descriptive

]:

	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

ssc_p (Secondary Education Percentage):
The percentage scores obtained in the secondary education level. It indicates the performance of individuals in their secondary education.

hsc_p (Higher Secondary Education Percentage):
The higher secondary education level. It reflects the academic performance of individuals in their higher secondary education.

degree_p (Undergraduate Degree Percentage):
The percentage scores attained in the undergraduate degree. It shows the academic achievement at the undergraduate level.

etest_p (Employability Test Percentage):
etest_p denotes the percentage scores obtained in the employability test. It may indicate the preparedness or competence of individuals for employment opportunities.

mba_p (MBA Percentage):
The percentage scores obtained in MBA (Master of Business Administration) programs. It represents the academic performance or achievement in MBA students

salary:
The salaries associated with the respective data entries. It reflects the monetary compensation received by individuals, presumably after securing placements or jobs.

This dataset provides a comprehensive overview of individuals' educational achievements (from secondary education to MBA) to corresponding salaries.

It allows for the analysis of the relationship between academic performance and salary levels and it can have purposes such as educational research,career planning, or organizational analysis

Insights:

The above table provides the average(Mean) score in SSC is 67 and the average has been increased in the entrance exam to 72 and these students are getting an average salary of 2,88,500 per year.

The most repeated(Mode) value in SSC is 62 and these students are getting a salary of 300000 which is above the average salary in the data

After sorting the data, we could see that Midpoint (Median) of the student’s performance in HSC,SSC,Degree,Entrance and MBA is close to average and hence he/she earns a salary of about 265000 per year which is below avergage

The students' performance is average in SSC,HSC,Degree and when it comes to Entrance exams they perform better to get into a MBA degree. And once getting into the MBA we can see a dip in the performance which resulted in a salary of 288500 per year

PERCENTILE

It tells the existing value of a range
ex:[“5”,”10”,”20”,”35”] in this range of values exist (5%,10%,15%)
How will you calculate PERCENTILE?or,what is the purpose of PERCENTILE?
1]Calculate the each part of PERCENTILE boundaries
original dataset=[10,9,8,7,6,5,4,3,2,1,]…….

sorted dataset=[1,2,3,4,5,6,7,8,9,10]……change
original dataset values to ascending values[,]

Why are we using PERCENTILE?

PERCENTILE provide into the distribution of data,
statistical analysis and clear decision-making processes
across.

]: descriptive

]:

	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
99%	212.86	87	91.86	83.86	97	76.1142	NaN
Q4:100%	215	89.4	97.7	91	98	77.89	940000

SSC Score Analysis:

• First quadrant(25%) of students scored 60.6 in SSC and next 25% that means second quadrant (50%) of students scored 67.0marks ,the increase percentage of 25% to 50% is 7%

• Second quadrant score(50%) is 67 and Third quadrant mark(75%) is 75.7 and the increase percentage between 50% to 75% is 8.7%

• Third quadrant(75%) of students scored 75.7 in SSC and the highest percentage 99% is 87 ,the difference between 75% to 995 is 2.4 %

• The Fourth quadrant (100%)of students scored 89.4 marks and the difference between last two highest percentages are 2.4%

HSC Score Analysis:

• First quadrant(25%) of students scored 60.9 in HSC and next 25% that means second quadrant (50%) of students scored 65.0marks ,the increase percentage of 25% to 50% is 5%

• Second quadrant score(50%) is 65 and Third quadrant mark(75%) is 73 and the increase percentage between 50% to 75% is 8%

• Third quadrant(75%) of students scored 73 in HSC and the highest percentage 99% is 91.8,the difference between 75% to 995 is 18.8%

• The Fourth quadrant (100%)of students scored 97.7 marks and the difference between last two highest percentages are 5.9%

Degree Score Analysis:

• First quadrant(25%) of students scored 61 in Degree and next 25% that means second quadrant (50%) of students scored 66 marks ,the increase percentage of 25% to 50% is 7%

• Second quadrant score(50%) is 66 and Third quadrant mark(75%) is 72 and the increase percentage between 50% to 75% is 6%

• Third quadrant(75%) of students scored 72 in Degree and the highest percentage 99% is 83.8 ,the difference between 75% to 99% is 11.8%

• The Fourth quadrant (100%)of students scored 91 marks and the difference between last two highest percentages are 7.2%

Entrance test Score Analysis:

• First quadrant(25%) of students scored 60 in Entrance and next 25% that means second quadrant (50%) of students scored 71 marks ,the increase percentage of 25% to 50% is 11%

• Second quadrant score(50%) is 71 and Third quadrant mark(75%) is 83.5 and the increase percentage between 50% to 75% is 12.5%

• Third quadrant(75%) of students scored 83.5 in Entrance and the highest percentage 99% is 97 ,the difference between 75% to 99% is 13.5%

• The Fourth quadrant (100%)of students scored 98 marks and the difference between last two highest percentages are 1%

MBA Score Analysis:

• First quadrant(25%) of students scored 57.9 in MBA and next 25% that means second quadrant (50%) of students scored 62 marks ,the increase percentage of 25% to 50% is 4.1%

• Second quadrant score(50%) is 62 and Third quadrant mark(75%) is 66.2 and the increase percentage between 50% to 75% is 4.2%

• Third quadrant(75%) of students scored 66.2 in MBA and the highest percentage 99% is 76.1,the difference between 75% to 99% is 9.9%

• The Fourth quadrant (100%)of students scored 77.8 marks and the difference between last two highest percentages are 1.7%

INTERQUARTILE RANGE[IQR]

About IQR (Inter quartile range)?

1] It represents the range of values within which the middle 50% of the data falls, making it less sensitive to outliers than the range

2] Calculated the difference between the third quartile (Q3) and the first quartile (Q1)

3]Quartile=Q1----25%-------Q2-----75%-----Q3-----MAX-----Q4

4]Calculate between the outlier range,50% data called central balance data.

Use of 1.5 in IQR (InterQuartile Range):

The 1.5 multiplier is commonly used in Tukey's fences, a method for identifying outliers. Tukey suggested that data points outside 1.5 times the IQR from the first and third quartiles could be considered potential outliers. This threshold is somewhat arbitrary but has become a widely accepted standard. A method for identifying potential outliers in a dataset. The idea is to establish an acceptable range within which most

data points are expected to fall, and anything outside this range is considered a potential outlier.

Outlier Detection: Identifies data points that deviate significantly from the rest of the dataset.

Data Cleaning: Helps remove or adjust outliers to improve data accuracy.

Quality Control: Used in manufacturing, finance, etc., to spot anomalies indicating errors or fraud.
Robust Estimation: Provides reliable estimates despite outliers' influence.

Data Visualization: Outliers can be highlighted in visualizations for better understanding.

Formula of IQR:
There are Two types of outliers:
1] LESSER RANGE
2]GREATER RANGE

Calculate between the outlier range,50% data called central balance data.

Lesser than quartile range: [IQR=Q3-Q1]
[Q3=75% Central balance data Q1=25%]
Q1-1.5rule*IQR

Calculate between outlier range,50% data called central balance data.
Greater than quartile range: [IQR=Q3-Q1]
[Q3=75% Central balance data Q1=25%]
Q3+1.5rule*IQR

A .The interquartile range.compare the 2 interquartile ranges.
B .Any outliers in either set

The five number summary for day & night is

	Min	Q1	Median	Q3	Max
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

Day

IQR=Q3-Q1
=82.5-56
=26.5

Lower outlier =Q1-1.5*IQR
=56-1.5*26.5
=16.25

Greater outlier=Q3+1.5*IQR
=82.5+1.5*26.5
=122.25

Comparing lower outlier value with min value ,there is no value lesser than 16.25,so there is no lesser outliers present in day & Comparing greater outlier value with max value ,there is value greater than 122.25,so there is no greater outliers present in day

Night
IQR=Q3-Q1
=89-78
=11

Lower outlier = Q1-1.5*IQR
=78-1.5*11
=61.5
Greater outlier=Q1+1.5*IQR
=89+1.5*11
=90.5

Comparing greater outlier value with max value ,there is value greater than 90.5,so there is greater outliers present in night and Comparing lower outlier value with min value ,there is value lesser than 61.5 ,so there is lesser outliers present in night, i.e, Present of lower and greater outliers in night.

Report for skew and Kurtosis

	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
99%	212.86	87	91.86	83.86	97	76.1142	NaN
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
kurtosis	-1.2	-0.60751	0.450765	0.0521433	-1.08858	-0.470723	18.5443
skewness	0	-0.132649	0.163639	0.244917	0.282308	0.313576	3.56975
Variance	3870	117.228	118.756	54.1511	176.251	34.0284	8.7343e+09
Standard deviation	62.2093	10.8272	10.8975	7.35874	13.276	5.83338	93457.5

SSC:

Referring ssc marks skewness value lies -0.132 which is less than 0 value ,it peak shows mode<median<mode, mode is occupied many places i.e., continuous values in one area, while comparing it using the graph position ,positive skewness

HSC:

In hsc marks in the table shows the value as 0.16 which lies greater than 0,it peaks shows mean<median<mode , while comparing it using the graph position ,negative skewness.

DEGREE:

In degree marks in the table shows the value as 0.20 which lies greater than 0,its peaks show mean<median<mode , while comparing it using the graph position ,negative skewness.
ENTRANCE TEST:

In Entrance test marks in the table show the value as 0.28 which lies greater than 0,it peaks shows mean<median<mode , while comparing it using the graph position ,negative skewness.

MBA:

In mba marks in the table shows the value as 0.3 which lies greater than 0,it peaks shows mean<median<mode , while comparing it using the graph position ,negative skewness.

SALARY:

In salary value in the table shows the value as 0.80 which lies greater than 0,it peaks shows mean<median<mode , while comparing it using the graph position ,negative skewness

KURTOSIS:

Kurtosis is a critical tool for understanding data distribution and outlier behavior.
It helps in making informed decisions about data preprocessing and modeling.

Types of Kurtosis

1. **Mesokurtic (Normal Distribution)**:
    - A kurtosis value close to 3.
    - Indicates a normal distribution with moderate tails.
    - Example: Bell curve.
2. **Leptokurtic (Heavy-Tailed)**:
    - A kurtosis value > 3.
    - Indicates a distribution with heavy tails and more outliers.
    - Example: Financial returns data.
3. **Platykurtic (Light-Tailed)**:
    - A kurtosis value < 3.
    - Indicates a distribution with light tails and fewer outliers.
    - Example: Uniform distribution.

How to Interpret Kurtosis in a Dataset?

Kurtosis = 3: Normal distribution.

Kurtosis > 3: Extreme outliers, heavier tails.

Kurtosis < 3: Few or no outliers, lighter tails.

In the above table we refer that all the values are less than 3 and the lies on platykurtic,which is less than 3

Since our data is Platykurtic, it's like saying the numbers are all playing it safe—there's nothing wild or extreme happening. This consistency can help us make reliable predictions or ensure stable processes

ssc_p (-0.60751)

The kurtosis is slightly negative, indicating a Platykurtic distribution.
The data for secondary school percentages (ssc_p) is more spread out, with fewer extreme outliers compared to a normal distribution.
hsc_p (0.450765)

Positive but close to zero, indicating a distribution that is slightly closer to normal.
The data for higher secondary percentages (hsc_p) has moderate variability, with no significant outliers.

degree_p (0.0521433)

Very close to zero, indicating it is almost normal.
Degree percentages have a balanced distribution with few or no extreme values.
etest_p (-1.08858)

Strongly negative kurtosis, indicating a Platykurtic distribution.
The employability test percentages (etest_p) are spread out with very few or no extreme data points, and the tails are lighter than normal.
mba_p (-0.470723)

Negative kurtosis indicates a Platykurtic distribution.
MBA percentages (mba_p) have fewer extreme values and are moderately spread out.
salary (18.5443)

A very high positive kurtosis indicates a Leptokurtic distribution.
Salary data contains extreme outliers, with some salaries being significantly higher than most others. This suggests heavy-tailed behavior in salary distribution.