DATA SCIENCE

UNIVARIATE

	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 mean, often referred to as the average
- It is calculated by summing up all the values in the dataset and then dividing by the total number of values.
- In school, the 10th-grade students attained an average mark 67%
- The identical group of students achieved an average mark in their 12th-grade examinations.66%
- The aforementioned students maintained an average mark throughout their degree program.66%
- The aforementioned students attained commendable marks in their entrance examinations.72%
- The same cohort of students experienced a decline in their marks during the MBA examination62%.

Median

- The mean value coincides with the median value; however, the divergence arises from the presence of outliers
- Which heavily influence the mean while being disregarded in the calculation of the median
- The difference between mean and median salary - 23655

Mode

- The mode represents the occurrence of a particular value with the highest frequency within a dataset
- indicating the most frequently repeated observation."
- The repetition of marks:

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

- The dataset reveals that students who achieve average marks typically earn salaries ranging between $288,655 and $300,000.

PERCENTILE

- **25th Percentile (60.6)** : This means that a quarter of the students scored 60.6
- **50th Percentile (67)** : This is the median score, where half of the students scored 67.
- **75th Percentile (75.7)** : Three-quarters of the students scored 75.7
- **100th Percentile(89.4)** : This value, 89.4, is typically considered as the highest score

Difference between 25th and 50th Percentile:

Difference=67−60.6=6.4

Difference between 50th and 75th Percentile:

Difference=75.7−67=8.7

Difference between 75th and 100th Percentile:

Difference=89.4−75.7=13.7

These descriptions provide a clearer understanding of how the scores increase progressively across the percentiles

IQR [INTER QUARTILE RANGE]

Why Multiply by 1.5?

- Tukey's Rule: Introduced by John Tukey, a prominent statistician, Tukey's rule states that outliers are typically defined as points that are less than Q1 - 1.5 * IQR or greater than Q3 + 1.5 * IQR.

Reasoning:

- **Symmetry**: The factor of 1.5 provides a symmetric range around the quartiles. This symmetry helps in applying the same criteria to both lower and upper bounds of the dataset.

- **Statistical Distribution**: While the normal (Gaussian) distribution is a key consideration, the use of 1.5 is not solely because it fits Gaussian distribution. It also balances sensitivity (the ability to detect outliers) with specificity (the ability to correctly identify non-outliers)

Example Sum of IQR

	MIN	Q1	MEDIAN	Q3	MAX
DAY	32	56	74.5	82.5	99
NIGHT	25.5	78	81	89	98

DAY

IQR		1.5*IQR	LOWER	GREATER
FORMULA	Q3-Q1	1.5*IQR	Q1-1.5*IQR	Q3+1.5*IQR
Q1=56 Q3=82.5 Q3-Q1=82.5-56=26.5 IQR=26.5		1.5*26.5=39.75	56-39.75=16.25	82.5+39.75=122.25
NO LOWER AND UPPER OUTLLIER				

NIGHT	MIN	Q1	MEDIAN	Q3	MAX
	25.5	78	81	89	98

NIGHT

IQR		**1.5*IQR**	LOWER	GREATER
FORMULA	Q3-Q1	1.5*IQR	Q1-1.5*IQR	Q3+1.5*IQR
Q1=78 Q3=89 Q3-Q1=89-78=11 IQR=11		1.5*11=16.5	78-16.5=61.5	89+16.5=105.5
NO LOWER AND UPPER OUTLLIER				

Histogram

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
kurtos	-1.2	-0.60751	0.450765	0.052143	-1.08858	-0.470723	18.544273

- Kurtosis is a measure of the tailedness of a distribution.
- **Different types Kurtosis**

➢ **Leptokurtic**: A distribution with positive kurtosis (excess kurtosis greater than 0) has heavier tails and a sharper peak than a normal distribution. This means it has more data in the tails and a more pronounced peak.

➢ **Mesokurtic**: A distribution with a kurtosis of exactly 3 (excess kurtosis of 0) has tails and a peak similar to those of a normal distribution.

➢ **Platykurtic**: A distribution with negative kurtosis (excess kurtosis less than 0) has lighter tails and is flatter compared to a normal distribution. This means it has less data in the tails and a less pronounced peak.

❖ Kurtosis of ssc_p values is **-0.60751** is **platykurtic**
❖ Kurtosis of hsc_p values is **0.450765** is **leptokurtic**
❖ Kurtosis of degree_p is **0.052143** is **leptokurtic**
❖ Kurtosis of etest_p is **-1.08858** is **Platykurtic**
❖ Kurtosis of mba_p is **-0.470723** is **Platykurtic**

Skewness

Skewness:

Skewness is a statistical measure that describes the asymmetry of the probability distribution of a real-valued random variable about its mean. In simpler terms, it measures the degree of asymmetry in the distribution of data points around their mean.

Positive Skewness:

Also known as right-skewed or right-tailed, this occurs when the tail on the right side of the distribution is longer or fatter than the left side. The mean and median of positively skewed data are typically greater than the mode.

Negative Skewness:

Also known as left-skewed or left-tailed, this occurs when the tail on the left side of the distribution is longer or fatter than the right side. The mean and median of negatively skewed data are typically less than the mode.

Interpretation

Skewness value of 0 indicates a perfectly symmetrical distribution.

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
skeweness	0.0	-0.132649	0.163639	0.244917	0.282308	0.313576	3.569747

- Skewness of ssc_p values is **-0.132649** is **Negative Skewness**
- Skewness of hsc_p values is **0.163639** is **Positive Skewness**
- Skewness of degree_p is **0.244917** is **Positive Skewness**
- Skewness of etest_p **0.282308** is **Positive Skewness**
- Skewness of mba_p is **0.313576** is **Positive Skewness**

PROBABILITY DENSITY FUNCTION

CODING:
```
def get_pdf_probability(data,startrange,endrange):
    from matplotlib import pyplot
    from scipy.stats import norm
     import seaborn as sns
```

- The code utilizes libraries like matplotlib and seaborn for visualization, and scipy.stats for statistical functions, to create a graph of the probability density function (PDF).

CODING:
```
sns.distplot(data,kde=True,kde_kws={"color":"blue"},color="green")
pyplot.axvline(startrange,color="red")
pyplot.axvline(endrange,color="red")
```

- Seaborn is employed for visualization to depict a bell-shaped curve, histograms, and probability density functions (PDFs)

CODING:
```
sample=data
sample_mean=sample.mean()
sample_std=sample.std()
print("mean=%.3f,standarddeviation=%.3f"%(sample_mean,sample_
std))
```

- Before calculating the PDF, it is necessary to compute the mean and standard deviation for the dataset.
- % operator is used for string formatting
- .3f specifies that floating-point numbers (f) should be displayed with three decimal places (.3).

CODING:

```
dist=norm(sample_mean, sample_std)
```

- **norm**: This is a function from the scipy.stats module that represents a normal (Gaussian) distribution
- Creating an instance of a normal distribution object (dist) that is parameterized by sample_mean (mean) and sample_std (standard deviation).

CODING:

```
values=[value for value in range(startrange,endrange)]
```

- Certainly! If you want to optimize the traditional "for" loop method of creating a list, you can integrate the list creation directly within the loop. This approach is more efficient and concise than initializing an empty list and appending elements to it iteratively.

CODING:

```
probabilities=[dist.pdf(value)for value in values]
prob=sum(probabilities)
```

- for more traditional "for" loop approach where the PDF values are computed within the loop itself, you would initialize an empty list and append the computed PDF values for each value in the range [startrange, endrange).