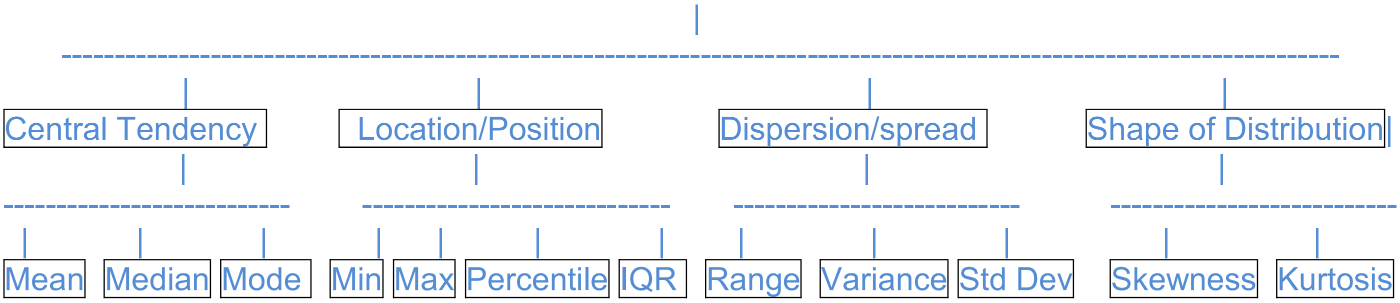
DATA SCIENCE- UNIVARIATE ANALYSIS

MEASURES IN UNIVARIATE ANALYSIS

Contents:

Measures in univariate analysis for Quantitative continuous variable

Univariate Analysis for Continuous values

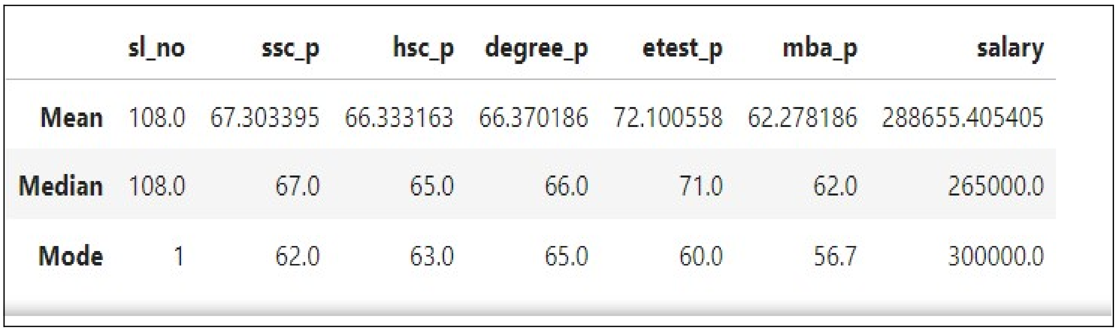


- Central Tendency
  - Mean
  - Median
  - Mode
- Location/Position
  - Min
  - Max
  - Percentile
  - IQR
- Dispersion/spread
  - Range
  - Variance
  - Std Dev
- Shape of Distribution
  - Skewness
  - Kurtosis

1) Measure of Central Tendency

These measures help to understand the center or average value of the data.

- **Mean**: The average value.
- **Median**: The middle value.
- **Mode**: The most frequent value.



	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

This table provides summary statistics for a placement dataset containing several variables related to students' academic and employment data. Here's an explanation of each row:

Mean:

The mean is the average value of each column in the placement data, representing the central tendency of the data.

- **ssc_p**: 67.30% is the average percentage obtained in secondary school.
- **hsc_p**: 66.33% is the average percentage in higher secondary school.
- **degree_p**: 66.37% is the average percentage obtained in the degree program.
- **etest_p**: 72.10% is the average score in the entrance test.
- **mba_p**: 62.28% is the average percentage obtained in the MBA program.
- **salary**: ₹288,655.41 is the average salary of the individuals in the dataset.

Median:

The **median** is the middle value when the data is arranged in ascending order. It's less affected by outliers compared to the mean.

- **ssc_p**: 67% is the middle value of secondary school percentages, meaning half the students scored below and half above this.
- **hsc_p**: 65% is the middle value of higher secondary school percentages.
- **degree_p**: 66% is the middle value of degree percentages.
- **etest_p**: 71% is the middle value of entrance test scores.
- **mba_p**: 62% is the middle value of MBA percentages.
- **salary**: ₹265,000 is the middle salary, showing that half the students earn less than this, and half earn more.

Mode:

The **mode** is the most frequently occurring value in each column.

- **ssc_p**: 62% is the most common secondary school percentage.
- **hsc_p**: 63% is the most common higher secondary percentage.
- **degree_p**: 65% is the most common degree percentage.
- **etest_p**: 60% is the most frequent entrance test score.
- **mba_p**: 56.7% is the most frequent MBA percentage.
- **salary**: ₹300,000 is the most common salary.

Real-Time Insights From This Data:

Academic Performance Distribution:

- The average scores across ssc_p, hsc_p, degree_p, and mba_p are fairly close, suggesting consistency in academic performance.
- The entrance test scores have the highest mean (72.10%), indicating that students tend to perform better in entrance exams than in degree or MBA exams.

Salary Insights:

- The average salary (₹288,655.41) is higher than the median salary (₹265,000). This suggests that a few higher salaries may be pulling up the average.
- The most common salary (₹300,000) is higher than both the mean and the median, indicating that most people earn around ₹300,000 but outliers with lower salaries reduce the median.

Outliers and Distribution:

- The gap between the mean and median in salary suggests skewness in the data, possibly due to some outliers (either very high or low salaries).
- The relatively close values of mean and median in academic scores indicate that there is a normal distribution for these variables without extreme outliers.

Common Traits:

- Many students seem to score around 65% in their degree and MBA percentages, as reflected by the mode in degree_p and mba_p.
- Students often achieve around 62% in their secondary school scores and 63% in their higher secondary scores, showing a pattern of slightly lower scores in early education, with improvement in later stages.

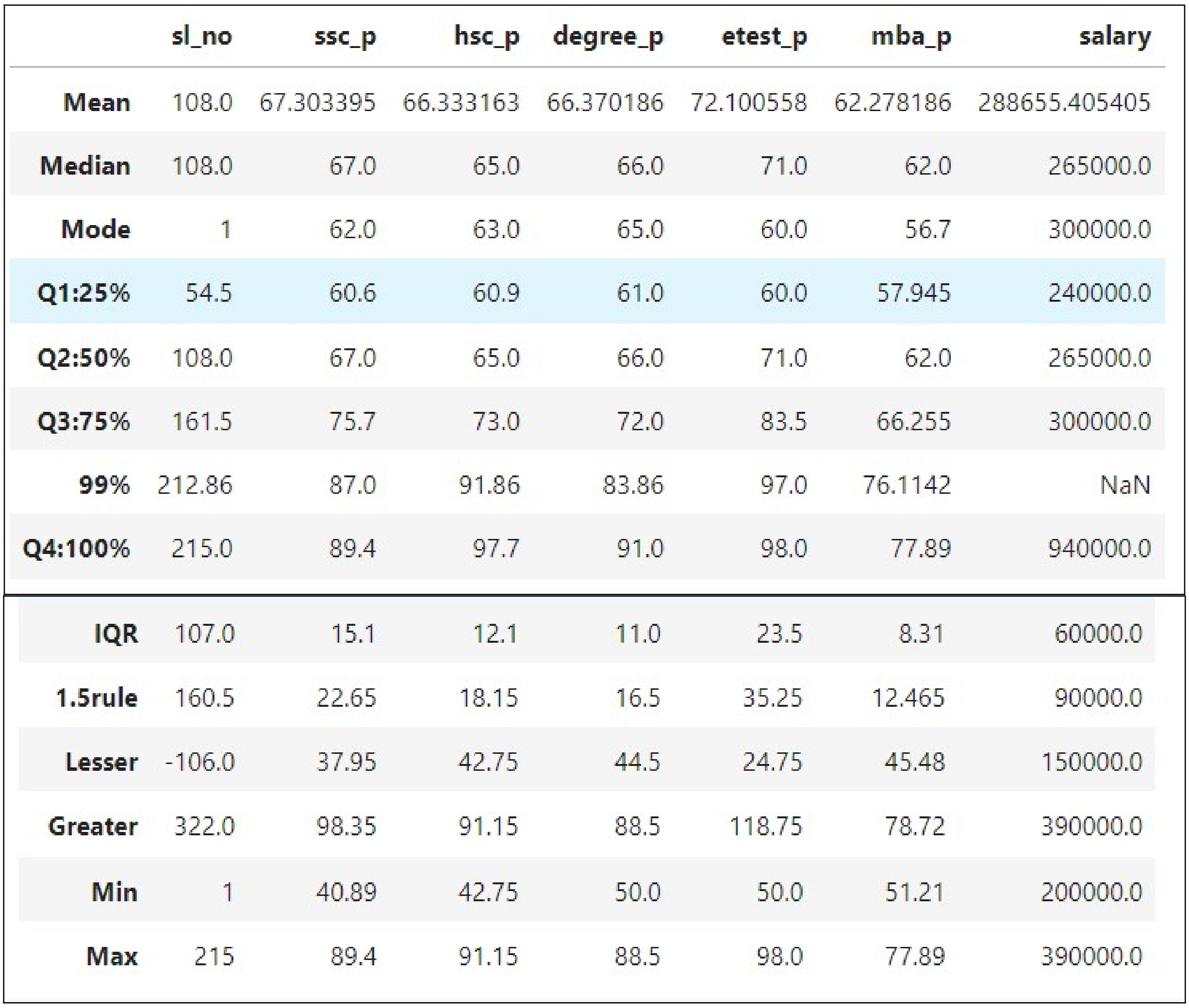
2) Measures of Location

These show where specific data points lie within the distribution.

- **Minimum (Min)**: The smallest value in the dataset.
- **Maximum (Max)**: The largest value in the dataset.
- **Percentiles**: Points that divide the data into 100 equal parts (e.g., the 25th percentile represents the value below which 25% of the data fall).
- **Quartiles**: Special percentiles (Q1: 25th percentile, Q2: 50th percentile/median, Q3: 75th percentile).
- **Interquartile Range (IQR)**: The range between the 25th and 75th percentiles, representing the spread of the middle 50% of the data.

Percentile:

A percentile is a statistical measure that indicates the relative standing of a data point within a dataset. It divides a set of data into 100 equal parts, so each percentile represents 1% of the data. Percentiles are often used to understand the distribution and variation within data, and they can give insights into how individual data points compare to the rest of the dataset.



	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	-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	42.75	50.0	50.0	51.21	200000.0
Max	215	89.4	91.15	88.5	98.0	77.89	390000.0

How Percentiles Work:

Nth Percentile: The Nth percentile is the value below which **N%** of the data falls. For example, the 90th percentile is the value below which 90% of the data points lie.

Example: If a student scores in the 80th percentile on a test, it means he performed better than 80% of the people who took the test.

Key Percentiles in Data Analysis:

- **25th Percentile (Q1)**: This is also called the **first quartile**, indicating that 25% of the data lies below this value. It's useful for identifying the lower boundary of the data.
  - Example: If the 25th percentile salary is ₹240,000, it means 25% of people earn less than ₹240,000.
- **50th Percentile (Median or Q2)**: This is the **middle** value (median) of the dataset, where 50% of the data falls below it. It divides the dataset into two equal halves.
  - Example: If the median salary is ₹265,000, then half the people earn below ₹265,000 and the other half earn above it.
- **75th Percentile (Q3)**: Known as the **third quartile**, this value indicates that 75% of the data lies below this point.
  - Example: If the 75th percentile salary is ₹300,000, then 75% of people earn less than ₹300,000.
- **99th Percentile**: This represents the **extreme high** end of the data. Only 1% of data points lie above this value.
  - Example: If the 99th percentile for salary is ₹900,000, it means that only 1% of people earn more than ₹900,000.
- **Q4: 100%:** This is the maximum value in the dataset.
  - Example: The highest salary is 940,000, and the highest MBA percentage (mba_p) is 77.89%.

Real-life Use Cases of Percentiles:

- **Exams**: Percentiles are often used to rank students. A student in the 90th percentile scored higher than 90% of the test-takers.
- **Income Distribution**: In salary analysis, percentiles can show income disparity. For example, the 10th percentile income could be ₹20,000 (very low earners), and the 90th percentile income could be ₹900,000 (high earners).
- **Health Statistics**: In medical fields, percentiles help assess children's growth. A baby in the 80th percentile for height is taller than 80% of other babies of the same age.

Key Points:

- **Higher percentiles indicate better performance** (or larger values) compared to the rest of the dataset.
- Percentiles give an idea of how data is distributed and where individual points stand in relation to the whole.

Inter Quartile Range (IQR):

It is a measure of statistical dispersion, indicating the spread of the middle 50% of a dataset. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1)

IQR= Q3−Q1

Key Points:

- **Quartiles** divide a dataset into four equal parts.
    - **First Quartile (Q1)**: 25th percentile
    - **Second Quartile (Q2)**: 50th percentile (median)
    - **Third Quartile (Q3)**: 75th percentile

- The IQR helps to identify outliers.
- Values that fall below Q1−(1.5×IQR) or above Q3+(1.5×IQR) are typically considered outliers.

Why the number 1.5 is used to detect outliers?

1. Empirical Basis

The value of 1.5 was chosen through empirical studies of various datasets. It has been found to work well across a wide range of data distributions to identify outliers without being too aggressive or too lenient.

- **Too Small a Multiplier**: If the multiplier is less than 1.5 (e.g., 1.0), many non-outliers (normal data points) may be incorrectly flagged as outliers. This would result in over-identification of outliers.
- **Too Large a Multiplier**: If the multiplier is much greater than 1.5 (e.g., 3.0), it might miss true outliers, especially if the data is skewed or has subtle anomalies. A large multiplier makes the method less sensitive.

2. Symmetry with Normal Distribution

The 1.5 multiplier roughly aligns with how outliers behave in a normal distribution, but it's not strictly tied to normality. In a normal distribution, most data points lie within 1.5 times the IQR from the median.

Specifically, in a normal distribution:

- About 68% of the data lies within 1 standard deviation (SD) from the mean.
- About 95% lies within 2 SDs.

The IQR contains the middle 50% of the data, and using 1.5×IQR ensures that the boundaries extend well beyond this central range (around 99% coverage for normal distributions). This catches extreme values without being overly strict.

3. Robustness to Skewed Distributions

The IQR and the 1.5 multiplier work well even with non-normal or skewed data. Unlike methods that rely on the mean and standard deviation (which are sensitive to extreme values), the IQR is a robust measure of spread because it focuses on the middle 50% of the data. The 1.5 multiplier scales this range to detect values that are unusually far from the majority of the data.

4. Simplicity and Widespread Use

The 1.5 multiplier is simple to apply and widely accepted in practice. It gives practitioners a consistent way to define what constitutes an outlier across different datasets. It strikes a balance between catching real outliers and avoiding flagging too many data points as outliers.

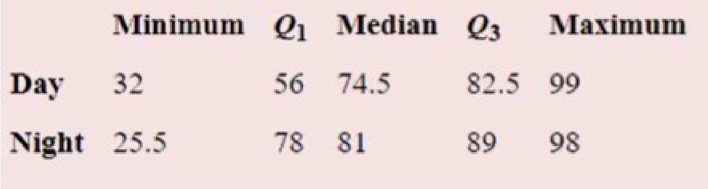
In summary, the value **1.5** is used because it:

- Provides a practical balance for identifying outliers across various distributions.
- Is effective for both normal and skewed datasets.
- Ensures robustness by using the IQR, which is not easily influenced by extreme values.

This is why 1.5 is the default choice in many outlier detection methods, such as boxplots.

How to find Lesser and Greater outlier for given dataset?

The 5 number summary for the day and night classes are



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

Finding Outlier for Day

IQR = Q3-Q1 = 82.5-56

IQR=26.5

Lesser Outlier = Q1-(1.5XIQR) = 56-(1.5X26.5)

=16.25

Greater Outlier = Q3+(1.5XIQR) = 82.5+(1.5X26.5)

=122.25

Finding Outlier for Night

IQR = Q3-Q1 = 89-78

IQR=11

Lesser Outlier = Q1-(1.5XIQR) = 78-(1.5X11)

=61.5

Greater Outlier = Q3+(1.5XIQR) = 89+(1.5X11)

=105.5

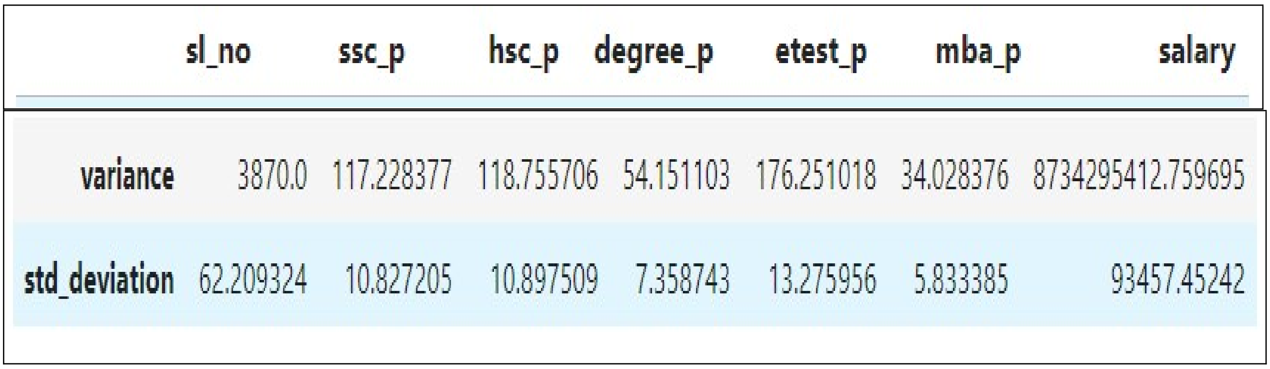
Here, we have to filter out the datas that are outside the bounds(lesser and greater) , which results in a more consistent dataset.

- Day data has no outliers
- Night Data has lesser range outlier (Min value=25.5) which is less than lesser bound 61.5, So 25.5 has to be replaced with 61.5.

3. Measure of Spread (Dispersion):

These measures describe how much the data varies or spreads out from the center, quantify how spread out the data points are, indicating the variability or consistency of the dataset.

- **Range**: The difference between the maximum and minimum values.
- **Variance**: The average of the squared differences from the mean.
- **Standard Deviation**: The square root of the variance, representing the typical deviation from the mean.
- **Coefficient of Variation (CV)**: The ratio of the standard deviation to the mean, often used to compare spread across datasets.



	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
variance	3870.0	117.228377	118.755706	54.151103	176.251018	34.028376	8734295412.759695
std_deviation	62.209324	10.827205	10.897509	7.358743	13.275956	5.833385	93457.45242

This table provides two important measures of dispersion: **variance** and **standard deviation** for various variables (e.g., **ssc_p**, **hsc_p**, **degree_p**, **etest_p**, **mba_p**, and **salary**). These metrics help quantify the spread of the data.

Variance:

Variance is the average of the squared differences from the mean. A higher variance indicates greater spread or variability in the dataset.

- **sl_no**: 3870 – High variance for serial numbers.
- **ssc_p**: 117.23 – Moderate variance for secondary school percentage.
- **hsc_p**: 118.76 – Similar to **ssc_p**, a moderate spread in higher secondary percentages.
- **degree_p**: 54.15 – Lower variance for degree percentages.
- **etest_p**: 176.25 – Higher spread in employment test scores.
- **mba_p**: 34.03 – Lower variance in MBA percentages.
- **salary**: 8734295412.76 – Extremely high variance, showing that salaries vary significantly.

Standard Deviation:

The standard deviation is the square root of the variance and is a commonly used measure to understand how much data deviates from the mean.

- **sl_no**: 62.21 – High standard deviation for serial numbers (which might not be meaningful in terms of analysis).
- **ssc_p**: 10.83 – Around 10% spread in secondary school percentage scores.
- **hsc_p**: 10.90 – Similar spread in higher secondary percentages.
- **degree_p**: 7.36 – Less spread for degree percentages.
- **etest_p**: 13.28 – Large spread for employment test scores.
- **mba_p**: 5.83 – Lower spread in MBA percentages.
- **salary**: 93457.45 – Extremely high standard deviation for salary, indicating a large variation in salaries.

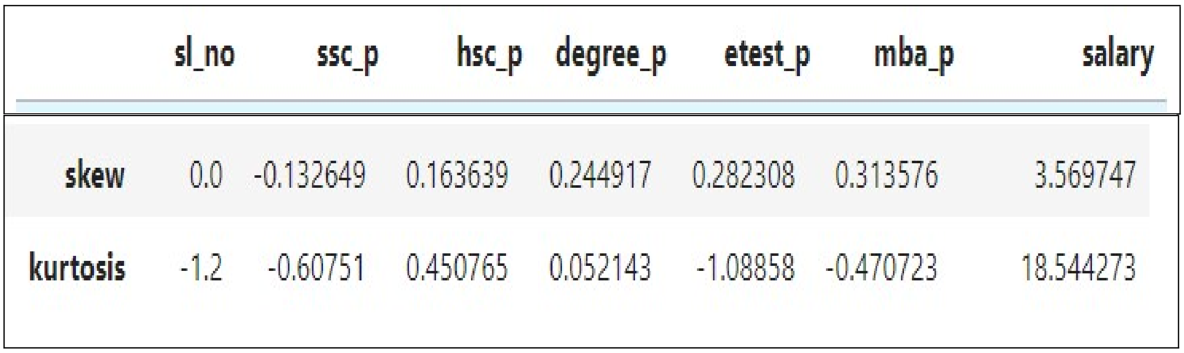
Summary:

- The **salary** variable shows both a very high variance and standard deviation, indicating a wide range of salaries in the dataset, with extreme differences between the lowest and highest salaries.
- Most of the percentage variables (**ssc_p**, **hsc_p**, **degree_p**) have moderate variability, with **etest_p** having slightly higher variability compared to others.
- **MBA percentages (mba_p)** have the least variability, showing that the MBA scores are more tightly clustered around the mean.

4. Measure of Shape (Distribution):

These measures describe the shape or symmetry of the data distribution, explains the general shape of the data distribution, indicating whether it is symmetric, skewed, or has outliers

- **Skewness**: Describes the asymmetry of the data distribution.
  - ◆ **Positive Skew**: Tail on the right side (right-skewed).
  - ◆ **Negative Skew**: Tail on the left side (left-skewed).
- **Kurtosis**: Describes the "tailedness" of the data distribution.
  - ■ **High Kurtois**: Indicates heavy tails and more outliers.
  - ■ **Low Kurtosis**: Indicates lighter tails and fewer outliers.



	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
skew	0.0	-0.132649	0.163639	0.244917	0.282308	0.313576	3.569747
kurtosis	-1.2	-0.60751	0.450765	0.052143	-1.08858	-0.470723	18.544273

This table shows the **skewness** and **kurtosis** values for different variables such as **ssc_p**, **hsc_p**, **degree_p**, **etest_p**, **mba_p**, and **salary**. These are measures of a dataset's distribution shape and outliers. Here's an explanation:

Skewness:

Skewness measures the asymmetry of the distribution.

- A skewness of **0** indicates a perfectly symmetrical distribution.
- Negative skewness means the tail on the left side of the distribution is longer or fatter than the right.
- Positive skewness means the tail on the right side is longer or fatter.

Values:

- **sl_no**: 0.0 – Symmetrical distribution.
- **ssc_p**: -0.13 – Slightly left-skewed (negative).
- **hsc_p**: 0.16 – Slightly right-skewed (positive).
- **degree_p**: 0.24 – Slightly right-skewed (positive).
- **etest_p**: 0.28 – Right-skewed.
- **mba_p**: 0.31 – Moderately right-skewed.
- **salary**: 3.57 – Highly right-skewed (indicating a small number of very high salaries).

Kurtosis:

Kurtosis measures the "tailedness" of the distribution.

- **Kurtosis = 3** represents a normal (Gaussian) distribution.
- **Kurtosis > 3** means the distribution has heavier tails (more extreme outliers) and is called leptokurtic.
- **Kurtosis < 3** indicates a lighter tail (fewer outliers), called platykurtic.

Values:

- **sl_no**: -1.2 – Platykurtic, indicating fewer outliers.
- **ssc_p**: -0.61 – Platykurtic, fewer outliers.
- **hsc_p**: 0.45 – Close to normal, very few outliers.
- **degree_p**: 0.05 – Very close to a normal distribution.
- **etest_p**: -1.09 – Platykurtic, few outliers.
- **mba_p**: -0.47 – Slightly platykurtic.
- **salary**: 18.54 – Highly leptokurtic, indicating many extreme outliers in salaries.

Summary:

- Most of the variables (such as **ssc_p**, **hsc_p**, **degree_p**) are close to being symmetrical and have a low number of extreme values (platykurtic).
- **Salary** is highly right-skewed with extreme outliers, suggesting a small number of individuals earning significantly higher salaries than the rest.