Understanding IQR

* **Interquartile Range (IQR):** The IQR is a measure of statistical dispersion and is calculated as the difference between the third quartile (Q3) and the first quartile (Q1).
  + **Q1 (First Quartile):** The 25th percentile, meaning 25% of the data is below this value.
  + **Q3 (Third Quartile):** The 75th percentile, meaning 75% of the data is below this value.

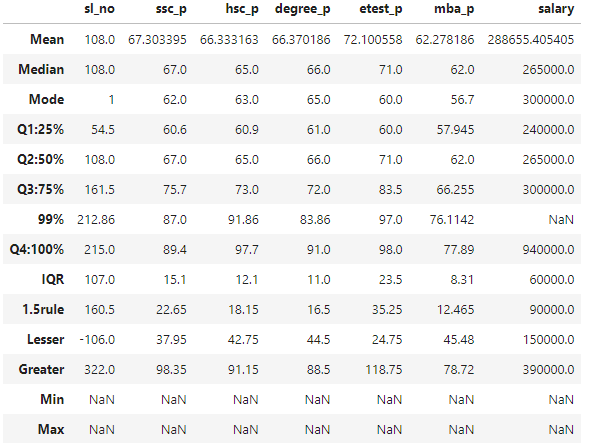
**Steps to Calculate the Lesser and Greater Outer Ranges**

1. **Calculate the Quartiles:**
   * **Q1 (First Quartile):** This is the 25th percentile of the data, meaning 25% of the data points are below this value.
   * **Q3 (Third Quartile):** This is the 75th percentile of the data, meaning 75% of the data points are below this value.
2. **Calculate the Interquartile Range (IQR):**
   * **IQR = Q3 - Q1**
   * This measures the spread of the middle 50% of the data.
3. **Calculate the "Lesser" Outer Range:**
   * **Lesser Range = Q1 - 1.5 × IQR**
   * This defines the lower bound. Any data point below this value is considered an outlier on the lower end of the distribution.
4. **Calculate the "Greater" Outer Range:**
   * **Greater Range = Q3 + 1.5 × IQR**
   * This defines the upper bound. Any data point above this value is considered an outlier on the upper end of the distribution.

The use of 1.5 times the Interquartile Range (IQR) as a threshold to identify outliers is a well-established statistical method. Here's an explanation of why the factor of 1.5 is used in both the "Lesser" and "Greater" ranges.

**Why Use 1.5 times the IQR?**

1. **Historical and Practical Basis:**
   * The 1.5 multiplier is not arbitrary but was proposed by John Tukey, a renowned statistician, as a simple rule of thumb to identify outliers. Tukey's work in exploratory data analysis (EDA) introduced this method as a way to detect data points that are unusually far from the central bulk of the data.
2. **Outlier Detection:**
   * **Outliers:** Outliers are data points that deviate significantly from the rest of the data. They can occur due to variability in the data, experimental errors, or other factors.
   * The IQR represents the range of the middle 50% of the data, capturing the central tendency without being influenced by extreme values.
   * By considering points that fall more than 1.5 times the IQR away from the quartiles, we are flagging data that is unusually far from where most data points lie.
3. **Balance Between Sensitivity and Robustness:**
   * **Sensitivity:** The 1.5 factor strikes a balance between being too sensitive (flagging too many points as outliers) and too robust (failing to identify significant outliers).
   * If the multiplier were much smaller (e.g., 1.0), it might classify too many data points as outliers, including those that are simply on the high or low end of the normal range. If the multiplier were much larger (e.g., 3.0), it might miss legitimate outliers that could be important for analysis.
4. **Why Subtract and Add 1.5 IQR?**
   * **Lesser IQR (Lower Bound):**
     + Q1−1.5×IQRQ1 - 1.5 \times IQRQ1−1.5×IQR
     + This formula identifies the lower bound. Any data point below this value is unusually low compared to the central data distribution and is flagged as a potential outlier.
   * **Greater IQR (Upper Bound):**
     + Q3+1.5×IQRQ3 + 1.5 \times IQRQ3+1.5×IQR
     + This formula identifies the upper bound. Any data point above this value is unusually high compared to the central data distribution and is flagged as a potential outlier.



"Lesser" and "Greater” columns indicate the lower and upper boundaries of potential outliers according to the 1.5 times the interquartile range (IQR) rule.

1. **Lesser Outer Range:**
   * These values are calculated as Q1−1.5×IQRQ1 - 1.5 \times IQRQ1−1.5×IQR. Any value below this is considered an outlier.
   * The "Lesser" row provides these boundary values:
     + **(10th std score) ssc\_p**: 37.9537.9537.95
     + **(12th std score) hsc\_p**: 42.7542.7542.75
     + **(College score) degree\_p**: 44.444.444.4
     + **(Entrance test score) etest\_p**: 24.4524.4524.45
     + **(MBA score) mba\_p**: 45.4845.4845.48
     + **Placed Salary Details**: 150000150000150000

Any values in the dataset below these thresholds could be considered outliers.

1. **Greater Outer Range:**
   * These values are calculated as Q3+1.5×IQRQ3 + 1.5 \times IQRQ3+1.5×IQR. Any value above this is considered an outlier.
   * The "Greater" row provides these boundary values:
     + **(10th std score) ssc\_p**: 98.3598.3598.35
     + **(12th std score) hsc\_p**: 91.1591.1591.15
     + **(College degree score) degree\_p**: 88.588.588.5
     + **(Entrance test score) etest\_p**: 118.75118.75118.75
     + **(MBA score) mba\_p**: 78.7278.7278.72
     + **Placed Salary Details**: 390000390000390000

Any values in the dataset above these thresholds could be considered outliers.

**Summary:**

* **"Lesser" Range**: Represents the lower bound of the data. Values below this can be considered lower outliers.
* **"Greater" Range**: Represents the upper bound of the data. Values above this can be considered upper outliers.

Identifying which individual rows have values in these columns that fall outside these "lesser" and "greater" thresholds, indicate potential outliers that may need to be handled during analysis.