**ASSIGNMENT I – DATASCIENCE UNIVARIATE ANALYSIS**

**Why 1.5 times IQR? Why not one or two or any other number?**

The IQR method of [outlier detection](https://builtin.com/data-science/how-find-outliers-examples) is a method that dictates that any data point in a [boxplot](https://builtin.com/data-science/boxplot) that’s more than 1.5 IQR points below the first [quartile data](https://builtin.com/data-science/how-to-calculate-quartiles) or more than 1.5 IQR points above the third quartile data is considered an outlier

The choice of using 1.5 times the Interquartile Range (IQR) as a threshold in identifying outliers in a dataset is a common convention in statistics.

The IQR is a measure of statistical dispersion, or spread, and is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of the data.

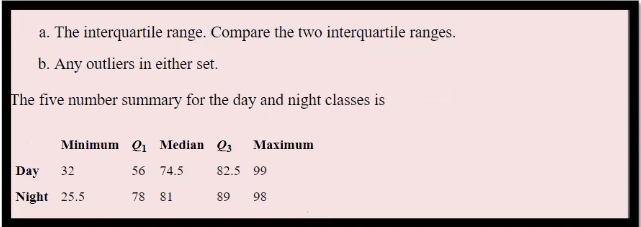
* The interquartile (IQR) method of outlier detection uses 1.5 as its scale to detect outliers because it most closely follows **Gaussian distribution**.
* As a result, the method dictates that any data point that's 1.5 points below the lower bound quartile or above the upper bound quartile is an outlier
* A commonly used rule says that a data point is an outlier if it is more than 1.5 ⋅ IQR ‍ above the third quartile or below the first quartile.
* Said differently, low outliers are below Q 1 − 1.5 ⋅ IQR ‍ and high outliers are above Q 3 + 1.5 ⋅ IQR

When it comes to detecting outliers using the IQR method, multiplying the IQR by 1.5 is a standard practice ***because it strikes a balance between sensitivity to outliers and robustness to noise in the data***.

**This multiplier of 1.5 is not a strict rule but a generally accepted standard** that provides a good compromise between detecting potential outliers and avoiding false alarms.

* Choosing a multiplier value higher than 1.5 would make the method more sensitive to outliers but might also lead to detecting too many false positives, while choosing a lower multiplier would make the method less sensitive to outliers, potentially missing some outliers in the data.
* Ultimately, the choice of 1.5 times the IQR as a threshold in outlier detection is a common convention that has been found to work well in practice for many datasets.
* However, in specific cases or for different types of data, researchers may choose to adjust this multiplier based on the characteristics of the data and the goals of the analysis.

**ASSIGNMENT II – CALCULATION OF IQR**



1. Compare the two interquartile ranges:

**DETAILED COMPARISON**

1. **First Quartile (Q1) Comparison**:
   * DAY: 56
   * NIGHT: 78

The first quartile in the NIGHT time classes is significantly higher than in the DAY time classes, indicating that 25% of the NIGHT time class data is above 78, while 25% of the DAY time class data is above 56.

1. **Third Quartile (Q3) Comparison**:
   * DAY: 82.5
   * NIGHT: 89

The third quartile in the NIGHT time classes is also higher than in the DAY time classes, indicating that 75% of the NIGHT time class data is below 89, while 75% of the DAY time class data is below 82.5.

1. **Interquartile Range (IQR) Comparison**:
   * DAY: 26.5
   * NIGHT: 11

IQR “DAY” = Q3-Q1 = 82.5-56

**DAY IQR = 26.5**

IQR”NIGHT” = Q3-Q1=89-78

**NIGHT** **IQR**= **11**

The interquartile range is much larger in the DAY time classes (26.5) compared to the NIGHT time classes (11). This indicates that the spread of the middle 50% of data points is greater in the DAY time classes than in the NIGHT time classes. The DAY time classes have more variability in the middle range of values.

**Implications**

* **Central Tendency**: The central tendency (measured by the quartiles) is higher for NIGHT time classes. Both Q1 and Q3 are higher in the NIGHT time classes compared to the DAY time classes, suggesting that the NIGHT time classes tend to have higher values overall.
* **Outliers**:
  + **DAY time classes**: No outliers were found.

1. Any outliers in the above dataset?

**Lesser Outliers – DAY**

**Minimum Value is 32**

**Maximum Value is 99**

Lesser Outliers Range = Q1-1.5\*IQR

56-(1.5\*26.5) = 56-(39.75) = **16.25**

Greatest Outliers Range = Q3+1.5\*IQR

82.5-(1.5\*26.5) = 82.5 + (39.75) = **122.25**

So, any value less than 16.25 or greater than 122.25 is considered an outlier. Since the minimum value is 32 and the maximum is 99, there are no outliers in the DAY time class data

**Lesser Outliers – NIGHT**

**Minimum Value is 25.5**

**Maximum Value is 98**

Lesser Outliers Range = Q1-1.5\*IQR

78-(1.5\*11) = 78-(16.5) = **61.5**

Greatest Outliers Range = Q3+1.5\*IQR

89-(1.5\*11) = 89+ (16.5) = **105.5**

So, any value less than 61.5 or greater than 105.5 is considered an outlier. Given the minimum value is 25.5 and the maximum is 98:

* 25.5 is an outlier because it is less than 61.5.
* There are no outliers on the upper end since the maximum value (98) is within the range.

**Summary:**

* **DAY time classes**: No outliers.
* **NIGHT time classes**: One lesser outlier at 25.5.