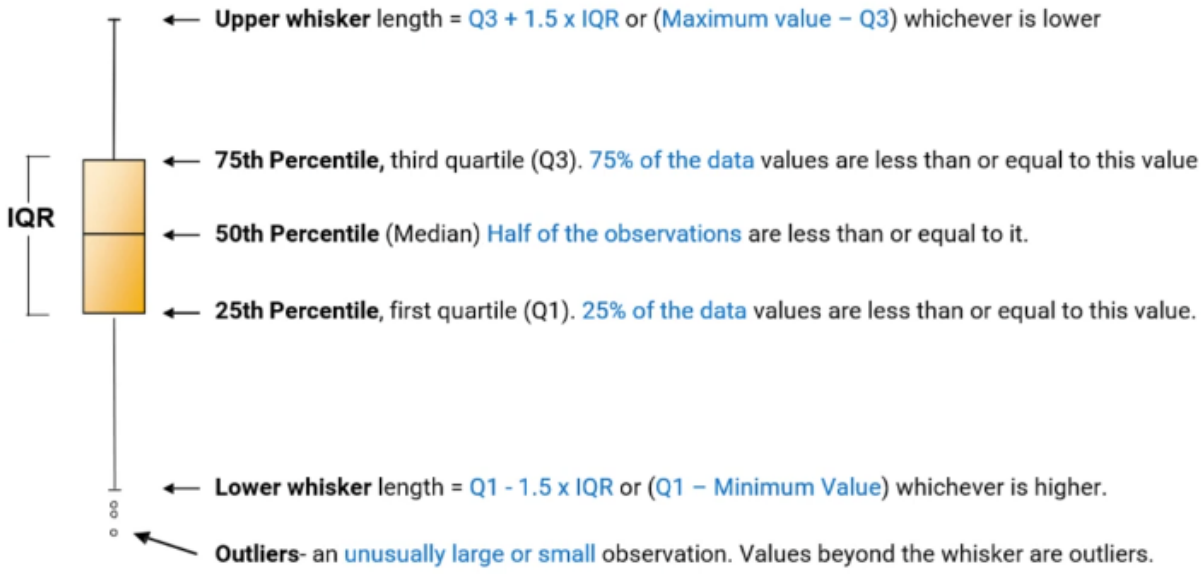
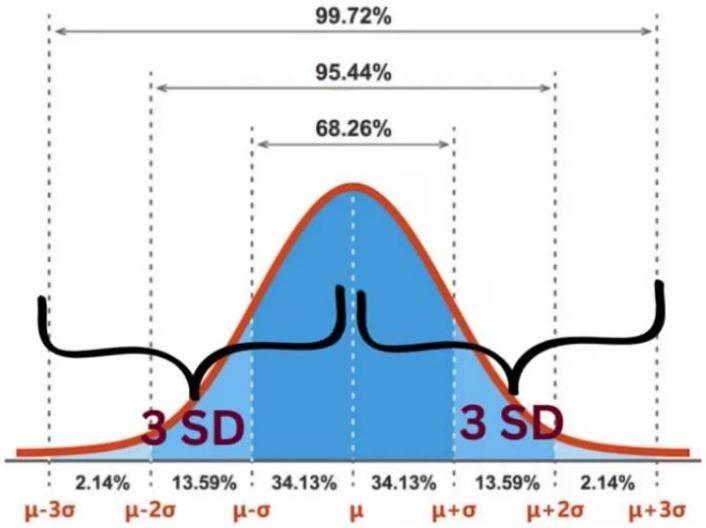
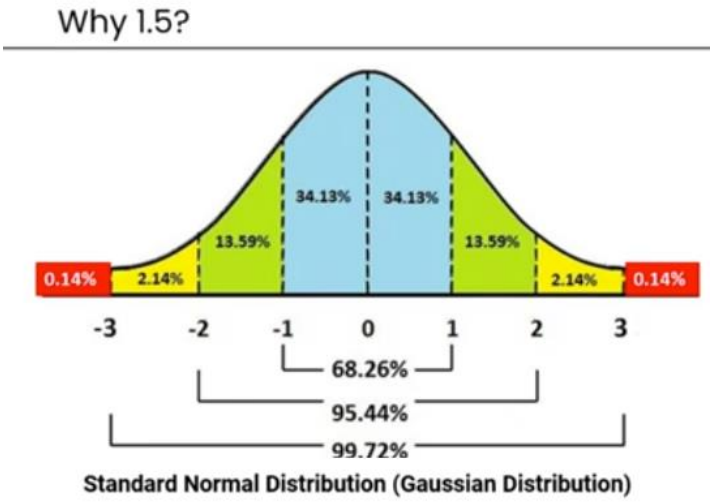
Why does multiply1.5xIQR?

The interquartile (IQR) method of outlier detection uses 1.5 as its scale to detect outliers because it most closely follows Gaussian distribution

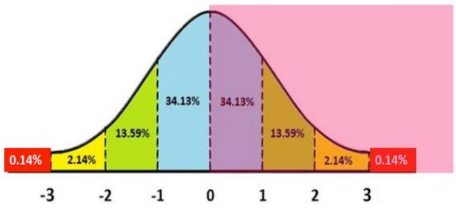
This decision range is the closest to what Gaussian Distribution tells us, i.e., 3σ . In other words, this makes the decision rule closest to what Gaussian distribution considers for outlier detection, and this is exactly what we wanted.



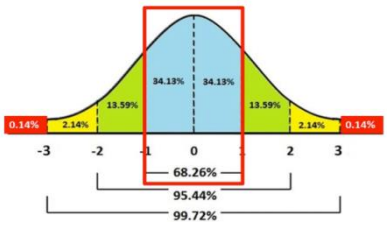




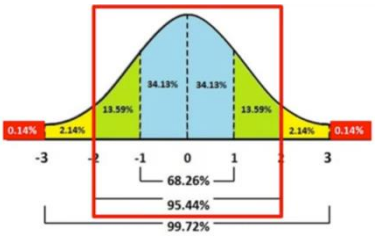
- Standard normalization is exactly symmetrical
- that means Left side of the distribution is exactly same as right side of the distribution



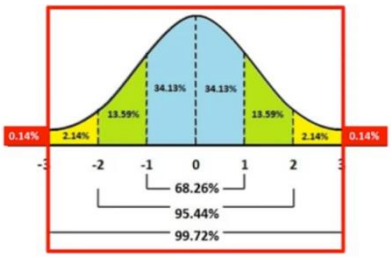
- The standard normal distribution is it consists of 68.26% of the data points when we talk about mean plus or minus one sigma.



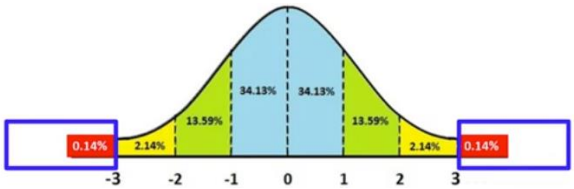
- It also consists of 95.44% of the data points, when we talk about mean plus or minus two sigma. here sigma is standard deviation of the data.



- next consists of 99.72% of the data points, when we talk about mean plus or minus three sigma.



- What is about the remaining 0.28% of the data points?
- It is evenly distributed on both side of the standard normal distribution. 0.14% on the left side and 0.14% on the right side of the distribution.
- here looking the data points which is useful for us



- We are looking for mean plus or minus here sigma data points which is useful for us.
- Which is going beyond that we are calling them as unusual observation

We are going to understand what is the z value corresponding to 25% and 75% of the probability; It is exactly coming at the middle of 0.67 and 0.68, so let's take that value as 0.675.

we have already seen Standard normal distribution is symmetrical. So we can say 25% on either side of the mean we are having the same z value. so it is +/- 0.675

We take the scale is equal to 1 then according to IQR method any data which lie beyond 2.02 sigma from the mean on either side shall be consider as an outlier, but as we know up to 3 sigma on either side on the mean the data is useful.

We can't take scale is equal to 1 because the decision range gets so small compare to 3 sigma that considers some data points as outlier even if they are not outlier. This is not desirable situation

Consider scale =2

We take the scale is equal to 2 then according to IQR method any data which lie beyond 3.3375 sigma from the mean on either side shall be consider as an outlier, but as we know up to 3 sigma on either side on the mean the data is useful

We can't take scale is equal to 2 because the decision range gets so s big compare to 3 sigma that considers some outlier as data points as outlier even This is not desirable condition.

Consider scale =1.5

We take the scale is equal to 1.5 then according to IQR method any data which lie beyond 2.7 sigma from the mean on either side shall be consider as an outlier, but as we know up to 3 sigma on either side on the mean the data is useful

If we take scale is equal to 1.5 this makes the decision range closes to what Gaussian distribution considers for outliers. This is exactly what we want.

So we consider 1.5 are a good scale to calculate the outlier.

Can be 1.5 is the exact value? No

The closest value of the scale which is giving three sigma is 1.7.

Consider scale =1.7

We take the scale is equal to 1.7 then according to IQR method any data which lie beyond 2.97 sigma from the mean on either side shall be consider as an outlier, but as we know after 3 sigma on either side on the mean the data is useful.

If we take scale is equal to 1.7 this makes the decision range he same as that of the Gaussian distribution considers for outliers detection. This is the desirable situation.

Instead of 1.7 why we are using the 1.5?

The IQR method is not a perfect method to detect the outlier therefore to keep the tolerance we are using 1.5 on safer side

Interquartile Range(IQR)

a. The interquartile range. Compare the two interquartile ranges.

b. Any outliers in either set.

The five number summary for the day and night classes is

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

Day shift

IQR= Q3-Q1 = 82.5 – 56 = 26.5

1.5xIQR=1.5*26.5 =39.75

Lesser outlier value = Q1-1.5*IQR = 56-1.5x39.75 = 16.25

Greater outlier value = Q3+1.5xIQR = 82.5+1.5x39.75 = 122.25

The day shit value is not less than 16.25 or greater than 122.25 values.so day shift don't have any potential outlier value.

Night Shift

IQR= Q3-Q1 = 89 – 78 = 11

1.5xIQR=1.5*11 =16.5

Lesser outlier value = Q1-1.5*IQR = 78-1.5x11= 61.5

Greater outlier value = Q3+1.5xIQR = 89+1.5x11= 105.5

The night shit no value is greater than 105.5, however minimum value is less than lesser outlier value, the 25.5 value is potential outlier value.