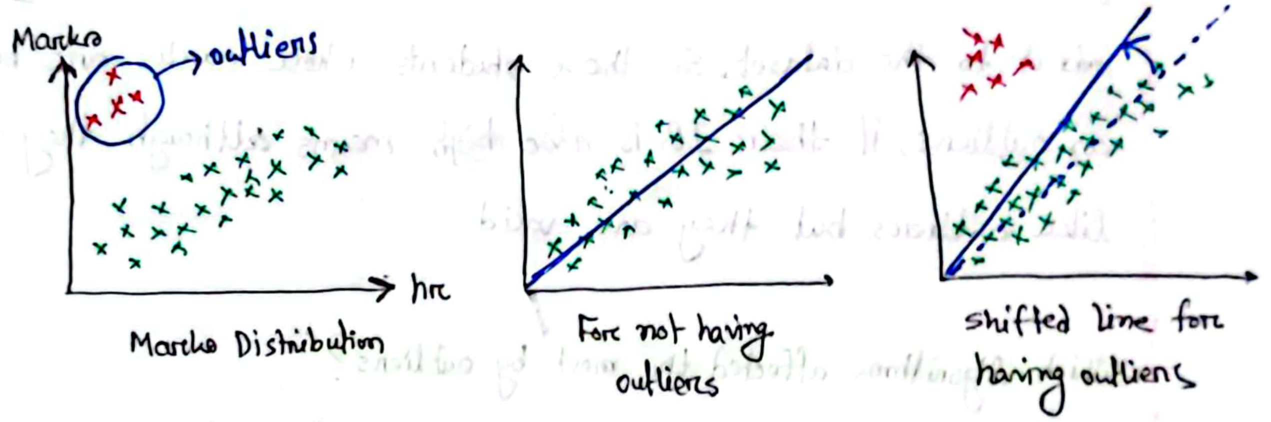
Outliers:

Let's check a data distribution of student no. of given hours vs No. of marks got.



Marks Distribution

For not having outliers

shifted line for having outliers

So, we can see we cannot get a proper line in the Linear Regression model instead a shifted line for having outliers in the data.

When outliers are really dangerous and really needs to remove?

Suppose, in Age Dataset Age comes 300.

So, we should remove those outliers whose values are unusual.

Age
20
34
18
300
245
70

outliers with unusual value

But in some cases outliers are important. For example, in Anomaly Detection case where we are working on Credit Card Fraud detection, in those cases we are actually working on to find/detect outliers.

So, we can't delete outliers always. We have to think based on the problem. Suppose, the first case we decided, student study hours vs Marks. Outliers are not very unusual and it's in the hieghest mark range.

In that case it is not justified to delete outliers there. To justify that outlier behaviour, we can take another column named "IQ" and add it to the dataset. So, those students whose marks are behaving as outliers, if their IQ is also high means although they are behaving like outliers but they are valid.

Which Algorithms affected the most by outliers?

→ Liner, Logistic Regression

→ Adaboost

→ Deep Learning

We are calculating weights in these algorithms. So, the algorithms which calculates weights, they are mostly affected by outliers.

There are algorithms which don't get affected by outliers →

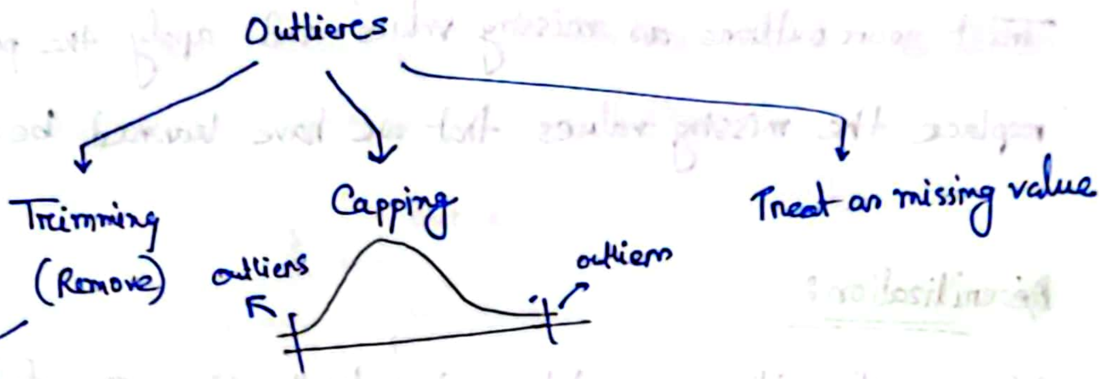
→ Decision Trees

→ Gradient Boosting

→ ~~ADA Boost~~ Xg boost

So, As we check all kind of algorithm in a dataset to find the best accuracy, So, it is a good practice to work on outliers.

How to treat outliers?



Cons

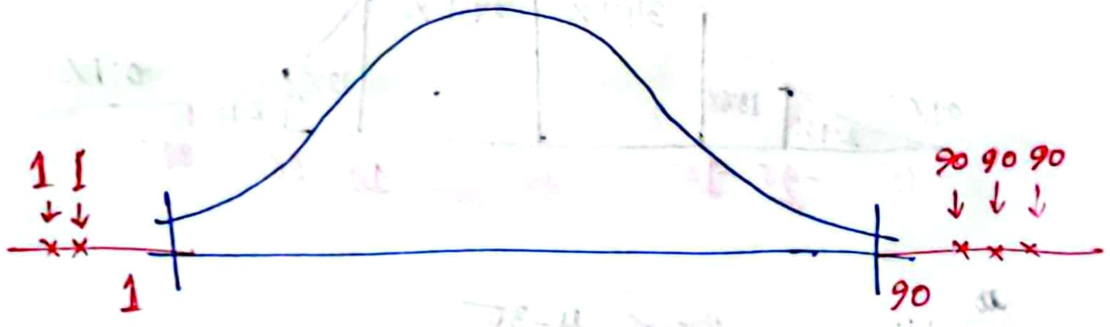
→ If there is too much outliers, data will be thinner after removing

Pros

→ Very fast

Trimming Method: You just remove your outliers from the data

Capping Method: In your distribution, you provide the min value to all the outliers who are smaller than min and you provide (replace) the max value with all the outliers who are greater than max value.



Treat outliers as missing values:

Treat your outliers as missing values (Nan) and apply the procedures to replace the missing values that we have learned before.

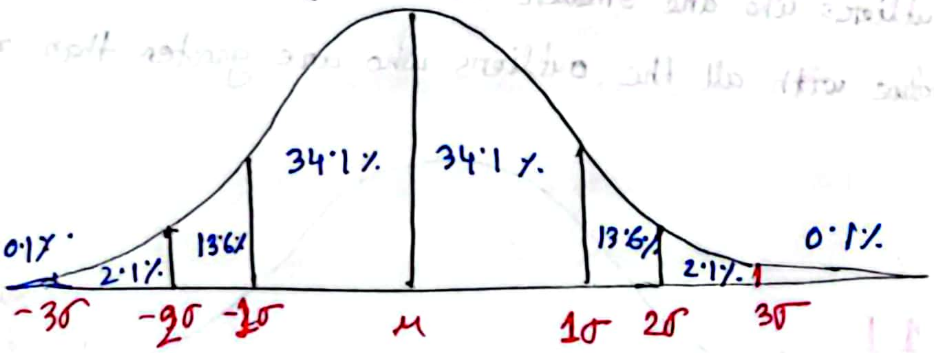
Descritization:

You can descritize your data and make the bins. So, what happens is, outliers can come within a bin range and can be treated as normal values.

"Trimming and Caping used the most among them"

"How to detect outliers?"

① **When your data is normally distributed** -

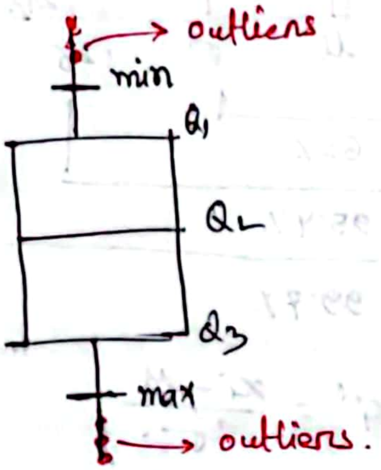


outliers $= < \mu - 3\sigma$

outliers $= > \mu + 3\sigma$

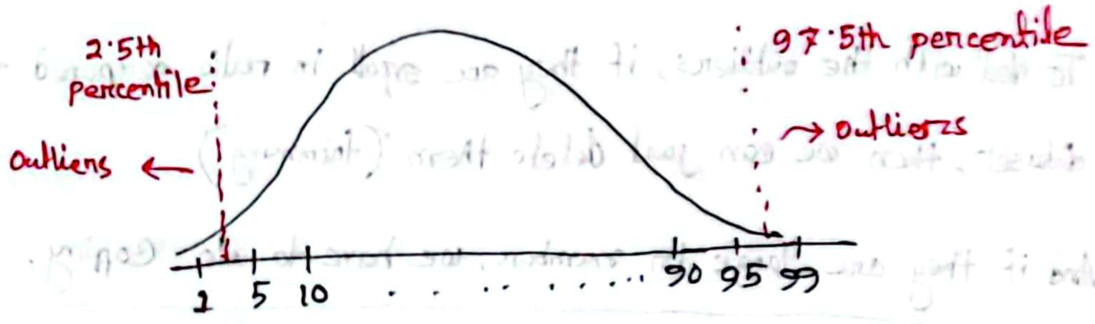
② When your data is skewed:

→ Plot Boxplot. The numbers < minimum would be outliers and the numbers > maximum are outliers.

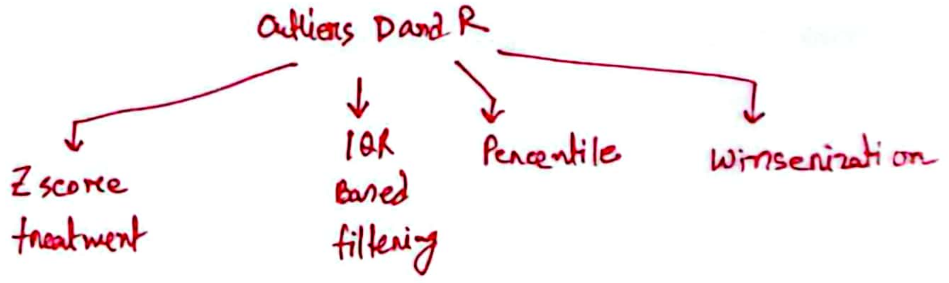


③ If your data is in some other distribution:

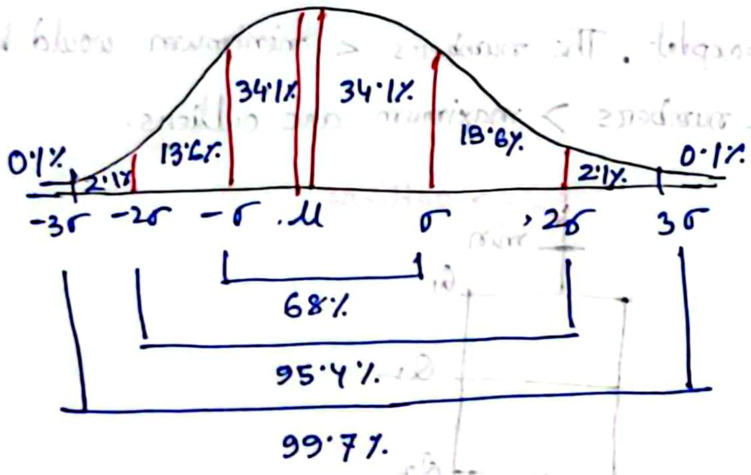
You can use the percentile based approach



Techniques for outlier detection and removal:



Outliers removal using Z score: (When data is normally distributed)



$$Z_{score} \to x_i' = \frac{x_i - \mu}{\sigma}$$

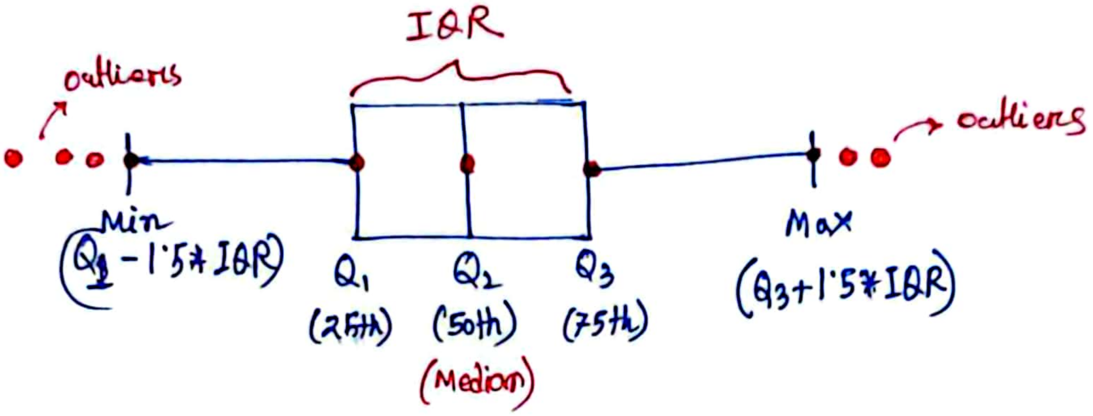
If Z score is between -3 to 3 range, than that would not be outliers. the out ranged values can be considered outliers.

To deal with the outliers, if they are small in ratio compared to the dataset, then we can just delete them (trimming)

And if they are large in number, we have to do capping.

If outliers values > maximum, then make it = maximum
if outliers value < minimum, then make it = minimum

Outlier removal using IQR Method: (When data is skewed)



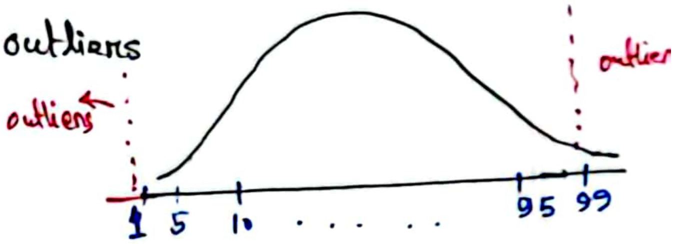
When your data distribution is skewed, use IQR method to find the outliers and remove them by trimming or cap them.

Outlier removal using Percentile method: (Any other distribution)

→ You can choose your minimum percentage threshold and maximum percentage threshold to check for outliers

Generall minimum is taken 1%.

and maximum is taken 99%



→ After finding the outliers, you can ~~yo~~ use trimming (if outliers are small in numbers) or capping (if outliers are large in number)

↳ Allso calle winsorization in percentile technique