Feature Scaling Using python:

pip install -U scikit-learn #This library is also called as sklearn.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns



df=pd.read_csv('')
df.info()

from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
```

Using these two libraries we would be able to perfrom the standardization and normalization.

normalization also known as minmaxscaling.

df.head()#prints the first five rows of the dataset

df.describe()#giveds statistical information about the data

To remove the null values:

```
new_df [ 'Age']=new_df['Age']. fillna(new_df[ 'Age']. mean ())


scaler=MinMaxScaler()
normalized_df=scaler.fit_transform(df_new)
```

an example with the normal array:

```
x_array=np.array([2],[3],[5],[6],[6])

scaler=MinMaxScaler()
normalized_arr=scaler.fit_transfor(x_array)
print(normalized_array)
```

internally the formula which is been applied is x'=(x-xmin)/(xmax-xmin)

same piece of code for the normalization as well

```
x_array=np.array([2],[3],[5],[6],[6])

scaler=StandardScaler()
normalized_arr=scaler.fit_transfor(x_array)
print(normalized_array)
```

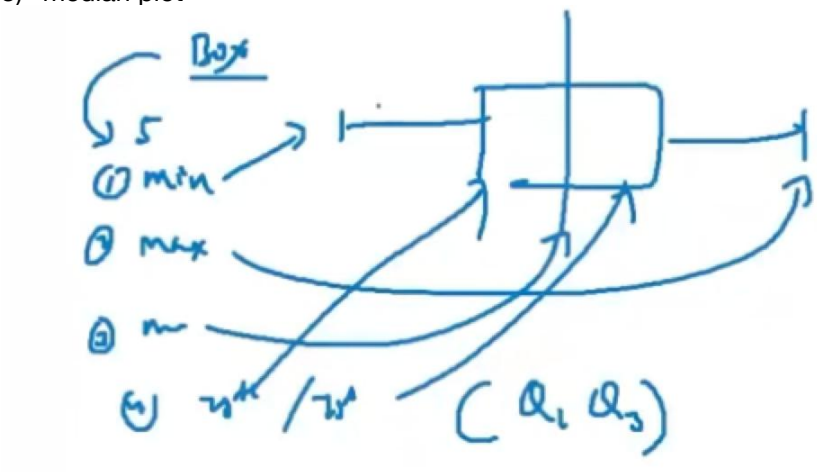
sd=root((x-u)square/N)

Outliers Treatment:

Outliers : The data which is an abnormal observation that deviate form the norm.Outliers do not fit in the normal behaviour of the data.

The outliers can be detected using these techniques:

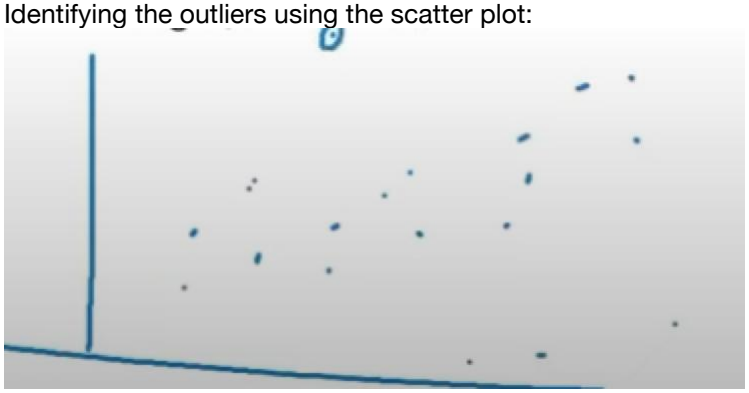
1) min plot
2) max plot
3) median plot



Histogram: recognizing the outlier using the histogram

x

Identifying the outliers using the scatter plot:



**The most classical way of recognizing the outliers is by using normal distribution:
-->we Know 99% of the data satndard devition lies in between the -3 to +3 if the data goes out of that then it would be directly classified as an outlier(Only can be used when the total data is normalized)

When to Treat Outliers:
When we are dealing with the time series forcasring consider this example:

you are having the dataset where the we have day wise noted temparatures, in such scenerio we should not get rid outliers, there maybe a day where the temparature would be very hig and some days very low,but this data emains as critical information.
When we are using the algorithms such as KNN,decision Tree,SVM, Bayes ,Ensemble Methods, these methods doesnt care about the outliers.

Example using python:

```
import numpy as np
import pandas pd
import matplotlib.pypliot as plt
import stastics

#This is called an 3 sigma technique, when it lied +_3 satandard deviation

#import the data from the drive and read the csv file using pandas


def find_anamolies(data):
#define a list to accumulate annmolies

anomalies=[]

#set upper and lower limit to 3 standard deviation
random_data_std=stastics.stdev(data)
random_data_mean=statistics.mean(data)

anamoly_cut_off=random_data_std*3

lower_limit=random_data_mean-anamoly_cut_off
upper_limit=random_data_mean + anamoly_cut_off

#Generative Outliers

for outlier in data:
  if outlier > upper_limit or outlier < lower_limit:
    anomalies. append (outlier)
return anomalies
```

why we are mulltiplying the data's standard deviation with 3:

Here, the standard deviation is multiplied by 3 to set a threshold for identifying anomalies. This is based on the empirical rule (also known as the 68-95-99.7 rule) in statistics, which states that for a normal distribution:

About 68% of the data falls within one standard deviation of the mean.

About 95% of the data falls within two standard deviations of the mean.

About 99.7% of the data falls within three standard deviations of the mean.

By multiplying the standard deviation by 3, you're setting a cut-off that will capture approximately 99.7% of the data within the range. Therefore, any data point outside of this range is considered an anomaly.