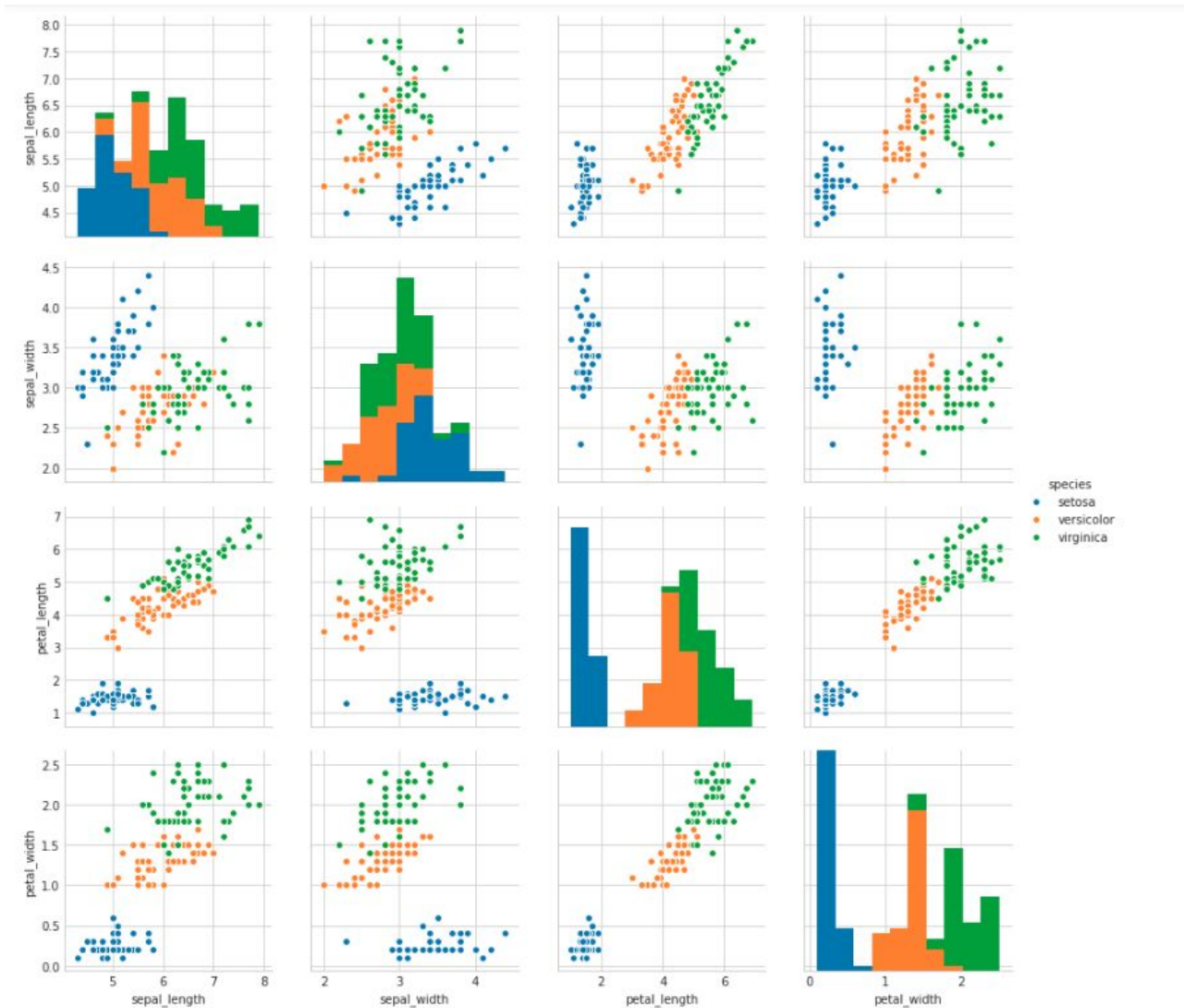


# EXPLORATORY DATA ANALYSIS

1 . Refer the EDA notebook for checking the steps before Pairplot shit.



The above pairplots are used for analysis of IRIS dataset. There are 6 pairplots which are substantial to us. As we can see the other 6 (from the diagonal histograms) are just mirrored pairplots. Now, we can analyse the pairplots and it can be clearly seen that the petal\_length vs petal width graph is clearly separating setosa and versicolor. So, We'll try to make an if else condition for that

## Finding the pair plot relations

So, we've found the correct pairplot (petal\_length vs petal width) to perform our EDA then we'll analyse it



Fig 1.1 For separating setosa

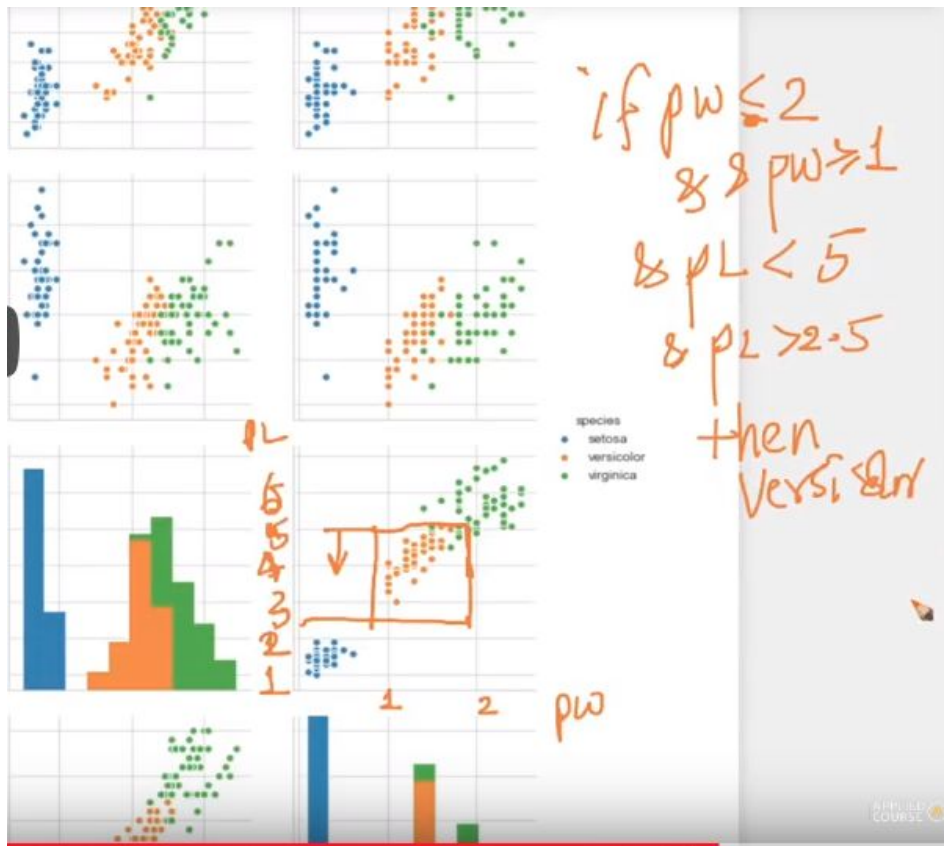


Fig 1.2 For separating versicolor

Here ,setosa can be seperated perfectly but in versicolor there can be some green points found but that's alright. In ML, model is rarely perfect . Things can lead to overfitting if we try doing it perfectly

### Limitations of pair plots

We'd 4 features here but what if we'd 10 or 100 features here. Then Pairplotting can get tough  
 So we use techniques like PCA and t-sne to analyse our model.

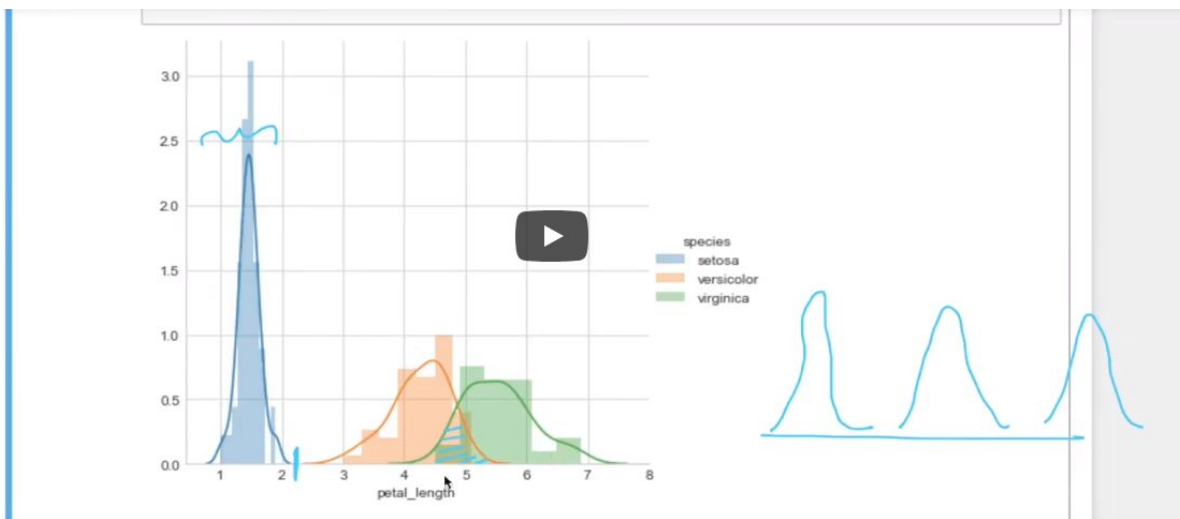
# UNIVARIATE ANALYSIS

## Definition

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; it takes data, summarizes that data and finds patterns in the data.

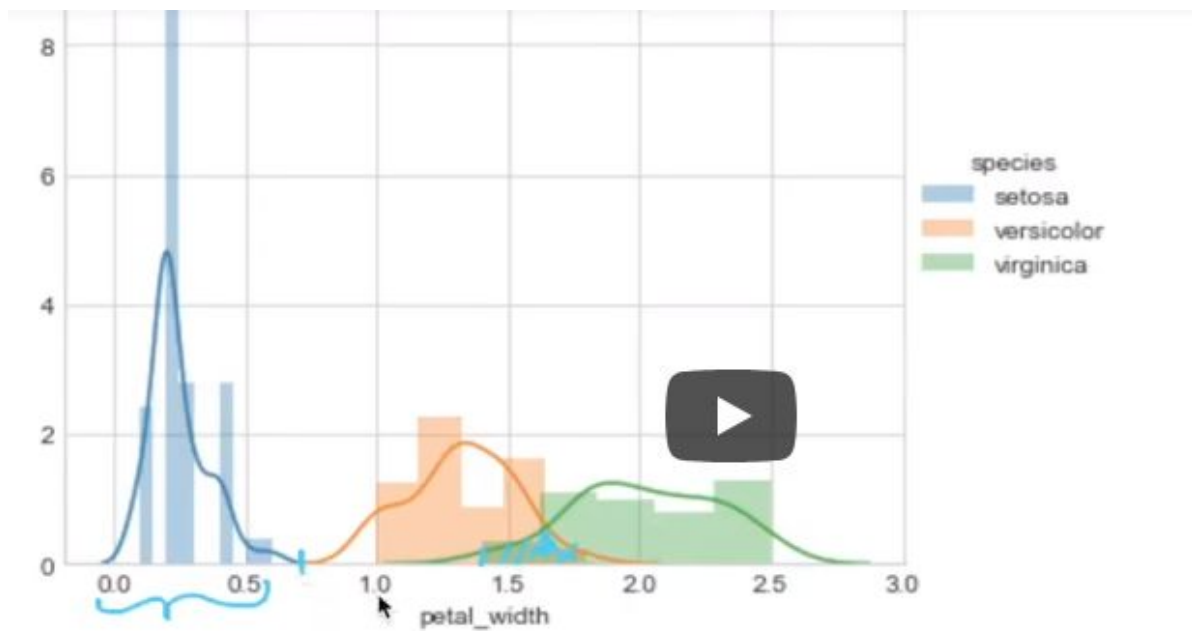
Here we did univariate analysis by using the 4 features we had in our dataset by plotting a histogram

Petal \_ length



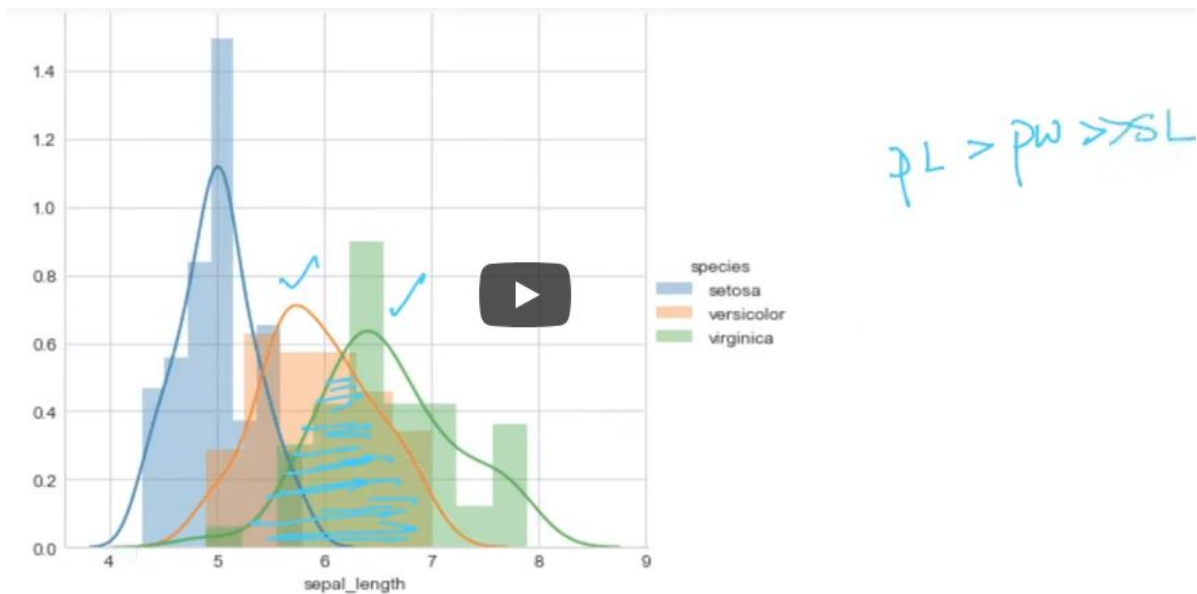
Here , we can observe that petal\_length is clearly seperating our setosa . There’s some overlap between versicolor and virginica but that’s Ok. As we can see , the ideal graph is the blue one to clearly distinguish.

## Petal\_width



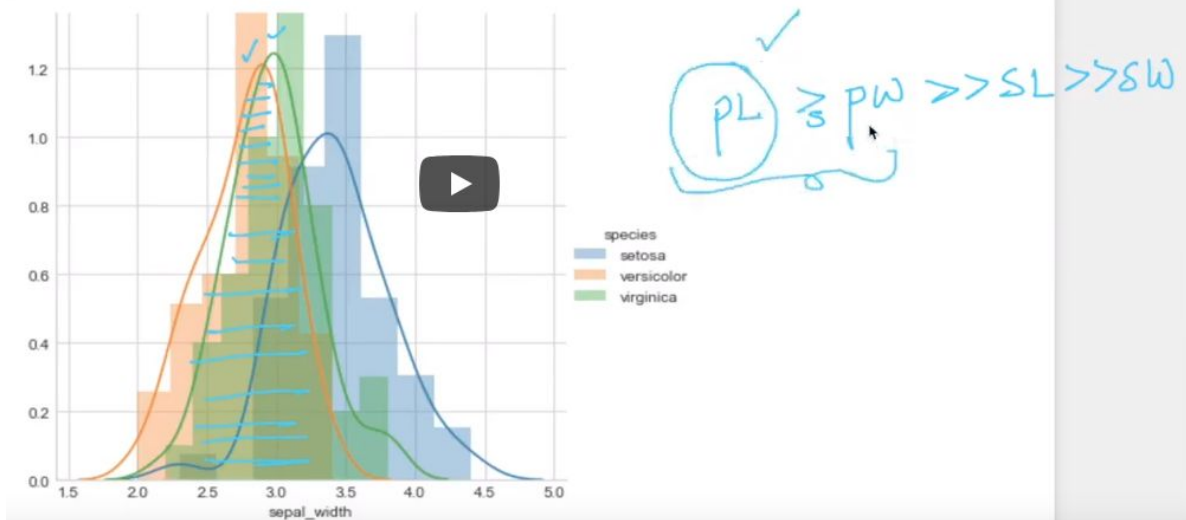
The distribution is fairly good but there's a slight overlap between the blue and the orange one.

## Sepal\_length



Okay ,here the overlap is huge . This is clearly not the feature we need a lot . So,  $PL > PW \gg SL$

## Sepal\_width



The overlap is shittier than anything here. Definitely not the feature to be used . Can be clearly inferred that  $PL > PW \gg SL \gg SW$ . So if we are taking one feature than PL . If two then PL and PW.

## MEAN, VARIANCE AND STD DEVIATION

### MEAN

#### (3.5) Mean, Variance and Std-dev

```
: #Mean, Variance, Std-deviation,  
print("Means:")  
print(np.mean(iris_setosa["petal_length"]))  
#Mean with an outlier.  
print(np.mean(np.append(iris_setosa["petal_length"],50)));  
print(np.mean(iris_virginica["petal_length"]))  
print(np.mean(iris_versicolor["petal_length"]))  
  
print("\nStd-dev:");  
print(np.std(iris_setosa["petal_length"]))  
print(np.std(iris_virginica["petal_length"]))  
print(np.std(iris_versicolor["petal_length"]))
```

```
Means:  
1.464
```

$$S_{-}pL = \{x_1, x_2, \dots, x_n\}$$
$$n = 50$$
$$\mu_{S_{-}pL} = \frac{x_1 + x_2 + \dots + x_n}{n}$$
$$= \sum_{i=1}^n x_i \times \frac{1}{n}$$

Mean has been mentioned here in the orange part. We are taking out the mean of all the three classes. But one question is why are we doing that?

```
Means:  
✓ S ✓ 1.464  
✓ V ✓ 2.41568627451  
✓ Ve ✓ 5.552  
✓ 4.26
```

We've calculated the means of the petal length (most significant feature) of all the three classes and we can clearly infer that the mean of petal length of setosa is smaller than the other means or setosa's petal length tends to be much smaller than the other ones.

Note: - The mean of setosa with outlier has increased the mean by a significant amt (2.41) . We need to take care of that .

## SPREAD



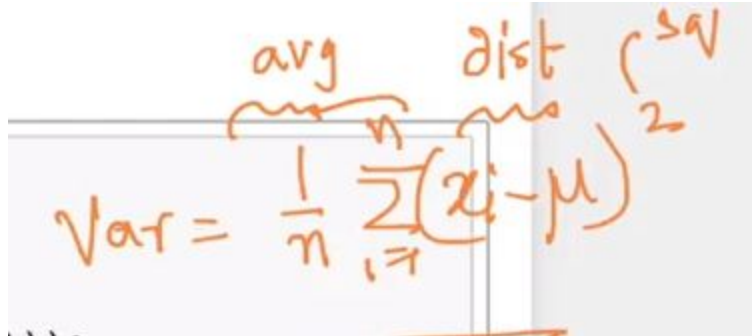
We know the means and we plotted it here but in spread we get the idea about how far our pts or petal lengths are distributed. For ex:- If our spread is small like in the blue one . We can get an idea that our points are closer to our central tendency (Mean).

If we want to do numerical analysis of spread then there's a thing called variance



## VARIANCE

It tells us about the spread of all data points from the mean



A handwritten formula for Variance on a whiteboard. The formula is 
$$\text{Var} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$
. Above the formula, there are handwritten annotations: 'avg' with a bracket over the denominator 'n', 'dist' with a bracket over the term '(x\_i - \mu)', and '(sq)' with a '2' superscript over the exponent. The word 'Var' is written in a large, bold font.

Where,

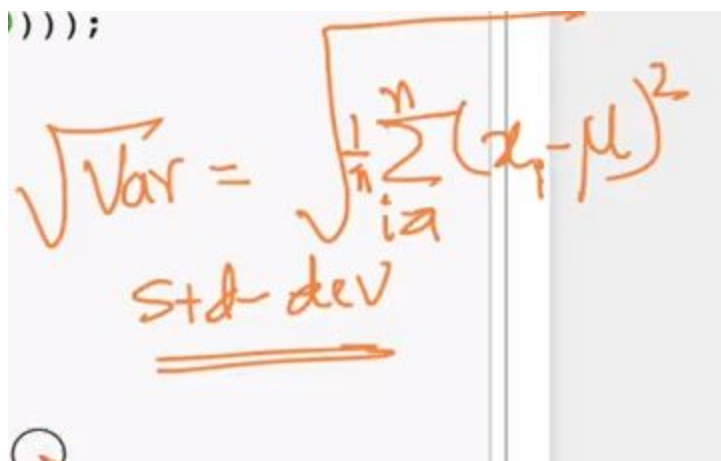
X - data points

$\mu$  - mean

n - No. of features( rows)

## STANDARD DEVIATION

It is root of Variance . It is basically telling the average deviation/distribution from the mean value.



A handwritten formula for Standard Deviation on a whiteboard. The formula is 
$$\sqrt{\text{Var}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$$
. Below the formula, the words 'Std dev' are written and underlined twice. The word 'Var' is written in a large, bold font.

Means:

1.464

~~2.41568627451~~

5.552

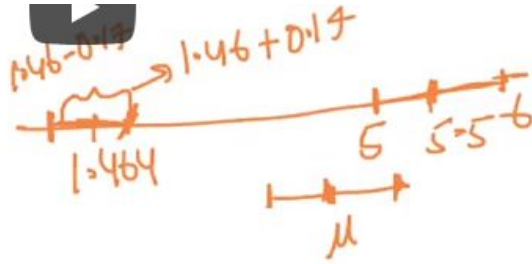
4.26

Std-dev:

0.171767284429

0.546347874527

0.465188133985



Std- dev is telling us about the distribution of points from the mean . From the drawn image it can be inferred from low std-dev of Setosa is telling us that the points are closer to mean without looking at the histogram.

## MEDIAN

The "median" is the "middle" value in the list of numbers.

⑦ numbers:

$$x = \{1, 1.2, 1.1, 2.1, 1.8, 1.6, 1.2\}$$

① Sort them in order

$$\{1, 1.1, 1.2, 1.2, 1.6, 1.8, 2.1\}$$

② pick the middle value

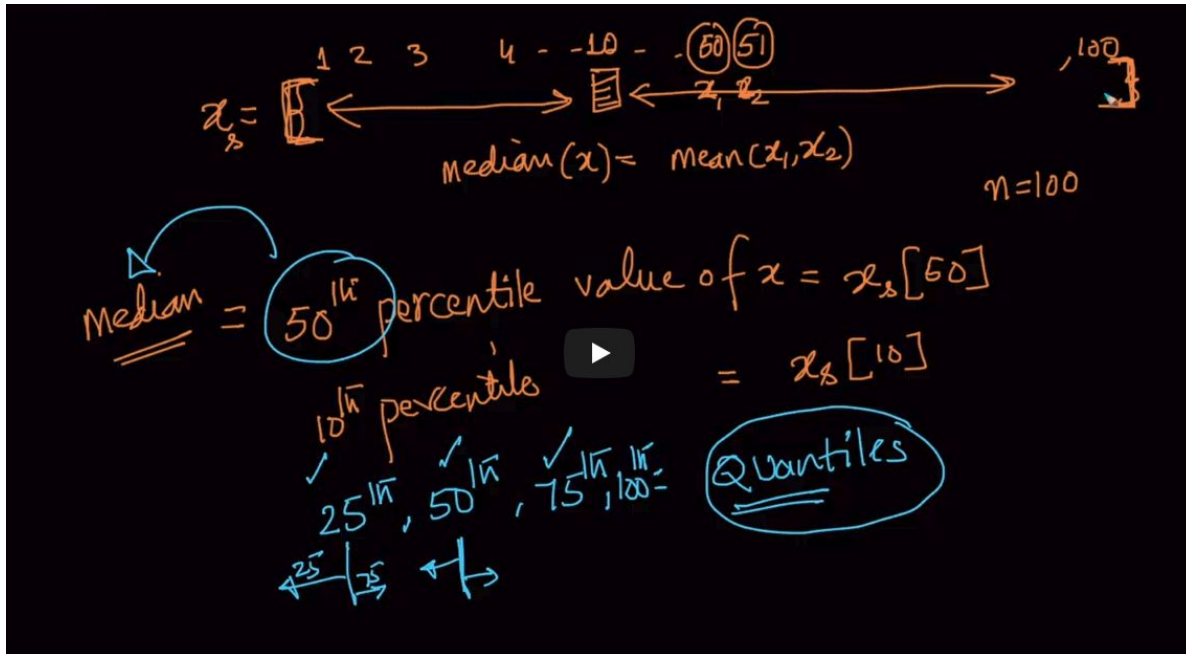
$$\text{Median}(x) = 1.2$$

For odd numbers -  $(n+1)/2$

For even numbers - avg(two middle numbers)

NOTE - Outliers don't affect the median values unless more than 50 % of data is corrupted

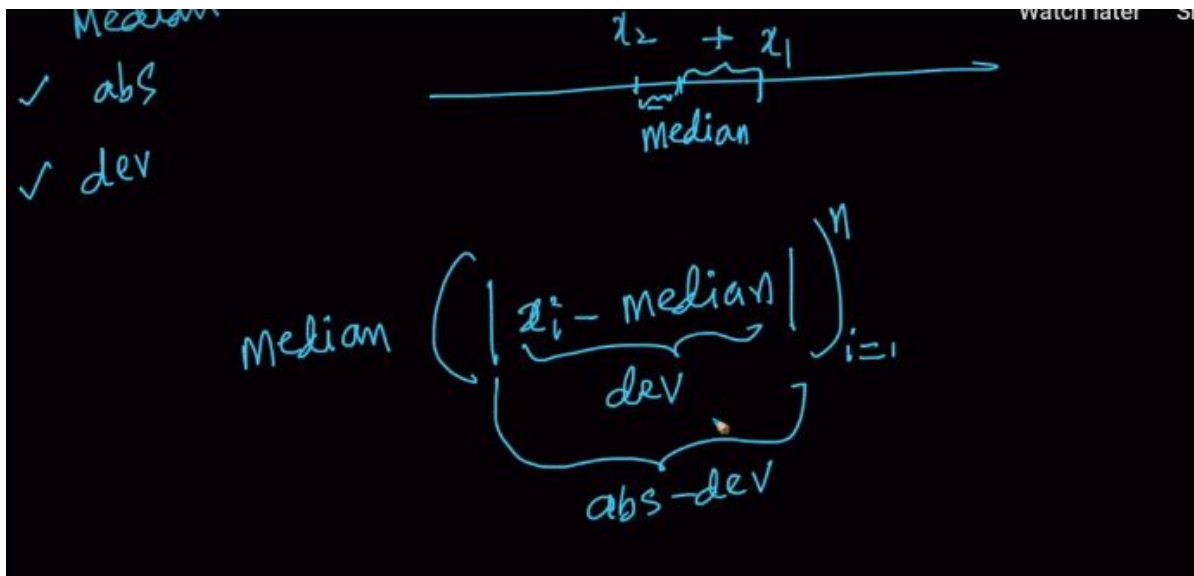
## PERCENTILES AND QUANTILES



Percentile - It means that how many percent are less than percentile value. For Ex- If 10 percentile then there are roughly 10 percent of points below that percentile value.

Quantiles are mentioned in the image (25th, 50th, 75th and 100th percentile)

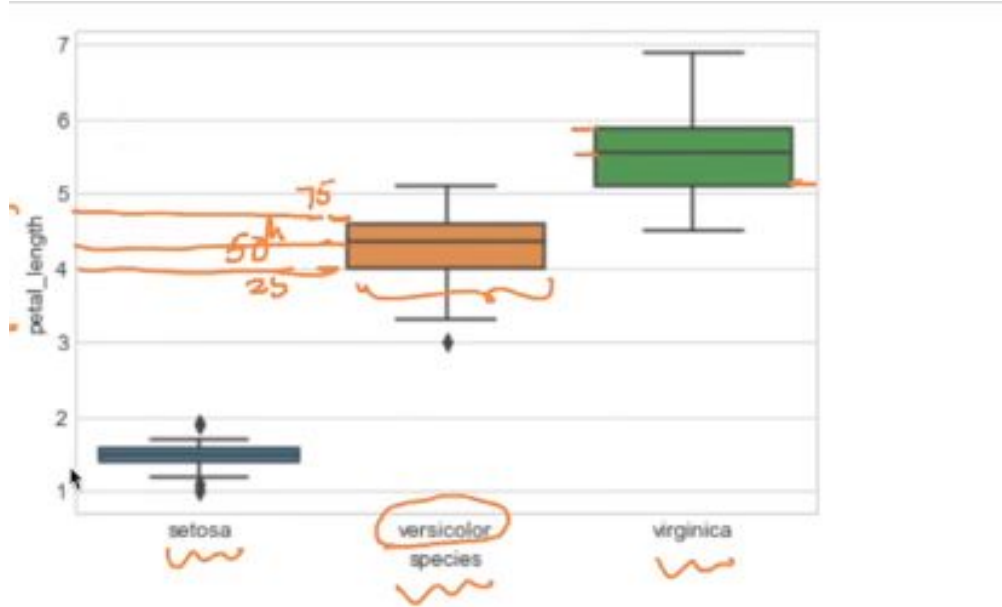
## MEDIAN ABSOLUTE DEVIATION



It's basically the median of distance of all the points from the median. This gives the idea of how far the points are spread out from the median. It is equivalent to STD- DEV.

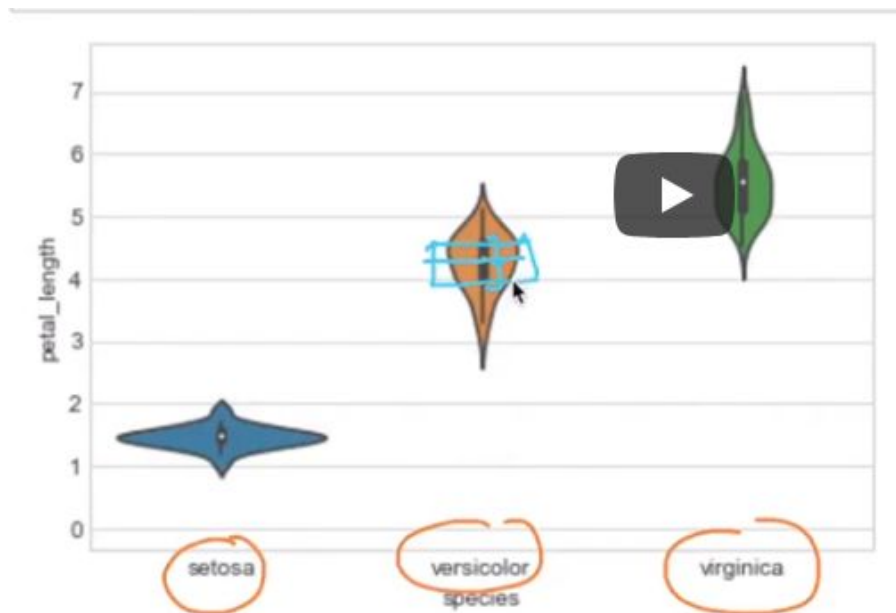
INTER QUARTILE RANGE( IQR) => 75th Percentile value - 25th percentile value

## BOX PLOTS AND WHISKERS



**Box plots** tells us about the 25th ,50th and 75th percentile value. Our Histograms didn't show us the value of percentile which tells us which values are less than that percentile. Whiskers are nothing but the T lines displayed (Normally selected as the min and max values)

## VIOLIN PLOTS



Explanation - The black portion in the plot is box plots and the curved part you are seeing here is Probability distribution or histograms . Refer the petal width histogram and remember that distribution was between 1 and 2