Exploratory Data Analysis

<https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>

MAD:

<http://www.statsmodels.org/devel/generated/statsmodels.robust.scale.mad.html>

Box Plot :

<https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule>

<https://youtu.be/FRlTh5HQORA>

<https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>

<https://www.varsitytutors.com/hotmath/hotmath_help/topics/box-and-whisker-plots>

Basic Terminology

* What is EDA?
* Data-point, Vector, Observation
* Data-set
* Feature/Variable/Input-variable/Dependent-variable
* Label/dependent-variable/output-variable/Classes/Class-label/Response label
* Vecot:2-D,3-D,4-D...n-D
* Balanced Dataset vs im-balaned dataset

2-D Scatter Plot

* y=mx+c
* Lineraly Separable

3-D Scatter Plot

* Plot.ly
* Drawbacks

Pair-plot

* Why ?
* How to use seaborn?
* Simple Model

Limitation of Pair-plot

* When no of feather increases then it is taught ex 4c2 for nc2
* The solution would be dimensionality reduction techniques Principal Component Analysis (PCA), t-SNE -- Linear Algebra

Histogram,PDF

* Histogram
* Probability Density Functions

Univariate Analysis

* What is a univariate analysis?
* How do we find out which feature is more useful to distinguish b/w classes

Mean, Variance and Std-Dev

* What is Mean?
* What is the problem with the Mean?
* What does the mean tell us?
* What is the Spread?

Variance

Std-dev

Median

* How to Compute the Median
* Why is median not corrupted by outliner

Percentiles and Quantiles

IQR(Inter Quartile Range) and MAD(Median Absolute Deviation)

* Compare

Box-Plot with Whiskers

* What is Box-Plot
* What is the problem it is going to solve
* What are Whiskers
* Calculation for Whiskers

Violin Plot

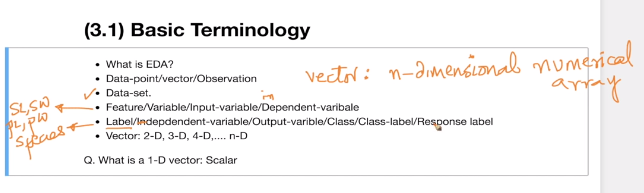
* What is the problem it is going to solve?
* Describe Violin Plot

Important Notes

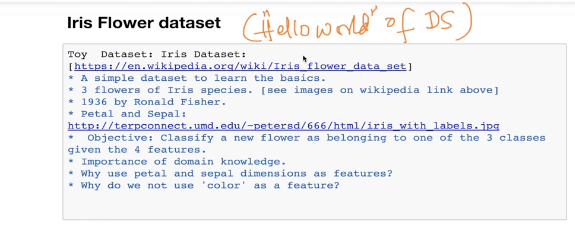
Multivariant Probability density, contour plot

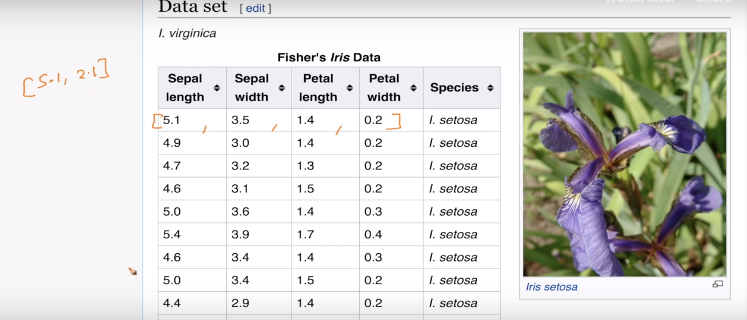
# CONTENT

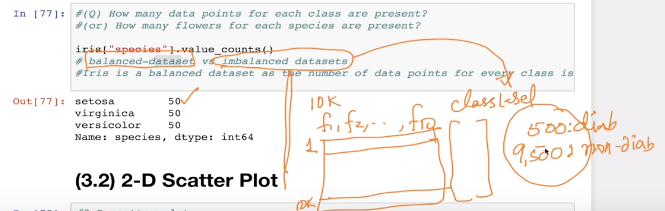
### Plotting for exploratory data analysis



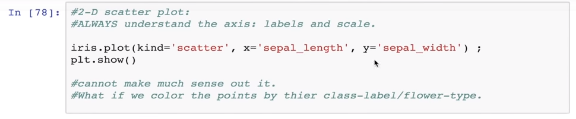
### Iris Flower dataset

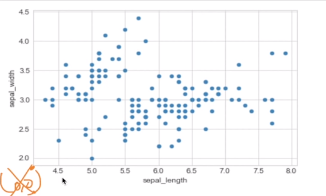


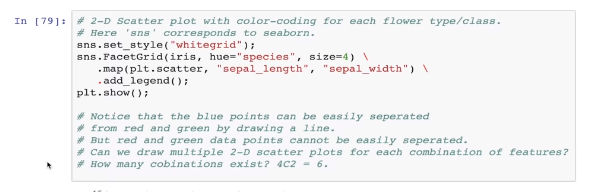


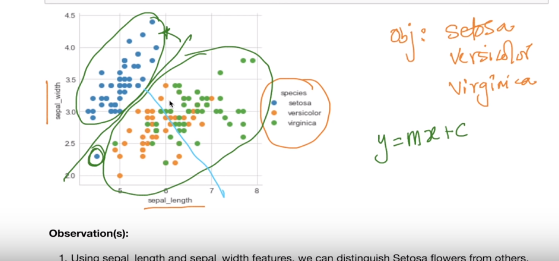


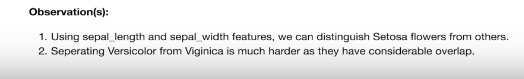
### 2-D Scatter Plot



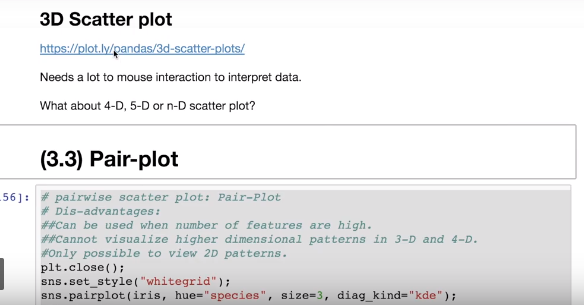


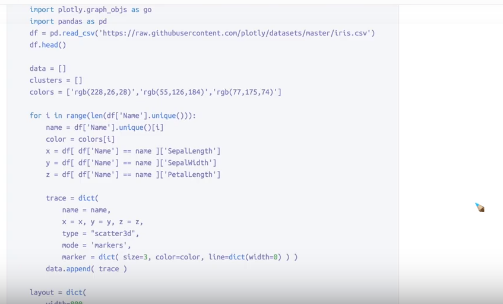


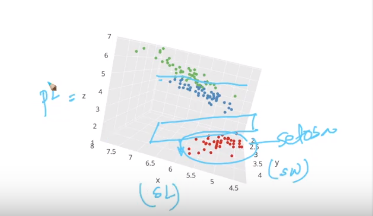




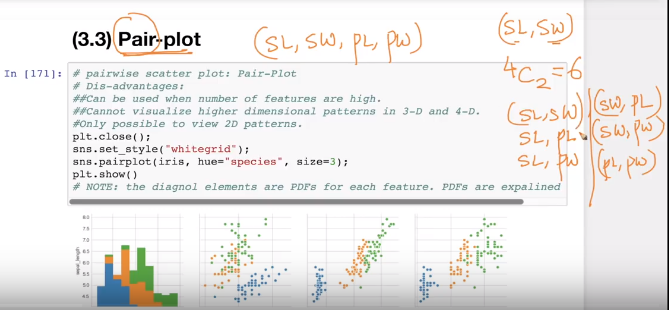
### 3-D Scatter Plot

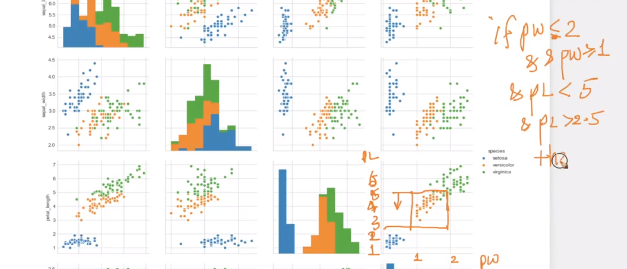




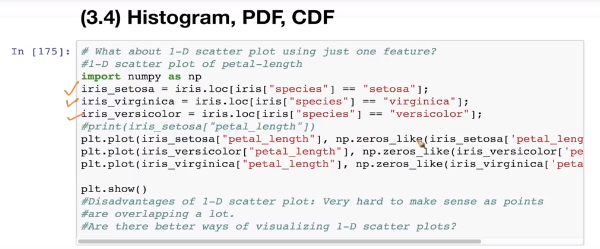


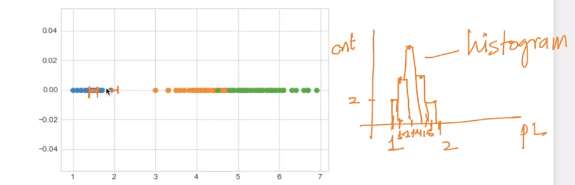
### Pair-Plot

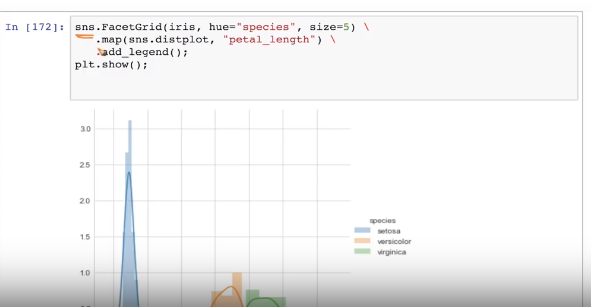




### Histogram, PDF



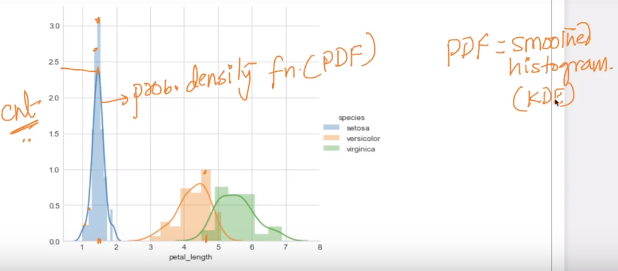




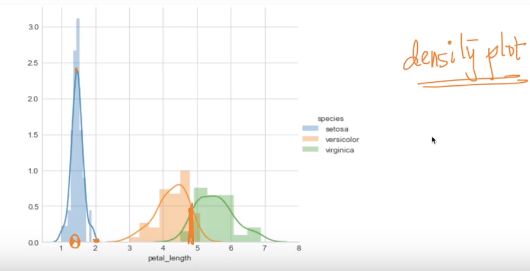
#### Histogram



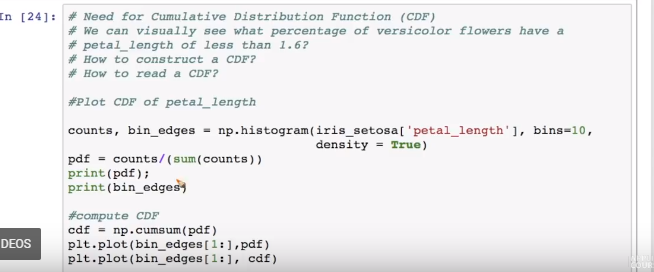
#### Probability Density Functions (PDF)

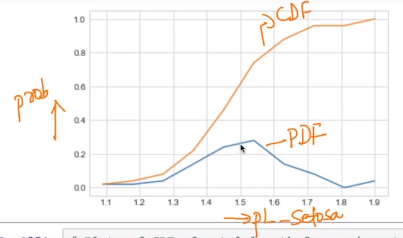


KDE: Kernal Density Estimation

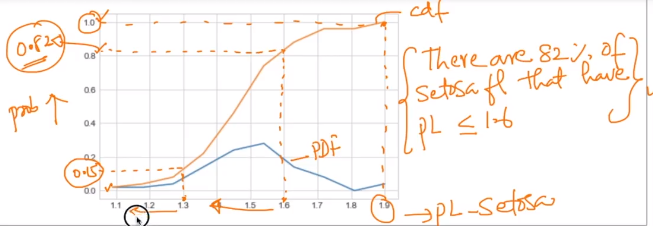


### Cumulative Distribution Functions (CDF)

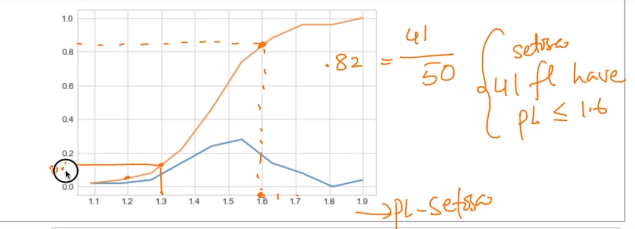


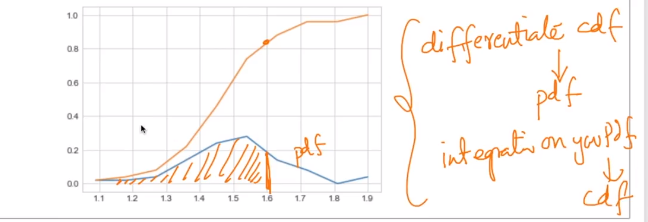


#### What does CDF convey?

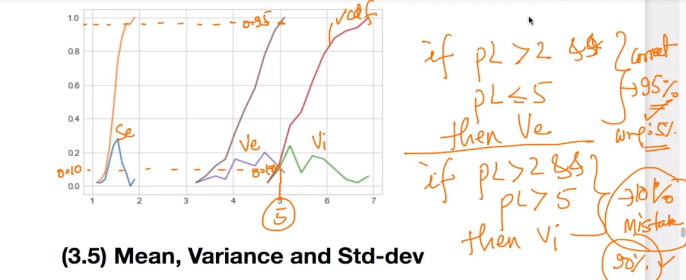


#### How to build a CDF?



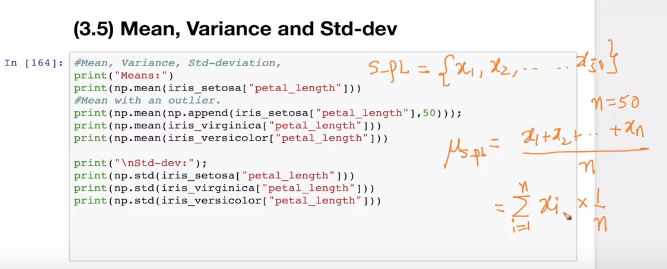


#### How CDF is useful?

We can find how accurate the model is using CDF as show

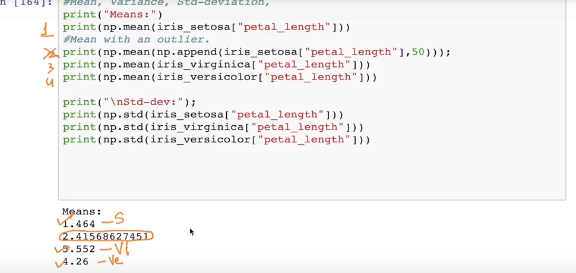
### Mean, Variance and Std-Dev

#### What is Mean?



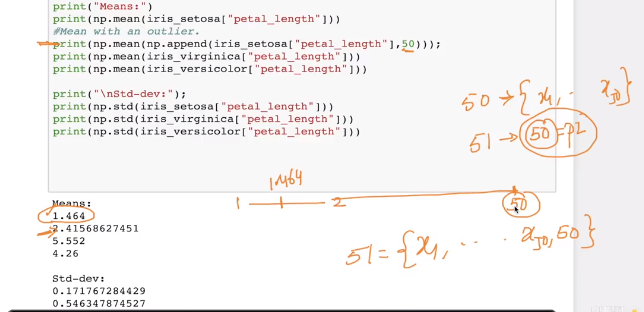
For a given 50 Petal lengths, we can identify the average. Mean is a simple average

#### How is it useful?



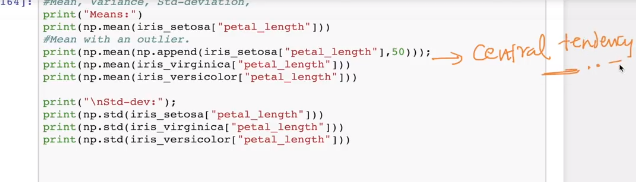
#### What is the problem with the Mean?

Outliers corrupt the mean a lot as shown below

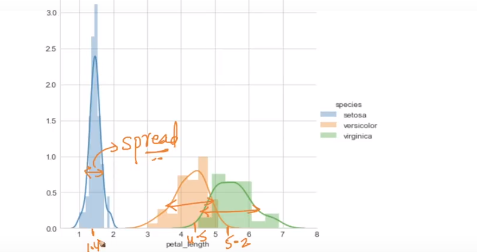


#### What does the mean tell us?

Average Tendency or Central Tendency



#### What is Spread?



Spread give intuitively of how widely these points are spread (i.e most of my points are spread b/w o.9 and 1.9)

There are numerical Approaches instead of plot.

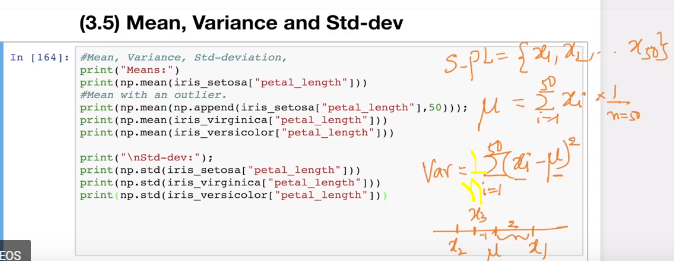
Spread can be thought of a Variance

### 



The Range of Setoja is 0.9 to 1.9

### Variance



Variance is often defined as the summation of across all the points and substracts mean from it.

(xi-𝝁) will give the distance from the point to 𝝁 that can be negative or positive so we square it which will always result in a positive number.

Intuitively it says what is the square distance of each of my points from my mean

Variance is nothing but the average squared distance from each point

### Std-dev



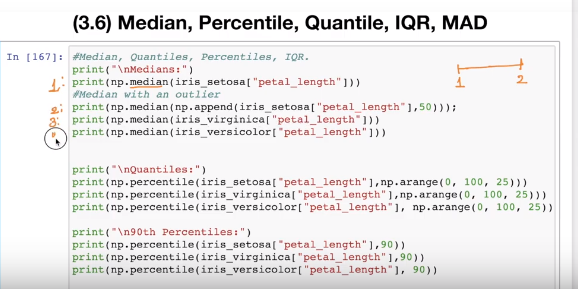
Spread is a simple English word to represent Std-dev

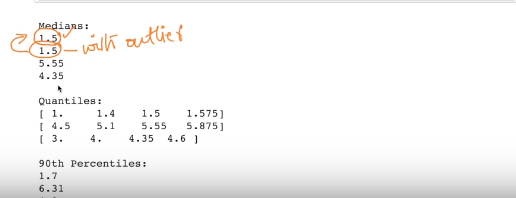
Spread typically says how far are my points from mean

As shown below

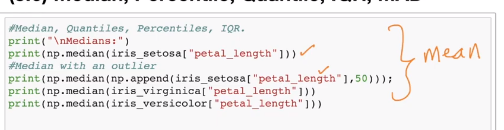


### Median

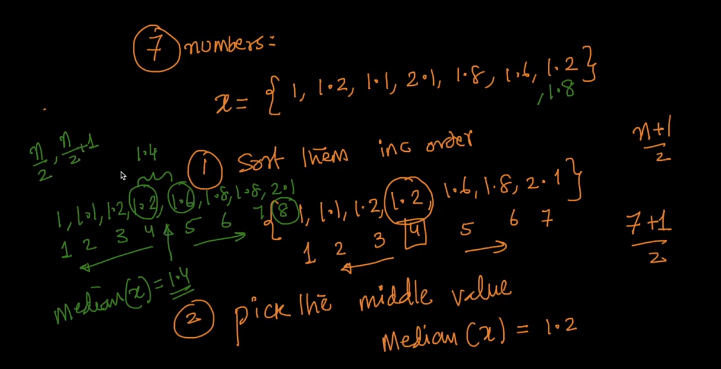




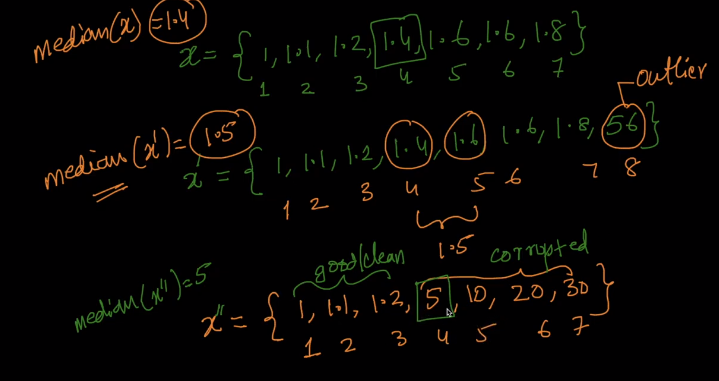
Median is almost equal to mean



#### How to Compute the Median



#### Why is Median not corrupted by an outlier?



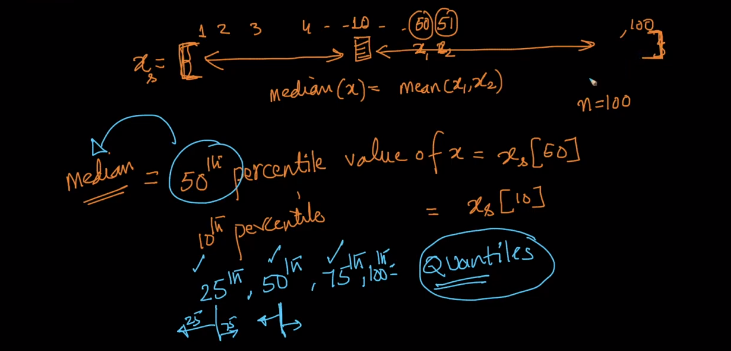
If more than 50% of points are corrected then the median is corrupted.

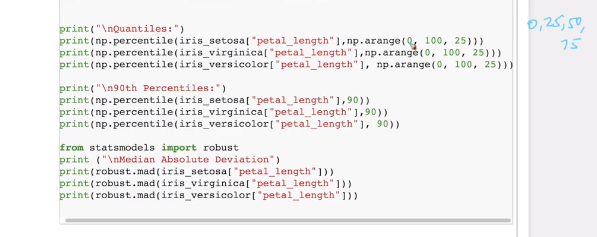
### Percentiles and Quantiles

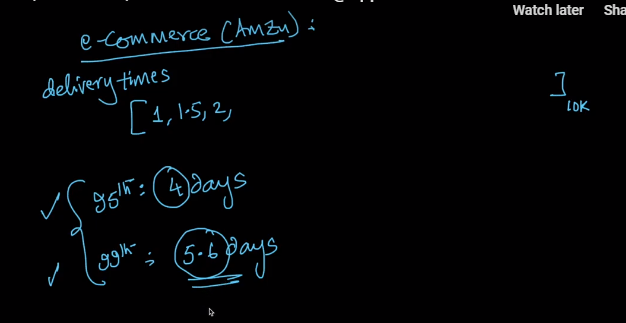
Percentile says where do you lie in the sorted list.

50th percentile means what percentage of value lies less than Xs [50] and what percentage of value lies greater than that value

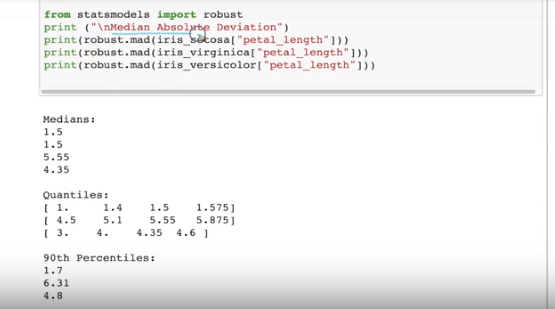
25th



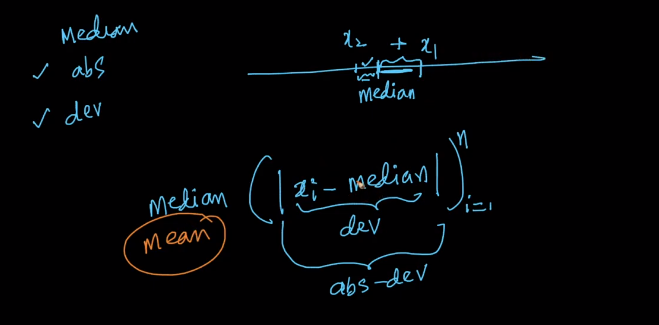


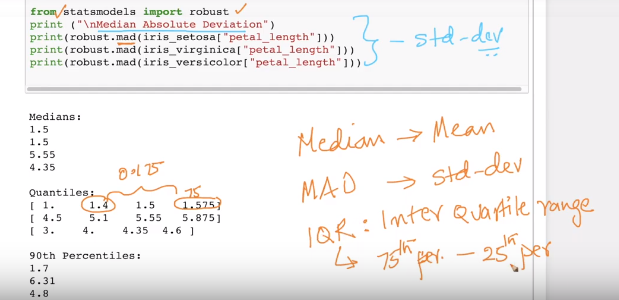


### IQR (Inter Quartile Range) and MAD (Median Absolute Deviation)







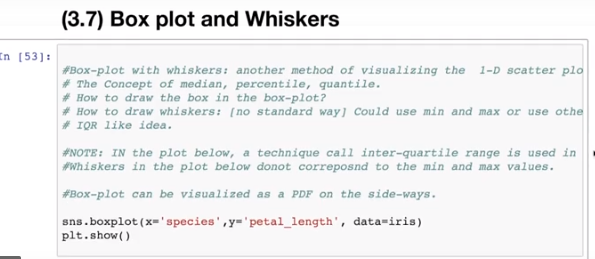


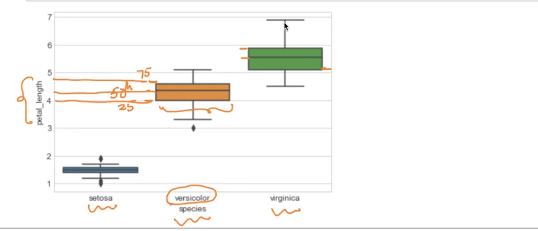
Inter Quartile Range = 50% of data lies

### Box-plot with Whiskers

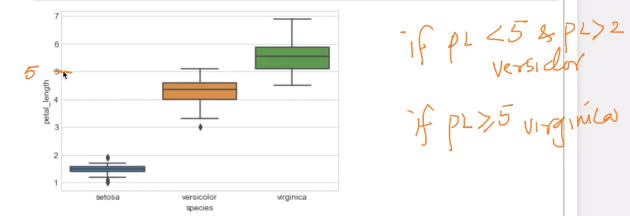
Why do we need Box-plot?

We have plotted Histograms and PDF earlier but one problem that we have is We don’t know 25thpercentile value or 50th percentile value. I do agree with you that some of the issues with CDF but are there are other plots without using lines …. Such a plot is Box plot

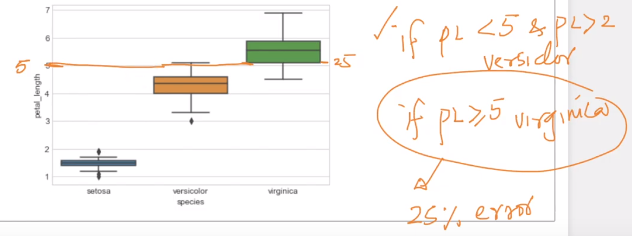




Now we will assume and write a rule to find the Virginia flowers

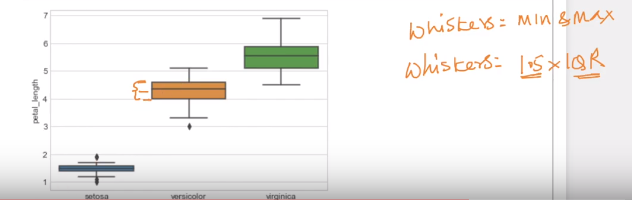


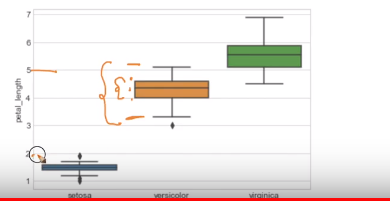
By just looking at the plot we can identify the error.



#### Whiskers:

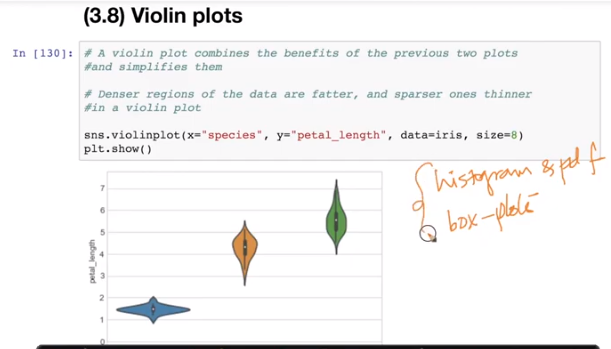
There is no standard way of drawing whiskers.Whiskers use min & max to plot whiskers. One way of drawing whisker is 1.5 times IQR.



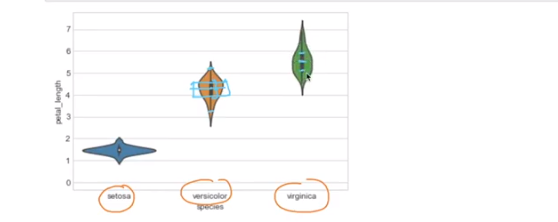


The other way of looking at these plots is 50% of the points line inbox and most of them lie between whiskers.

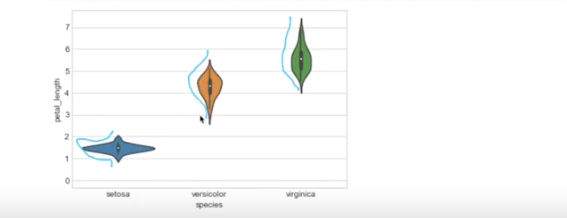
### Violin Plot



Violine plot gives the best of both worlds (i.e histogram, PDF and box-plot)

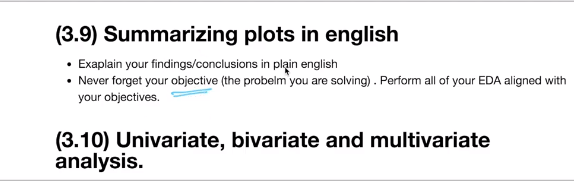


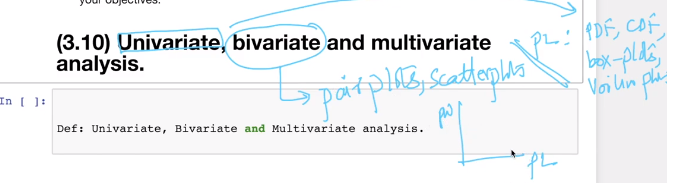
Box top as you see give the top line 75%, Median (50%), 25%, whiskers



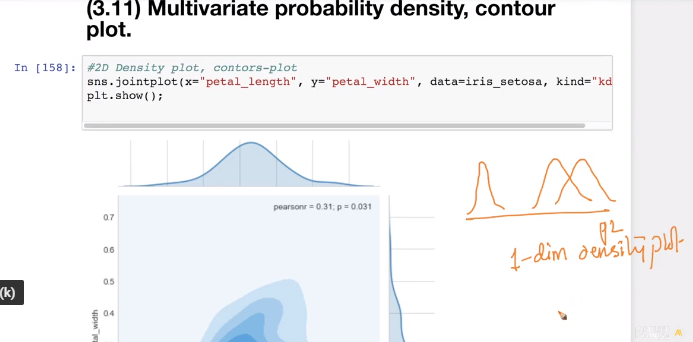
We can see a Bell Curve or PDF

### Important Notes



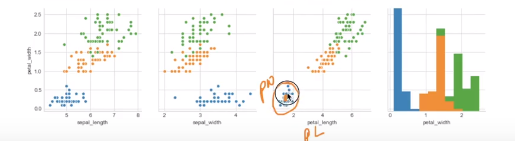


### Multivariate probability density, contour plot

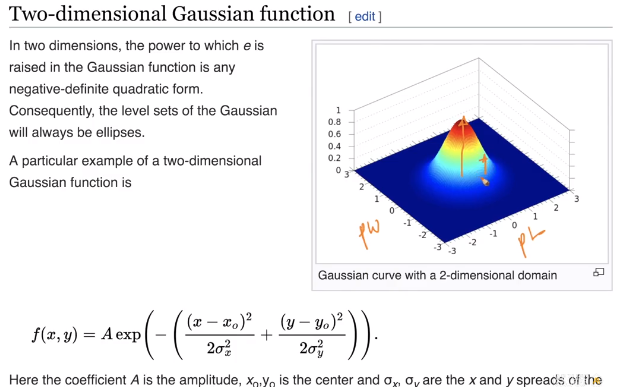


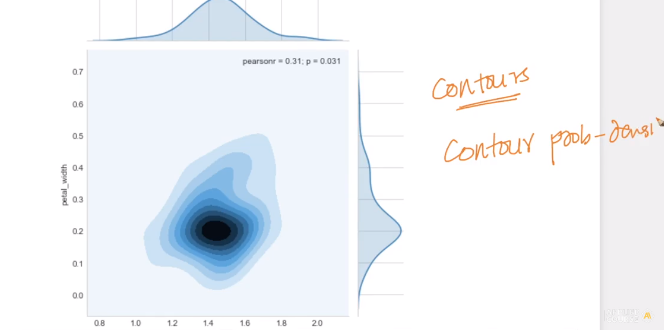
We have seen 1-D density plots but can we do a 2-D Density plot …

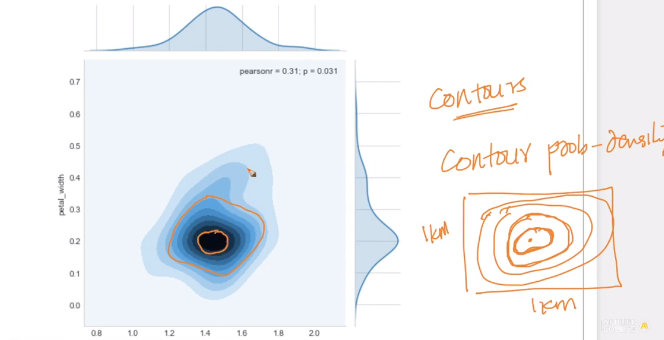
Let's look at p\_l and p\_w pair plots to connect dots.



We observe there are more points in the center region and there are lesser as we go away from center region.







What circle means all points are at the same height in that circle

