**Drawback of 1-D Scatter Plot:**

We can’t make sense of number of points in a region because points are overlapping a lot.

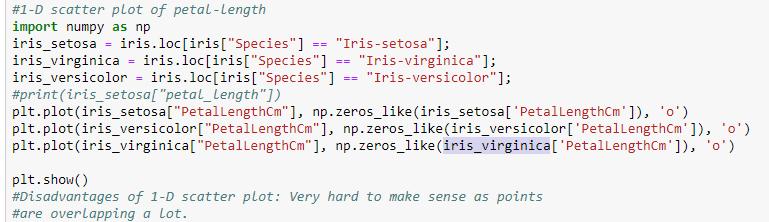
Ex:

plt.plot(iris\_setosa["petal\_length"], np.zeros\_like(iris\_setosa['petal\_length']), 'o')

plt.plot(iris\_versicolor["petal\_length"], np.zeros\_like(iris\_versicolor['petal\_length']), 'o')

plt.plot(iris\_virginica["petal\_length"], np.zeros\_like(iris\_virginica['petal\_length']), 'o')

here we are taking **x**  axis a **petal\_length** and y as 0, because we want scatter plot of only petal\_length (ie 1-D)

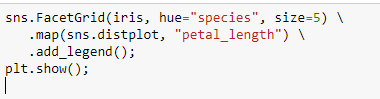




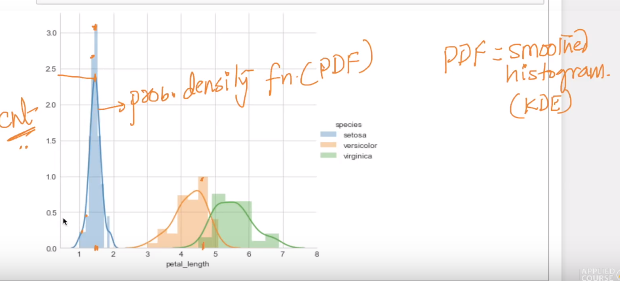
As you can see from above graph we can’t find out the no of points for particular length as they are overlapping each other.

So to overcome this we use histogram as on **x**  axis we keep  **petal\_length**  and on **y**  axis we keep  **probability of counts** in that particular petal\_length region.

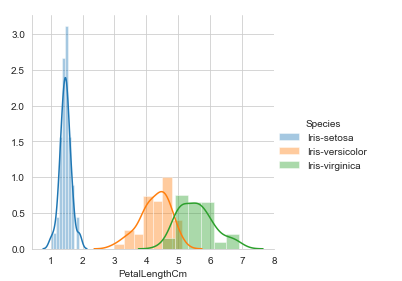


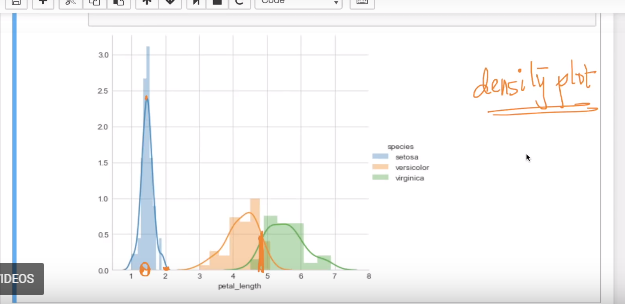


In above graph we divided in graph in small part of 0.2. an on y-axis we can see the probability of counts of points present in that region, ex: between petal length of 1 and 1.2 there are 2.



The above plot is also known as density plot (in case of 1-D) as it is telling the density of points in that region.





As you can see at 1.5 it is most dense for setossa iris

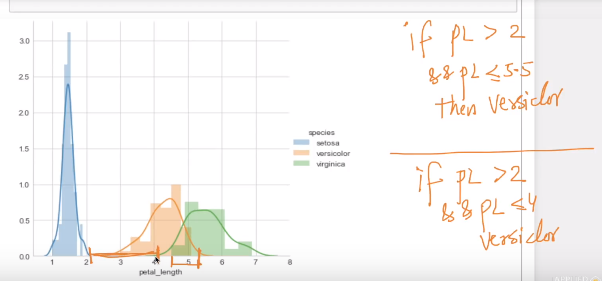
Now on the basis of above graph analysis we can create a basic model to classify setosa as on the basis of boundary of PDF.

If PL <= 2

Then it’s setosa;

But if try to build model to classify versicolor, virginica, there will be a lot of wrong decisions will be made as we are telling if PL <= 5.5 then it’s versicolor but b/w 4 – 5.5 there are virginica also

And if we take end boudary for versicolor as starting boundary for verginica ie 4.5, saying if PL <= 4 it’s versicolor, then it is also wrong because lot of versicolor flowers are there after 4.5 too.



Now One thing we can do build model as took threshold boundary as point where both versicolor and virginica PDF’s are intersecting ie at 4.7, now we can build model as.

If PL <= 2

It’s setosa

Else

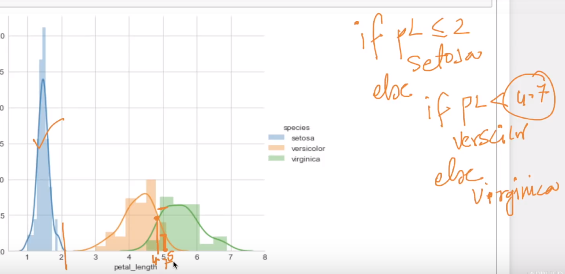
If PL < 4.7

It’s versicolor

Else

It’s verginica

Now let’s say we have PL 5, so by this model we can say it’s verginica it may be versicolor but by seeing below graph the histogram at 5 says there are more verginica than versicolor



**What is PDF ?**  
As per plot: PDF is a smoothed form of a histogram.  
Definition: PDF is a function which takes a value (X), Likely as a normal function in programming, with a logic, integration of f(x) dx with an interval a to b. a= may be a negative value and b may be a positive value.  
X may be continuous random variable (or) Gaussian variable (or) Discrete Random variable.