Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Distribution Plots

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Distribution plots display a single continuous feature and help visualize properties such as deviation and average values.

3 main distribution plot types:

- ○ Rug Plot
- ○ Histogram
- ○ KDE Plot

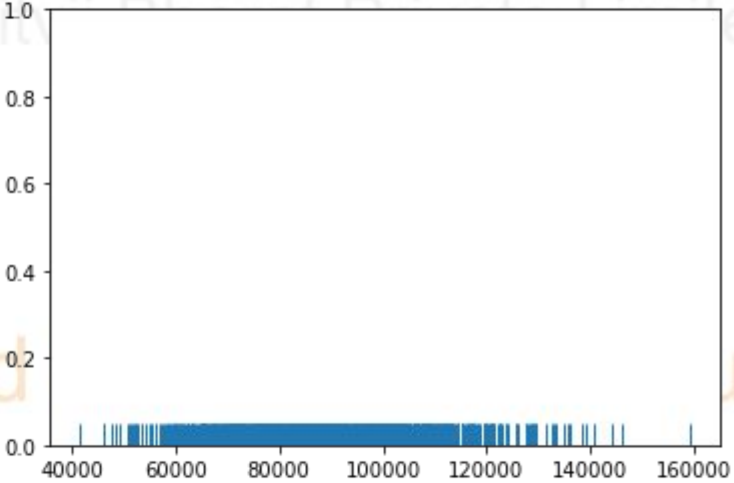
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
[illegible] Bharat Private Limited
Led by : Shreyas Shukla

Let’s explore the distribution of employee salaries.

- One way is a rug plot which is the simplest distribution plot and merely adds a dash or tick line for every single value.
- The y-axis does not really have a meaning.

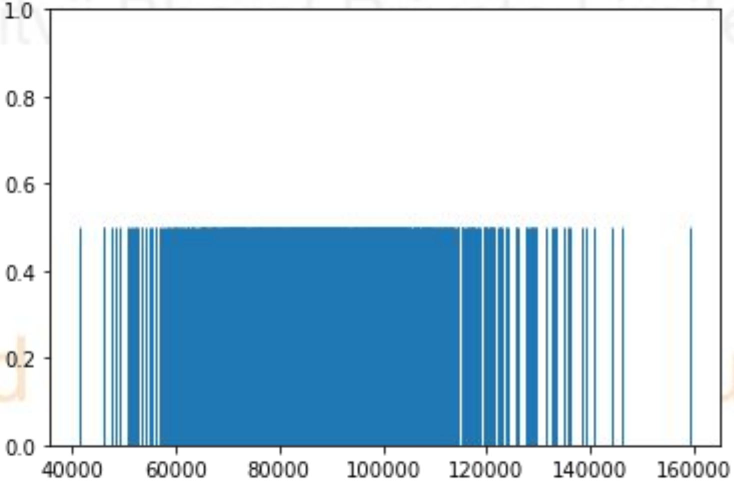
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
IHUB DivyaSampark, IIT Roorkee
and
Ritvi ed
Led ukla

1. Adds a tick for every salary value
2. Optionally adjust height of ticks
3. Y-axis not interpretable



1. Highest salary near $160,000
2. Many salaries between $60k - $120k

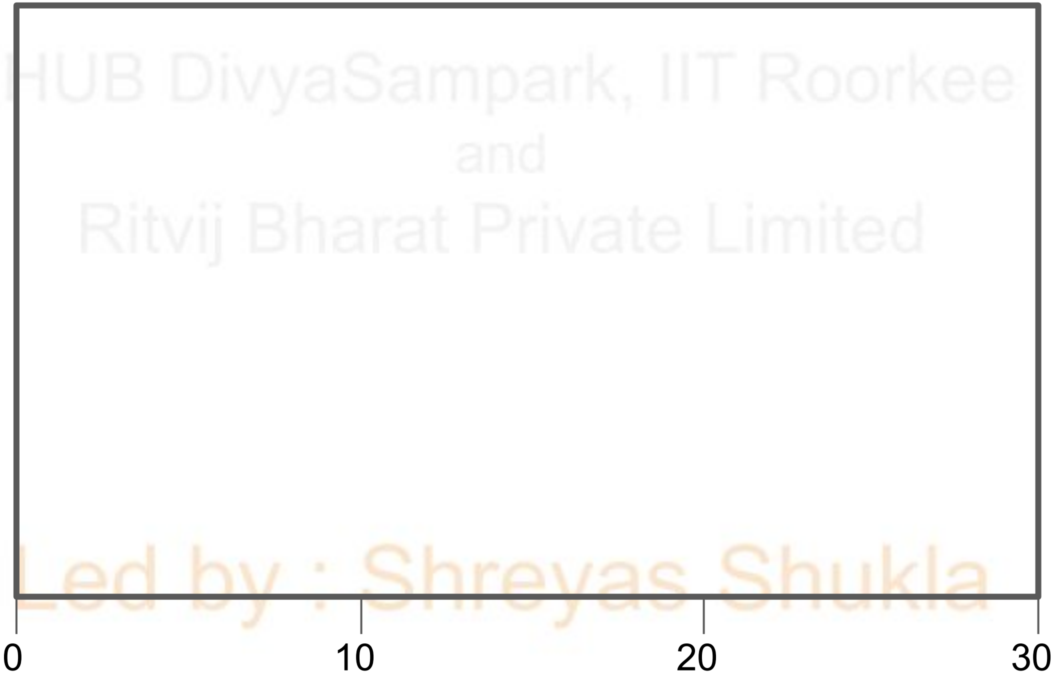
Many ticks could be right on top of eachother, we can't tell!If we count how many ticks there are per various x-ranges, we can create a histogram.



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

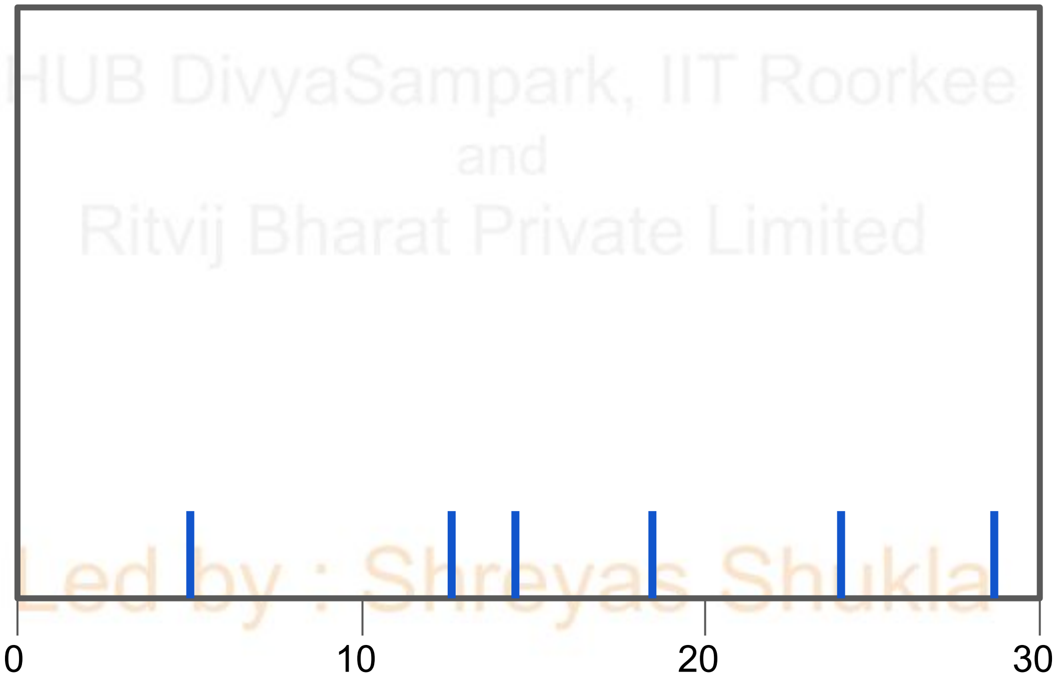
Let’s explore a simple example



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

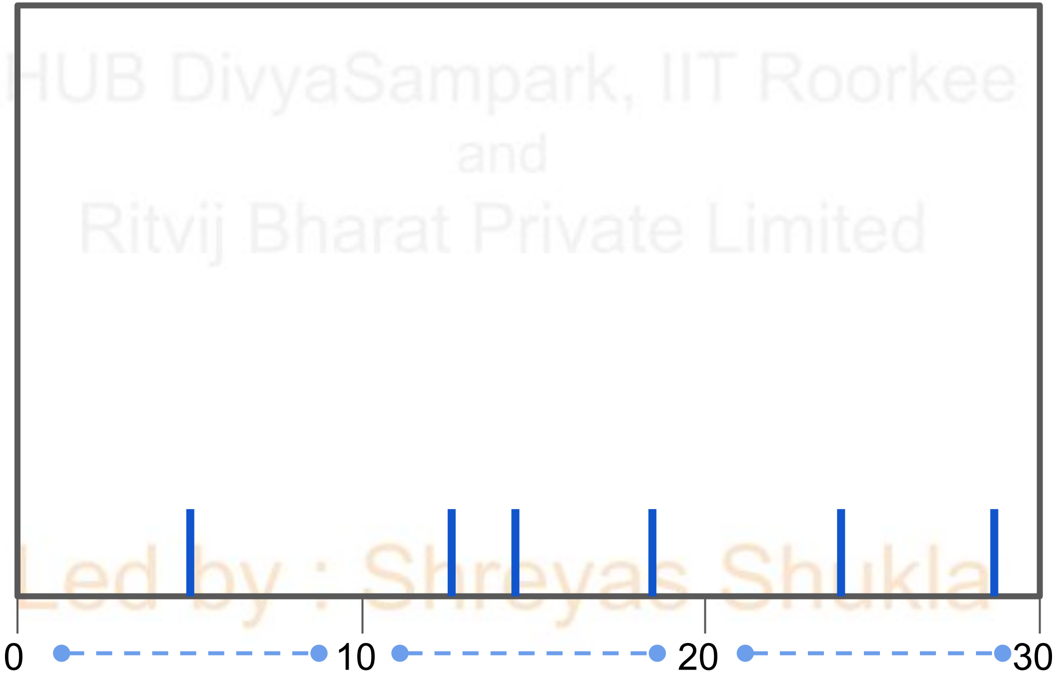
We place the rug plot ticks



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

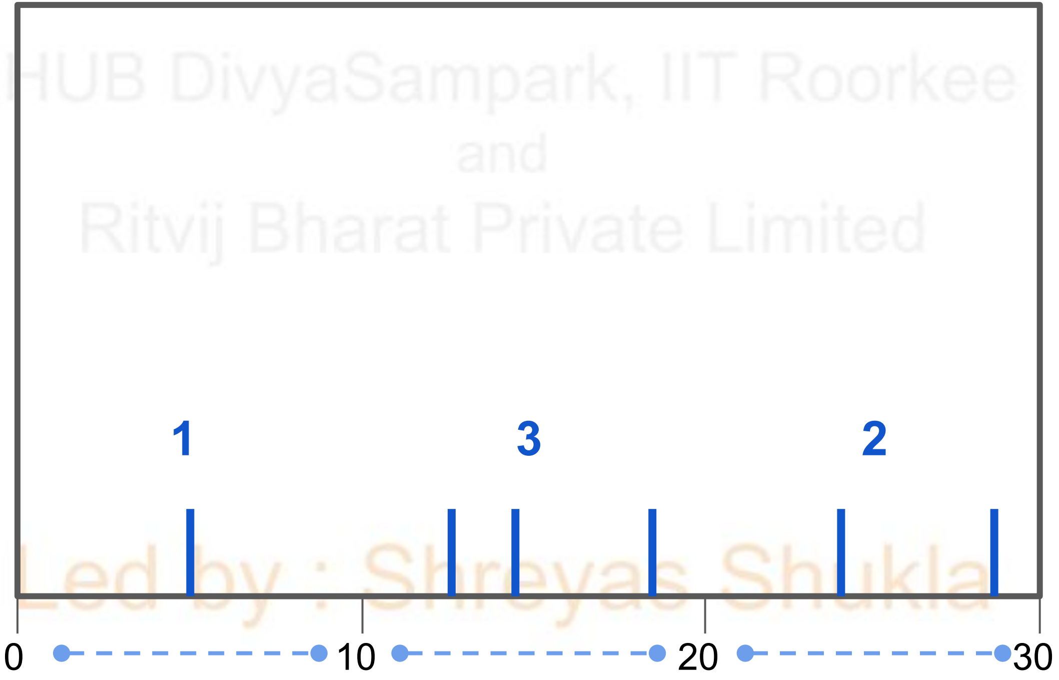
Choose a number of “bins”, we’ll pick 3



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

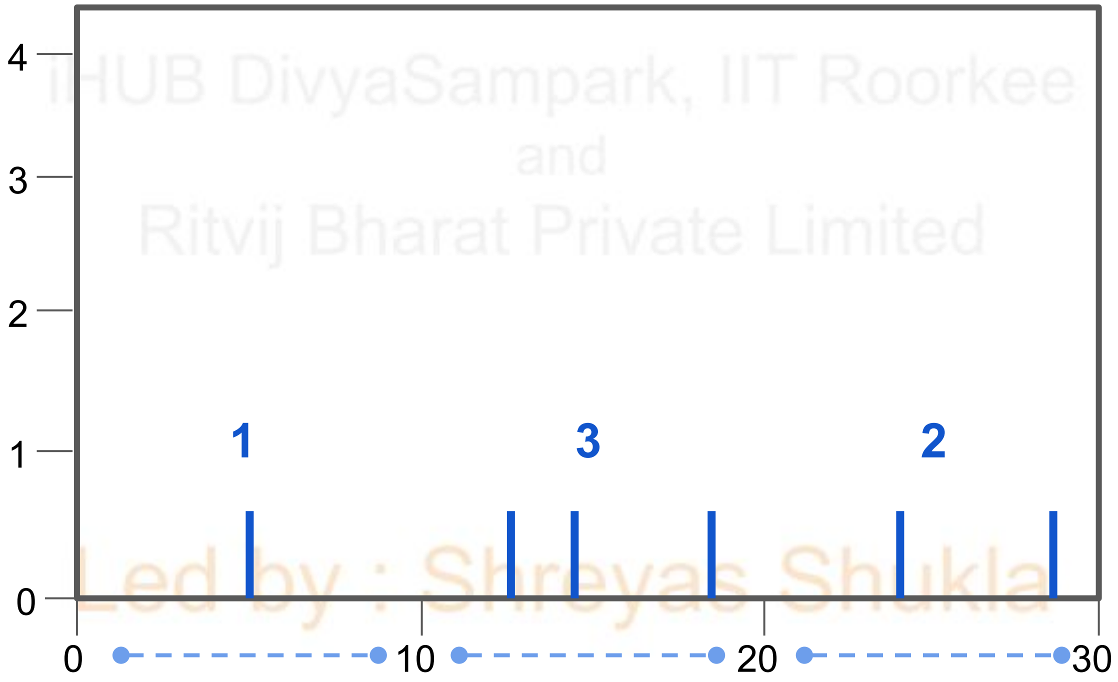
Count ticks per bin



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

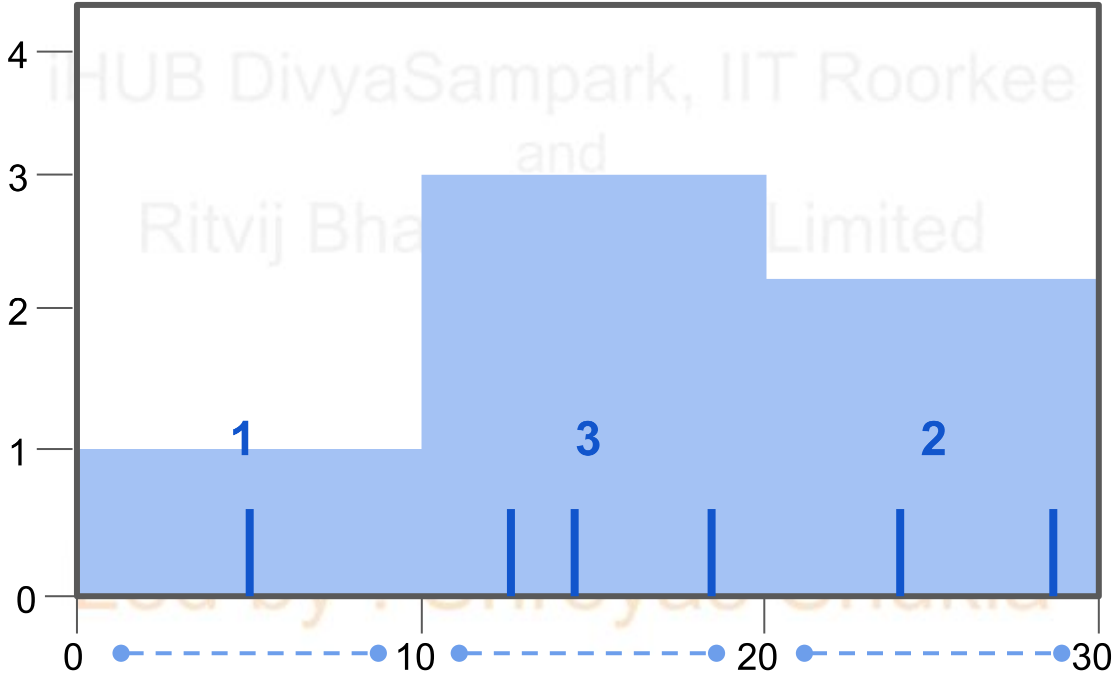
Create a bar as high as count



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

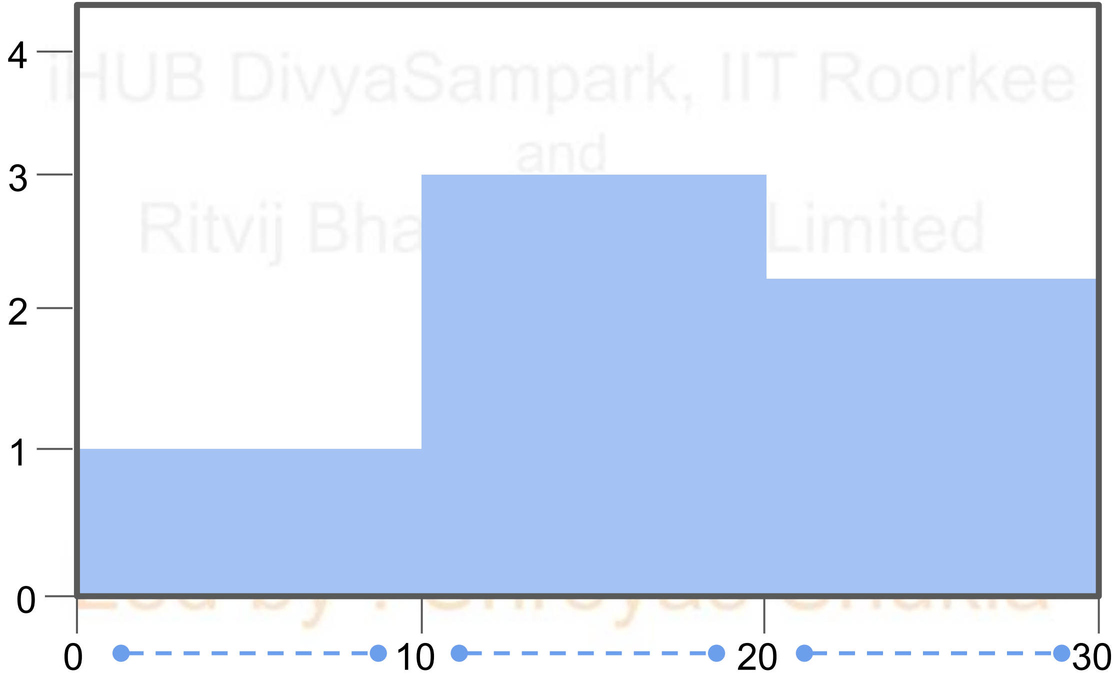
Create a bar as high as count



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bha... Limited

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

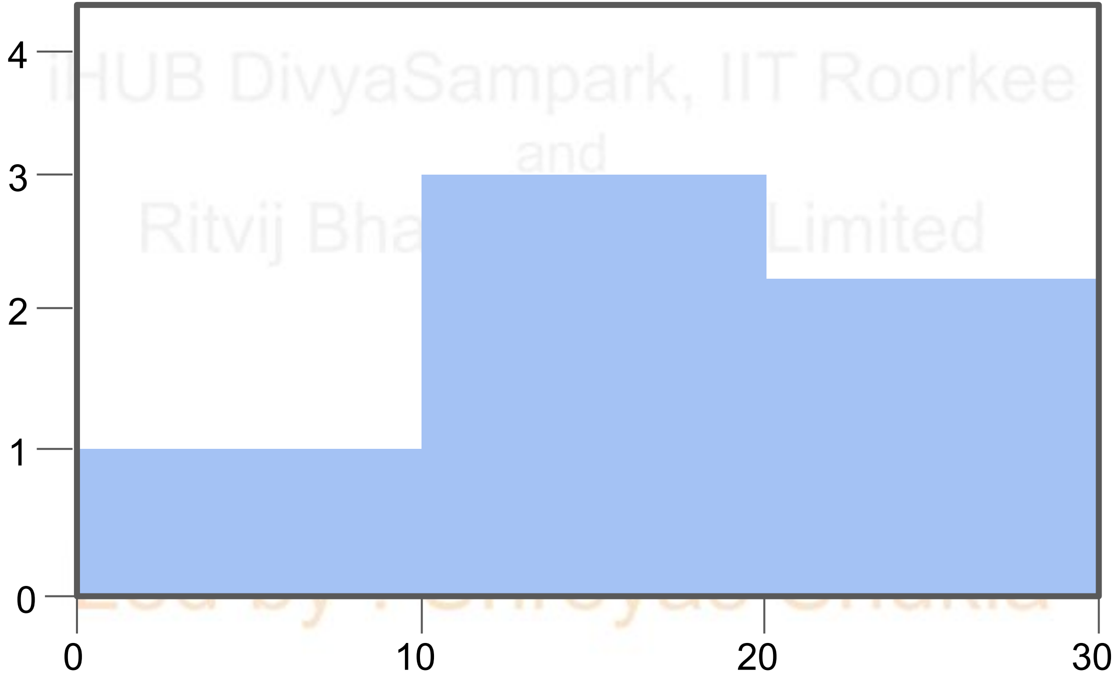
Create a bar as high as count



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bha... Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

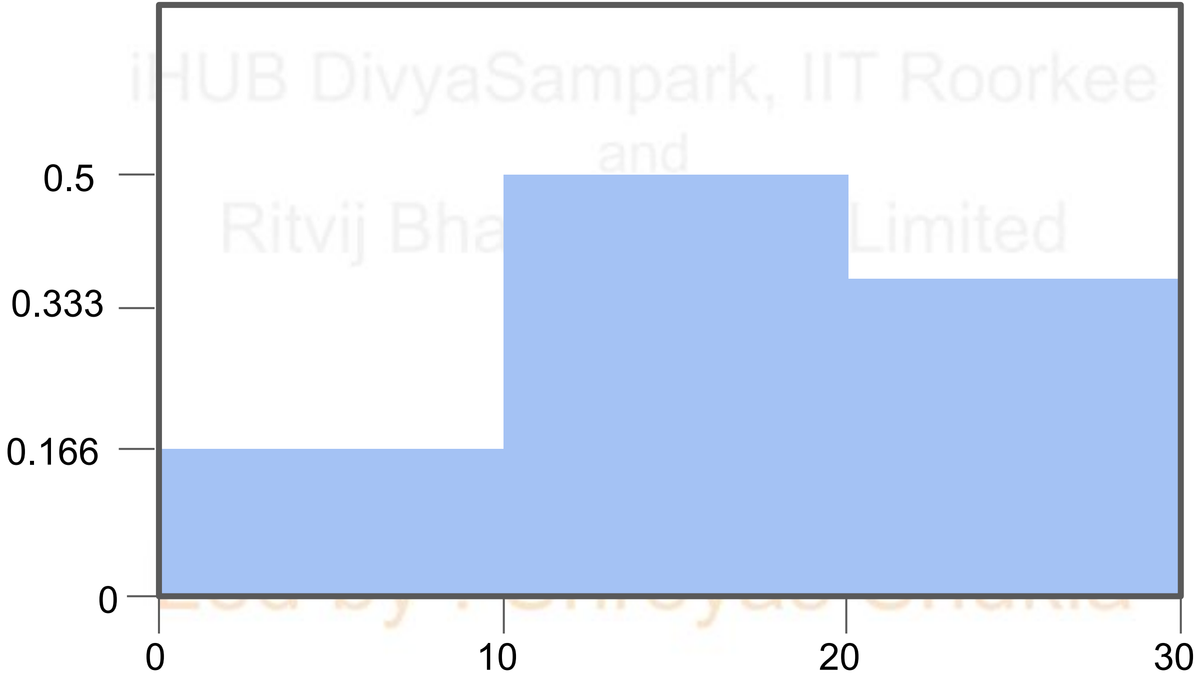
Histogram is complete



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bha... Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

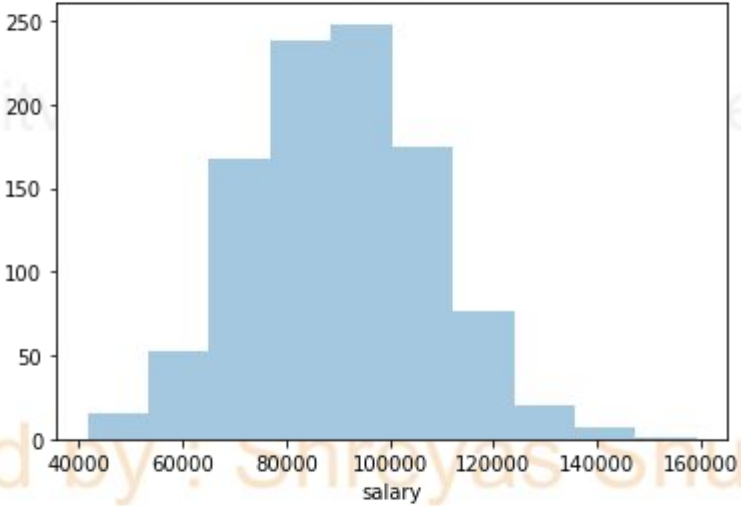
Y-axis can also be normalized as percent



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bha... Limited
Led by : Shreyas Shukla

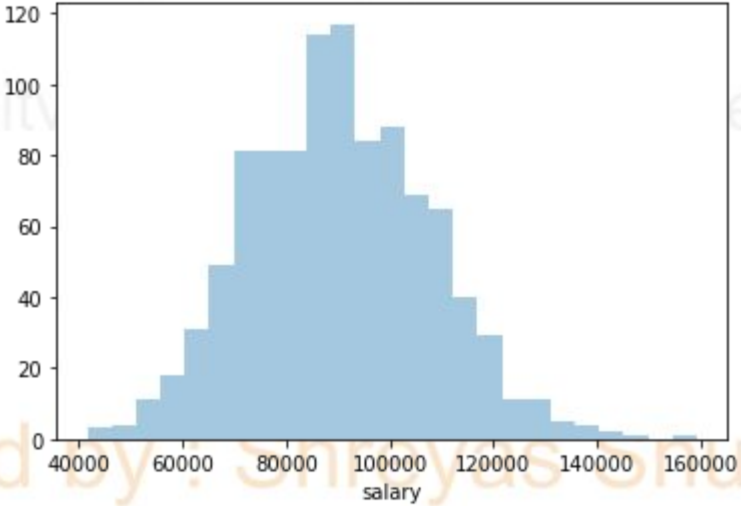
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
Rit ed
Led by: Shreyas Shukla

Changing number of bins shows more detail instead of general trends.



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
Rit ed
Led by: Shreyas Shukla

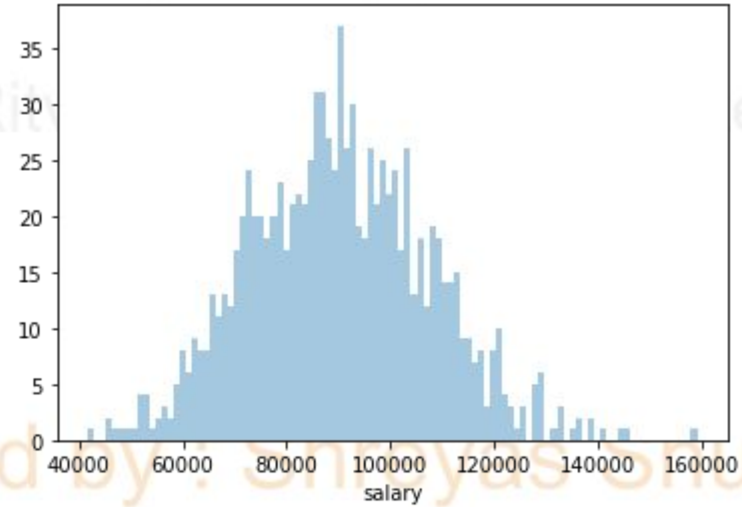
Changing number of bins shows more detail instead of general trends.



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Changing number of bins shows more detail instead of general trends.

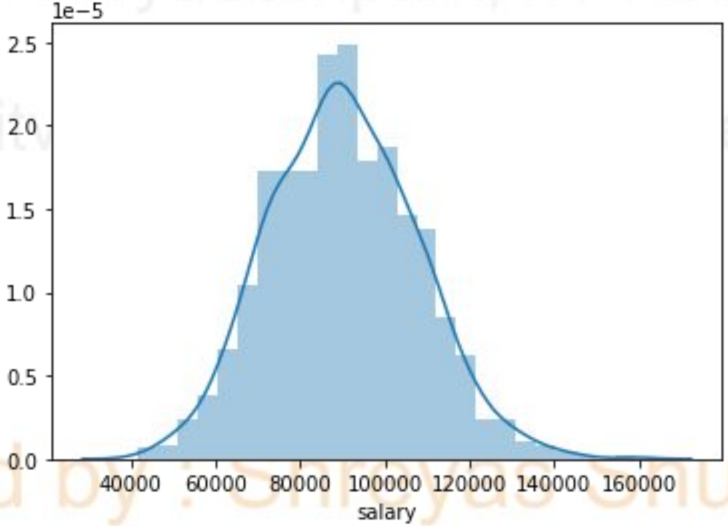
iHUB DivyaSampark, IIT Roorkee



Led by: Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
Led by : Shreyas Shukla

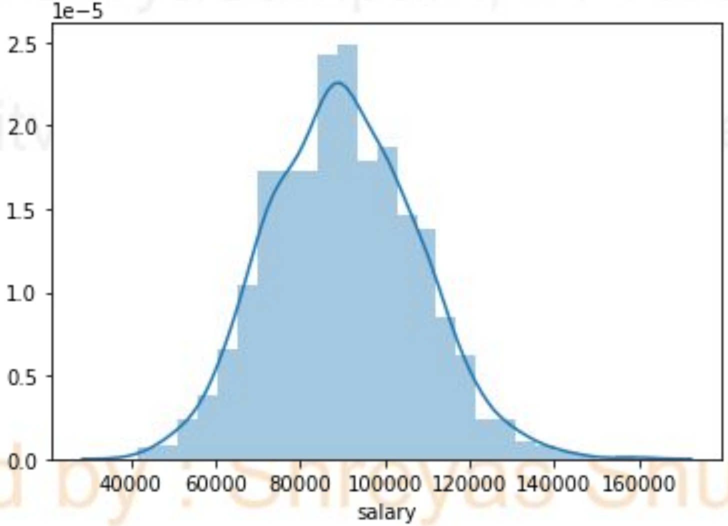
Seaborn also allows us to add on a KDE plot curve on top of a histogram.



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Let’s explore what a KDE plot is and how it is constructed.

iHUB DivyaSampark, IIT Roorkee



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
IHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

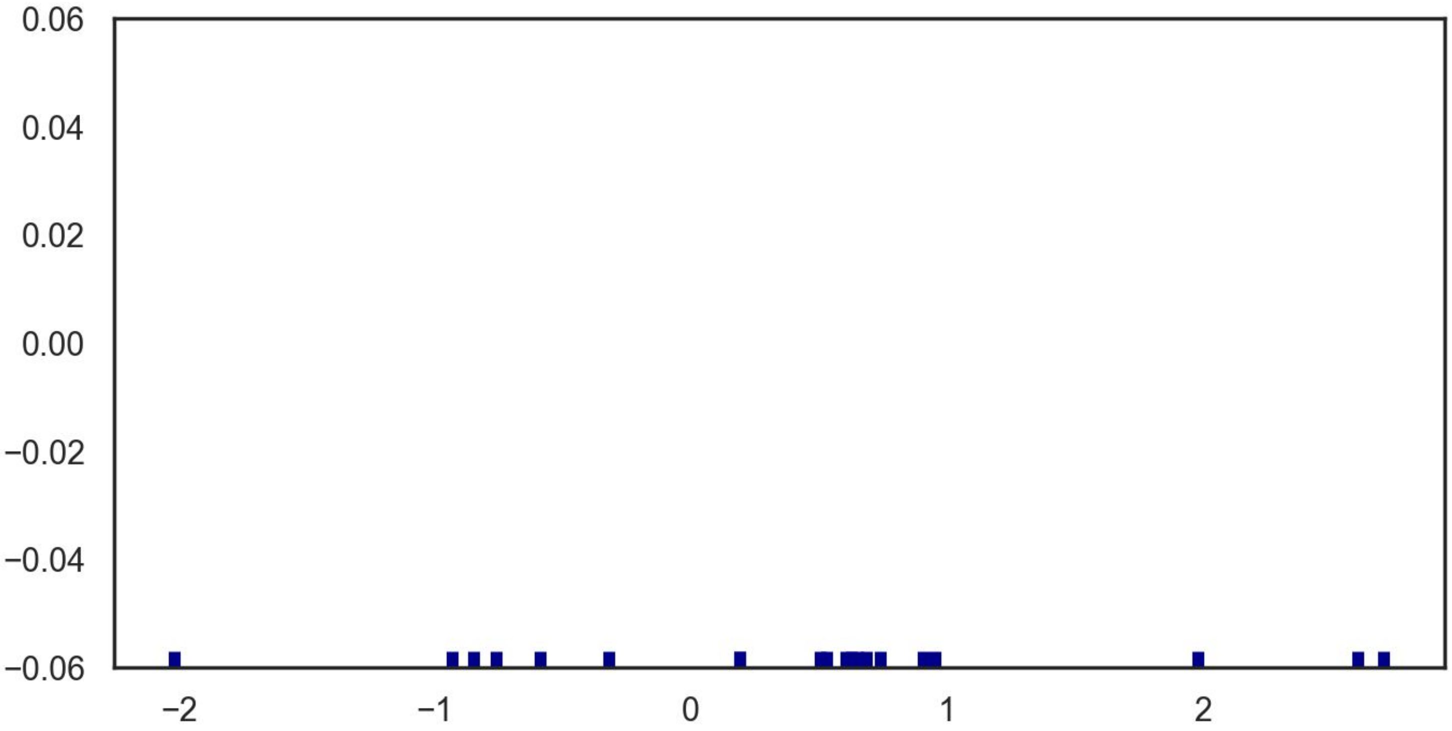
Kernel Density Estimation (KDE) is a method of **estimating** a probability density function of a random variable.

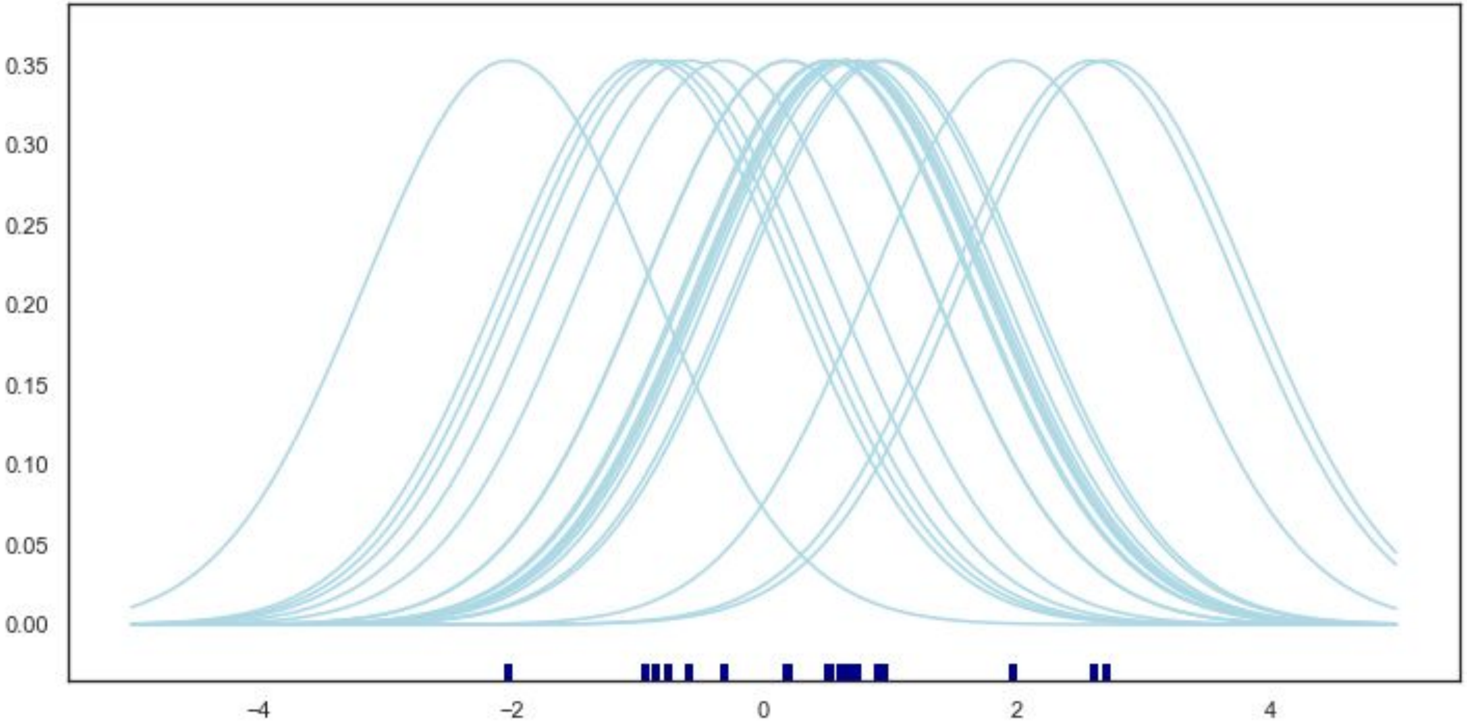
In simpler terms, it is a way of estimating a continuous probability curve for a finite data sample.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

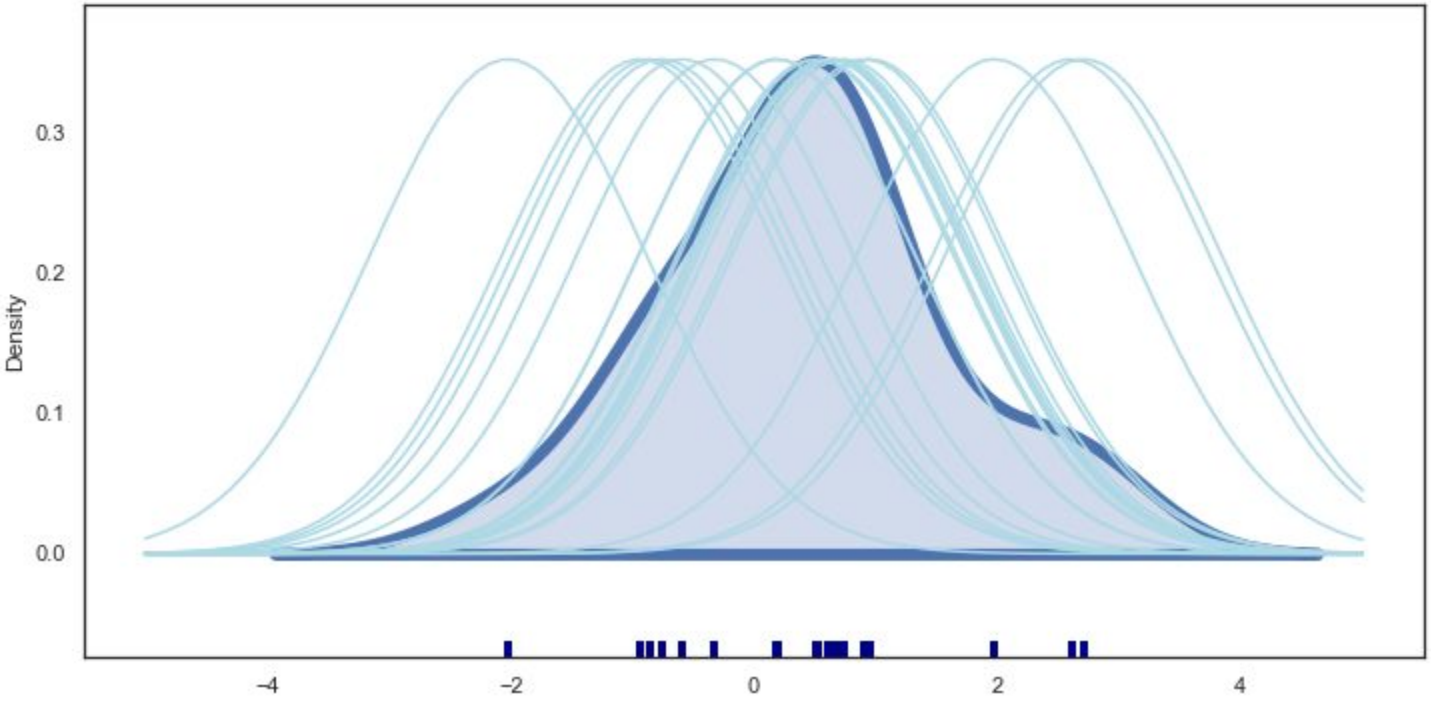
KDE plots are best understood by visualizing their “construction”.

Let’s start with a rug plot….


0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−2
−1
0
1
2

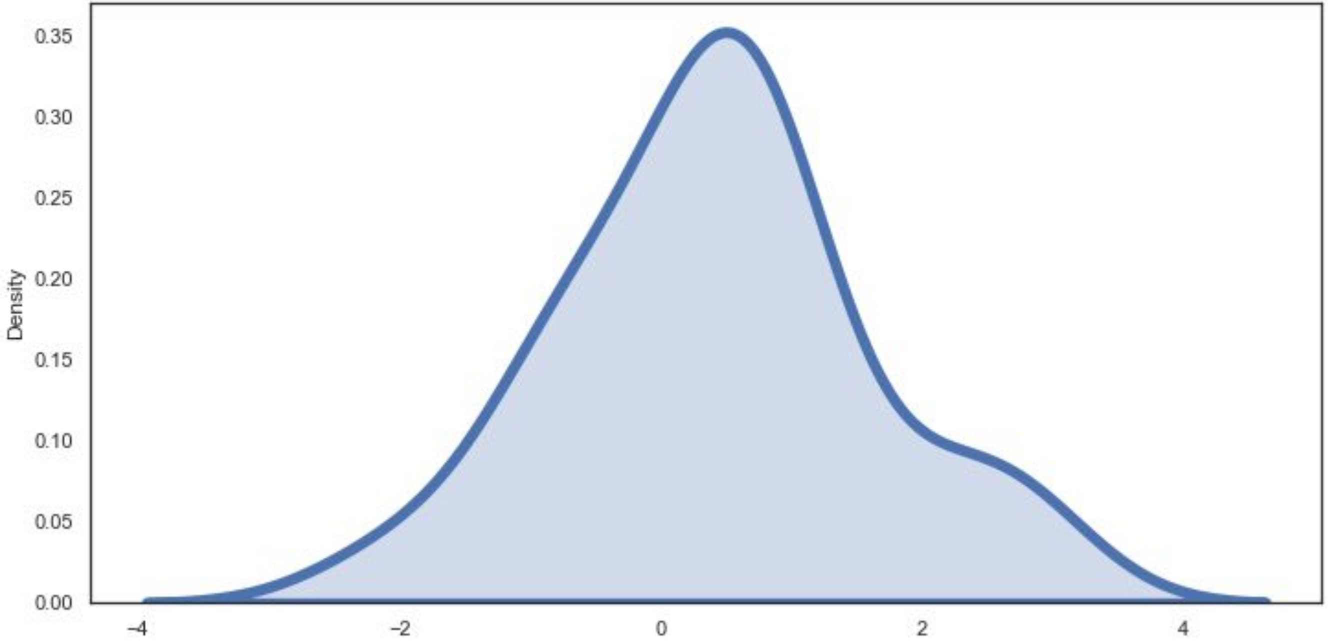

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−2
0
2
4

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

You can change the kernel and bandwidth used which can make your KDE show more or less of the variance contained in the data.

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

Let’s Code !!

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Categorical Plots

Statistical Estimation within Categories
Part One: Understanding the Plots

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

The categorical plots display a statistical metrics **per** a category.

- For example mean value per category or a count of the number of rows **per** category.
- It is the visualization equivalent of a groupby() call.

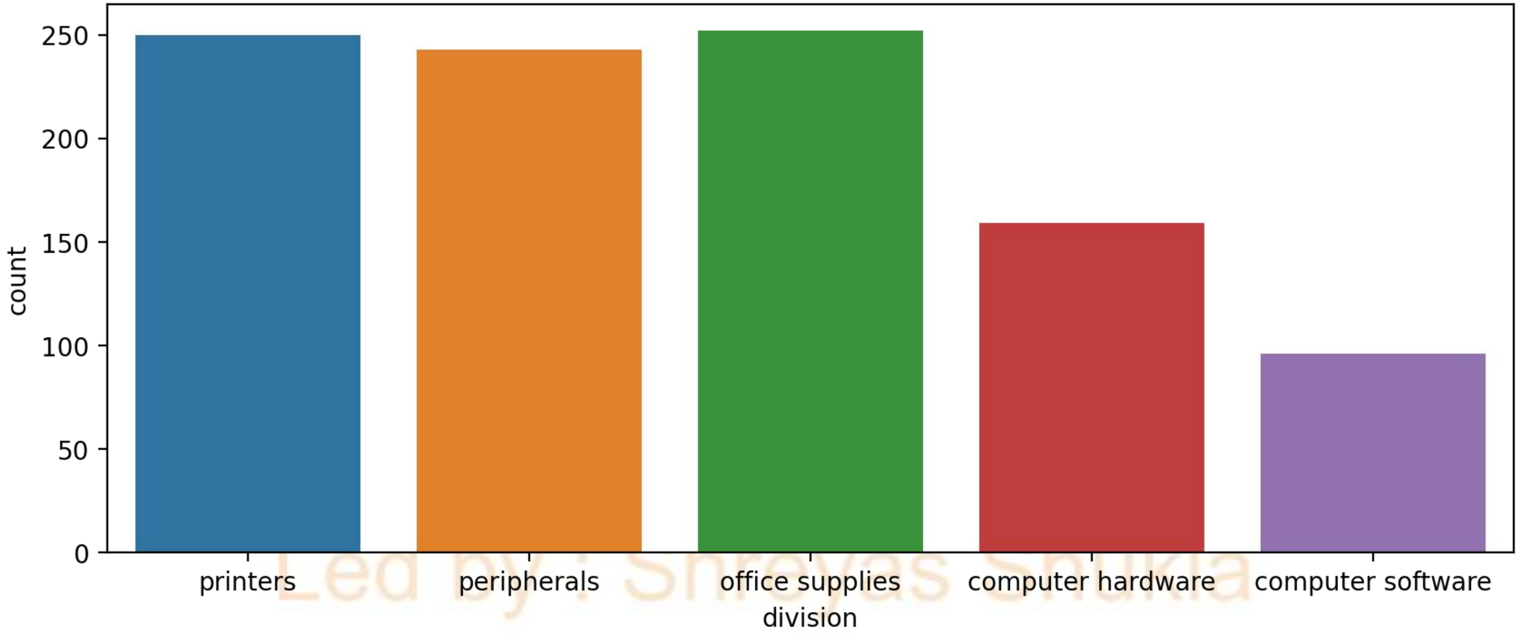
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ravij Bharat Private Limited
Led by : Shreyas Shukla

Two main types of plots for this are:

- countplot()
  - Counts number of rows per category.
- barplot()
  - General form of displaying any chosen metric per category.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

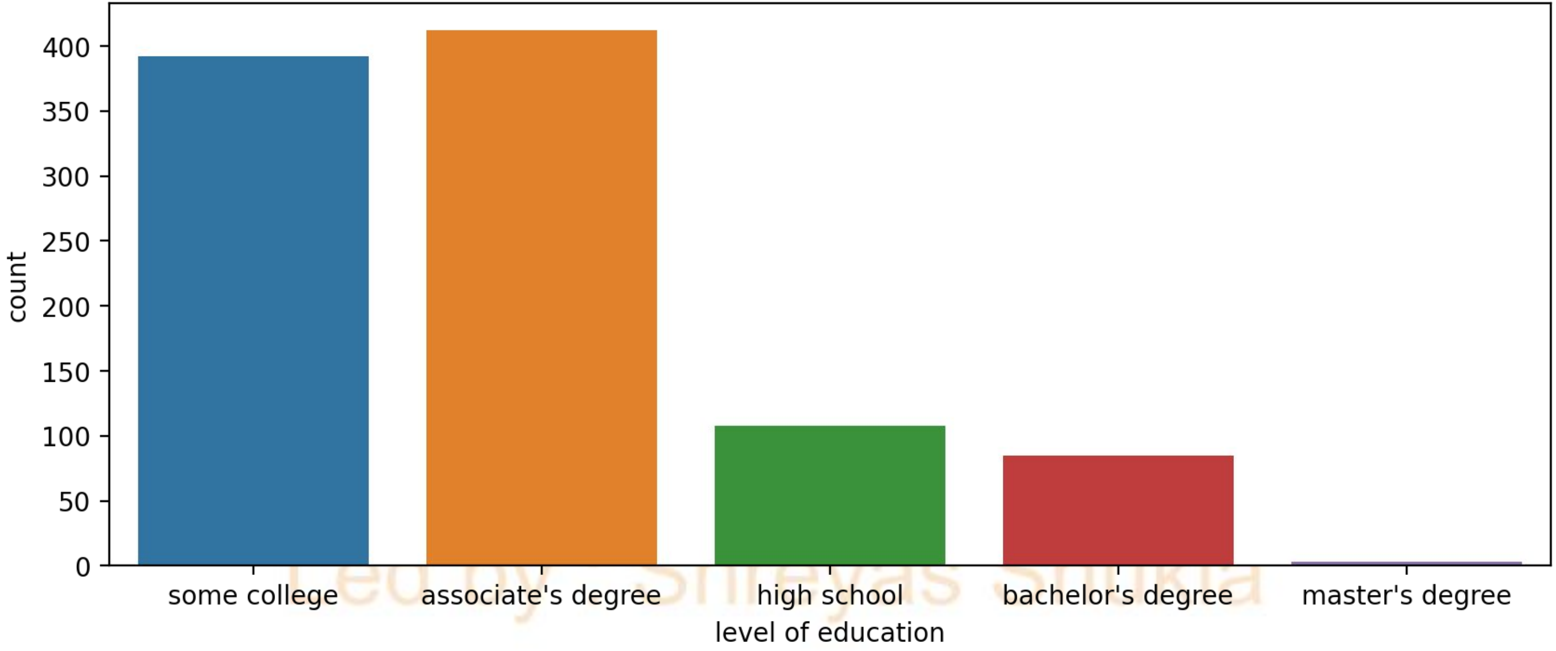
Countplot for corporate divisions



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

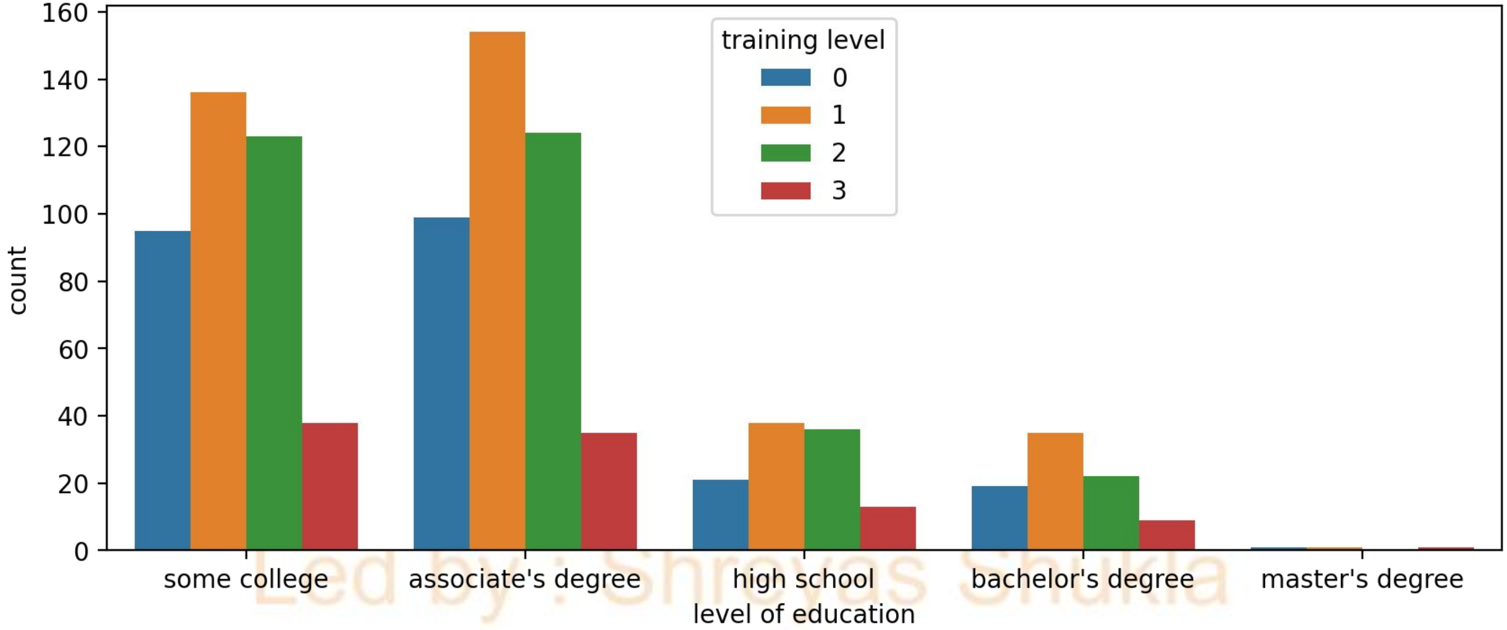
Countplot for education level



Led by Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Countplot with additional hue separation



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

The barplot is the general form that allows you to choose any measure or estimator for the y axis.

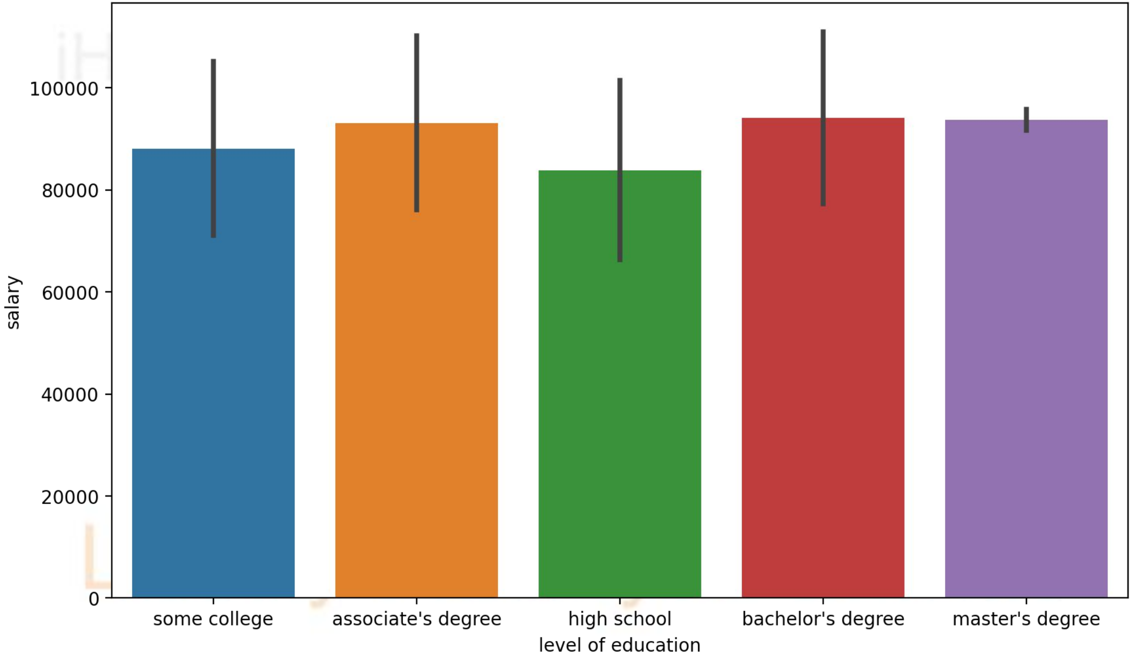
We could plot the mean value and standard deviation per category instead.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Prithvij Bharat Private Limited
Led by : Shreyas Shukla

Caution!

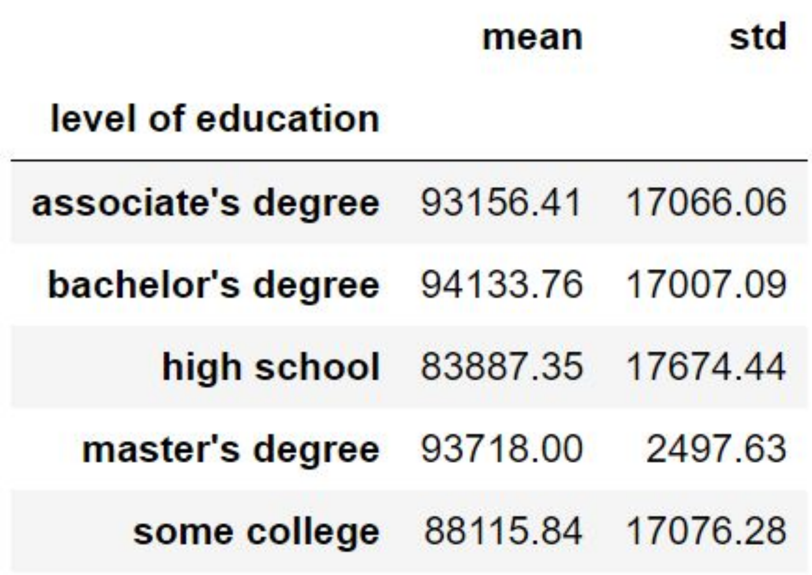
- Be very careful with these plots, since the bar is filled and continuous, a viewer may interpret continuity along the y-axis which may be incorrect!
- Always make sure to add additional labeling and explanation for these plots!

Barplot showing mean and SD bar



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

A simple table is probably better.



	mean	std
level of education		
associate's degree	93156.41	17066.06
bachelor's degree	94133.76	17007.09
high school	83887.35	17674.44
master's degree	93718.00	2497.63
some college	88115.84	17076.28

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Let’s explore coding out these plots with seaborn!

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

Let’s Code !!

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Categorical Plots

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

We’ve explored distribution plots for a single feature, but what if we want to compare distributions across categories?

For example, instead of the distribution of everyone’s salary, we can compare the distributions of salaries **per** level of education.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

We will first separate out each category, then create the distribution visualization.

Let’s explore what plot types we have available….

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
[illegible] Bharat Private Limited
Led by : Shreyas Shukla

Distribution within Categories

- ○ Boxplot
- ○ Violinplot
- ○ Swarmplot

Let’s explore understanding these plots on the previous salary dataset.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

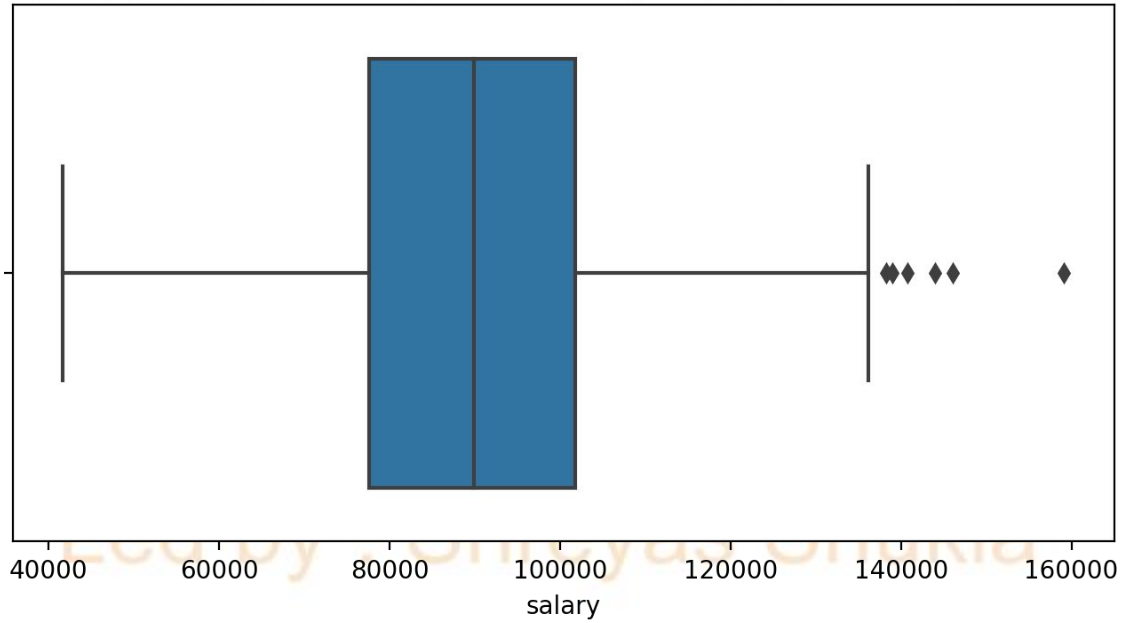
The Boxplot displays the distribution of a continuous variable. It does this through the use of quartiles.

Quartiles separate out the data into 4 equal number of data points :

- ○ 25% of data points are in bottom quartile.
- ○ 50th percentile (Q2) is the median.

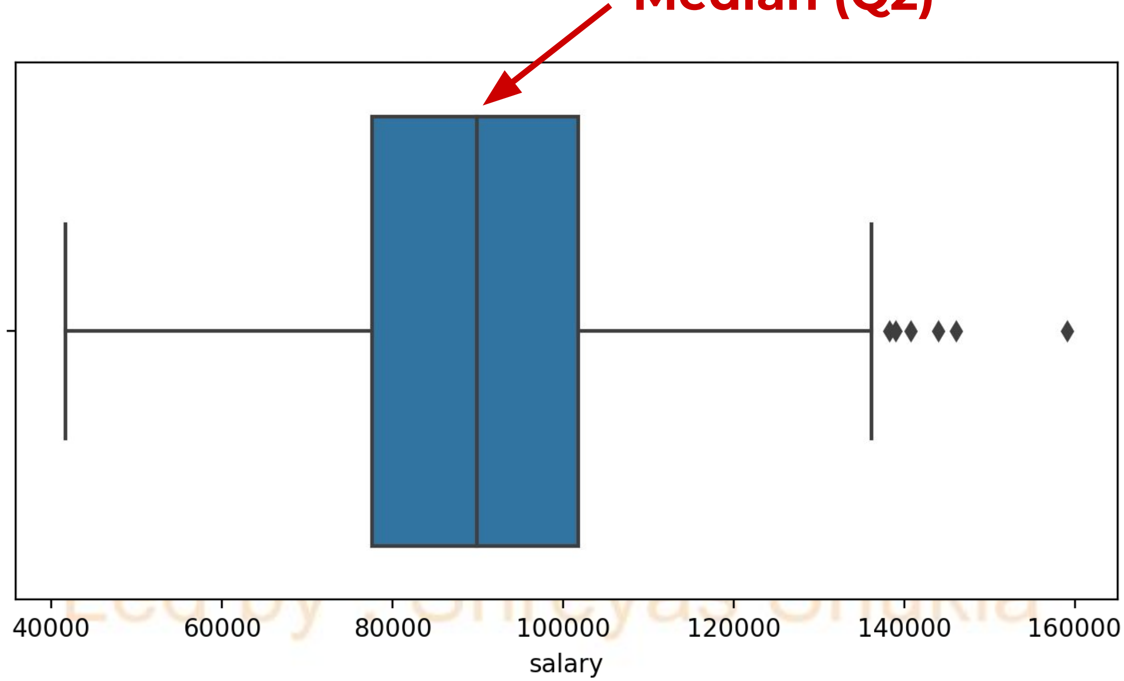
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Boxplot on single feature:



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Median is 50th percentile. Median splits data in half

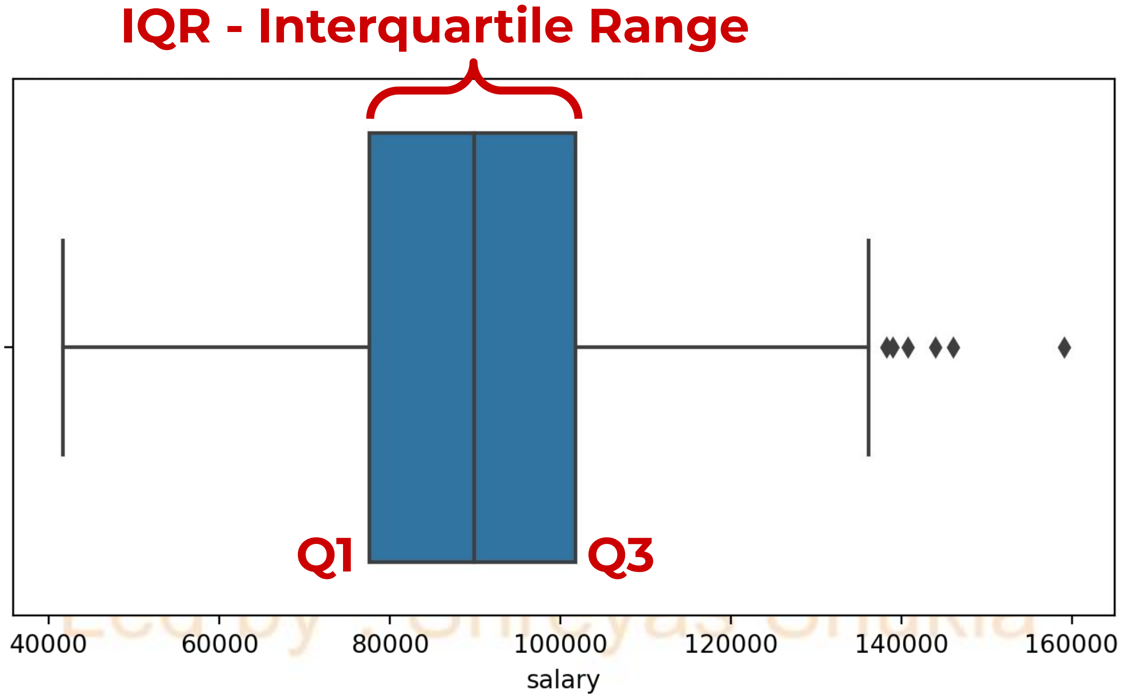


Median (Q2)

Led by Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 – 18th Oct 2024)

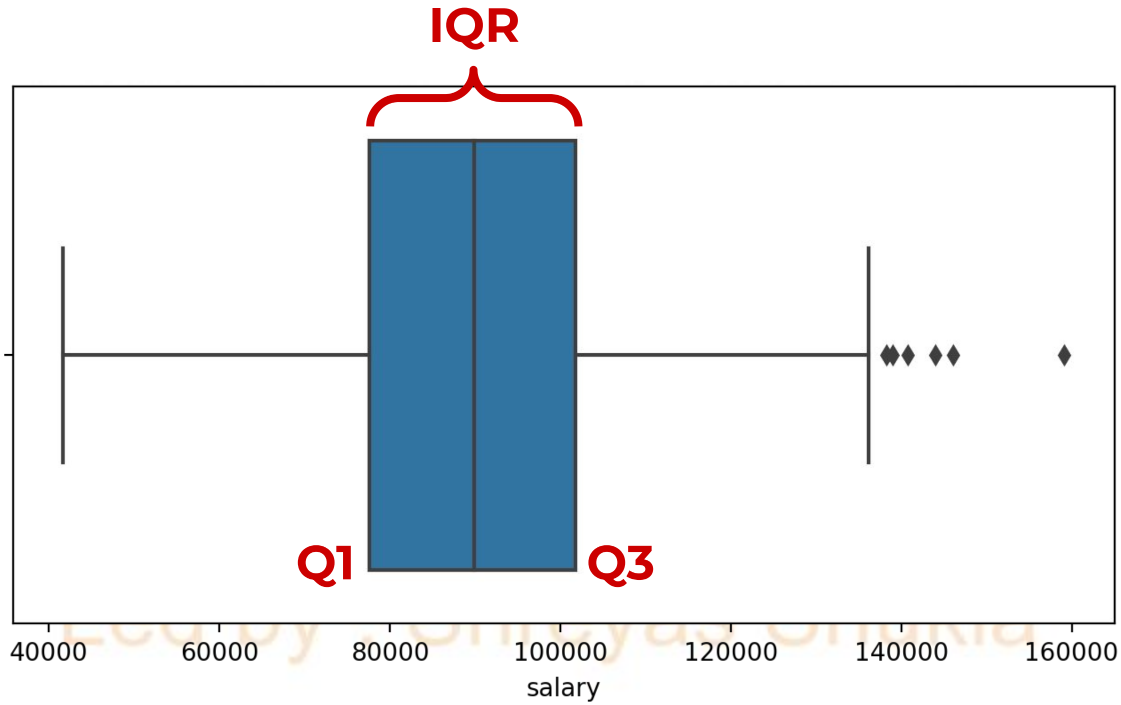
- IQR defines the box width
- 50% of all data points are inside the box



Led by Shreyas Shukla

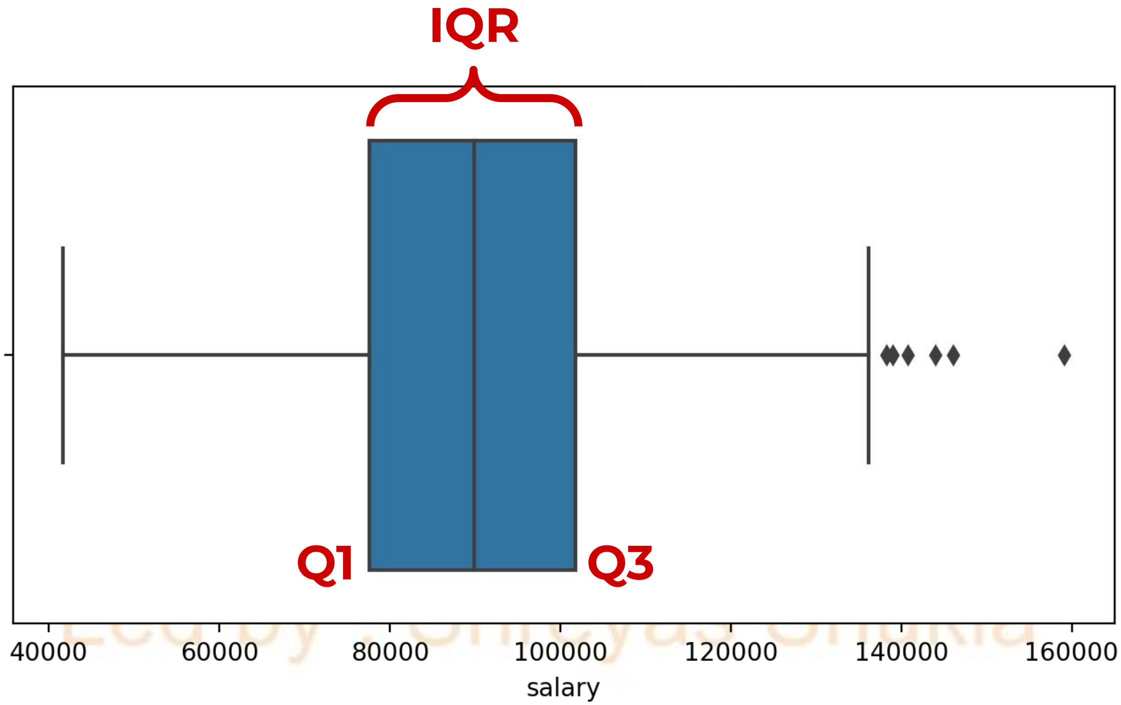
Mastering Machine Learning with Python
(27th Aug 2024 - 16th Oct 2024)

Q1 is the 25th percentile below which are 25% of data points



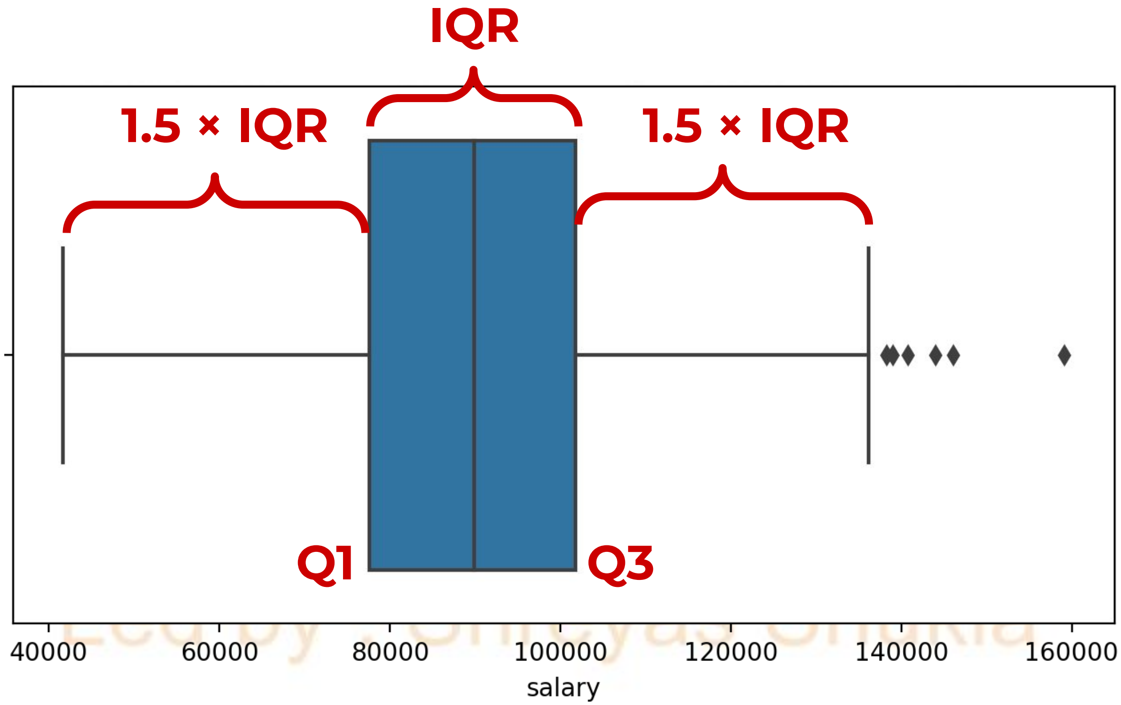
Mastering Machine Learning with Python

Q3 is the 75th percentile. 25% of all points are above Q3



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

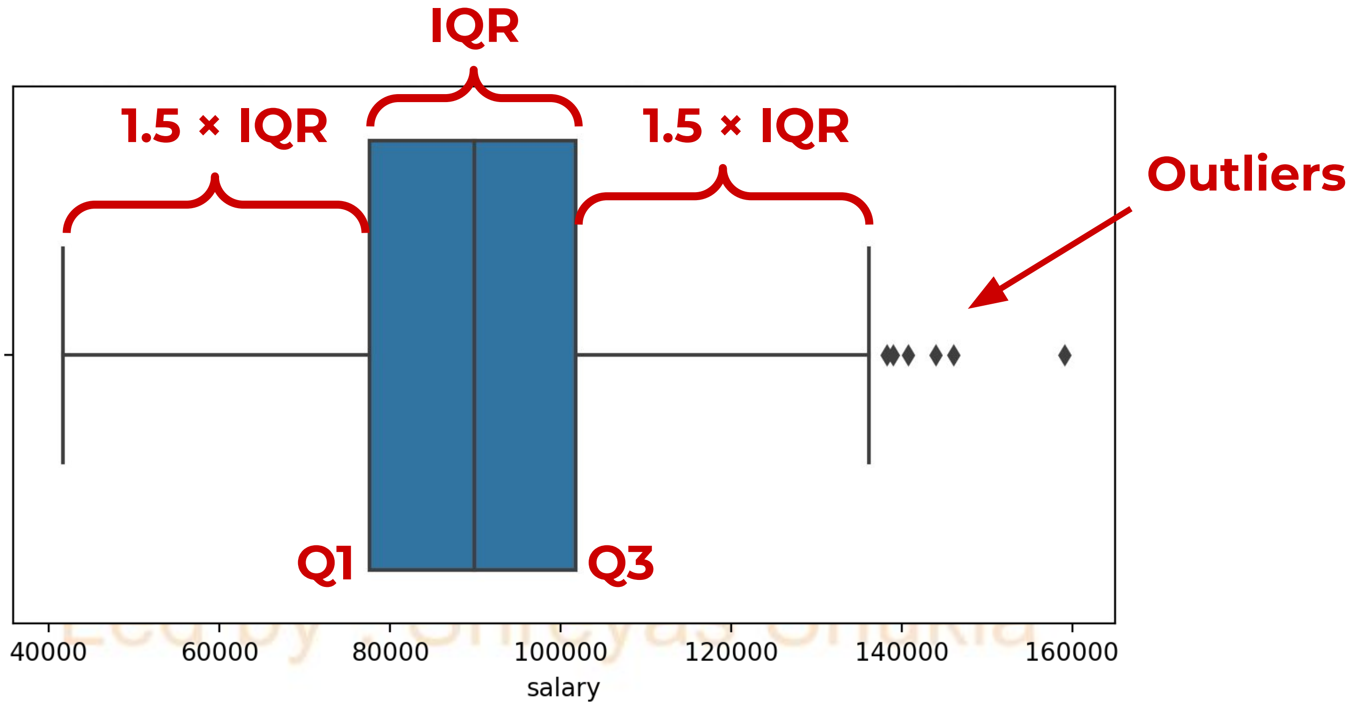
The “whiskers” are defined by 1.5 × IQR



Led by Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

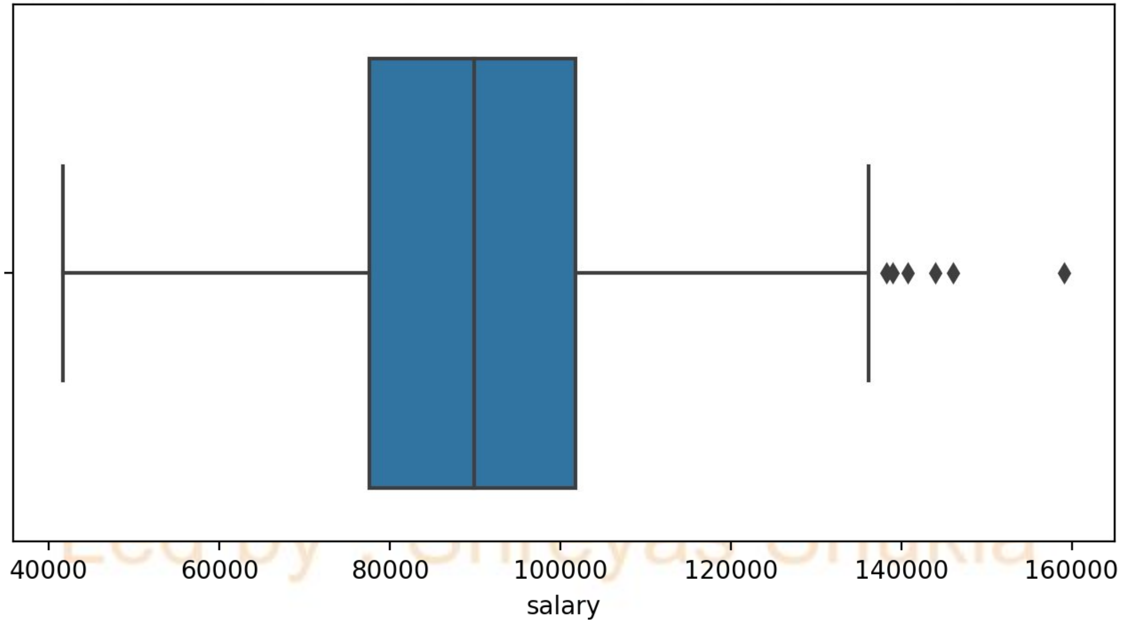
Outside of the whiskers are outliers



Led by Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

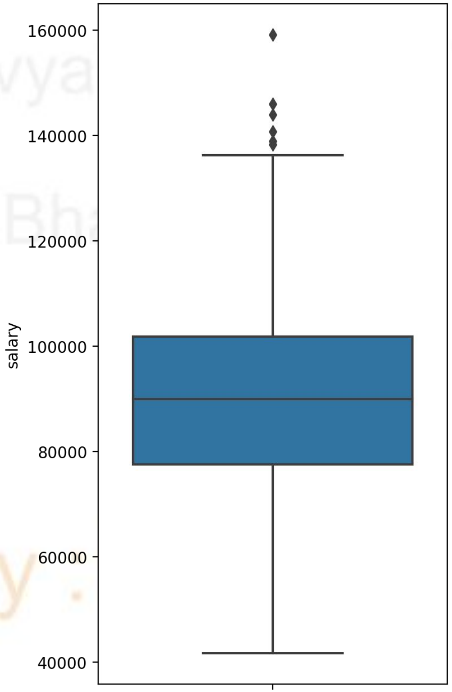
Boxplot gives statistical distribution information in a visual format:



Led by Shreyas Shukla

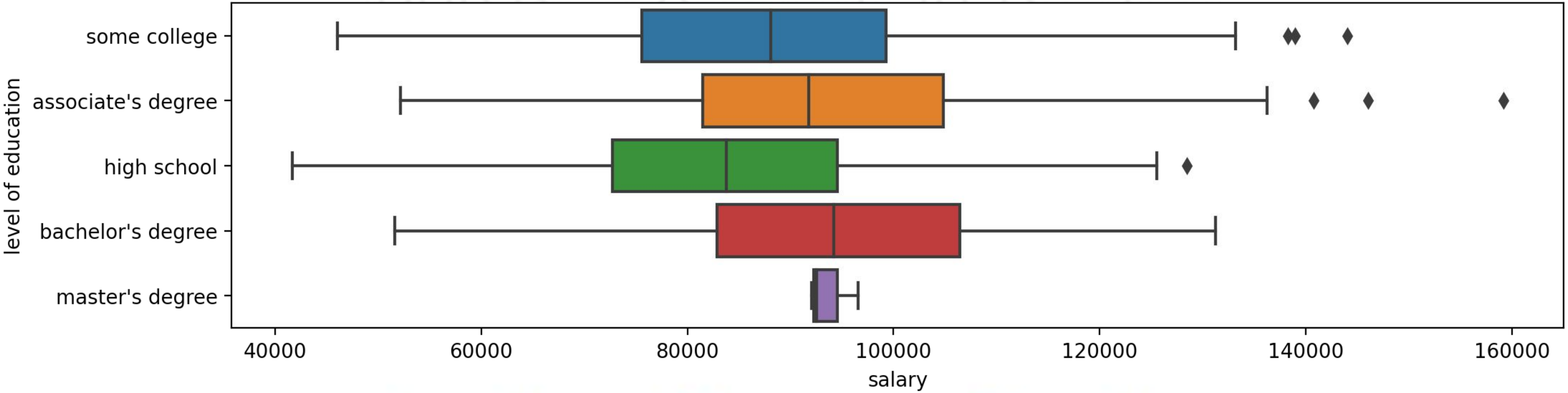
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB Divya T Roorkee
Ritvij Bh Limited
Led by . Shukla

Boxplot can be oriented vertically or horizontally.



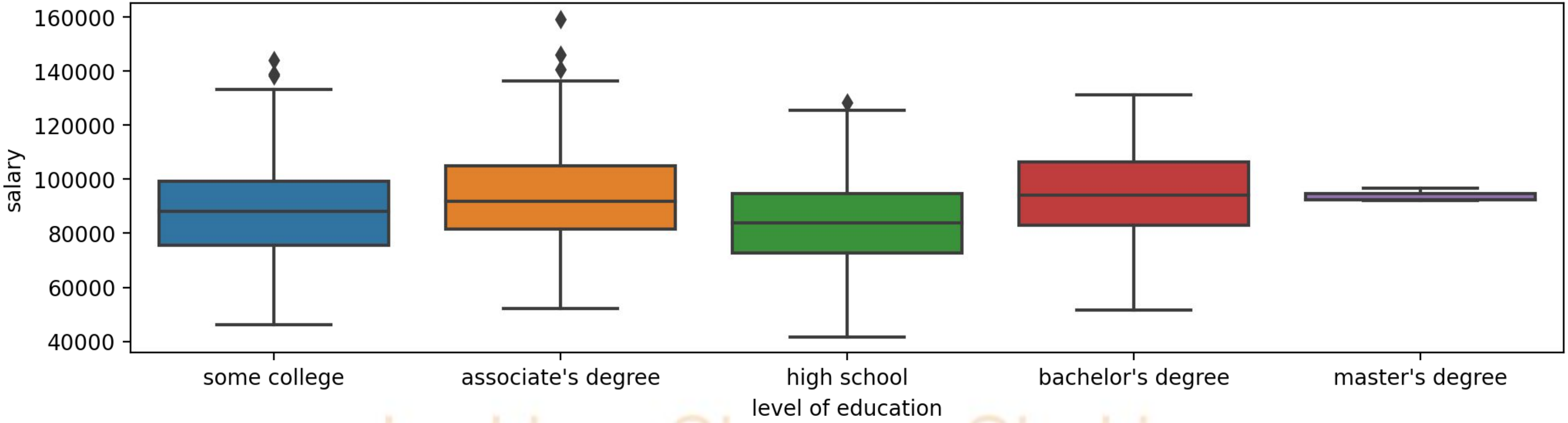
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

We can create a box plot **per** category!



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

The violin plot plays a similar role as the box plot. It displays the probability density across the data using a KDE.

We can imagine it as a mirrored KDE plot.

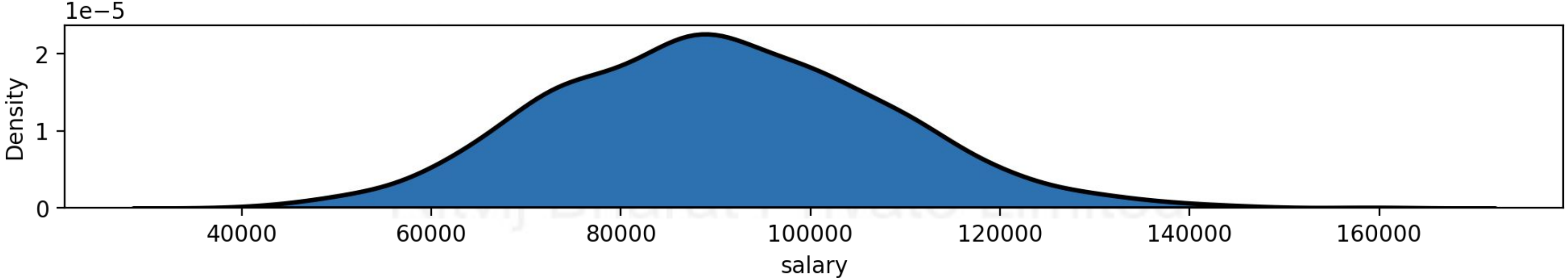
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

The violin plot plays a similar role as the box plot. It displays the probability density across the data using a KDE.

We can imagine it as a mirrored KDE plot.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

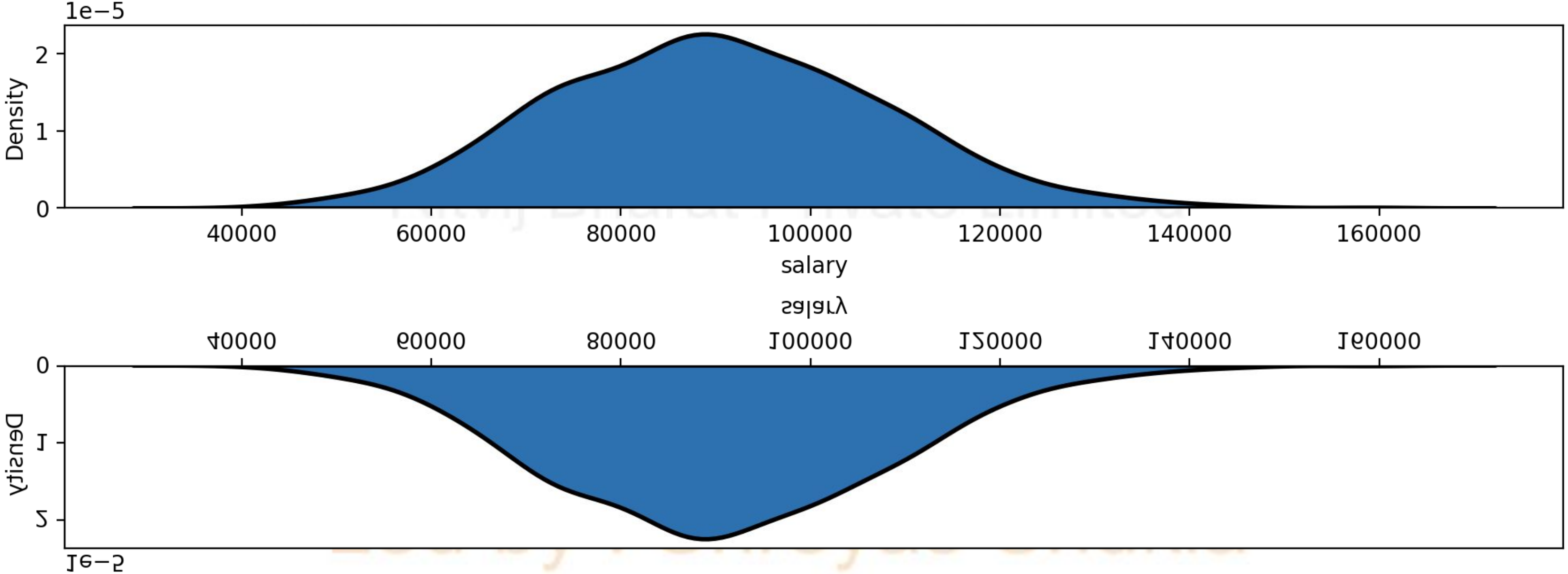
We take the KDE of a single feature:



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

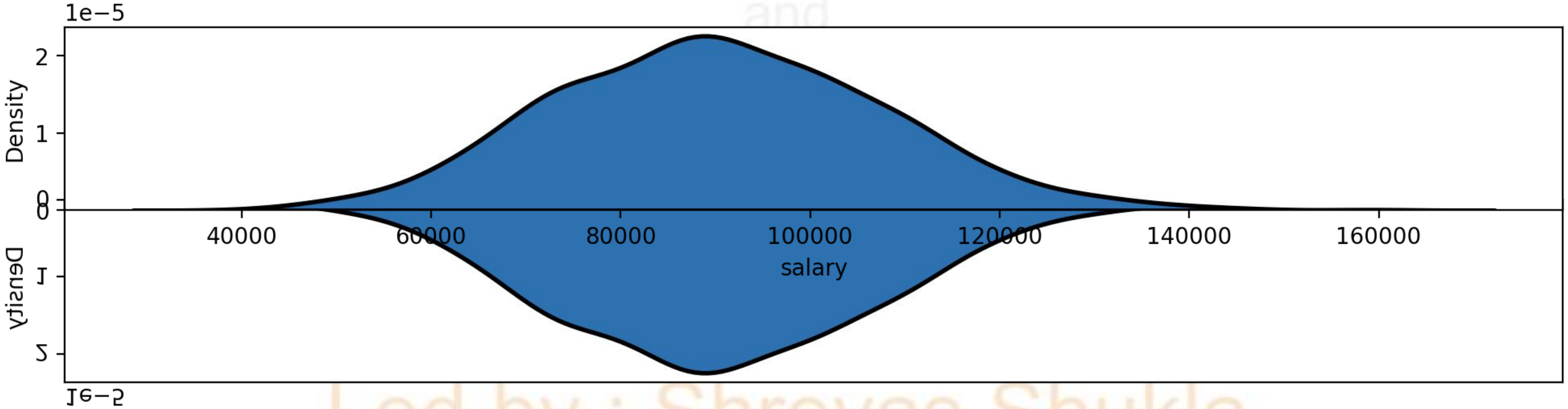
We could then “mirror” it:



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Then combine it to get the violin plot:

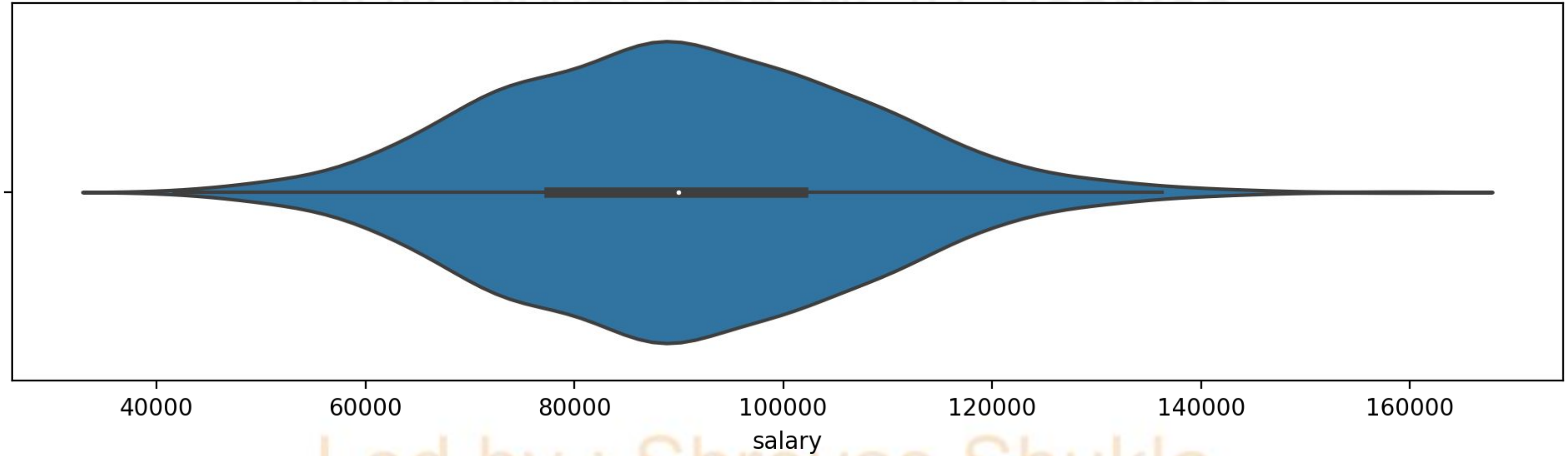
iHUB DivyaSampark, IIT Roorkee
and



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

Then combine it to get the violin plot:

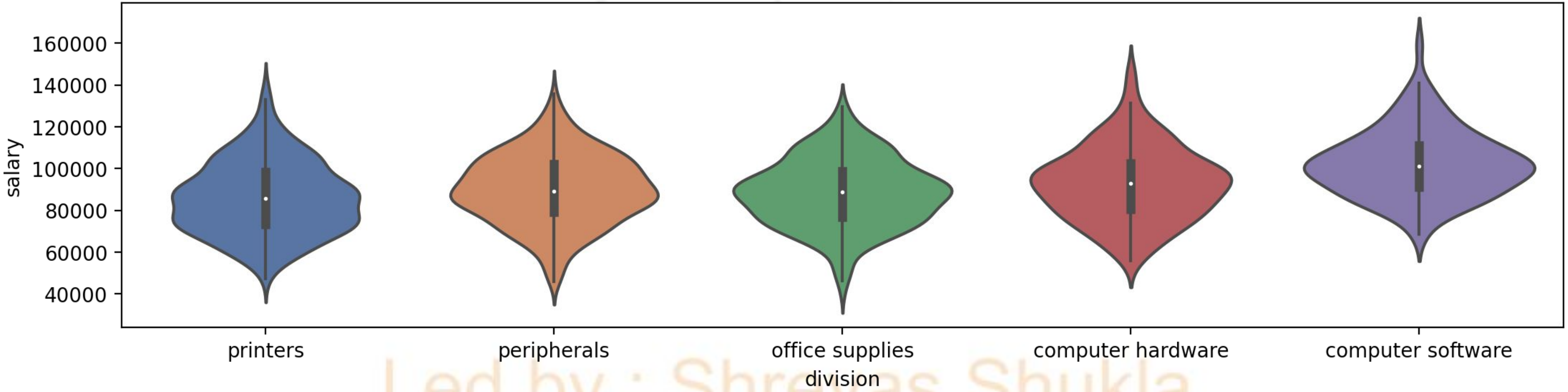


Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

The violin plots can then be created **per** category:

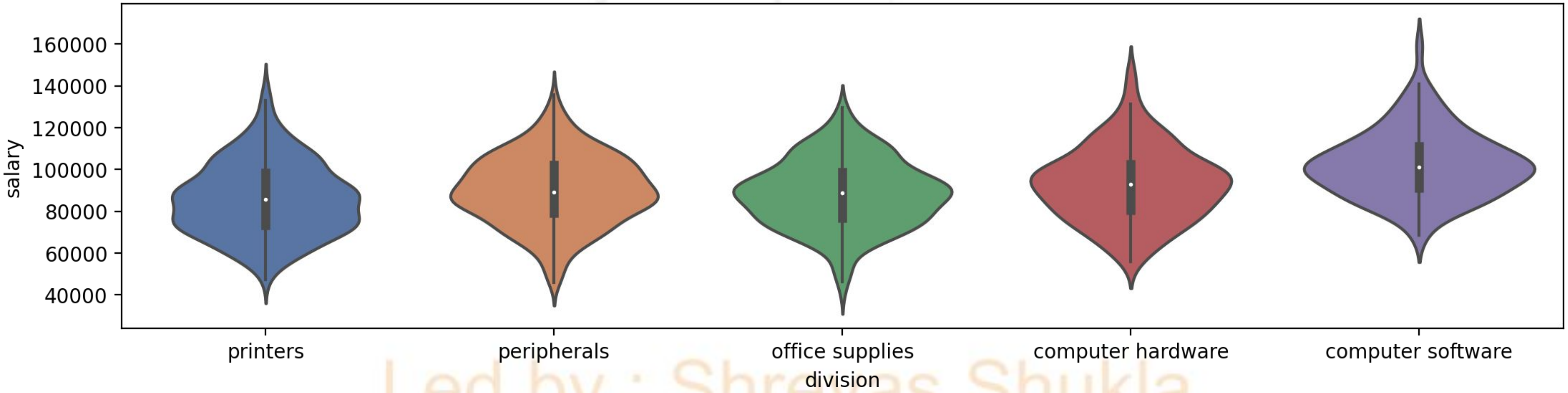
iHUB DivyaSampark, IIT Roorkee



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

A few more less common categorical distribution plot, that is, swarmplot .

iHUB DivyaSampark, IIT Roorkee
and
Rkvij Bharat Private Limited

Let’s quickly explore this plot type...

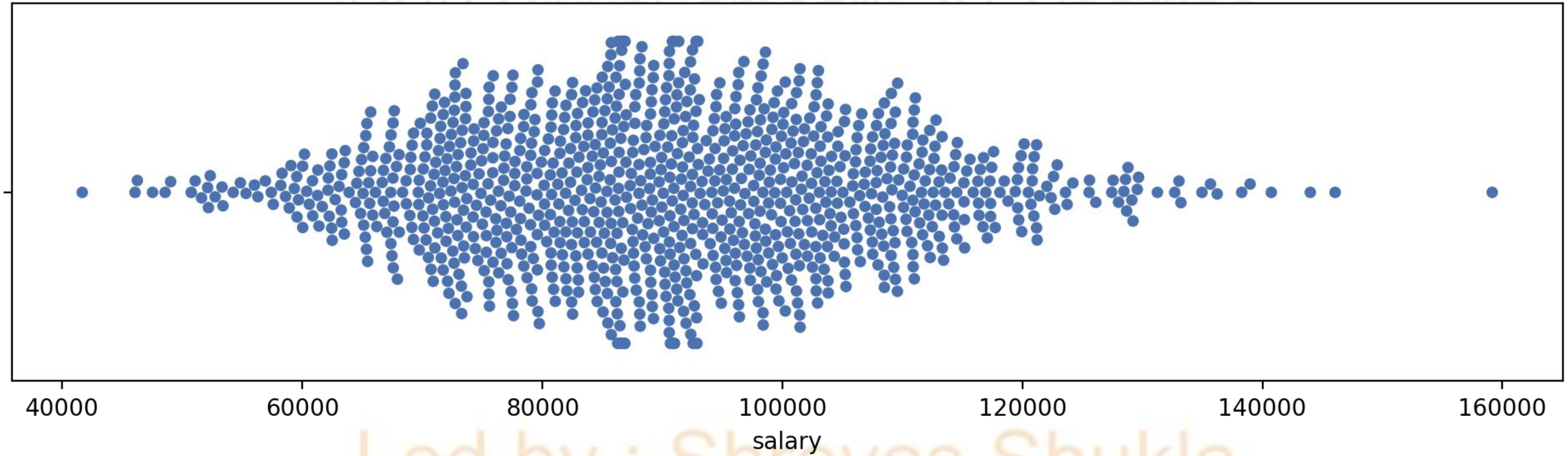
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

The swarmplot simply shows all the data points in the distribution.

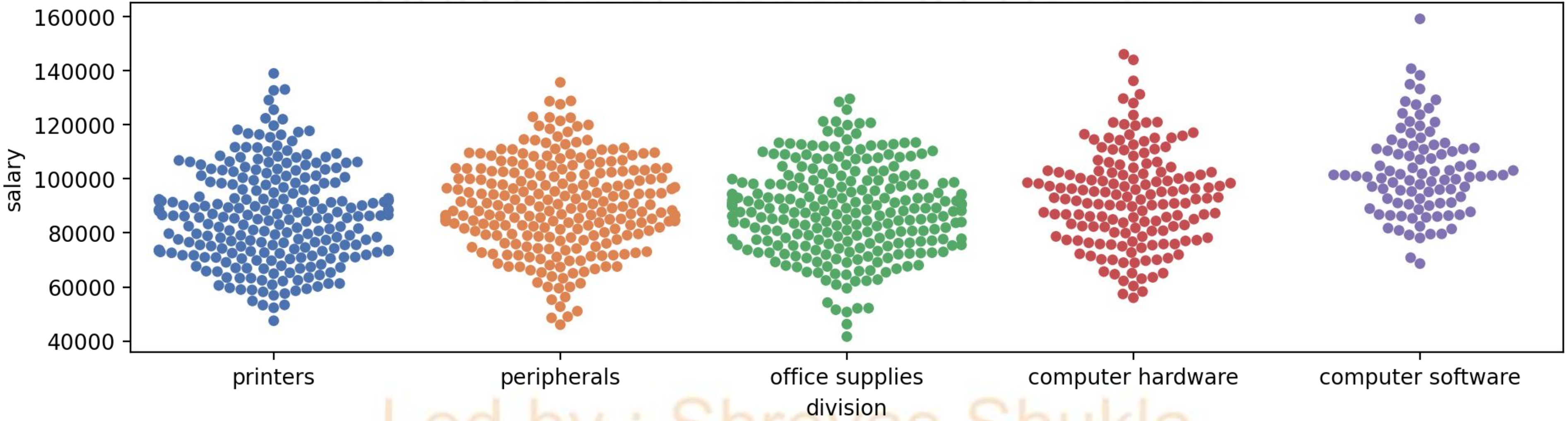
(For very large data sets, it won’t show all the points, but will display the general distribution of them.)

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)



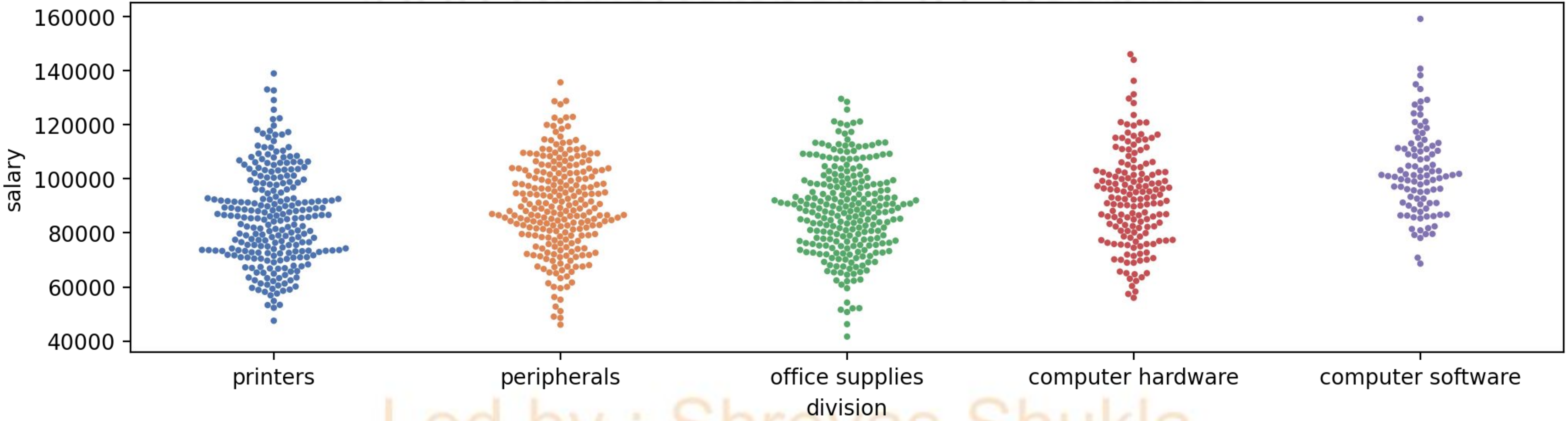
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)



Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

That is it!!

Let’s code !!