Exploratory Data Analysis - EDA

1. PDF
2. CDF
3. Univariate & Bivariate
4. Percentile & Quantile
5. Box plot
6. Multivariable Probalility, Contourplot

1. Random Variable:

- It means set of possible outcomes of an experiment.
- A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes.
- Types of Random Variable are -
- Discrete Random Variable
- Continuous Random Variable

1.1 Discrete Random Variable-

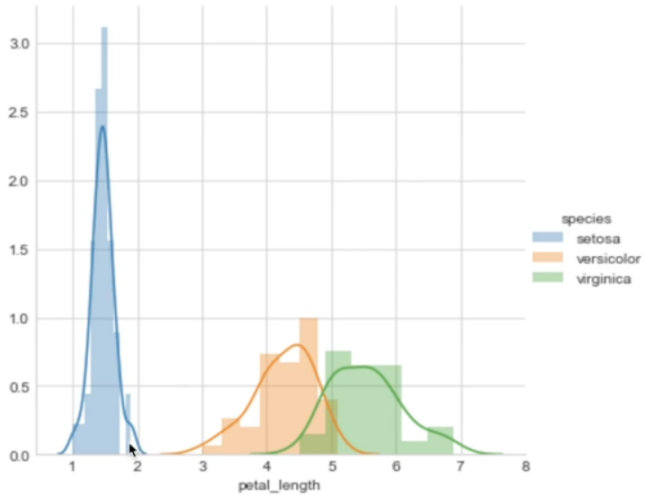
- Random Variable which has a finite set of outcomes or values is called a discrete random variable
- Eg - A random experiment of tossing a coin results in either heads or tails, Rolling of a dice

1.2 Continuous Random Variable-

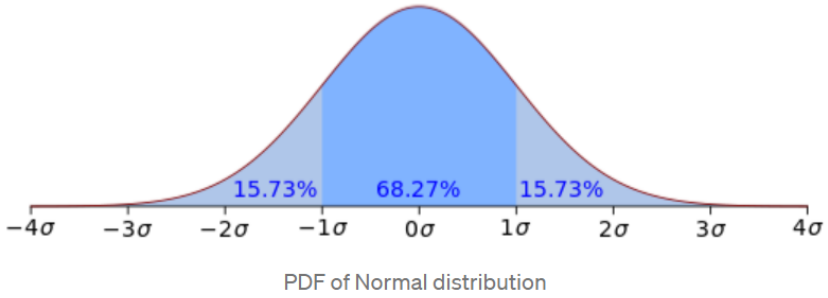
- Random variable which can take any real value is called a continuous random variable.
- Eg- The probability of height of a group of students

2. Probability Density Function - PDF

- It is smoothed form of Histogram
- Its a way of representing the range of possible values of a continuous random variable.
- Eg- If one wanted to calculate the probability that a temperature between 70-75 degrees will be reached
- PDF of a random variable is the plot between the random variable and the frequency of that random variable. I
- It gives the probability distribution of the random variable.



68–95–97 rule:

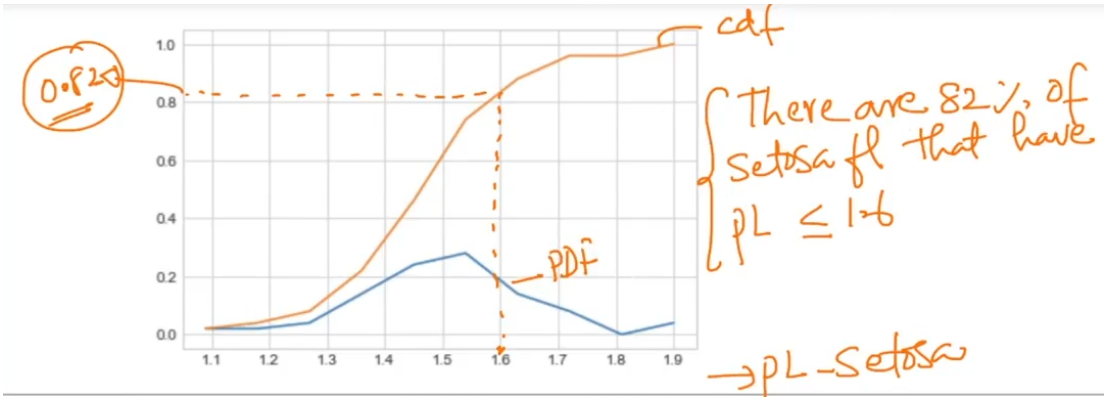


PDF of Normal distribution

- Another important insight from PDF is about the percentage spread of the data.
- If the data is normally distributed, about 68% of actual data lies in the range of (-1*standard_deviation,+1*standard_deviation).
- About 95% of data lies in the range of (-2*standard_deviation,+2*standard_deviation).
- About 97% of data lies in the range of (-3*standard_deviation, +3*standard_deviation)

2. Cummulative Distribution Function - CDF

- CDF is the probability that a random variable, X, will take a value less than or equal to x.
- CDF always lies between 0 and 1.



3. Percentile -

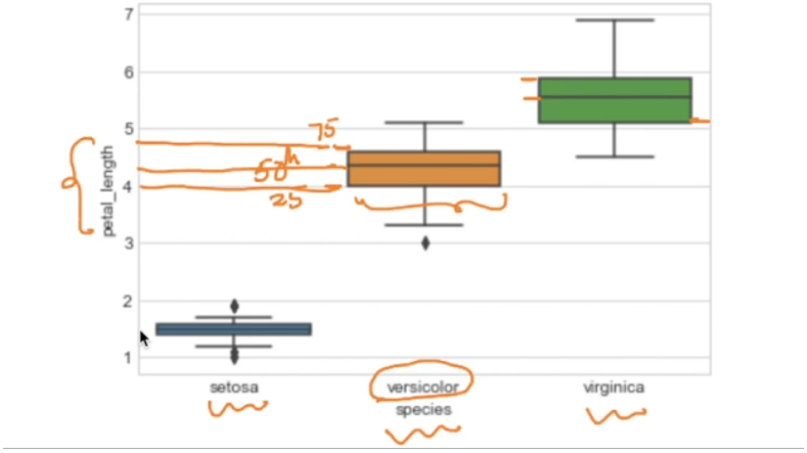
Percentile will gives us a number that describes the value that a given percent of the values are lower than.

4. Quantile -

- 25%, 50%, 75% and 100% are called quantile.

5. Box-Plot -

- A box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles of 25th , 50th and 75th Percentile value.
- Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles.
- There is no standard way of defining whiskers, one way is min and max value.



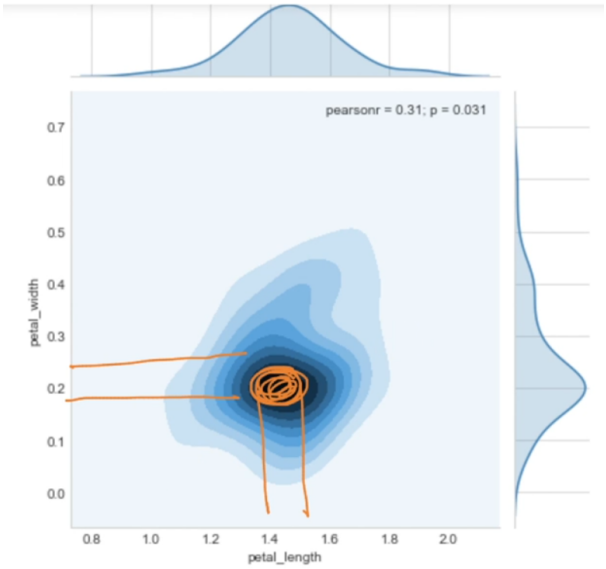
6. Violen Plot -

A violin plot is a method of plotting numeric data. It is similar to box plot with a rotated kernel density plot on each side.

- The darker region of Violen plot is also have 25th, 50th and 75th percentile.
- The side of this plot showing PDF.

7. Contour Plot -

- Contour plots (Level Plots) are a way to show a 3-D surface on a 2-D plane.
- Eg Used to show terrain of surface.



Blog

Refer - https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/ (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/)

Below are the steps involved to understand, clean and prepare your data for building your predictive model:

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

1. Variable Identification

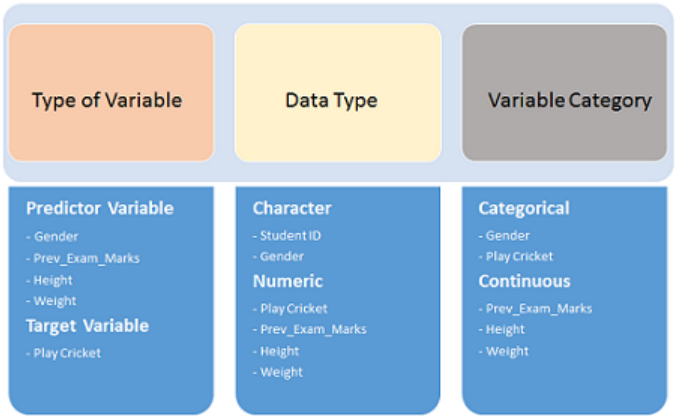
First, identify **Predictor** (Input) and **Target** (output) variables. Next, identify the data type and category of the variables.

Let's understand this step more clearly by taking an example.

Example:- Suppose, we want to predict, whether the students will play cricket or not. Here you need to identify predictor variables, target variable, data type of variables and category of variables.

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0

Below, the variables have been defined in different category:



2. Univariate analysis:-

It provides summary statistics for each field in the raw data set (or) summary only on one variable.

- Eg- CDF, PDF, Box plot, Violin plot.

3.Bivariate analysis:-

Performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding realtionship between them.

- Eg- Box plot,Voilin plot.

4.Multivariate analysis:-

It is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2.

- Eg- Pair plot and 3D scatter plot.