Machine learning Udemy

Python Basics

Pandas: Pandas is an open-source Python Library providing high-performance data loading ,manipulation and analysis tool using its powerful data structures.

Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyze.

Introduce Data frames and series that allow you to slice and dice rows and columns of information.

NumPy array :

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

Operations using NumPy

Using NumPy, a developer can perform the following operations −

* Mathematical and logical operations on arrays.
* Fourier transforms and routines for shape manipulation.
* Operations related to linear algebra. NumPy has in-built functions for linear algebra and random number generation.

Scipy

**SciPy** is a library that **uses NumPy** for more mathematical functions. **SciPy uses NumPy** arrays as the basic data structure, and comes with modules for various commonly **used** tasks in scientific programming, including linear algebra, integration (calculus), ordinary differential equation solving, and signal processing.

Scikit learn :

The machine learning library we'll use throughout this course is scikit\_learn, or sklearn, , and it generally takes Numpy array as its input

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines.

So, a typical thing to do is to load , clean , and manipulate your input data using pandas . Then convert your pandas dataframe into a NumPy array as its passed into some Scikit\_Learn function . That conversion can happen automatically

from scipy import stats

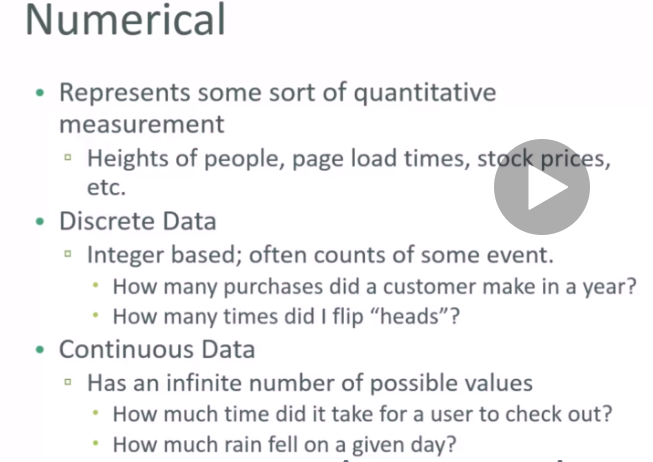
stats. Describe(NumPy array)

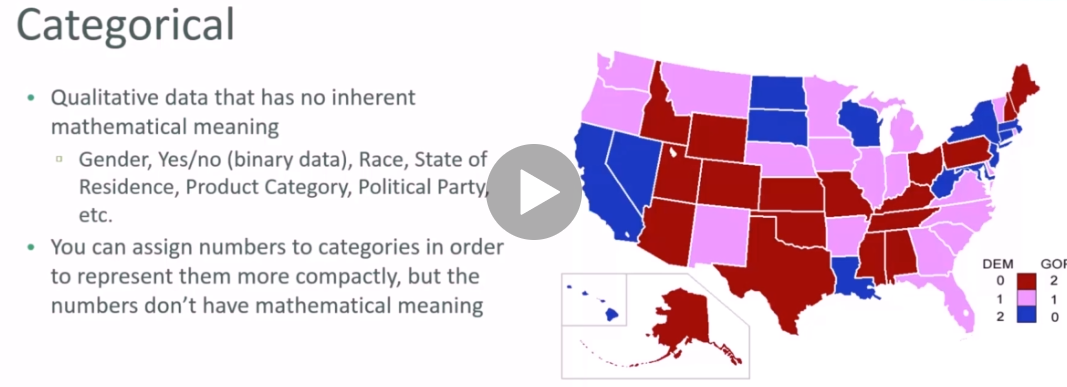
Types of Data:

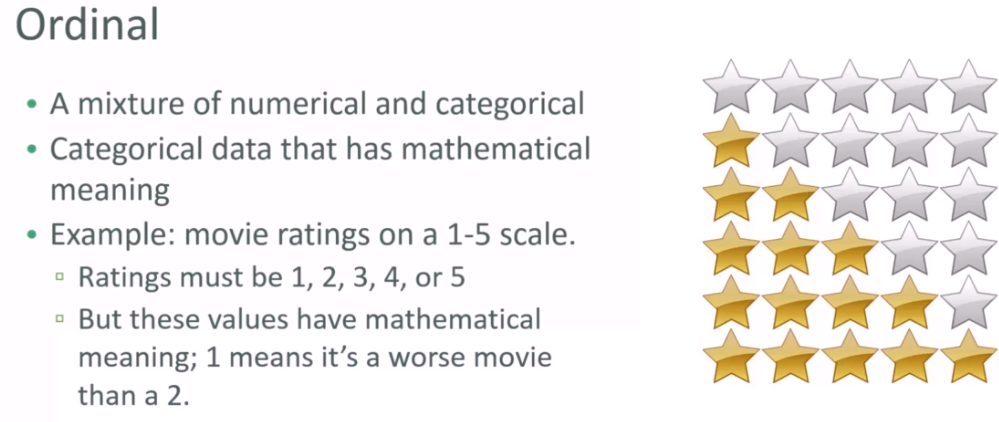
Numerical

Categorical

Ordinal







Quiz:

Are the following types of data numerical, categorical , or ordinal?

1. How much gas in your gas tank?

* Numerical and continuous

1. A rating of your overall health where the choice are 1,2,3, or 4 , corresponding to

‘poor’, ‘moderate’, ‘good’ , and ‘Excellent’

* Ordinal

3)The race of your classmates

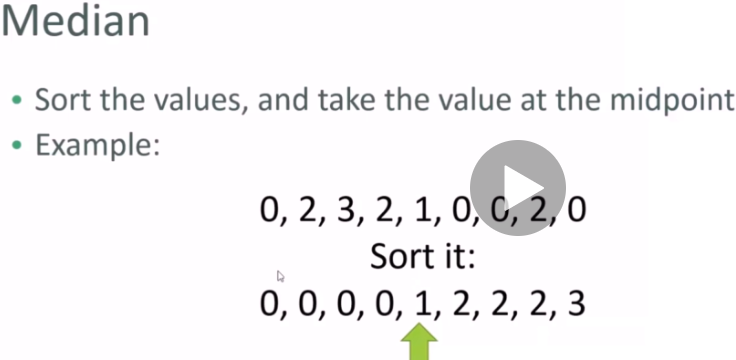
* Categorical

4)Ages in year

* Numerical

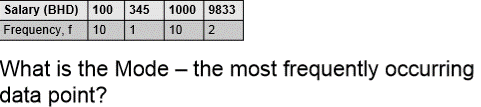
5) Money spent in a store

* Numerical continuous









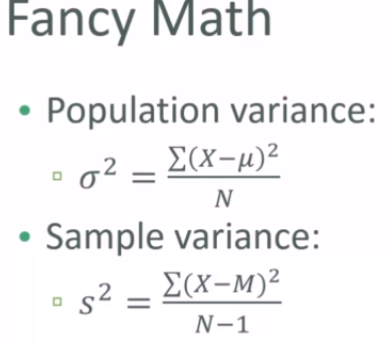
**Standard deviation and variance :**

Variance measures how “spread-out ” the data is.

Variance is simply is average of the squared difference of each data point from the mean.

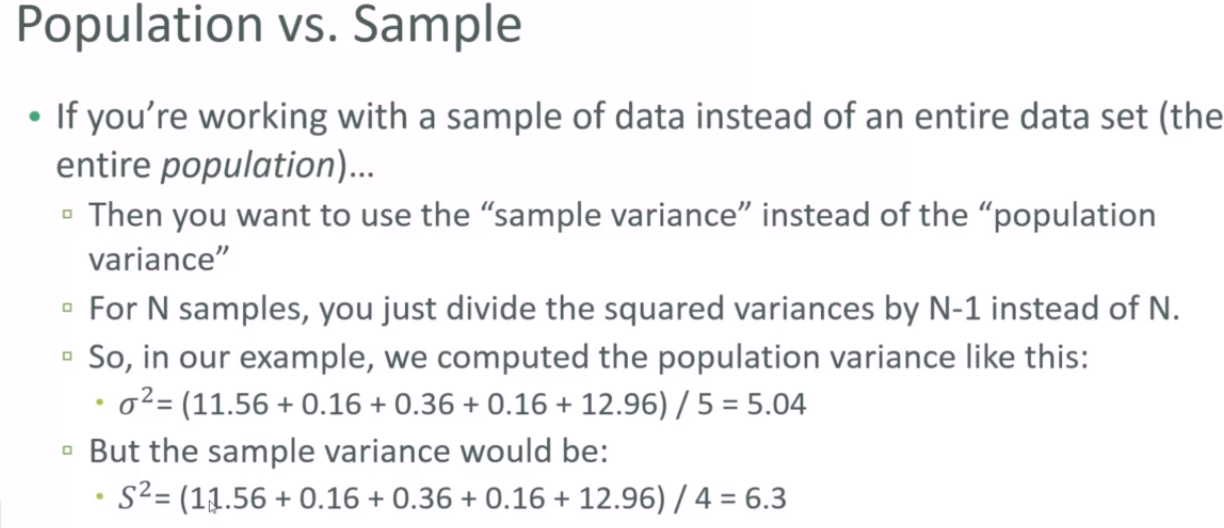




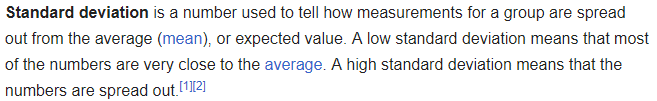


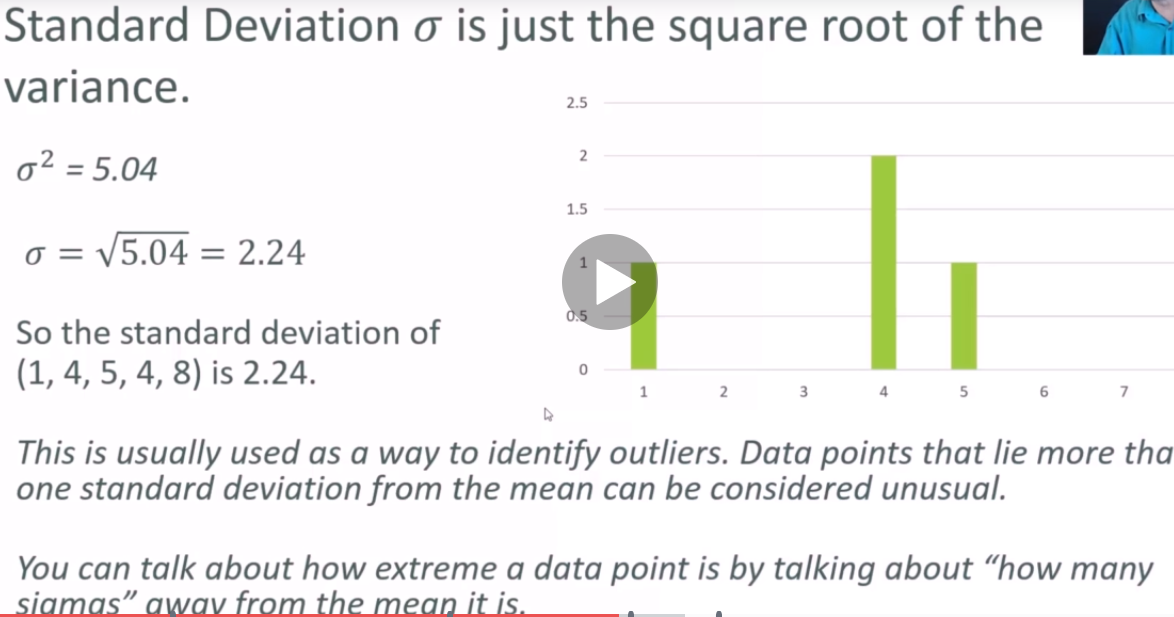
**Reasons for using 'N-1' samples**

This is called Bessel's correction and it is required for any sample statistics to prevent bias.



**Standard deviation**

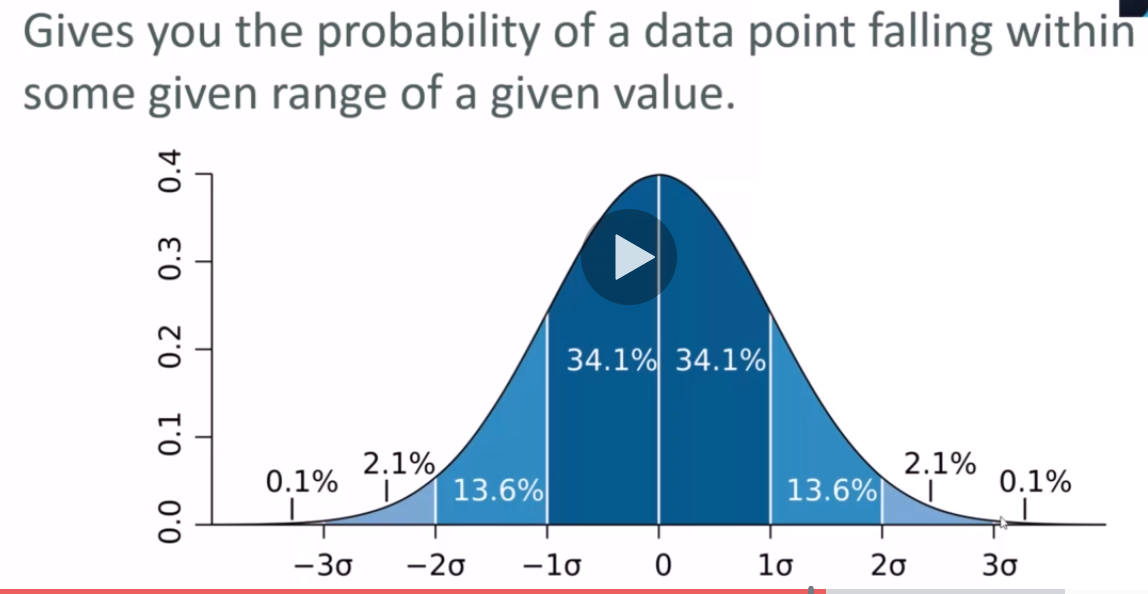




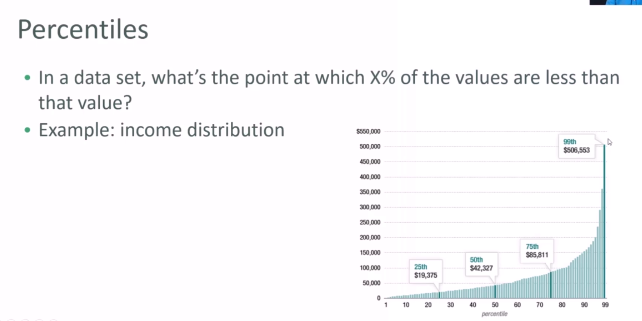
Probability Density function and Probability mass function

Probability Density function gives probability of a data continues point falling with in the given range of values (mean and std)

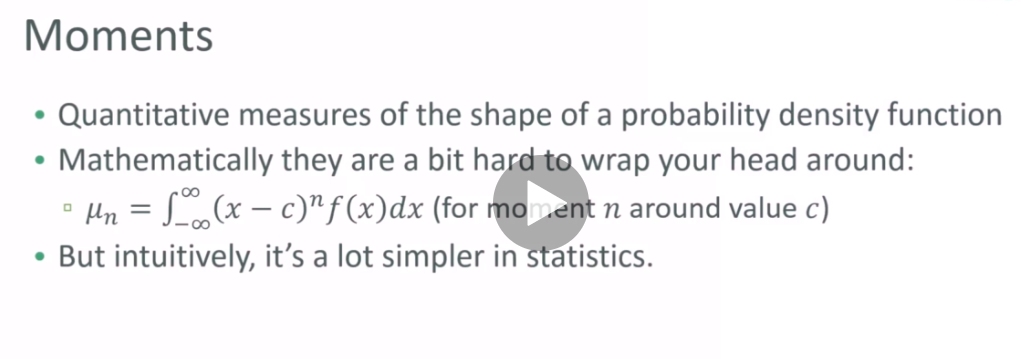
Here you can say that probability of falling some value between mean and 1 standard deviation is 34.1.%



Percentiles and Moments



We can find the data point at 50% and say that 50% of the datapoint from the data are less than that

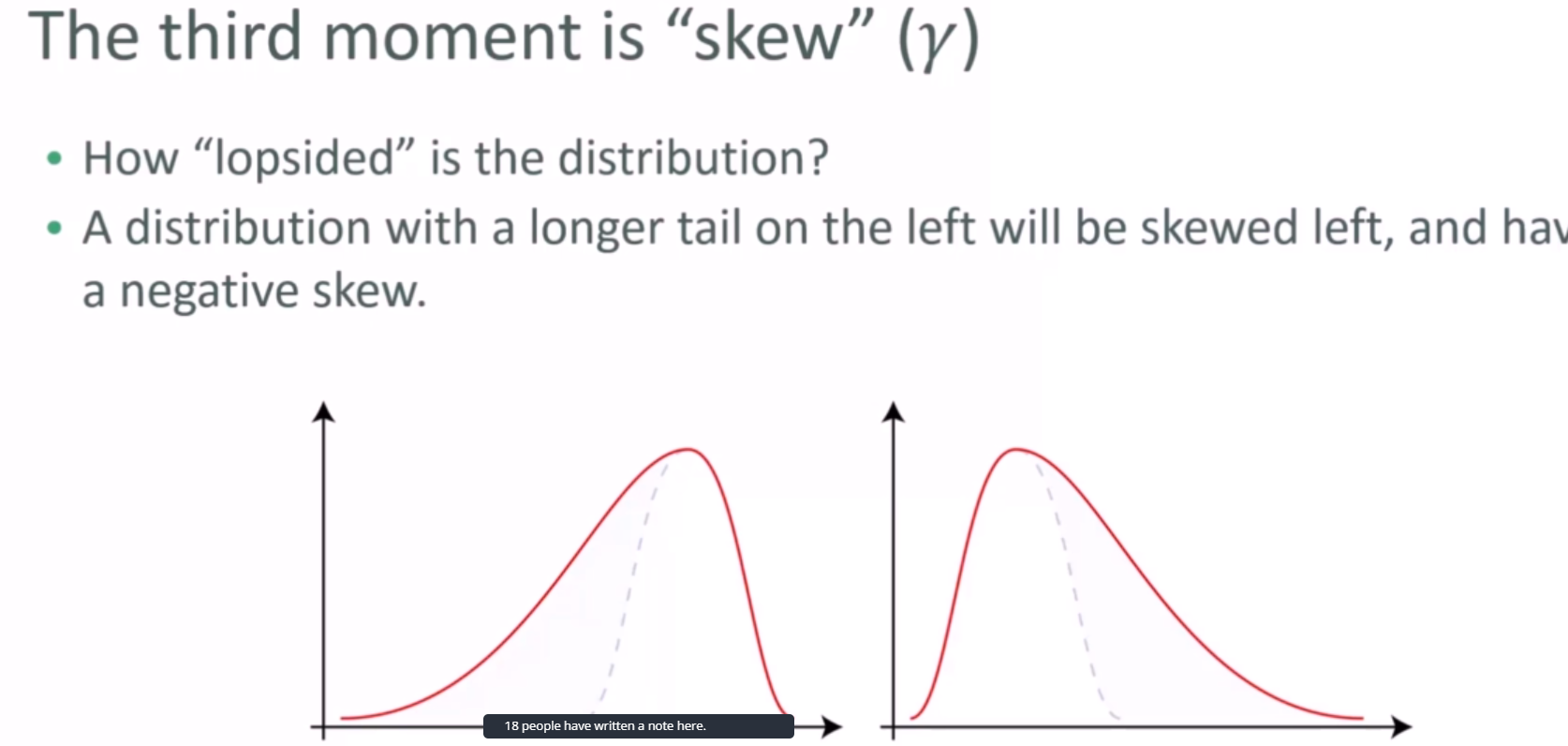


1)First moment is the mean

2) Second moment is the variance

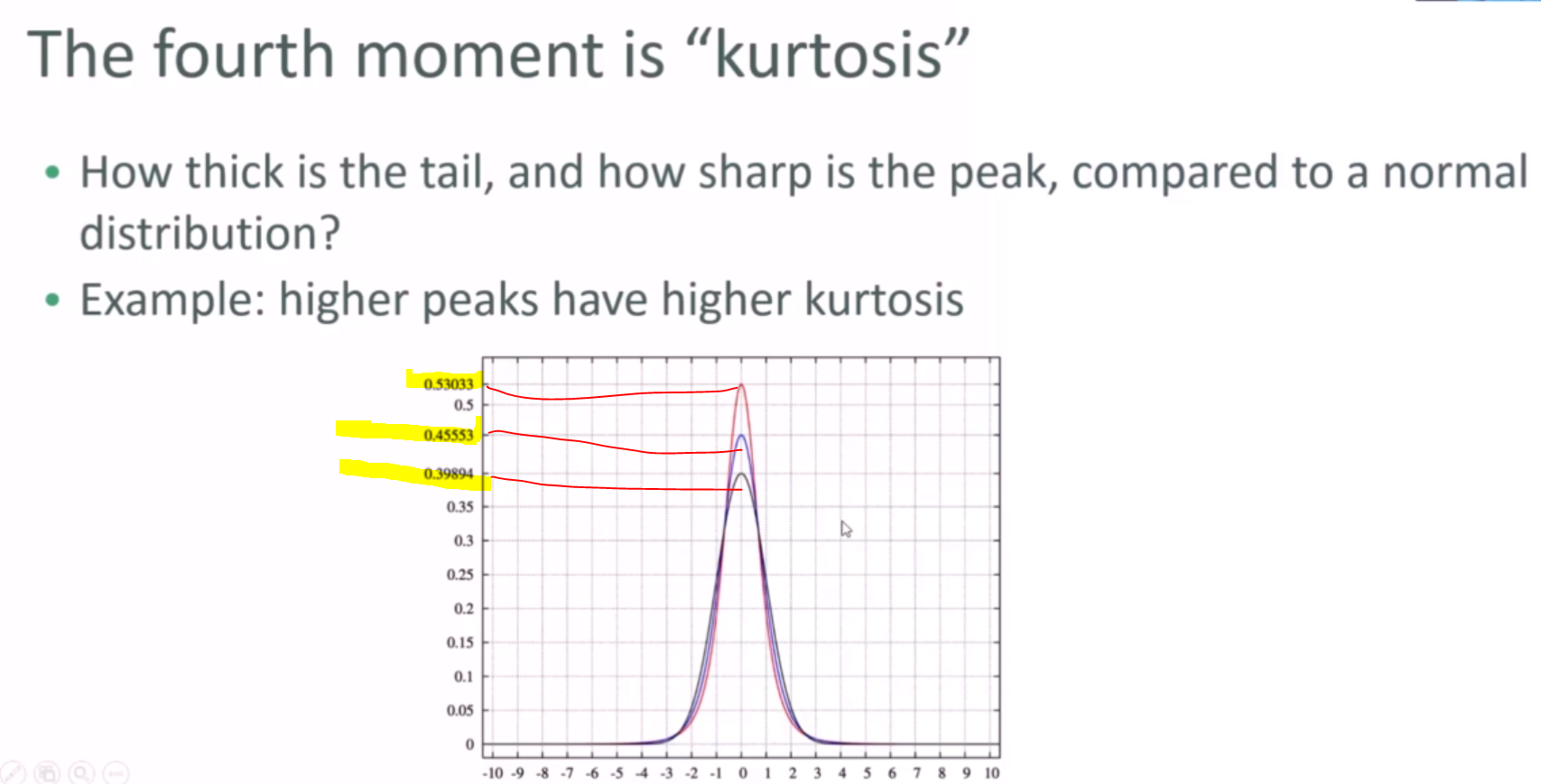
3) Third moment is ‘Skew’

Stretching of tail in one side or another side



Skewness refers to distortion or asymmetry in a symmetrical bell curve, or normal distribution, in a set of data. If the curve **is** shifted to the left or to the right, it **is** said to be **skewed**.

1. Fourth moments is ‘kurtosis’



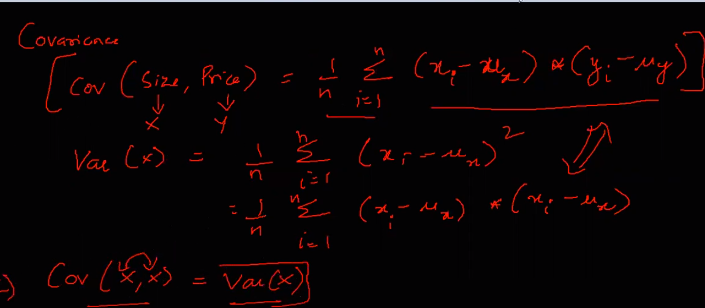
Covariance and correlation :

Covariance : This gives what is the relation between two attributes

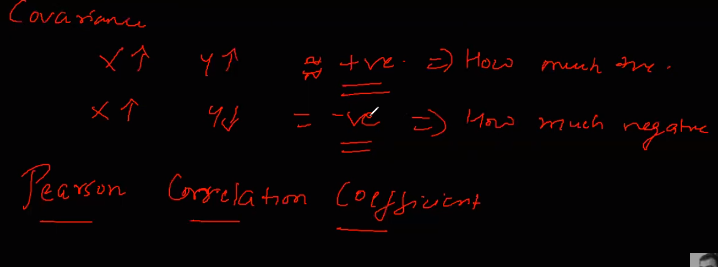
+ve / -ve related

Use np.cov for calculating covariance

* If both attributes are varying in same direction, then covariance will be +ve
* If both attributes are varying in opposite direction, then covariance will be -ve



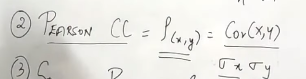
But it doesn’t tell how much they are correlated and this we use Pearson correlation coeffient.



Correlation coefficient :

Use np.corrcoef for calculating covariance

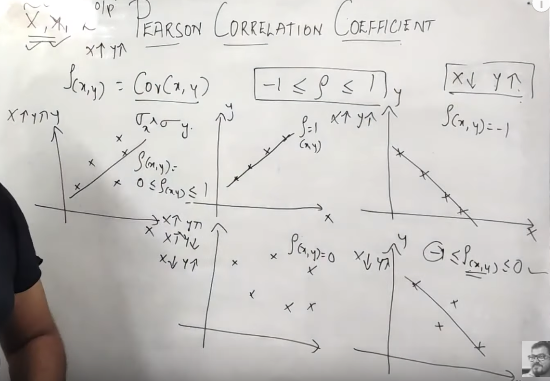
Covariance gives only direction like positively or negatively related to each other, but it does not give magnitude such that how much they are related (strong or week correlation in any direction .

 Covariance of x and y/ standard deviation x and y

Pearson correlation coeffect : Gives both magnitude and direction of how attributes are related .

* Strength – How much they are related value range (-1 to +1) 0: no relation .
* Direction relationship :

This is used for feature selection . Let’s consider we have x,y,z features x,y(independent ) and z is (Dependent ) if correlation between x and y is 1 then it mean both x and y are same and we can select any of x or y for creating machine earning algorithm .

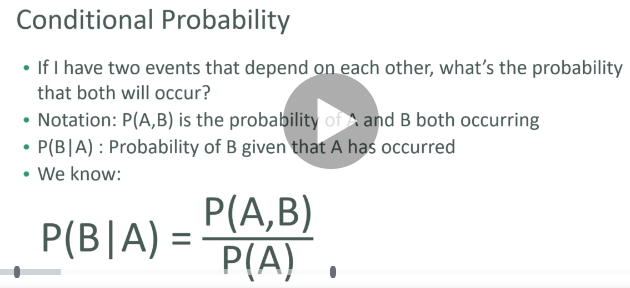


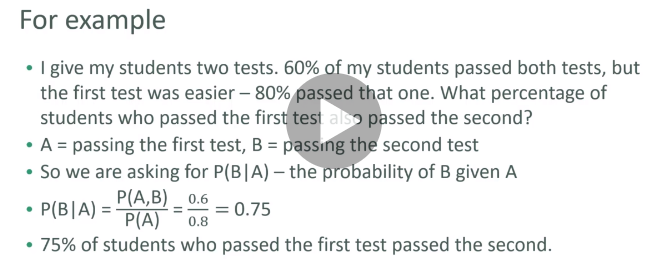
Probability:

The ratio of the number of outcomes favorable for the event to the total number of possible outcomes is termed as **probability**. In other **words**, a measure of the likelihood of an event (or measure of chance) is called **probability**. ... Sample space is the possible outcomes of the experiment.

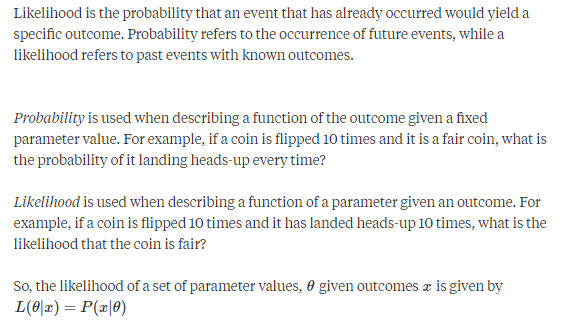
Rolling dice getting any one face is : 1 /6

Condition probability :

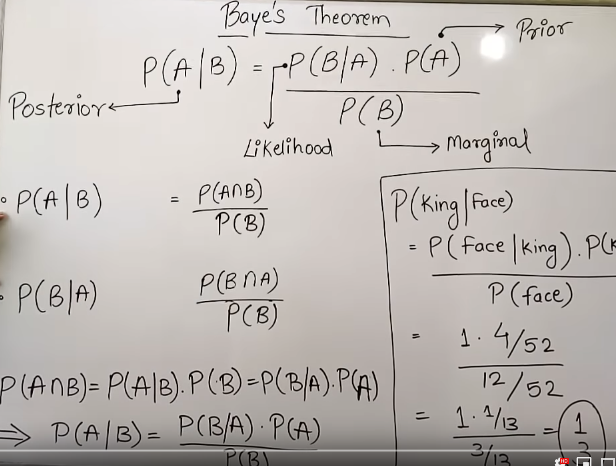




Bayes’s theorem :



A is hypothesis and B is already occurred

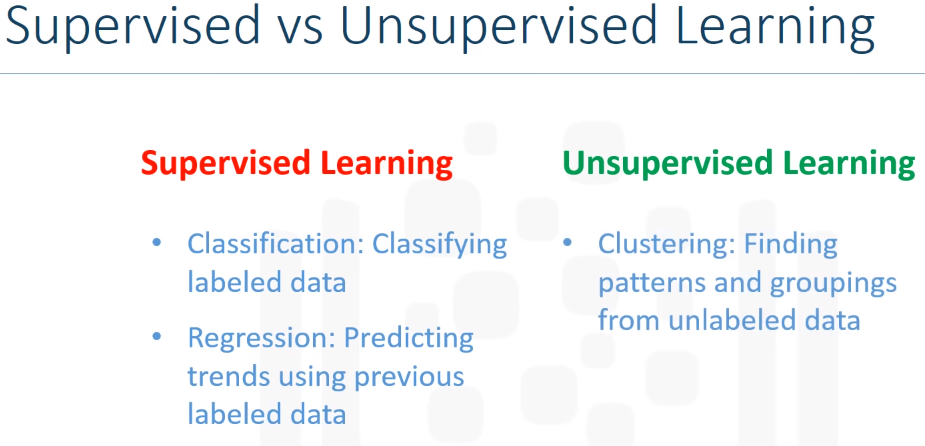


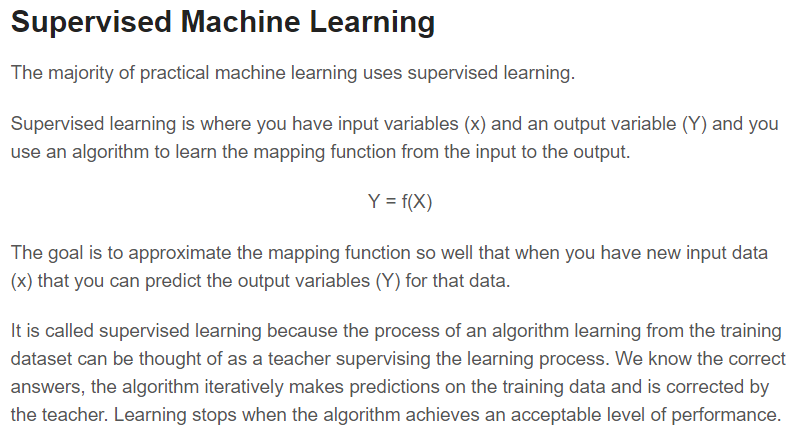
Machine learning :

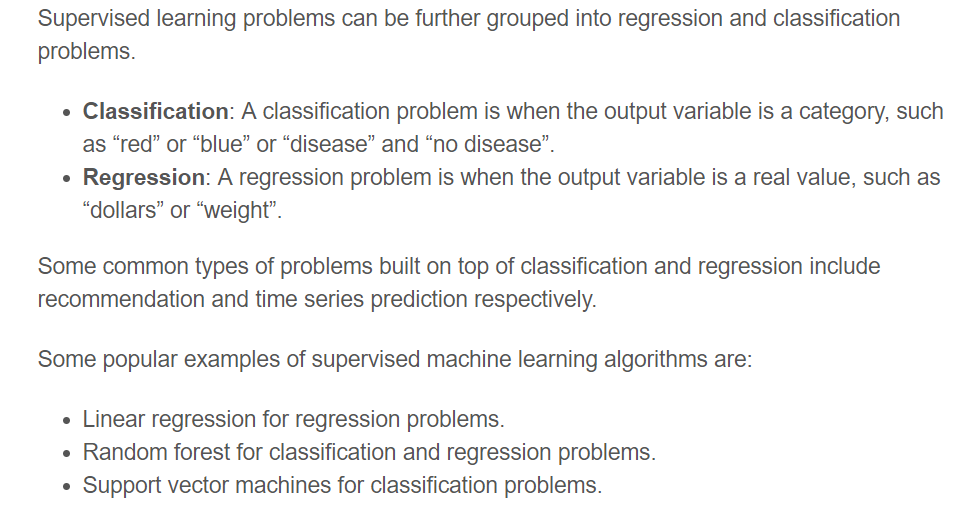
We need to do two things with this Data:

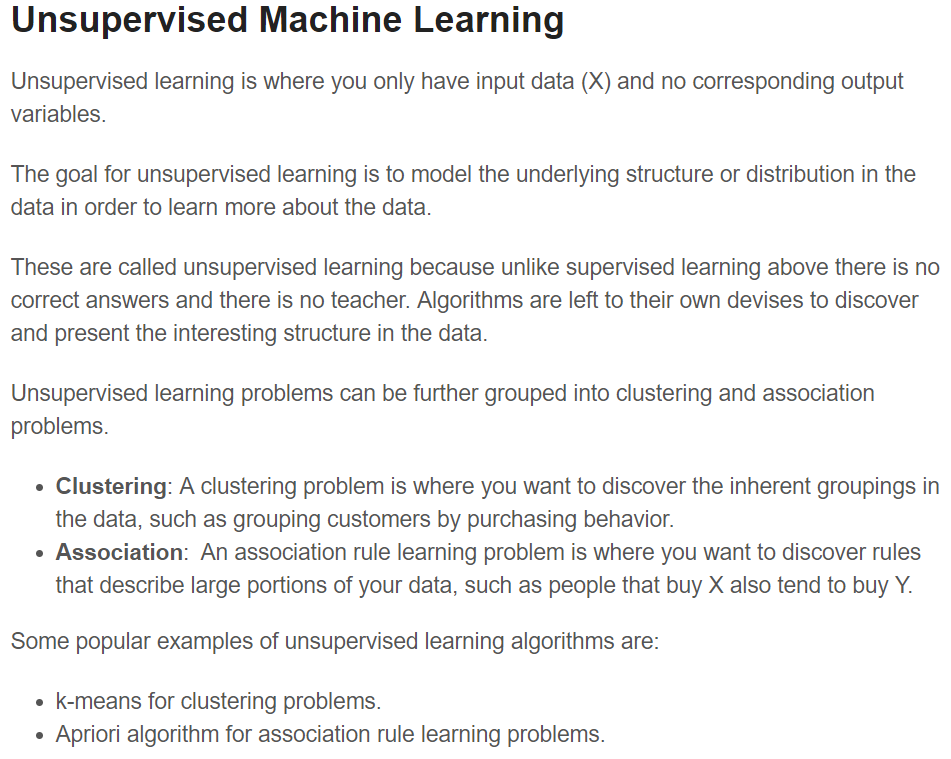
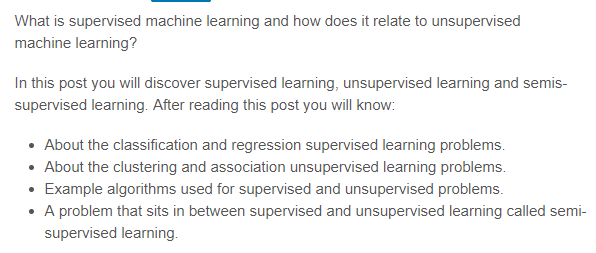
1)Estimate the parameter for the machine learning methods: In machine learning lingo, estimating parameter is called "Training the algorithm".

2)Evaluate how well the machine learning methods work: In machine learning lingo, Evaluating method is called "Testing the algorithm".







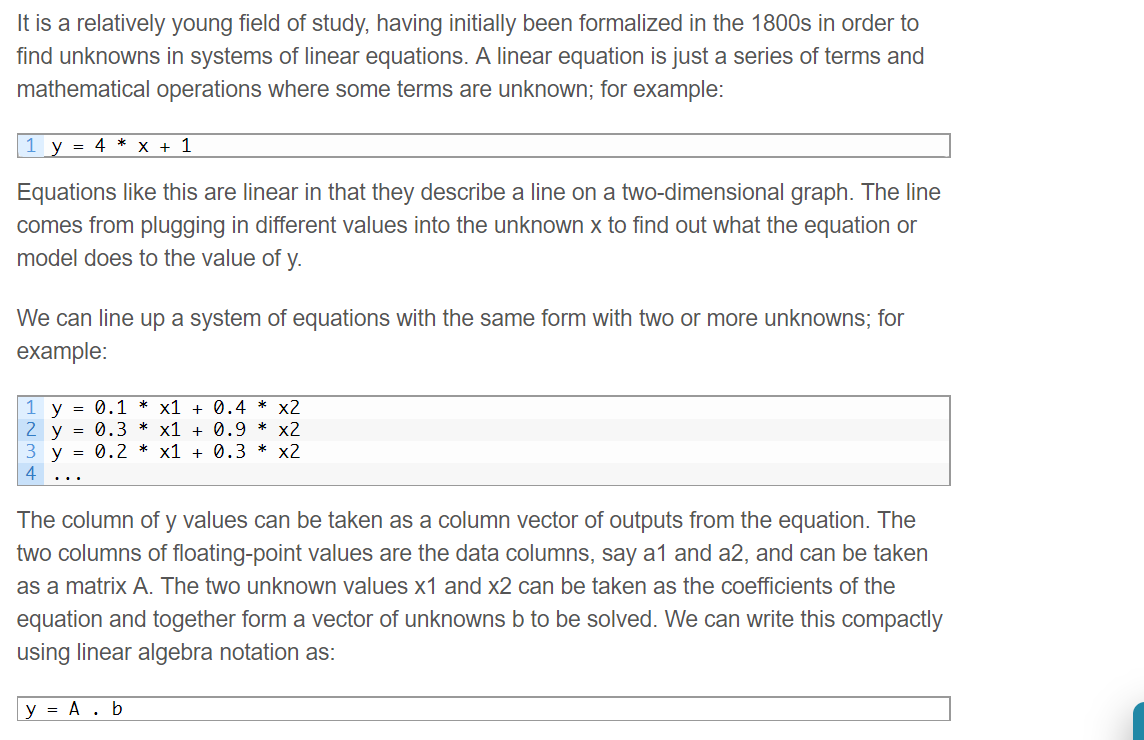


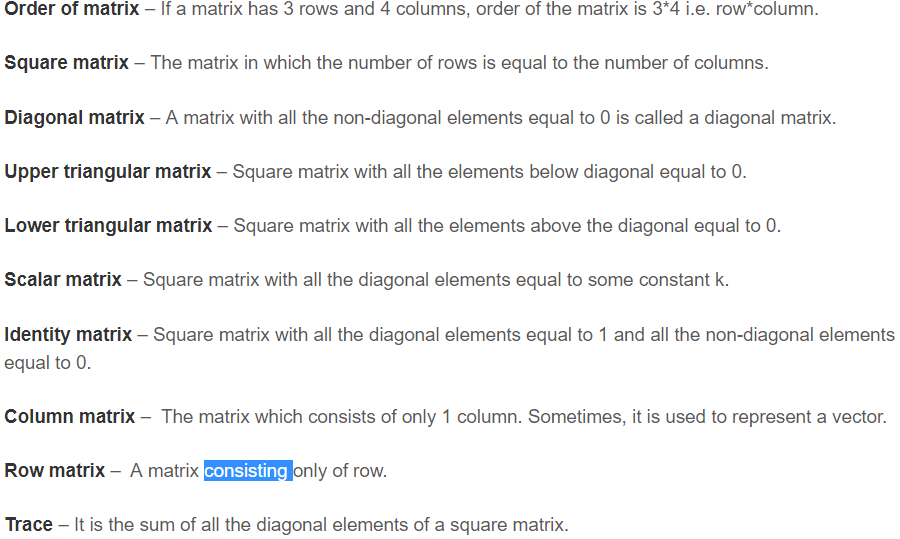
Linear regression :

Linear: arranged in or extending along a straight or nearly straight line.

Eg:"linear movement"

Graphing Linear and Non-linear Functions. ... The word 'linear' means something having to do with a line. On a Cartesian Plane, a linear function is a function where the graph is a straight line. The line can go in any direction, but it's always a straight line. A non-linear function has a shape that is not a straight line



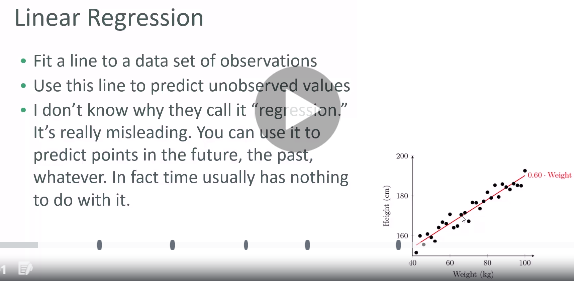


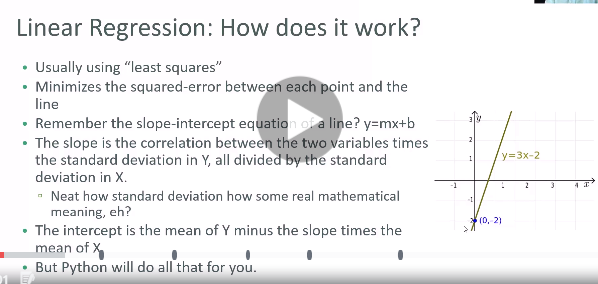
Victor: Are two dimentional line passing through origin having x and y values

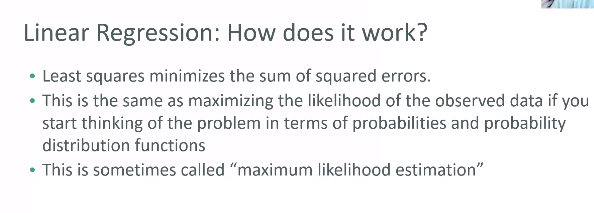
Every victor represents a pair of numbers and 3D victor is represnted by x,y and z values

Each vector defines a pair of values

Victor scalar multiplication is multipling both variable with same number to strech or skrech the victor line

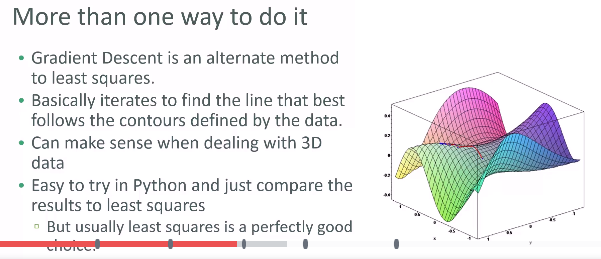


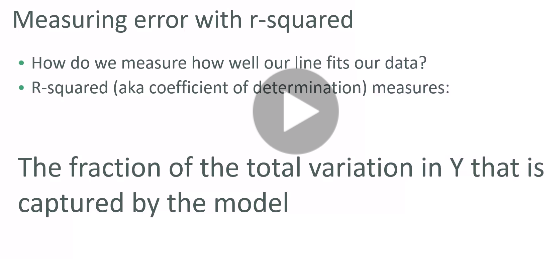


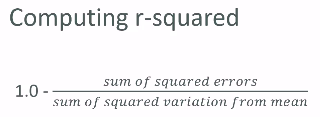


contour mean line

Gradient decent used for higher dimensional data





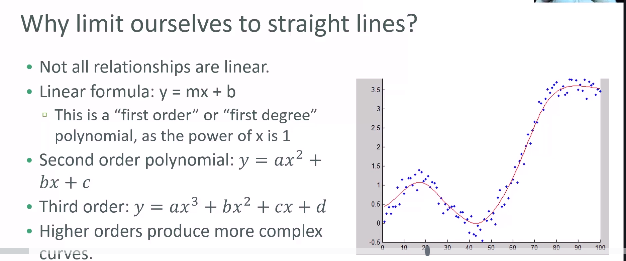




R-square show’s amount of variance captured by your model

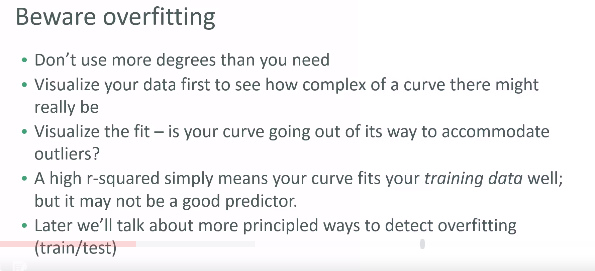
You can use R-square to do quantitative measure how good is your regression is for set of given data.

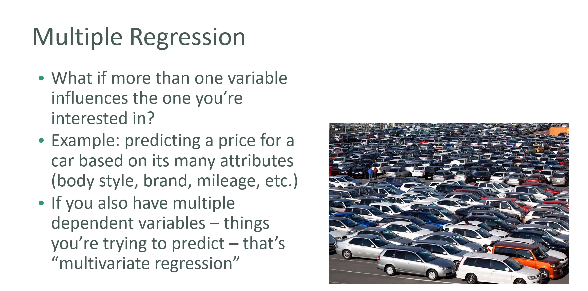
Polynomial regression :

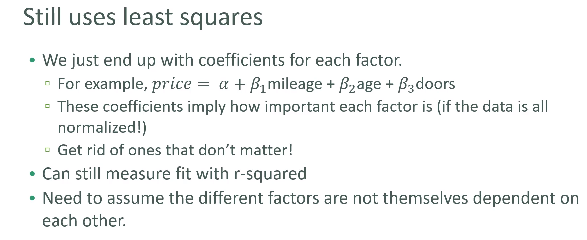


Higher the order of polynomial more complex curve and more complex relationship you will get

Fitting curve to higher degree is not always good it tends to do overfitting







Data visualization using Matplotlib

Outliers detection :

We have mean value , standard deviation value and data points in the data set .

Any data point which which lie outside this mean and standard deviation range is considered as unusual.