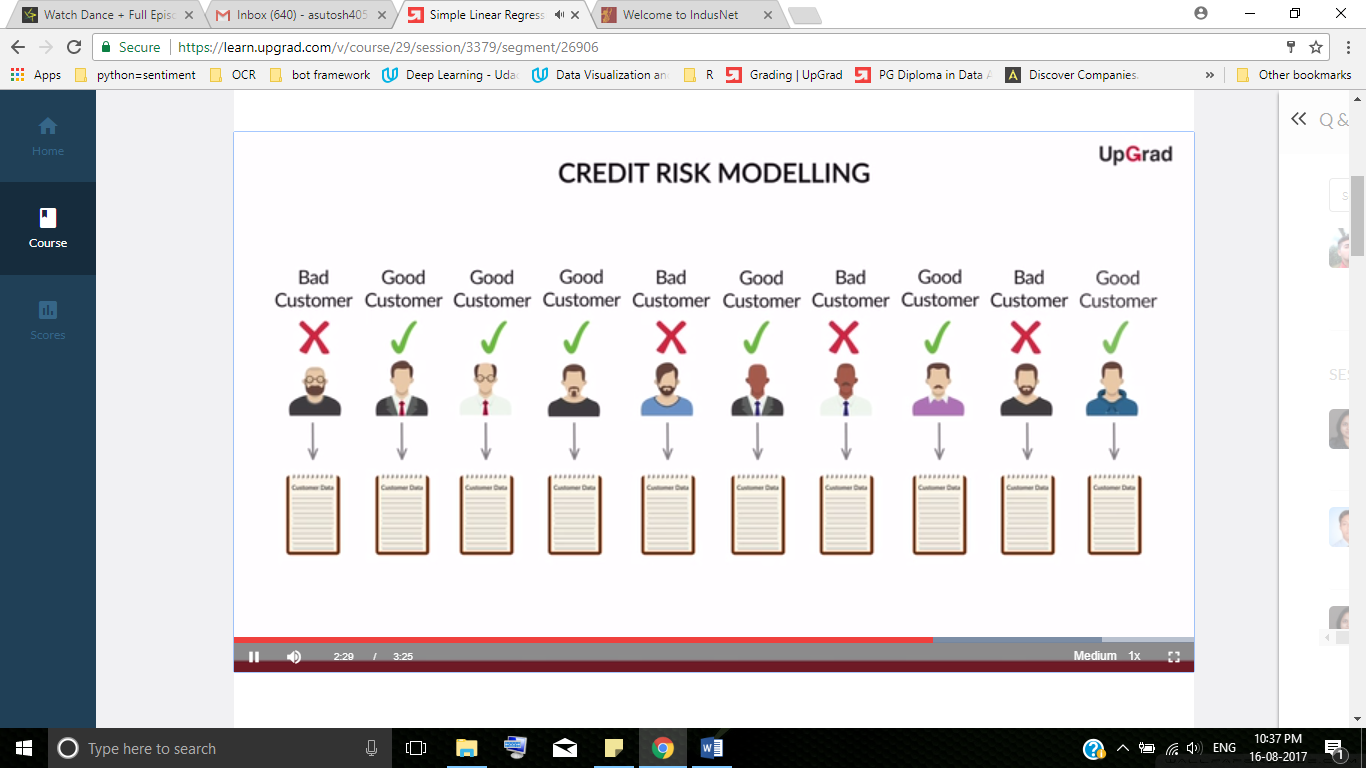
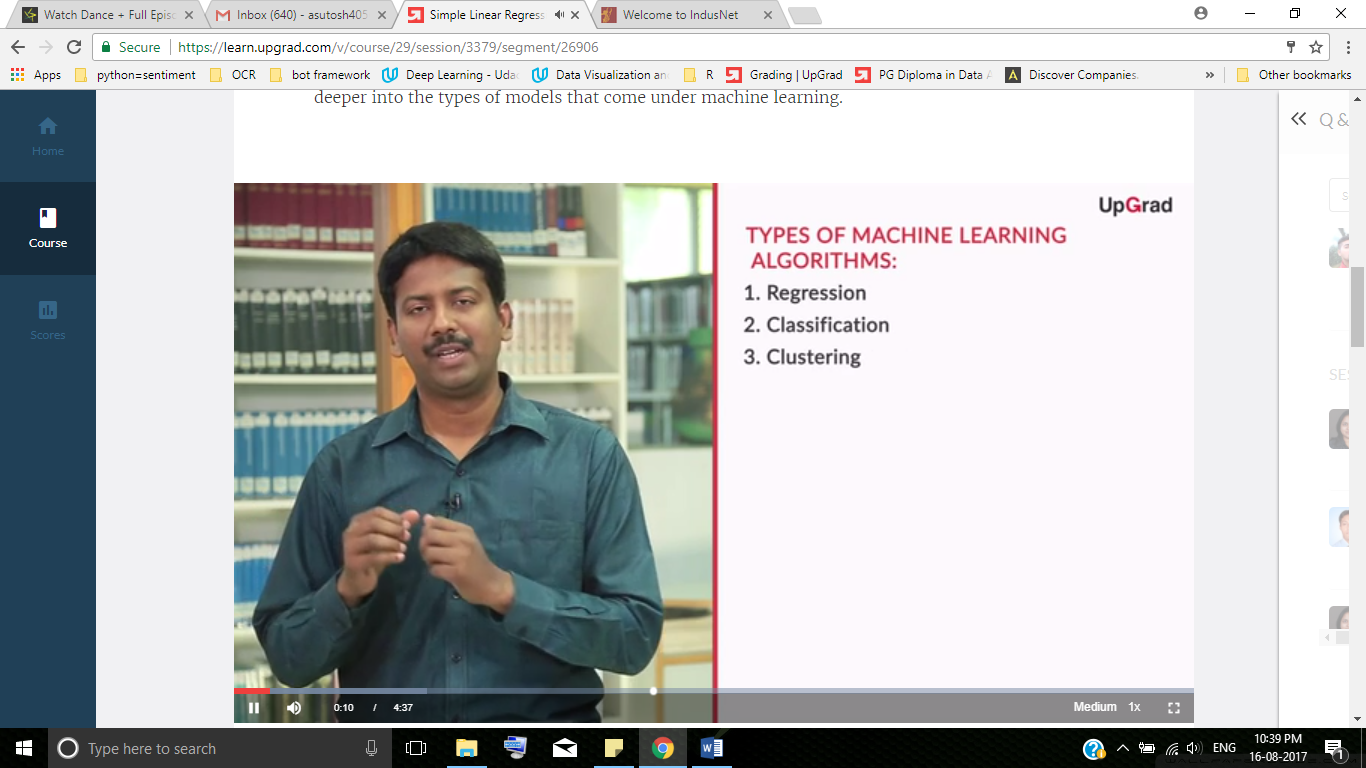
Modelling uses machine learning algorithms, in which the machine learns from the data just like humans learn from their experiences.

Machine Learning Example:

1. Credit Risk Modelling





Regression:

1. Predicting the prices or volume
2. Output value to be predicted is a continuous or a numeric variable

Classification Problem:

1. Email Spam and Email ham prediction
2. Output variable o be predicted is a categorical variable

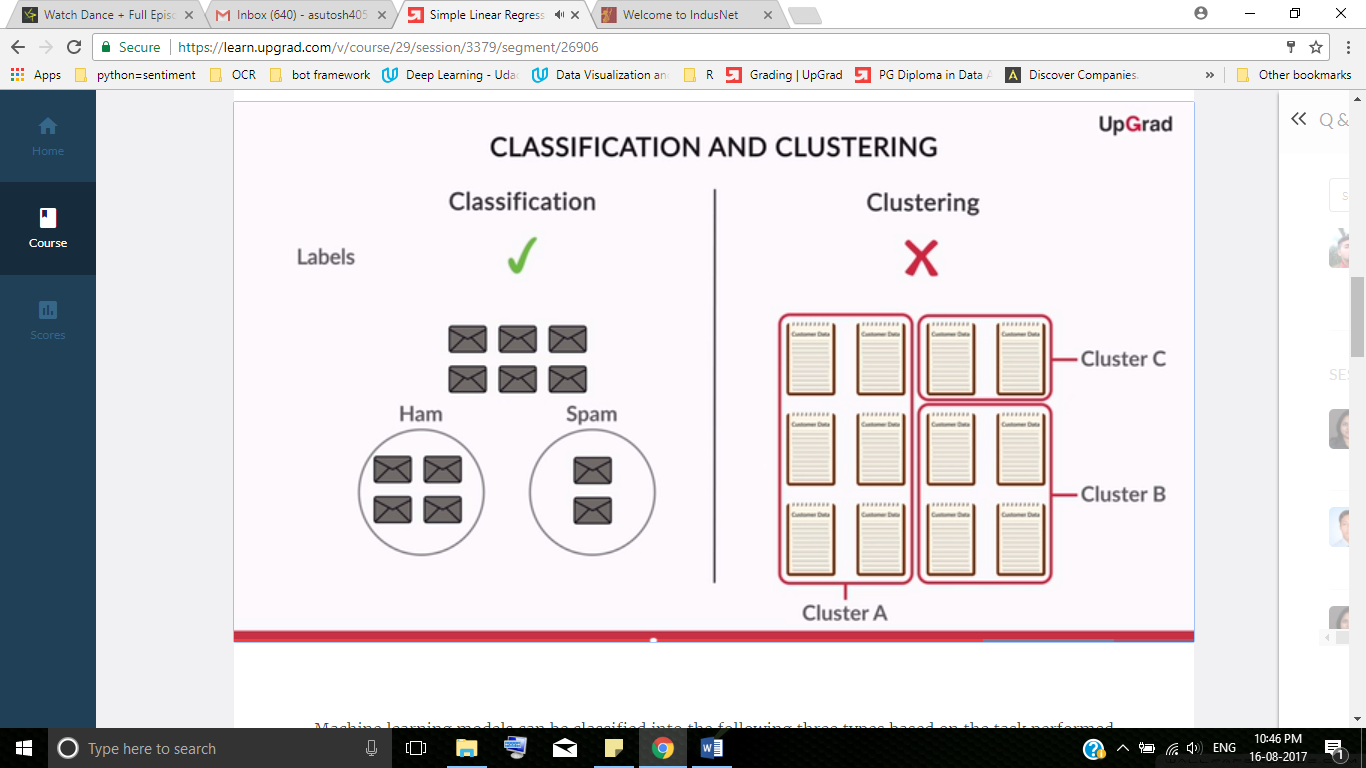
During credit risk analytics, you need to predict which customers would default on their credit card payments and which customers won’t. What type of model will be used to solve this problem?

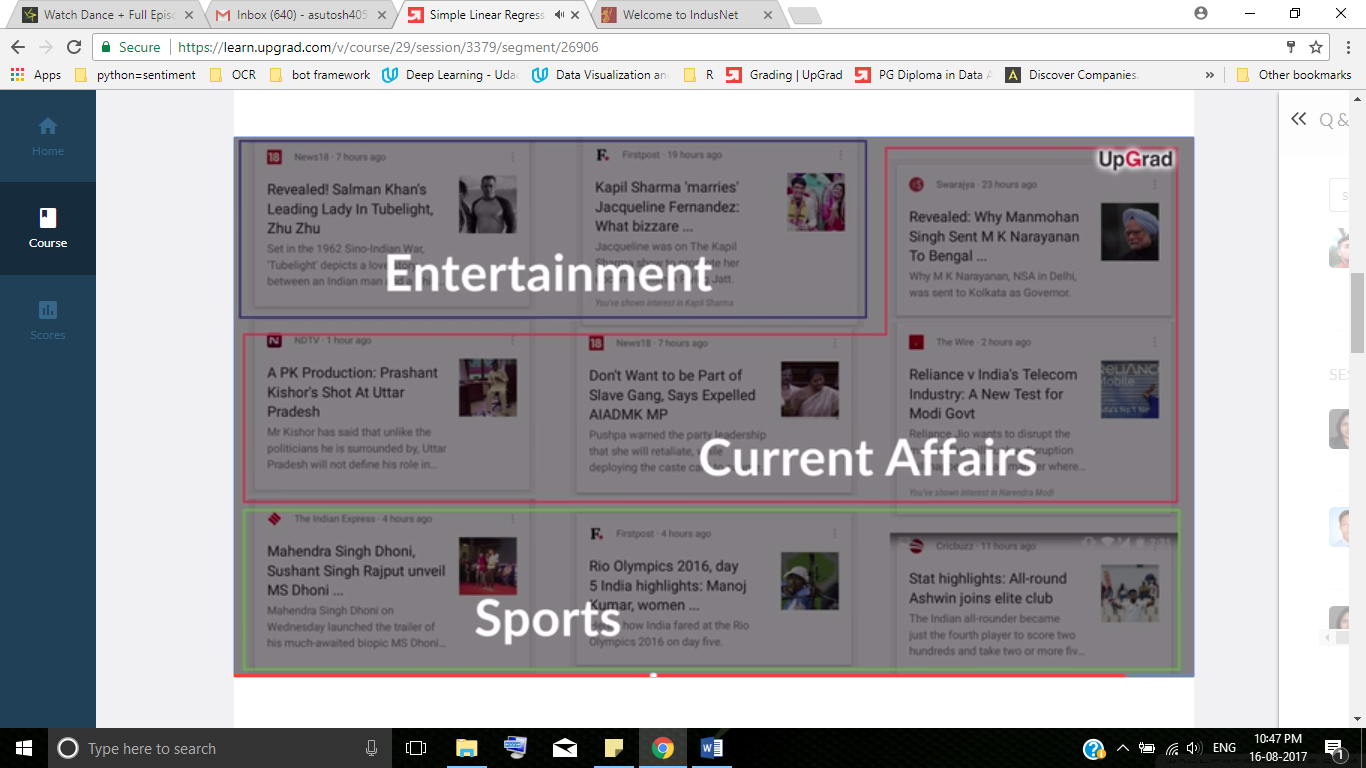
* Regression
* Classification

**Clustering**

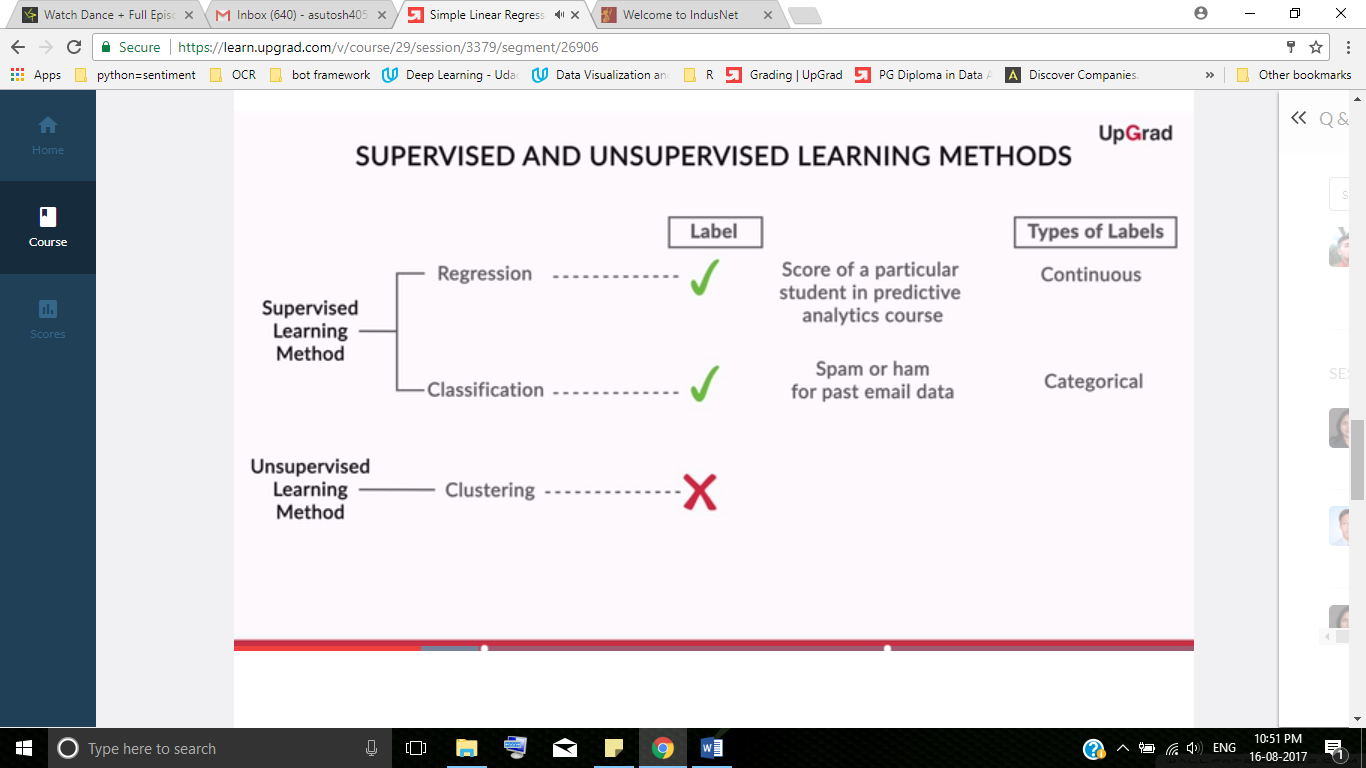


Classification vs clustering



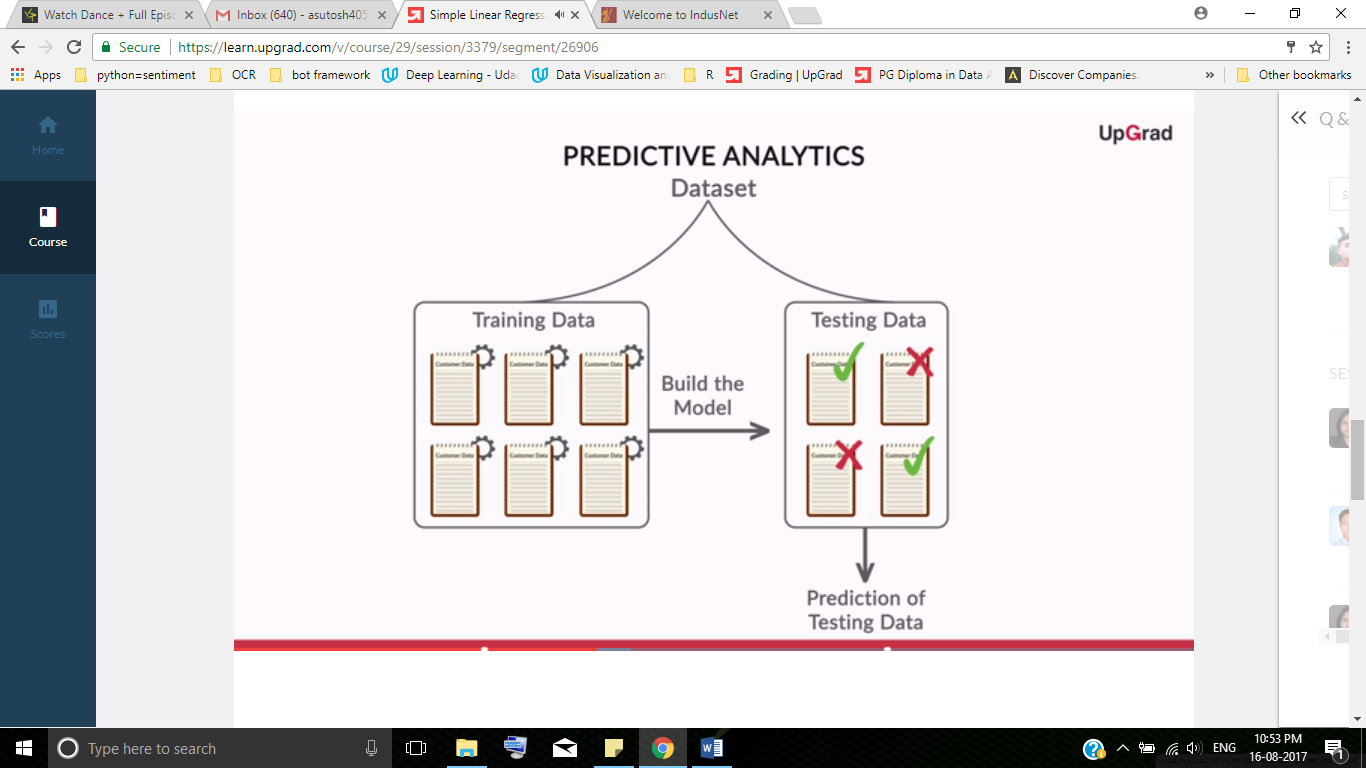


**supervised and unsupervised learning methods**



Dataset is divided into 2 parts. Namely,

* Training data
* Test Data



Consider the spam/ham example. Do you think the algorithm will/will not be able to predict if a given mail is spam or not, if the training data set contains data of only the ham mails?

Yes

No

Why do you think the algorithm will/will not be able to predict if a given mail is spam or not, if the training data set contains data on only the ham mails?

*The training data set should be abundant and diverse in nature for model building. Otherwise, your model won’t be trained properly since your model has not seen the spam. Therefore, it hasn't learnt about the characteristics of spam, so it won't be able to tell spam apart from ham. So here, your algorithm will not be able to predict very accurately.*

So, you can classify the machine learning models into two broad categories as follows:

1. **Supervised learning methods**
   1. Past data with labels is used for building the model
   2. Regression and classification algorithms fall under this category
   3. The past data is divided into training and testing data sets for building the model
2. **Unsupervised learning methods**
   1. No pre-defined labels are assigned to past data
   2. Clustering algorithms fall under this category

<http://www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm>

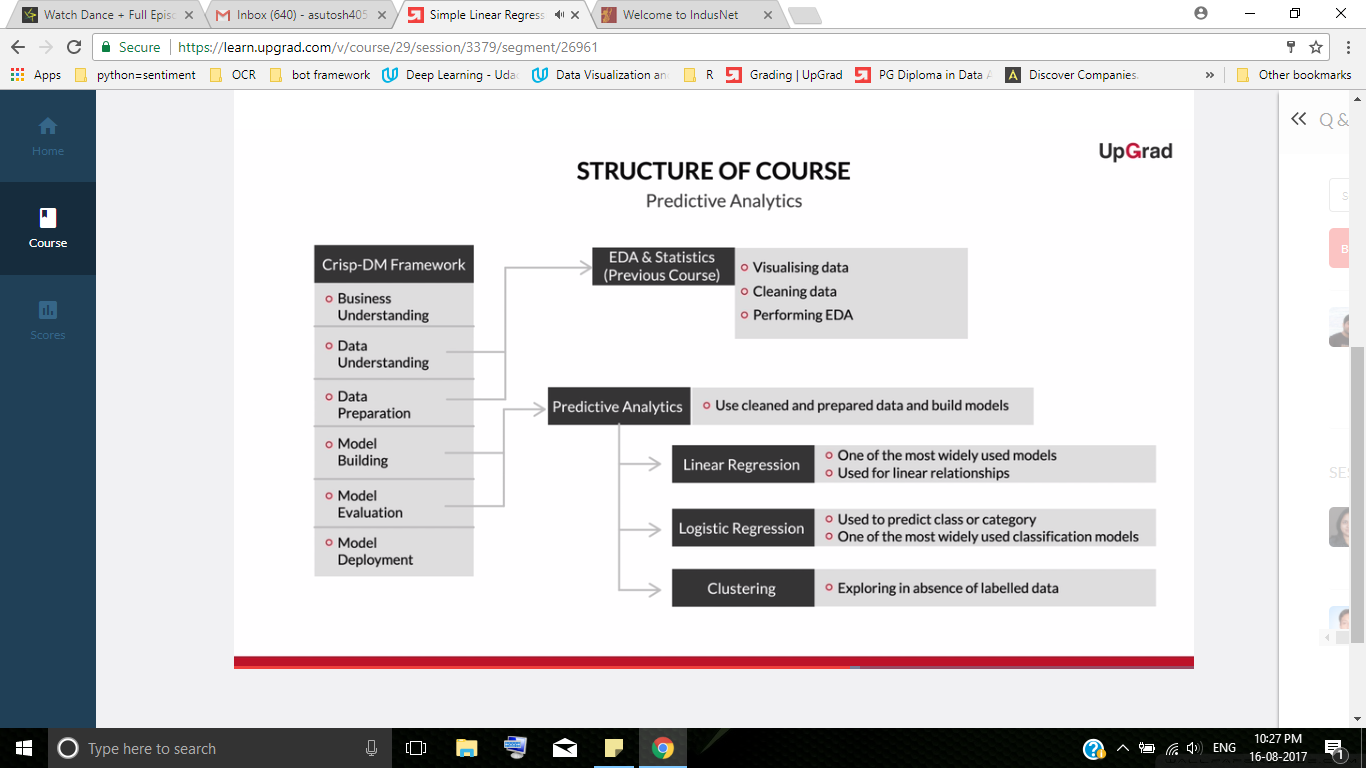
Supervised learning is the most common technique for training neural networks and decision trees. Both of these techniques are highly dependent on the information given by the pre-determined classifications. In the case of neural networks, the classification is used to determine the error of the network and then adjust the network to minimize it, and in decision trees, the classifications are used to determine what attributes provide the most information that can be used to solve the classification puzzle.

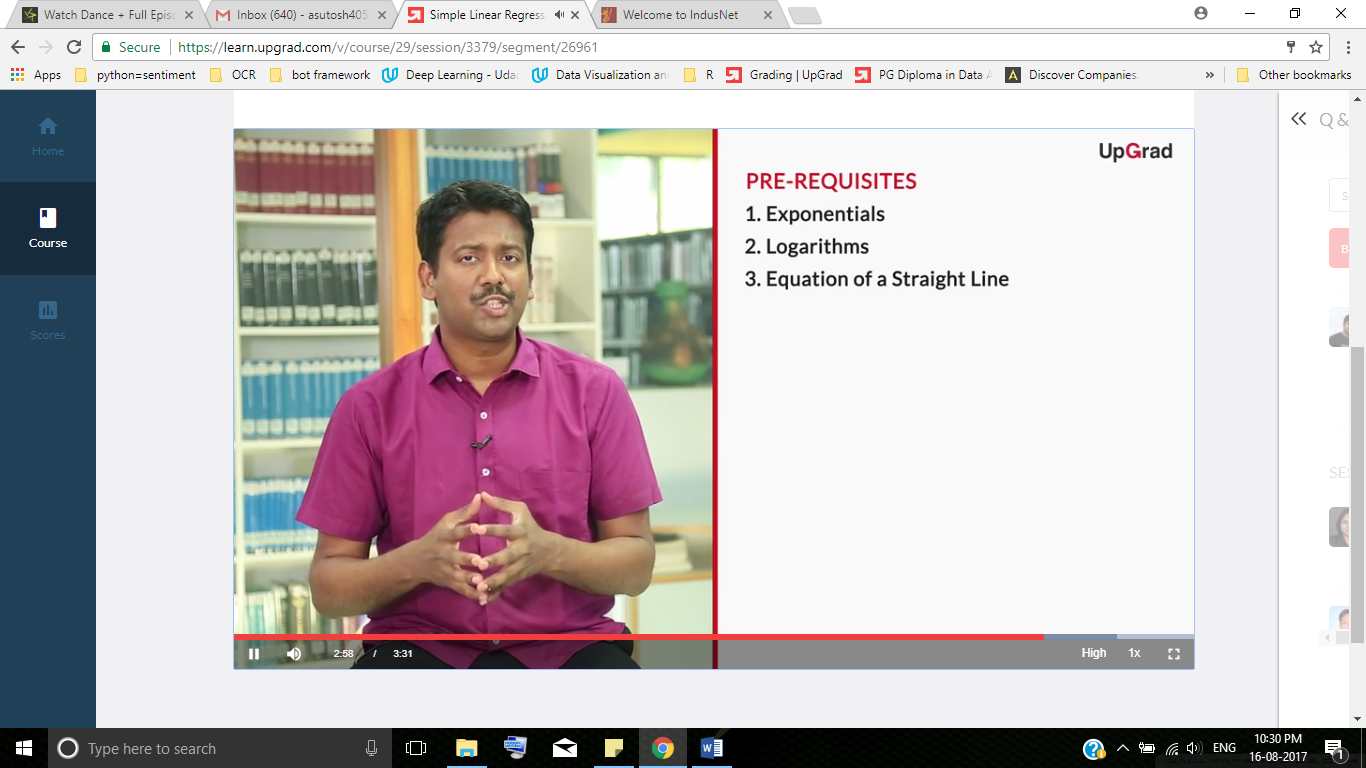
Speech recognition using hidden Markov models and Bayesian networks relies on some elements of supervision as well in order to adjust parameters to, as usual, minimize the error on the given inputs.

Unsupervised learning: Reinforcement learning and Clustering algorithms

A form of **reinforcement learning** can be used for unsupervised learning, where the agent bases its actions on the previous rewards and punishments without necessarily even learning any information about the exact ways that its actions affect the world. In a way, all of this information is unnecessary because by learning a reward function, the agent simply knows what to do without any processing because it knows the exact reward it expects to achieve for each action it could take. This can be extremely beneficial in cases where calculating every possibility is very time consuming (even if all of the transition probabilities between world states were known). On the other hand, it can be very time consuming to learn by, essentially, trial and error.

**Linear Regression**



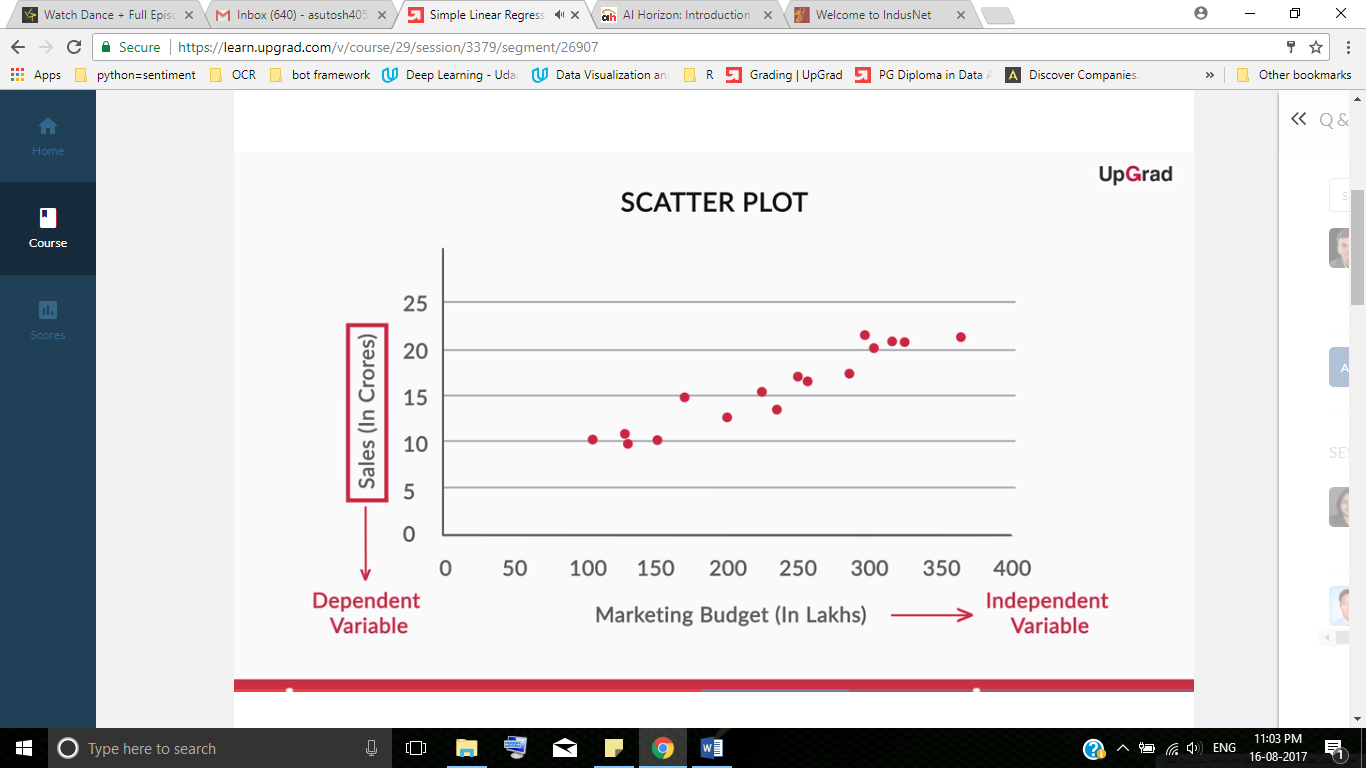


Two types of linear regression:

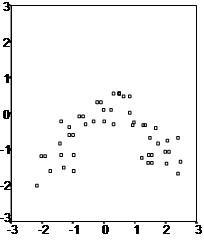
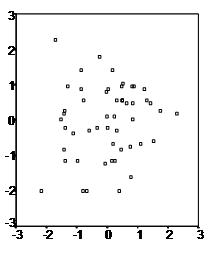
* Simple linear regression
* Multiple linear regression

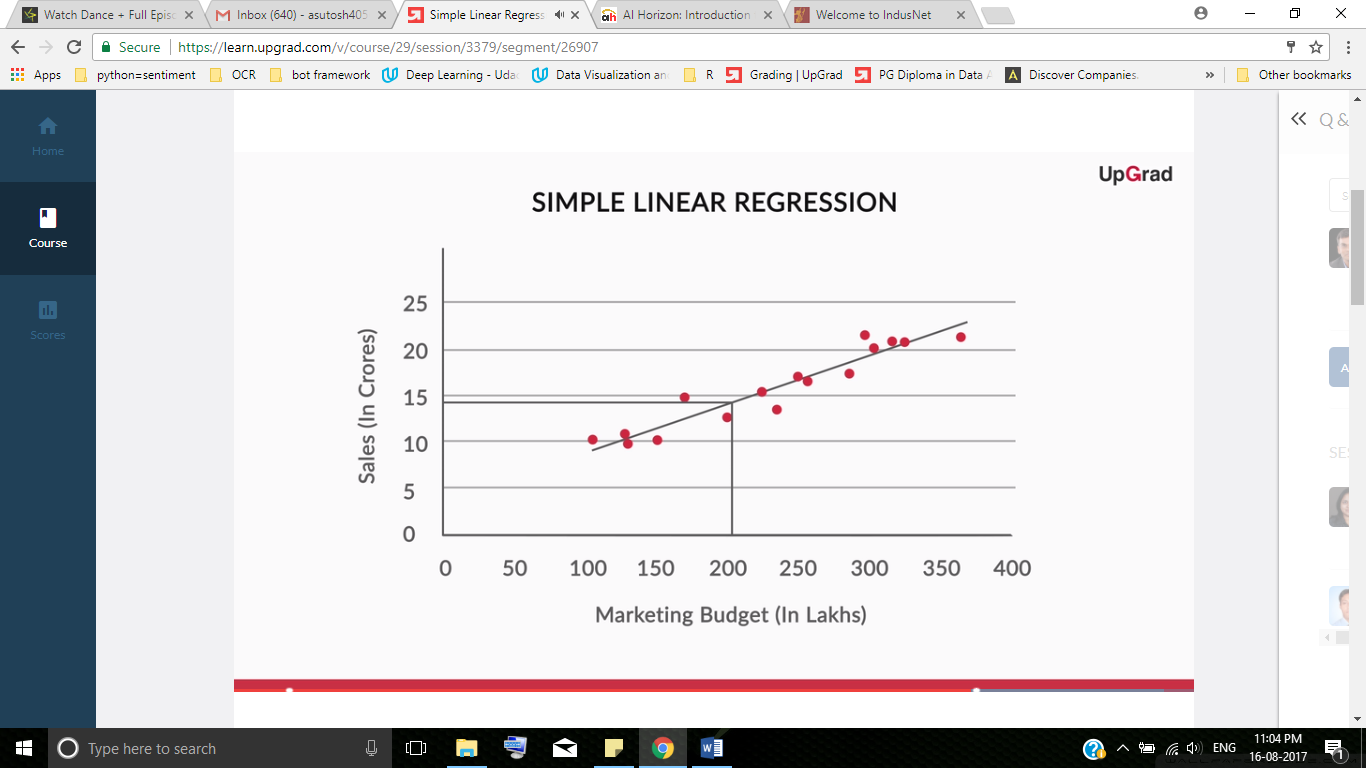
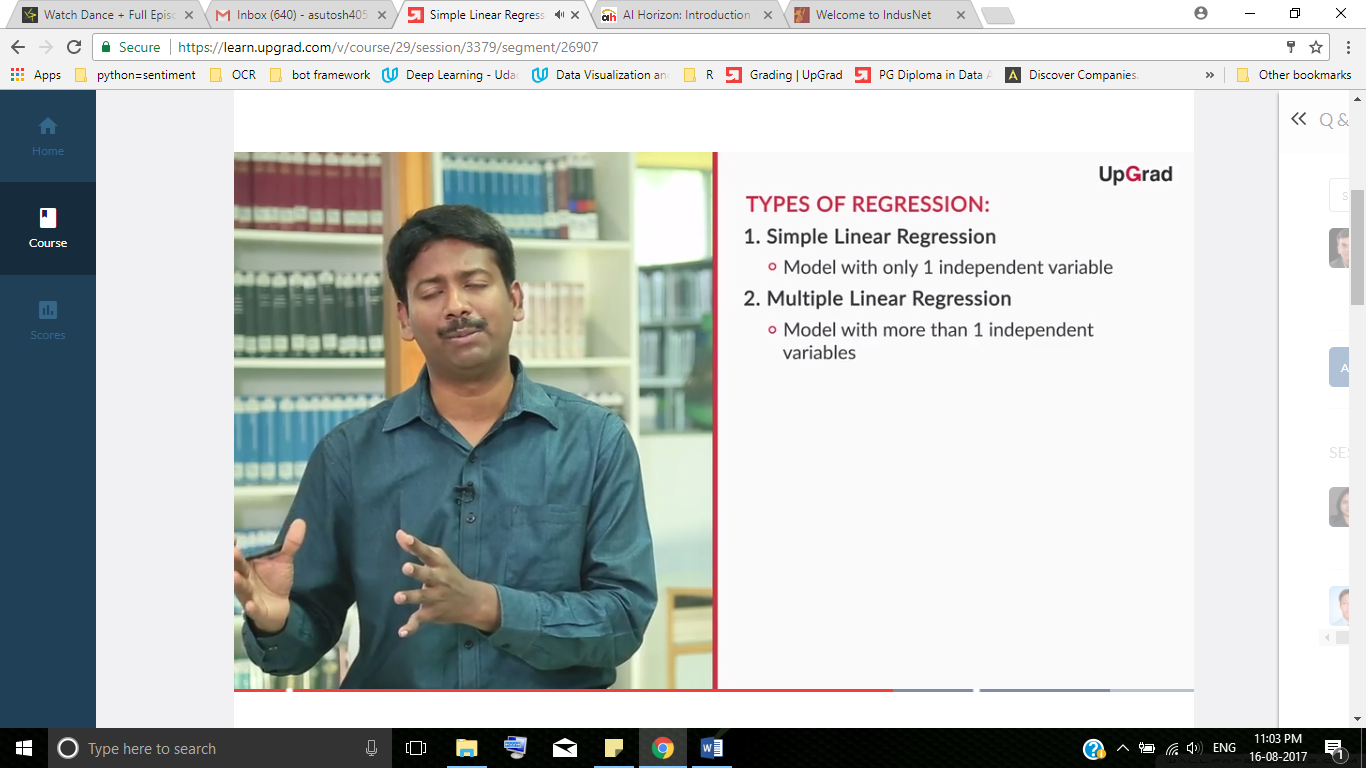
**Regression Line**

Let’s now go deeper into linear regression.



Two examples depict two cases, where no and little linearity is present.

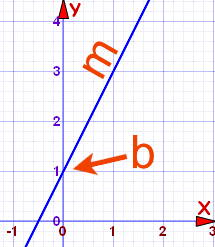
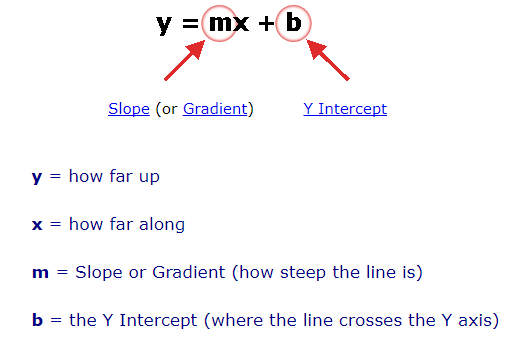
[](http://www.statisticssolutions.com/wp-content/uploads/2010/01/linearregression01.jpg)

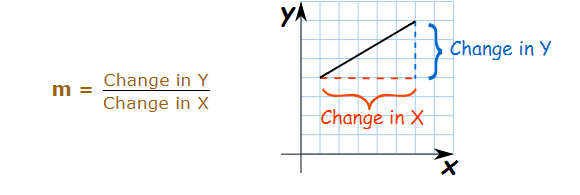


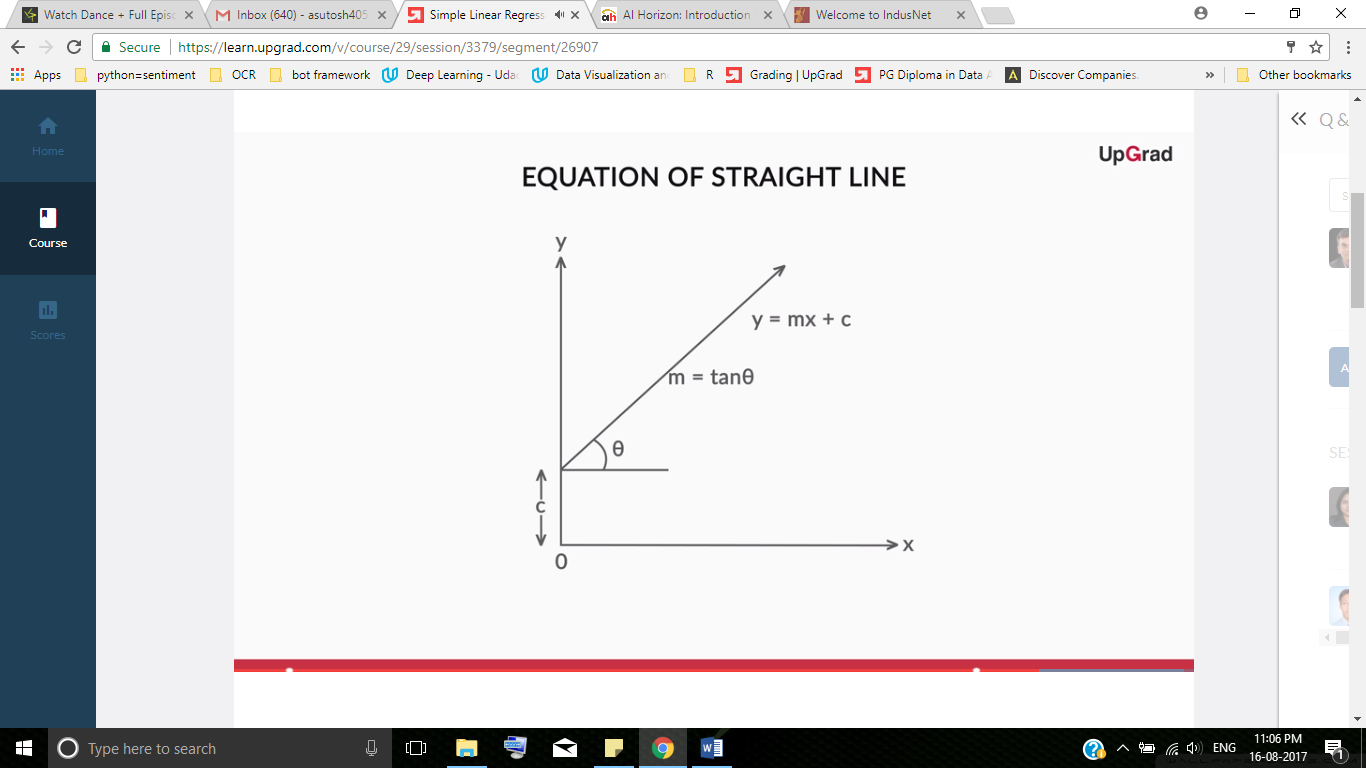
**Equation of a Line**

**y = mx + b**

### **What does it stand for?**





M indicates how strong is x related to y

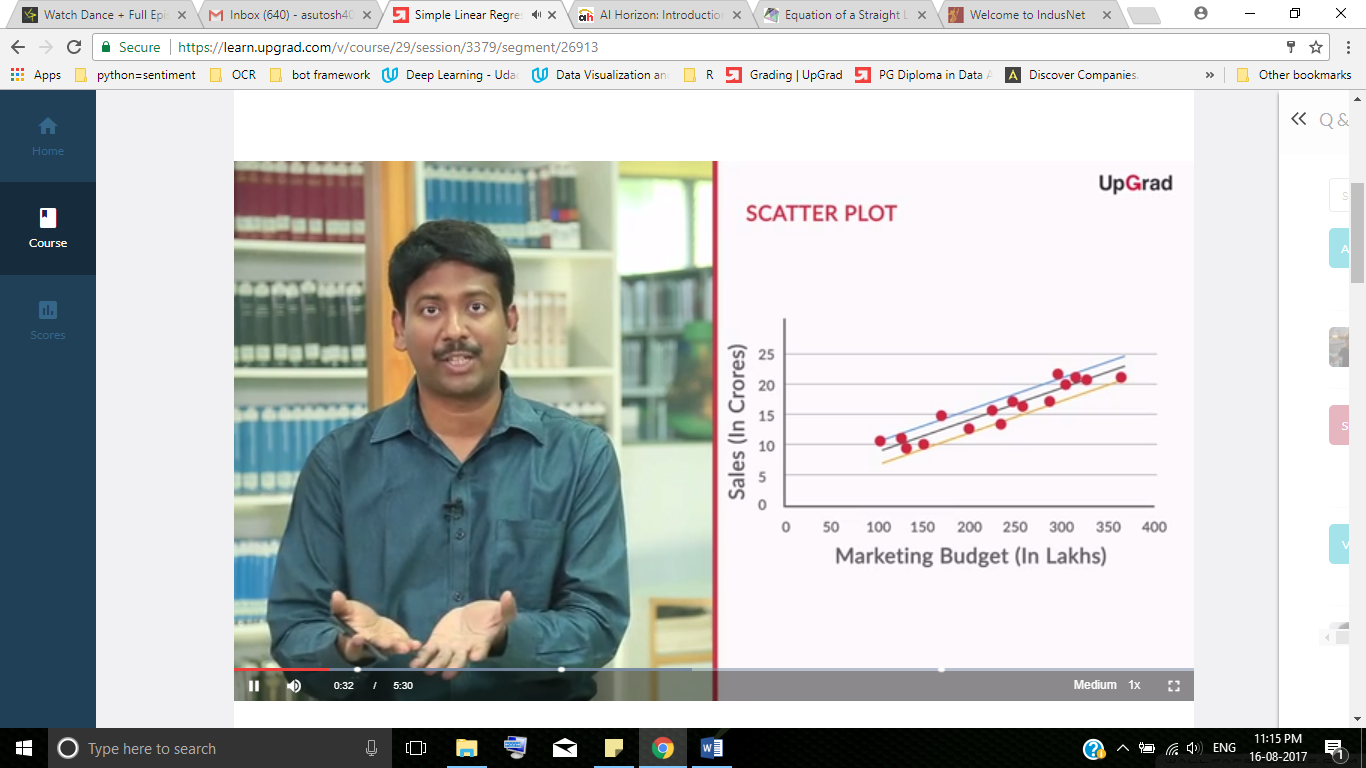
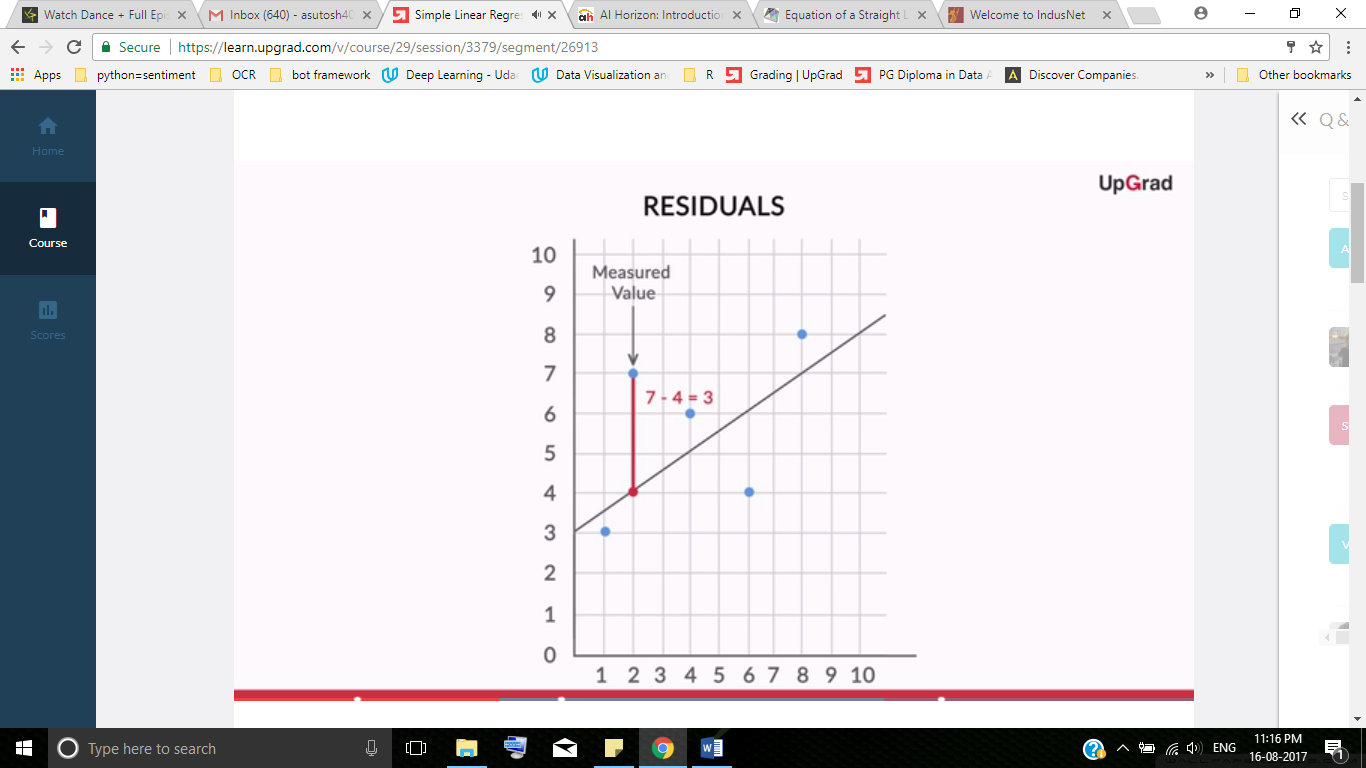
<https://www.mathsisfun.com/equation_of_line.html>

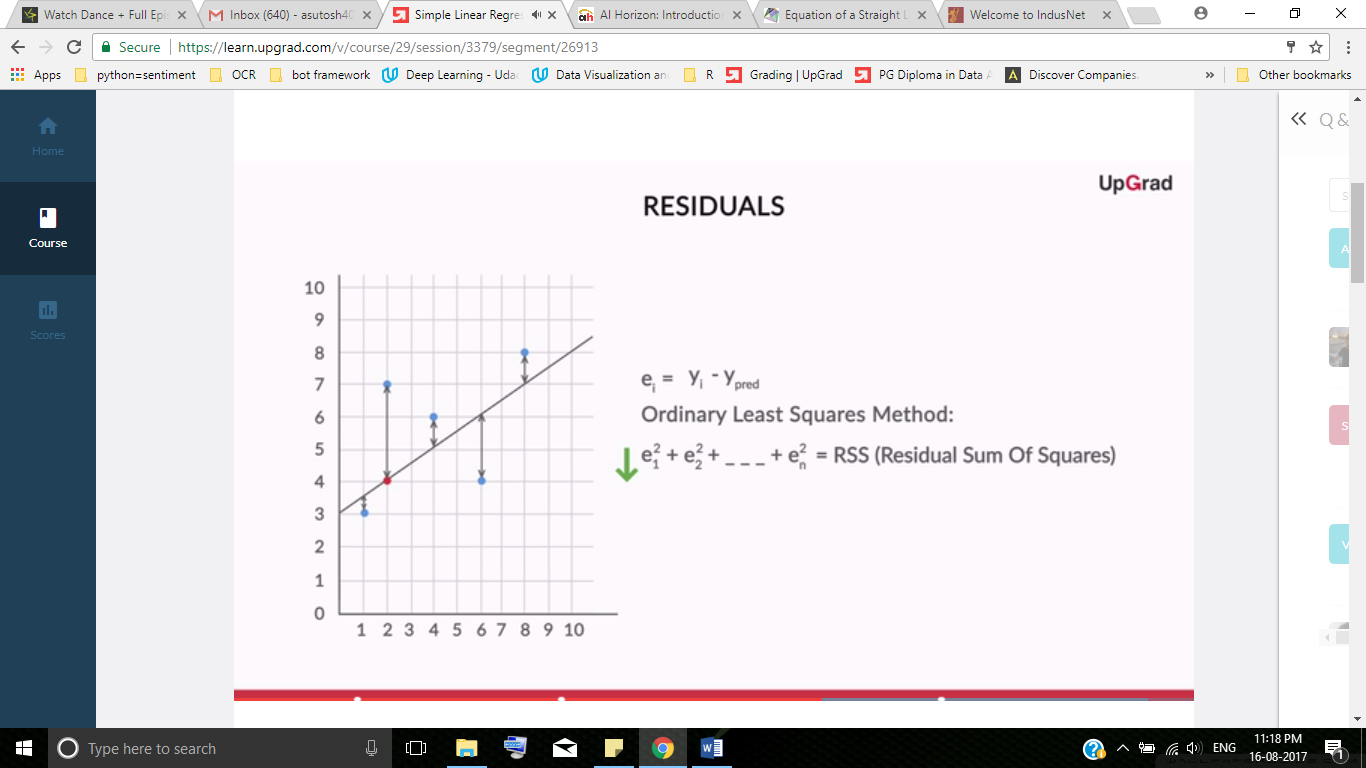


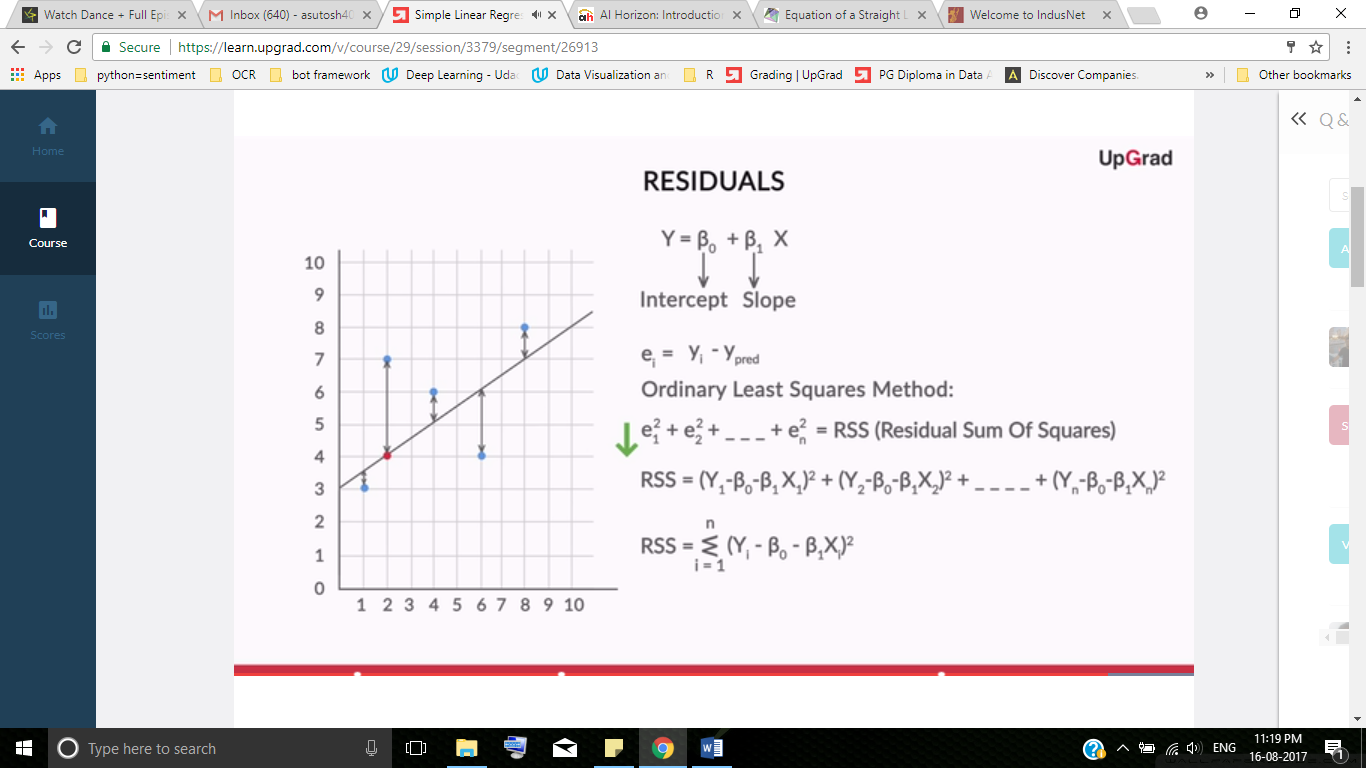
A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.  
The independent variable is also known as the **predictor variable**. And the dependent variables are also known as the **output variables**.

**Best Fit Line**

In regression, there is always a notion of a best-fit line — the line which fits the given scatter-plot in the best way.

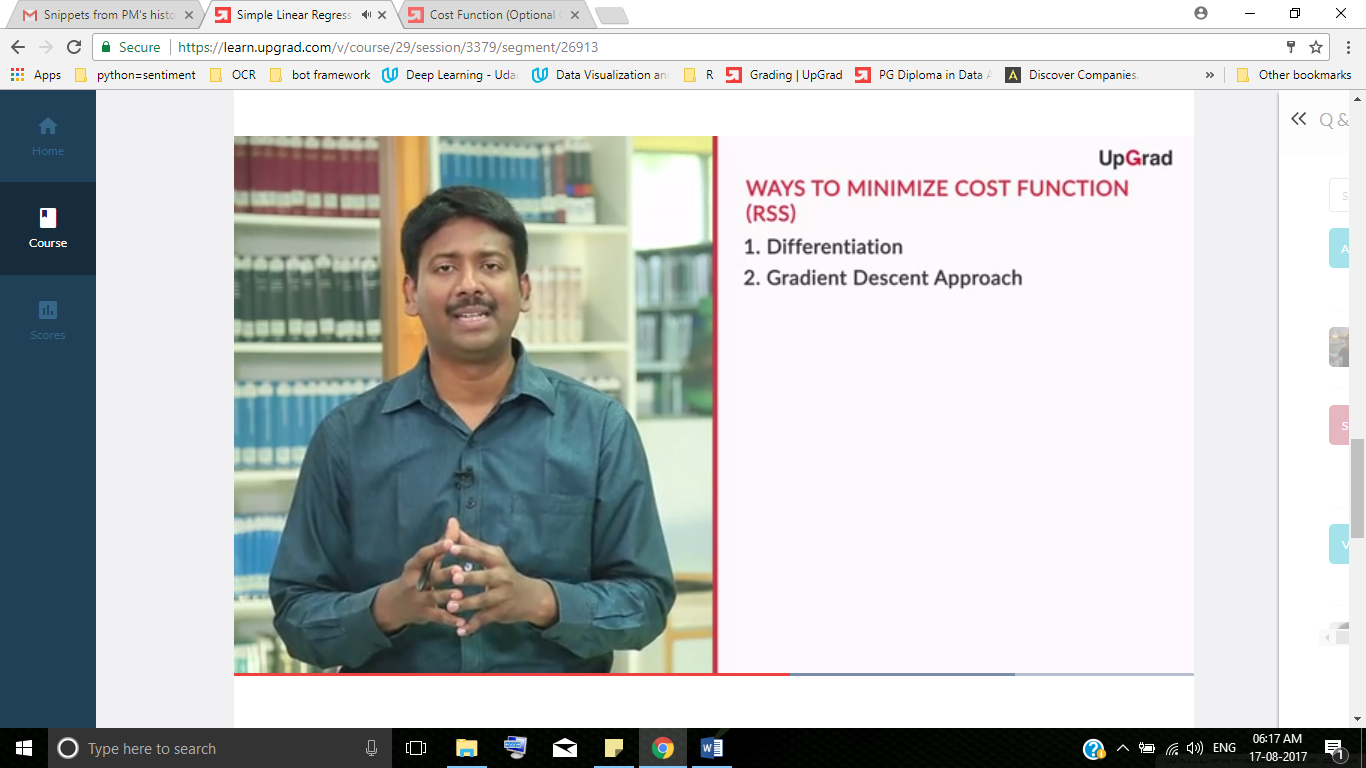
 





Let’s reiterate what you have learnt till now:  
1. You started with a scatter plot to check the relationship between sales and marketing budget.  
2. You found residuals and RSS for any given line passing through the scatter plot.  
3. Then you found the equation of the best-fit line by minimising the RSS`s and found the optimal value of β₀ and β₁.

Since now you know that the best-fit line is obtained by minimising a quantity called **Residual Sum of Squares (RSS)**, this is the best time to be introduced to what is known as the **cost function**.

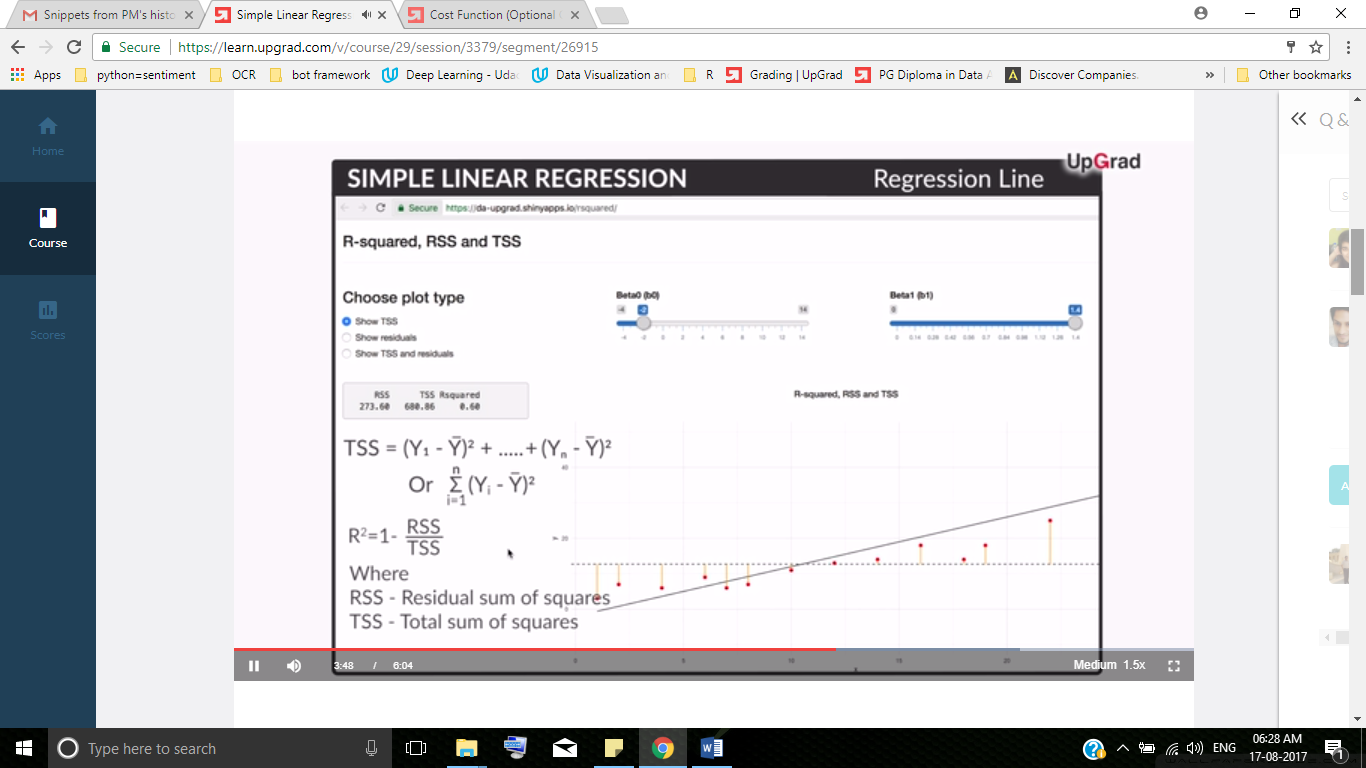


**Strength of Simple Linear Regression**After determining the best fit line, there are a few critical questions you need to answer, such as:

1. How well does the best-fit line represent the scatter-plot?
2. How well does the best-fit line predict the new data?

Are the above questions answered well by the RSS?

RSS is an absolute quantity; and is dependent on the scale of the datapoint.

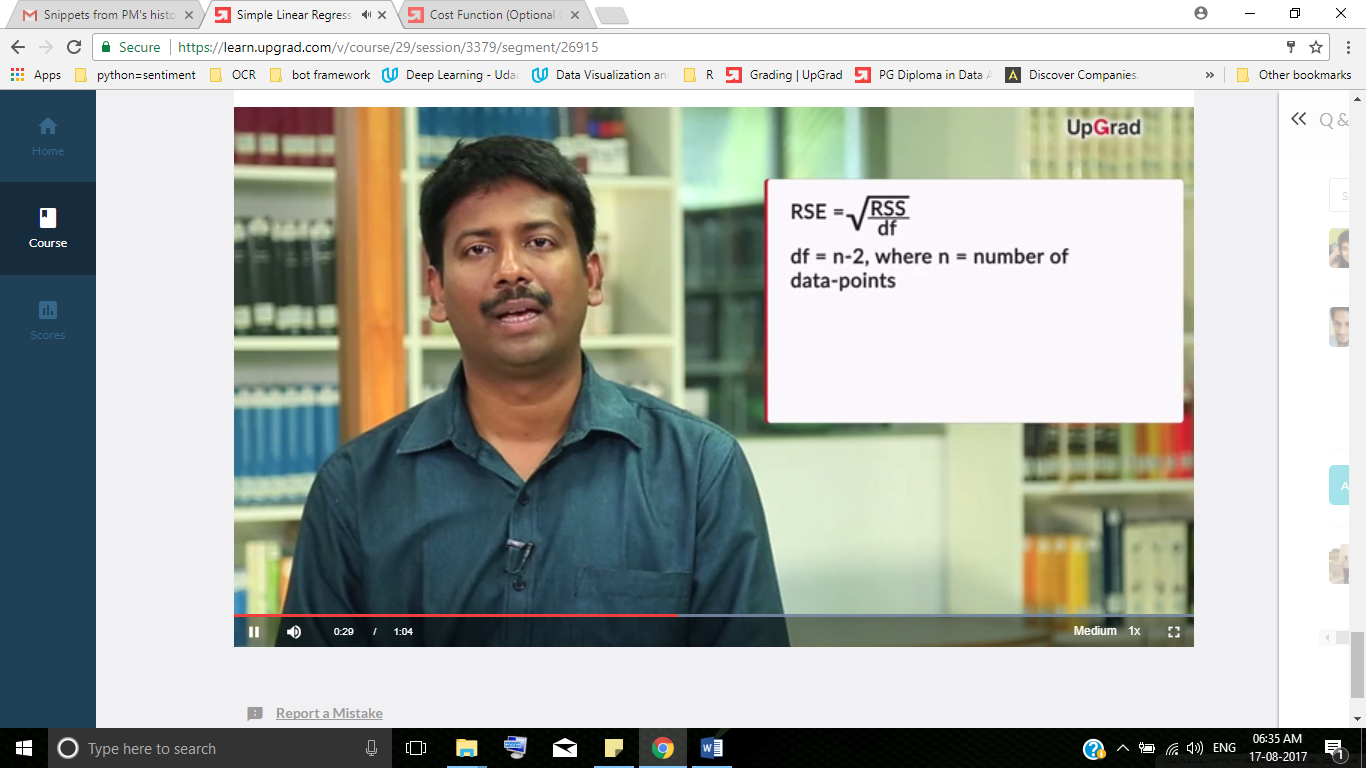


Higher the R-square; good.

Example: R-square is 0.60; means we are able to explain 60% of the variance in the dataset



Apart from R², there is one more quantity named **RSE (Residual Square Error)** which is linked to RSS.



**Correlation & R²**

Recall that **correlation** is a common way to understand how two (continuous) variables vary with each other.   
In general, in simple linear regression, the **R² value is equal to the square of correlation between the two variables.**For the model to perform well, it needs to be evaluated on a 'test' or unseen data set. You can evaluate your model by R² value for training and testing data set. **The R² value for the testing data set should be close to the R² value for the training data set, for the model to be performing well.**

**🡪 Correlation, Collinearity, Multicollinearity**<http://whatis.techtarget.com/definition/correlation>

**Collinearity**If in multiple regression analysis, one of the predictors is linearly associated/dependent on other predictor, then this issue is known as collinearity.

For example, let’s consider the linear model

Y = αx1 + β1x1 + β2x2 ... (1)

If predictor x1 can be expressed as linear combination of x2, say, x1 = 3\*x2

Then this is known as collinearity among the predictors. Note that there will be perfect (or very high) correlation between the predictors as opposed to the assumption of linear regression model (All predictors are assumed to be independent).

Essentially it means that one of the independent variables is not really necessary to the model because its effect/impact on the model is already captured by some of the other variables. This variable is not contributing anything extra to the predictions and can be removed. If we have true collinearity (perfect correlation as in the example above), the one of the predictor is automatically deleted or sown as warning in the Softwares or R/Python scripts.

Note: Collinearity implies correlation but not vice- versa.

In short, multicollinearity:

* can make choosing the correct predictors to include more difficult.
* interferes in determining the precise effect of each predictor, but...
* doesn’t affect the overall fit of the model or produce bad predictions

Summary  
So, what did you learn in this session?   
1. Machine learning models can be classified into following two categories on the basis of learning algorithm:  
(a) **Supervised learning method:**Past data with labels is available to build the model  
 **(i) Regression:**The output variable is continuous in nature  
 (ii) **Classification:**The output variable is categorical in nature  
(b) **Unsupervised learning method:**Past data with labels in not available  
 **(i) Clustering:** No pre-defined notion of labels is there  
2. Past data set is divided into two parts during supervised learning method:   
(a) **Training data**is used for the model to learn during modelling  
(b) **Testing data**is used by the trained model for prediction and model evaluation  
3. Linear regression models can be classified into two types depending upon the number of independent variables:   
(a) **Simple linear regression:** When the number of independent variables is 1  
(b) **Multiple linear regression:** When the number of independent variables is more than 1  
4. The equation of the best fit regression line Y = β₀ + β₁X can be found by minimising the cost function (RSS in this case, using the Ordinary Least Squares method) which is done using the following two methods:  
(a) **Differentiation  
(b) Gradient descent method  
5.** The strength of a linear regression model is mainly explained by R², whereR² = 1 - (RSS / TSS)  
(a) **RSS:** Residual Sum of Squares  
(b) **TSS:** Total Sum of Squares

<http://www.statsdirect.com/help/basics/p_values.htm>