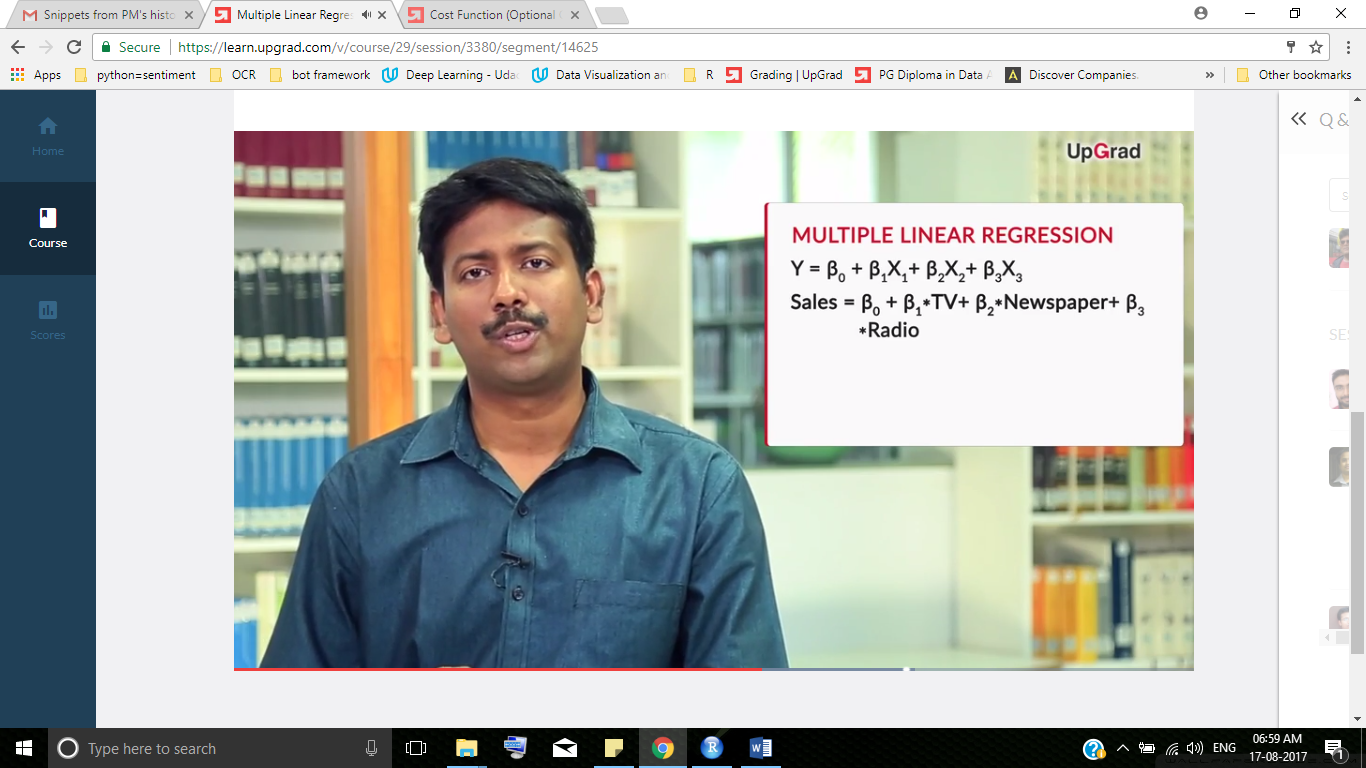
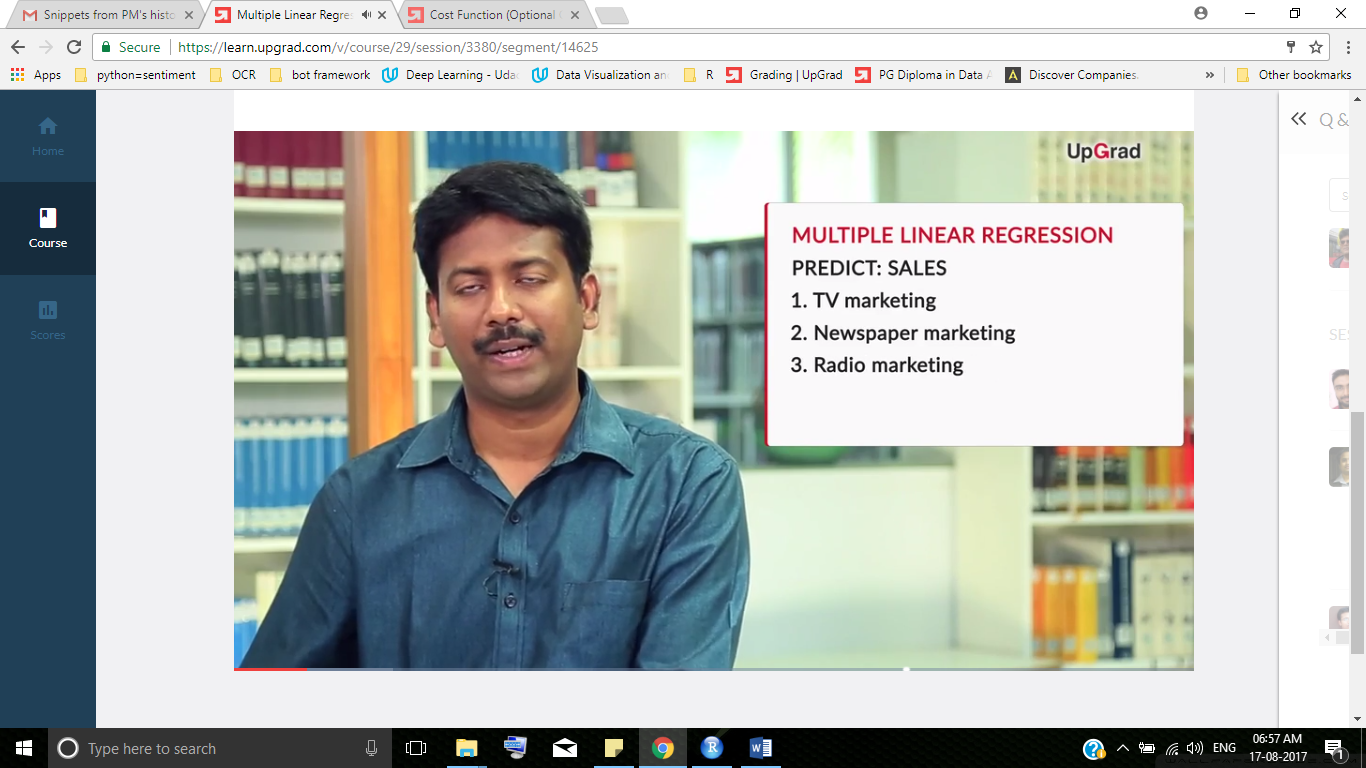
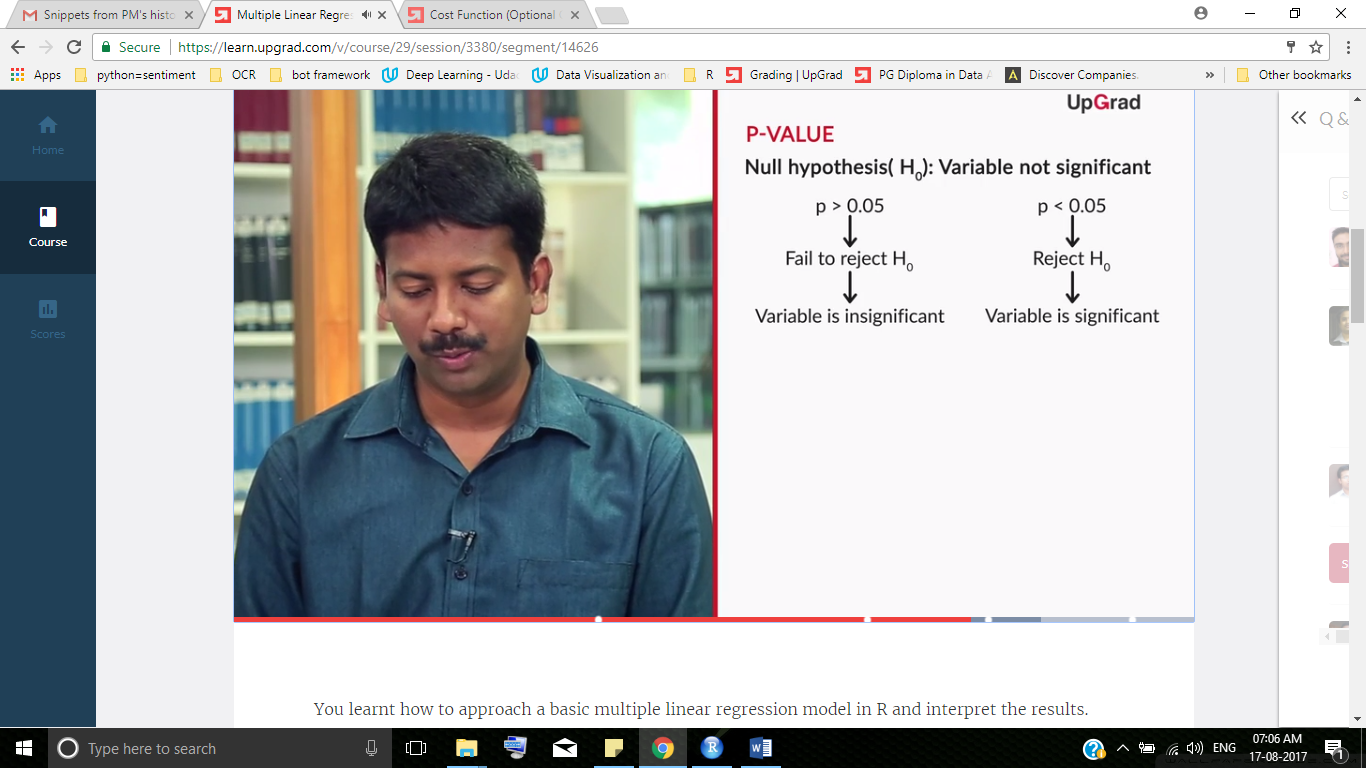
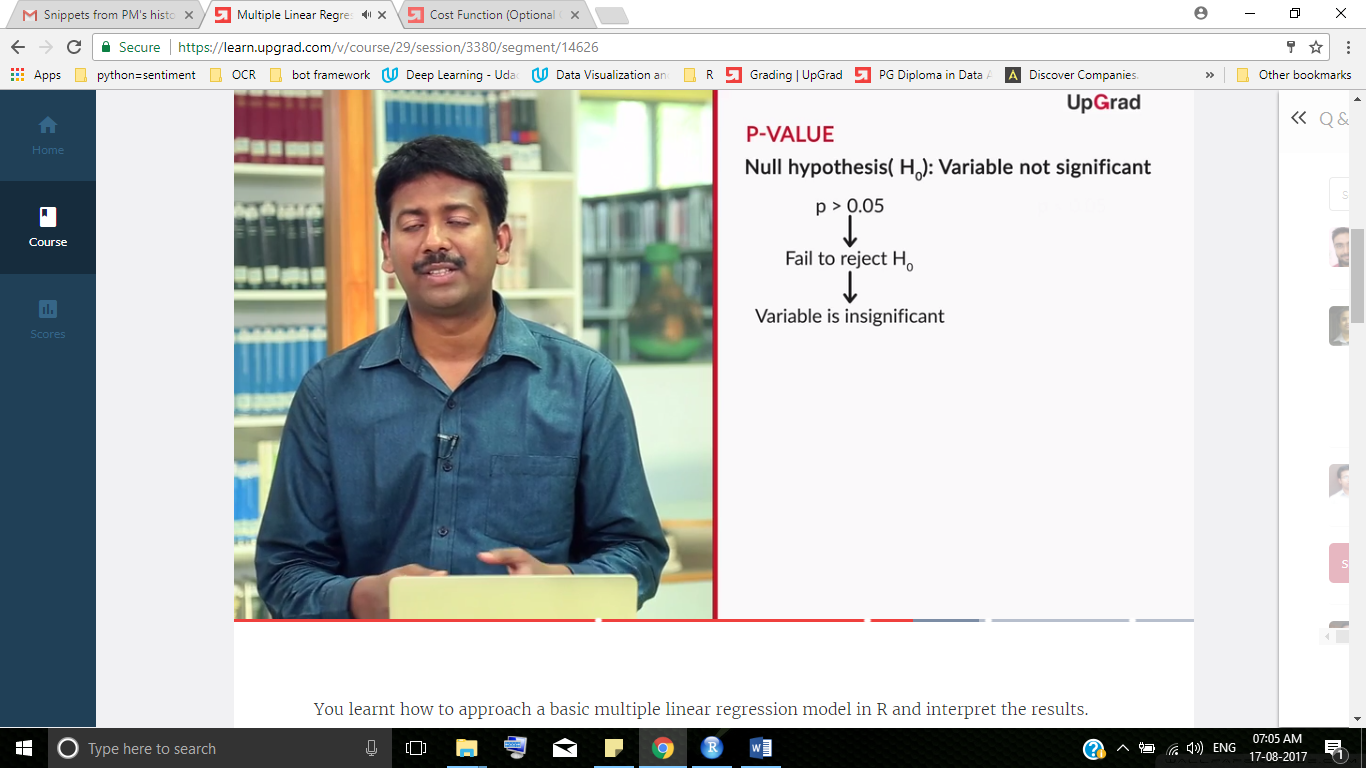
**Multiple Linear Regression**

The name multiple linear regression itself gives you a fair idea about it. It explains the relationship between two or more independent variables and a response variable by fitting a straight line.



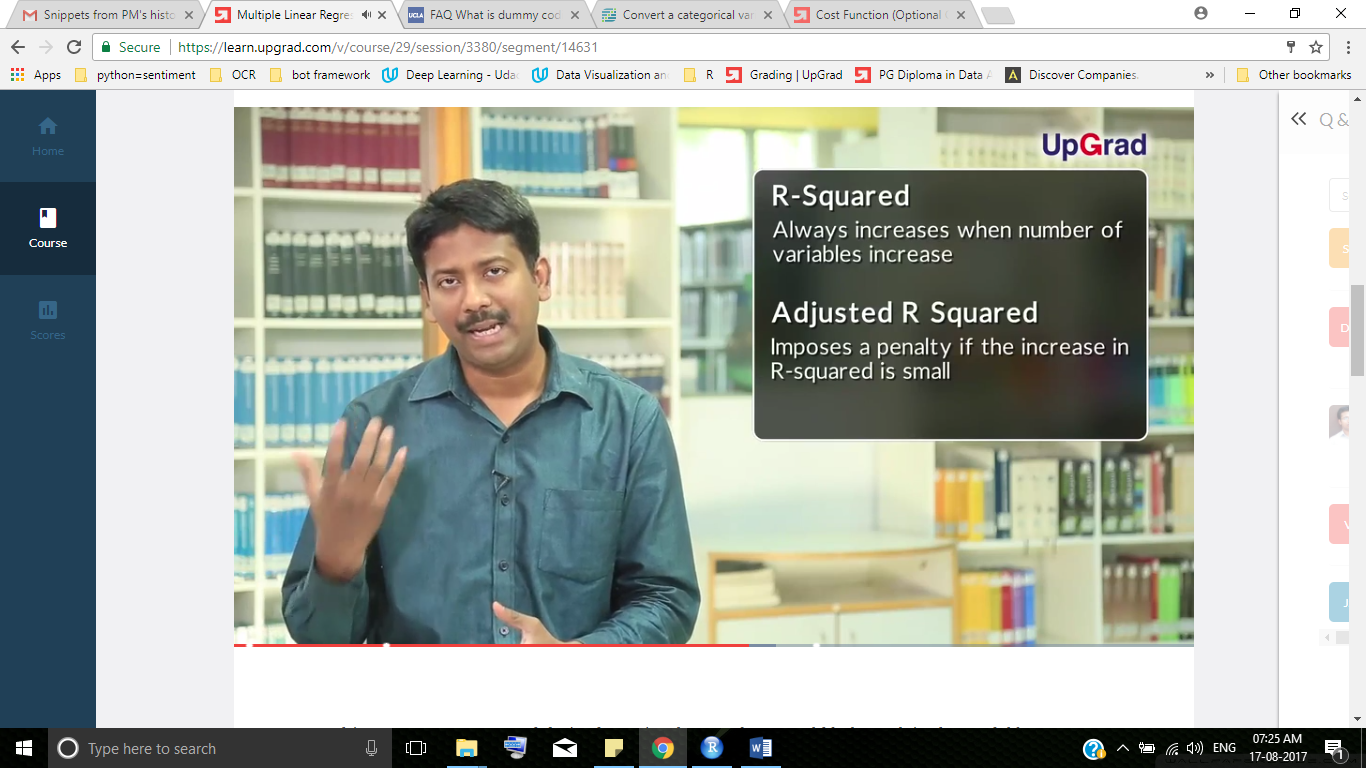
*A model with TV, newspaper and radio together was able to explain the sales variable better.*



You built the model which can now be used to predict the sales, given the TV and Radio marketing budget spend in a market. You understood that **not all variables are useful to build a model.** Some independent variables are insignificant and add nothing to your understanding of the outcome/ response/ dependent variable. In the case of the sales prediction problem, the variable Newspaper marketing was insignificant. Thus, the final model had only two significant variables, i.e. Radio marketing and TV marketing.

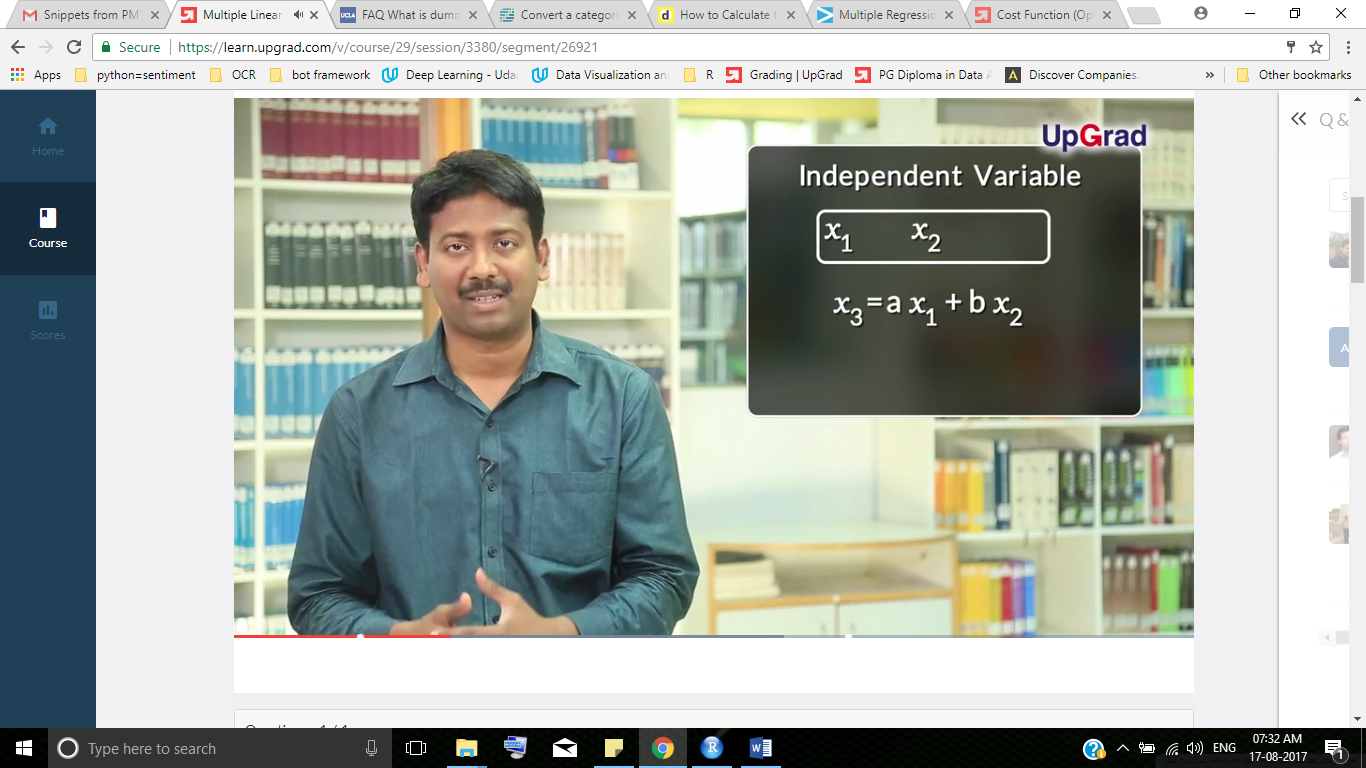
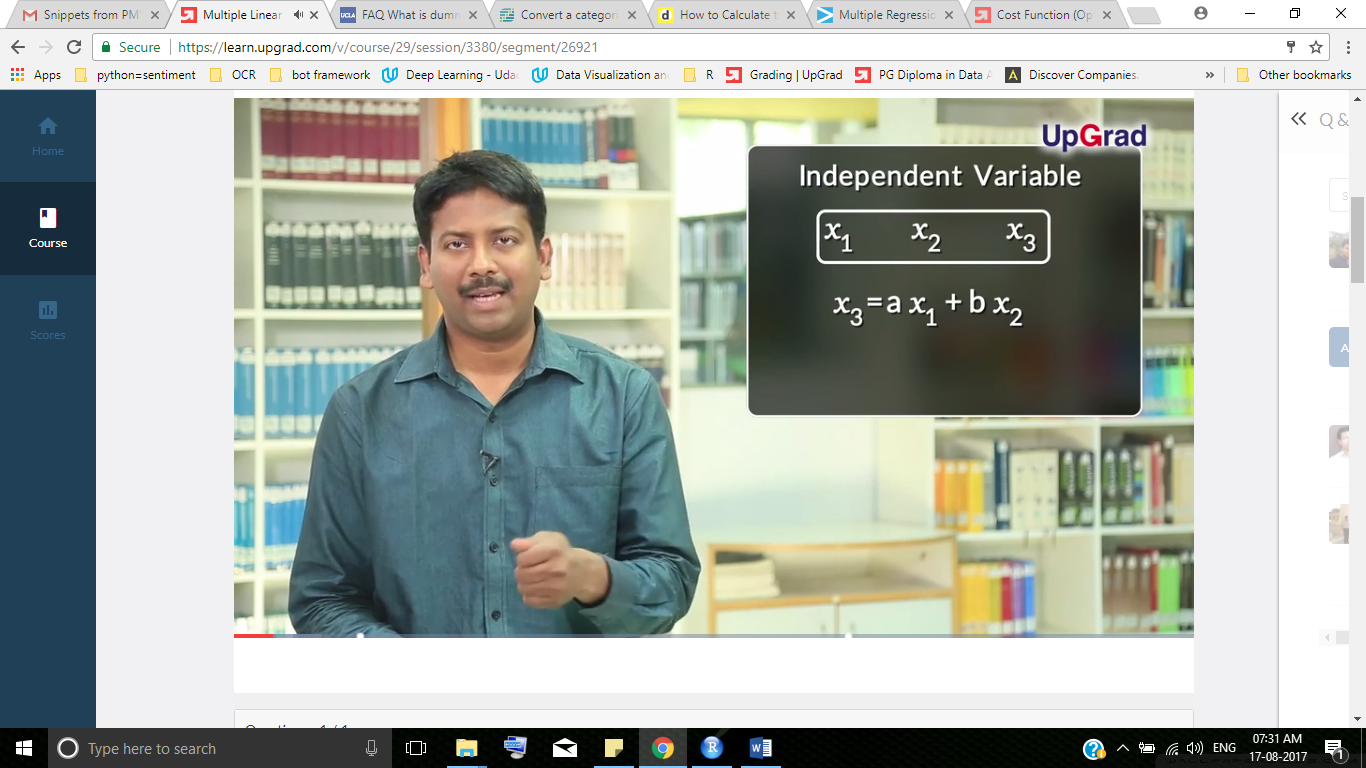
**Let us summarise the steps for model building that you learnt in this lecture:**

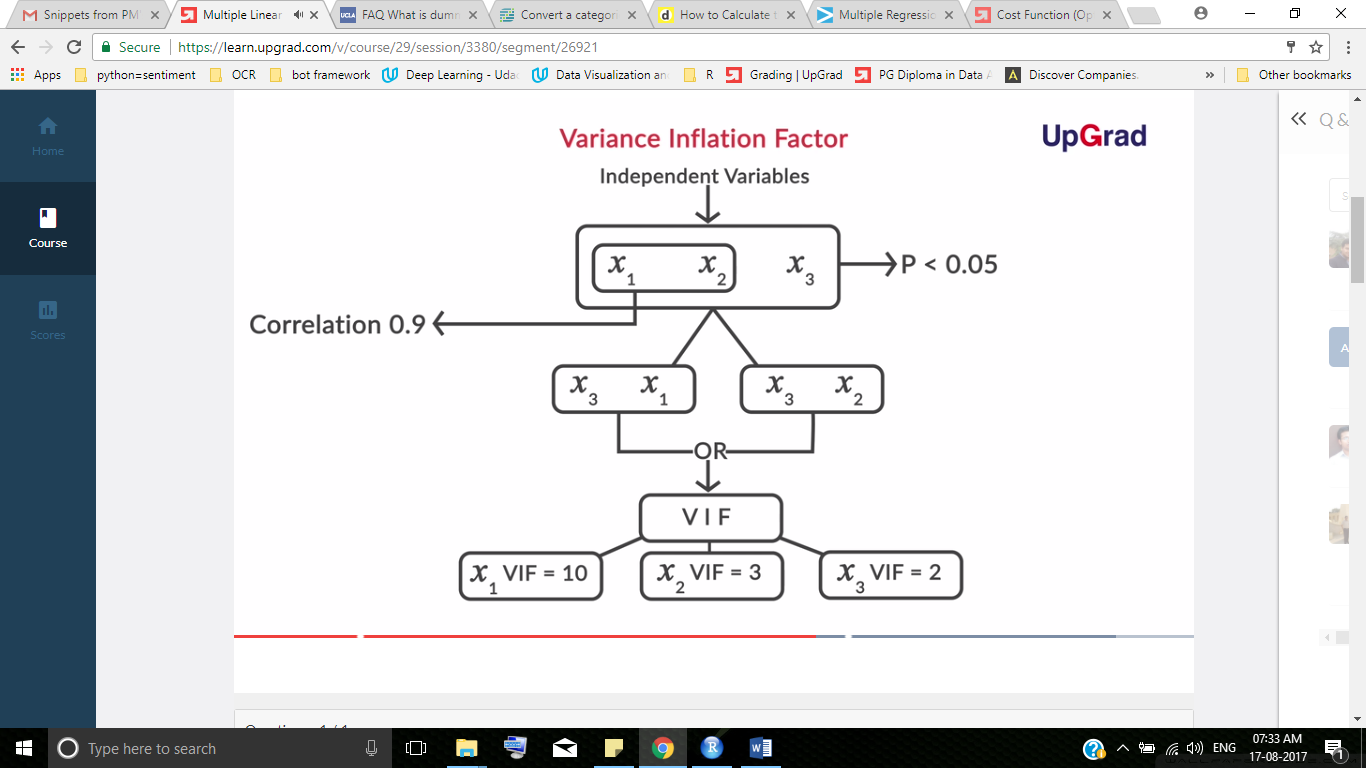
1. Build a primary model "model\_1" taking into consideration all the independent variables
2. Analyse the summary of "model\_1".
3. Remove the insignificant variable on the basis of the p-value. The p-value should be <0.05 for a variable to be significant.
4. Build the final model with the remaining variables.



**adjusted R-squared is a better metric than R-squared to assess how good the model fits the data.**

# Multicollinearity





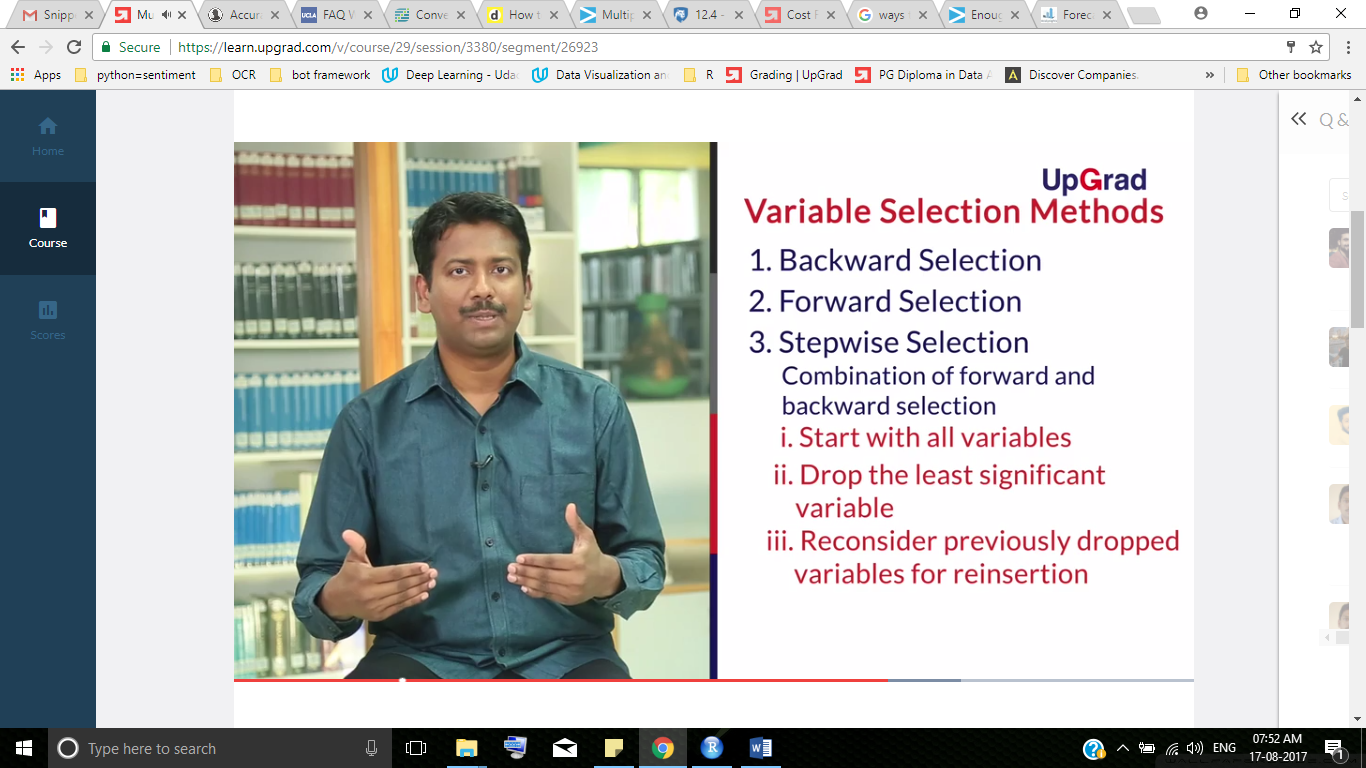
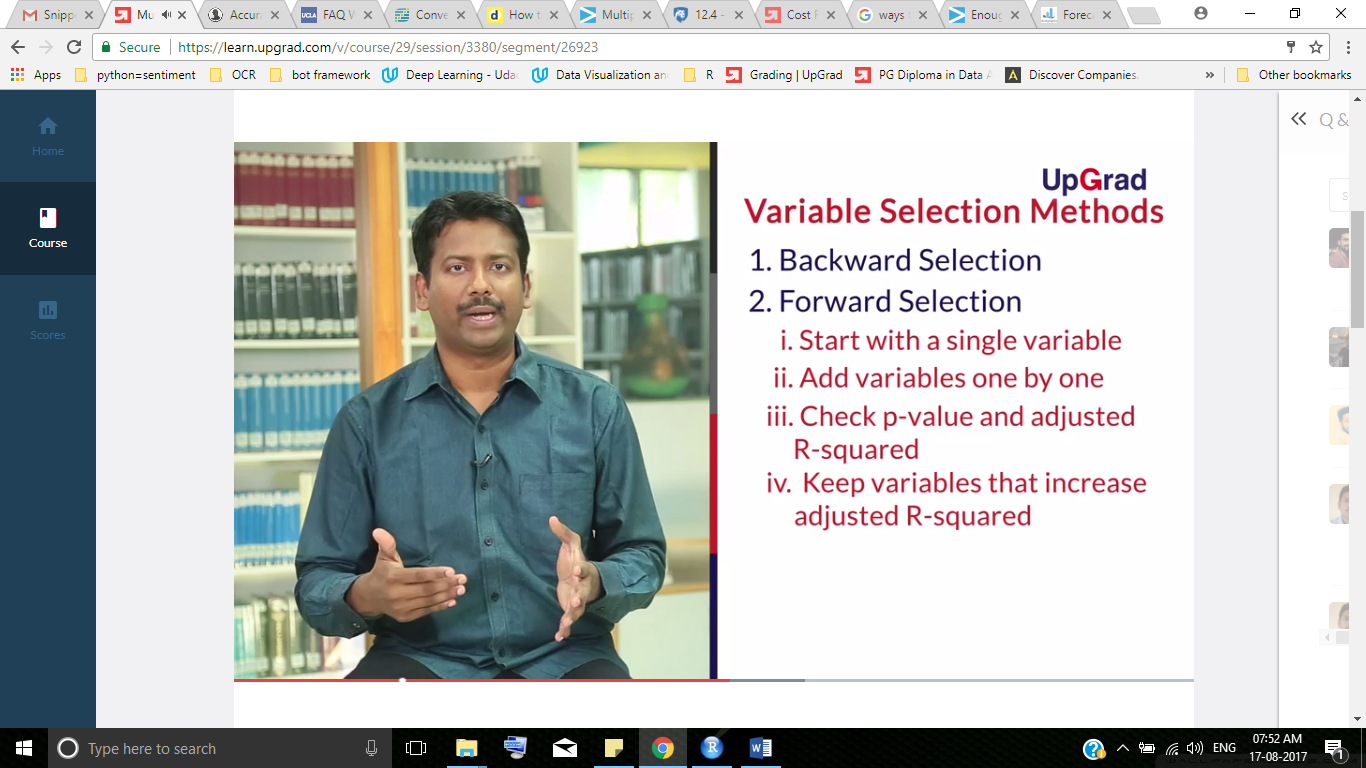
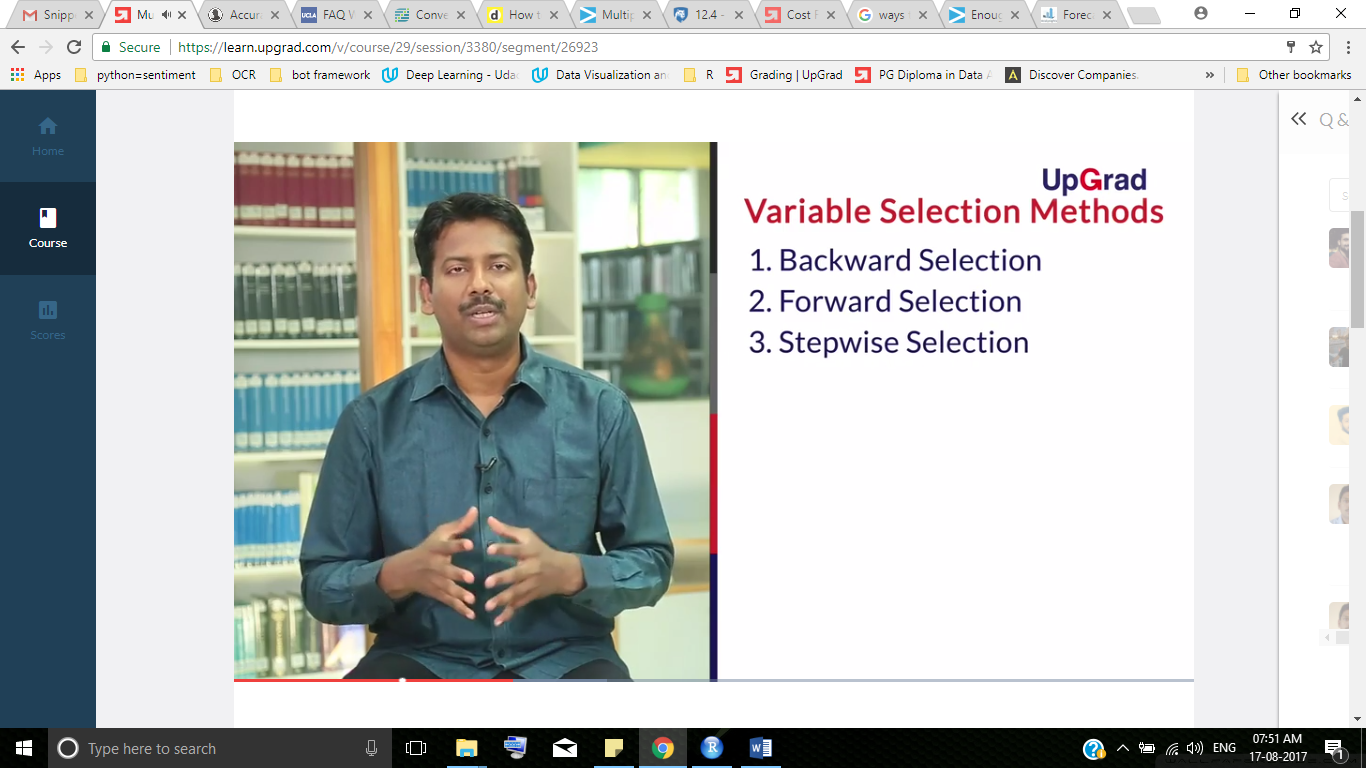


Since one of the major goals of linear regression is identifying the important explanatory variables, it is important to assess the impact of each and then keep those which have a significant impact on the outcome. This is the major issue with multicollinearity.**Multicollinearity makes it difficult to assess the effect of individual predictors**. A variable with a high VIF means it can be largely explained by other independent variables. Thus, you have to check and remove variables with a high VIF after checking for p-values, implying that their impact on the outcome can largely be explained by other variables. Thus, removing the variable with a high VIF would make it easier to assess the impact of other variables, while making little difference to the predicted outcome.

You learnt about the concept of multicollinearity, which essentially means that some of the independent variables also have a dependence on each other. To remove such variables, you learnt the concept of variance inflation factor (VIF).

To summarise: **the higher the VIF, the higher the multicollinearity.** But remember — variables with a high VIF or multicollinearity may be statistically significant (\*\*\*) or p<0.05, in which case you will first have to check for other insignificant variables before removing the variables with a higher VIF and lower p-values. We can take VIF = 2 as the threshold, but in real business scenarios, it will depend on the case requirements.

The model should predict the test set house prices accurately. Thus, it is desired that the **R-squared between the predicted value and the actual value in the test set should be high.**





Let us recall all the steps we used throughout the linear regression model building process:

1. Once you understood the business objective, you prepared the data, followed by EDA and the division of data into training and test data sets.
2. The next step was the selection of variables for the creation of the model. Variable selection is critical because you cannot just include all the variables in the model; otherwise, you run the risk of including insignificant variables too.
3. This is where forward selection, backward selection, and stepwise selection come into the picture. But in linear regression, we focused on the backward selection method. You used stepAIC to quickly shortlist some variables which are significant to save time.
4. However, these significant independent variables might be related to each other. This is where you need to check for multicollinearity amongst variables using variance inflation factor (VIF) and remove variables with high VIF and low significance (p>0.05).
5. The variables with a high VIF or multicollinearity may be statistically significant (\*\*\*) or p<0.05, in which case you will first have to check for other insignificant variables (p>0.05) before removing the variables with a higher VIF and lower p-values
6. Continue removing the variables until all variables are significant (\*\*\*) or p<0.05, and have low VIFs.
7. Finally, you arrive at a model where all variables are significant and there is no threat of multicollinearity.
8. The final step is to check the model accuracy on the testing data.

# Dummy Variables

In the previous example on Advertising, we dealt with data that had only numeric attributes. The Housing data set, on the other hand, also contains some categorical attributes such as air conditioning (yes/no), main road (yes/no), furnishing status (unfurnished, furnished, semi-furnished), etc.

To be continued …