**ML Task Two Group Assignment (23 Marks – 5% of course work)**

**Instructions:**

1. Upload your group **Notebook** on your git account **deadline Wednesday 20th May** before Midnight
2. Some few useful links:

<https://towardsdatascience.com/handling-missing-values-with-pandas-b876bf6f008f>

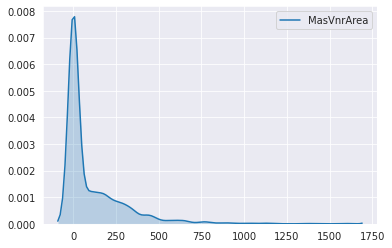
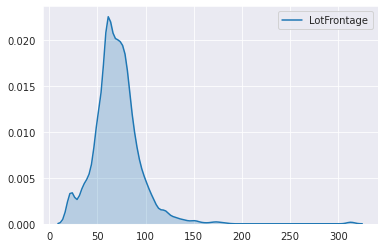
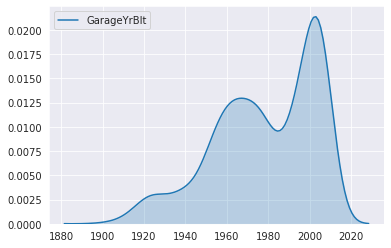
<https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html>

**Question:**

Using the provided *House\_Price\_data*:

* 1. Prepare the data to form a matrix indicate how you dealt with: NaN Values (Note mere deletion attracts lesser marks), Infinite value errors **(5 marks)**

**The following numerical columns/features contain NaN values: LotFrontage, MasVnrArea and GarageYrBlt. Replaced NaN values with the median as they are skewed.**

****

* 1. Perform **PCA** and filter out 2 **Principal Components (PC) (5 marks)** **iii.** Determine the percentage of **information carried** by the above 2 Principal

Component **(1 mark)** **iv.** If we were to capture **90% variance**, how many PCs will be needed? Provide code line **(1 mark)**

* 1. Plot a 3D plane of Best Fit **(10 marks)**
  2. Write down the **general linear regression equation** for this challenge using **only** two PCs **(1 mark)**