**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)**

**Divided the dataset into numerical and object features based on type. Columns of type ‘int64’ and ‘float64’ were considered numerical. Columns of type ‘object’ were considered as object columns.**

**Used SimpleImputer to replace null values in numeric columns with the median value in each numerical column. Also use SimpleImputer to replace null values in object columns with the most frequent value in each object column.**

**After dealing with nulls, One Hot Encoded nominal features and Label Encoded ordinal features. Converted the now prepared data (excluding target) into a matrix.**

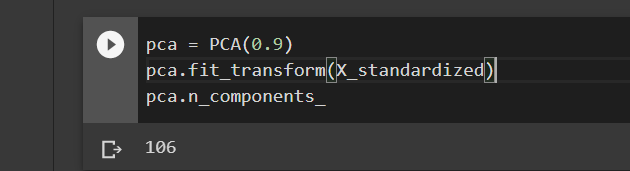
* 1. Perform **PCA** and filter out 2 **Principal Components (PC) (5 marks)**
  2. Determine the percentage of **information carried** by the above 2 Principal

Component **(1 mark)**

**PC1 – 7% and PC2 – 3%**

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

**106 PCs would be needed.**

****

* 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)**

**y\_hat = w1 X1 + w2 X2 + b**

**y\_hat = 17073.5 X1 + 1120.12 X2 + 180921.2**