**Linear Regression Assignment**

**Instructions**

1. Dataset for the assignment - [dataset link](https://www.kaggle.com/datasets/camnugent/california-housing-prices)

2. Create an ipython notebook and answer the questions based on California housing dataset.

3. Make sure you use markdown cells for writing headings, subheadings and observations in the ipython notebook.

4. You are free to use the libraries according to the requirements.

5. You can find the required resources [here](https://docs.google.com/document/d/1-u11nveLFmawUGCKTxRJzAqMBOwCZn51dnAYkVGZrH0/edit)

**Section 1 - Loading the dataset**

1.1 Load the dataset into a dataframe named df.

1.2 Print the number of rows and columns in the dataframe.

1.3 Print the first 5 rows of the dataframe.

**Section 2 -Performing Exploratory Data Analysis on the dataset**

2.1 Print name of all the columns and type of each column (numerical or categorical)

2.2 For all the numerical columns plot the histogram distribution and write your observations about them. (Except latitude and longitude)

2.3 For all the numerical columns plot the box plots and check if there are outliers in any of the column. (Except latitude and longitude)

2.4 Plot a heatmap and check whether there is a correlation between different numerical features. Also check the correlation of different features with the target feature median\_house\_value.

2.5 Plot a barplot to check how many houses are in different subcategories of the feature ‘ocean\_proximity'.

**Section 3- Preprocessing the dataset**

3.1 Find if there are any NaN values in the dataset.

3.2 Impute the NaN values with the median value of the column.

3.3 Remove the outlier datapoints from the dataset.

3.4 Encode the categorical variable ocean proximity using Label encoder or One hot encoder

**Section 4. Splitting the dataset**

4.1 Store the values of median\_house\_value feature in a variable named ‘target’ and then drop the feature from the dataframe. After this step you should have shape of df as 20640 x 9

4.2 Split df into X\_train and X\_test with test ratio 0.2. Similar create y\_train and y\_test

4.3 Plot histogram distribution of y\_train and y\_test and check whether they follow similar distribution.

**Section 5. Training the model**

5.1 Define a Linear regression model and fit(train) the model only on train data.

5.2 Compute the mean\_squared\_error and R square error on train\_data and test\_data for the model and write your observations on the performance of the model.

**Section 6. Feature importance and Feature selection**

6.1 Find top 3 features which are most useful in predicting the median\_house\_value.

6.2 Drop all other features and train your model using only these 3 features and compute the R square and mean\_squared\_error.

6.3 Compare the performance of the original model (using all the features) with the current model.