**FOUNDATIONS OF DATA SCIENCE**

**LAB ASSIGNMENT - 4**

**Dataset:**

The uploaded file – ‘Housing.csv’ is the dataset needed for this lab assignment. It contains house sales prices along with the different factors that affect pricing such as the number of bedrooms, bathrooms, area of the house etc.

**Perform the following operations:**

1. Import the data into a dataframe
   1. Display the head of the dataframe in order to get an idea of all the different columns.
   2. Display information about the different datatypes present in the dataframe.
   3. Display a statistical summary of the dataframe.
2. Data Preprocessing
   1. Drop any columns that do not seem useful or are unnecessary.
   2. Check for missing values in the columns and handle them appropriately.
3. Exploratory Data Analysis
   1. Display the number of houses with unique floor values.
   2. Show a boxplot to check whether more price outliers are present in houses with or without a waterfront view.
   3. Use a regplot to ascertain the correlation between ‘sqft\_above’ and ‘price’. Are they positively or negatively correlated?
   4. Find the other feature that is most correlated with price.
4. Developing the Model
   1. Fit a Linear Regression model to predict the price using the feature 'long' and caculate the R^2. Comment on the goodness-of-fit.
   2. Fit a Linear Regression model to predict the price using the feature 'sqft\_living’ and caculate the R^2. Comment on the goodness-of-fit.
   3. Fit a Linear Regression model to predict the price using the features – ‘floors’, ‘waterfront’, ’lat' , 'bedrooms' , 'sqft\_basement' ,'view', ‘bathrooms’, ‘sqft\_living15’, ‘sqft\_above’, ’grade’, ‘sqft\_living’. Calculate the R^2 score and comment on the goodness of fit.
5. Pipelining
   1. Create a pipeline object using a list of the following model constructors – StandardScaler(), PolynomialFeatures(include\_bias=False), LinearRegression().
   2. Fit the object to predict by the price using the same list of features mentioned in Q.4c. Calculate the R^2 score and comment on the goodness of fit.
6. Splitting the Dataset
   1. Split the data into training and testing datasets with a 85/15 split.
   2. Display the number of training and testing samples.
7. Create and fit a Ridge regression object using the training data, setting the regularization parameter to 0.1 and calculate the R^2 using the test data.
8. Perform a second order polynomial transform on both the training data and testing data. Create and fit a Ridge regression object using the training data, setting the regularisation parameter to 0.1. Calculate the R^2 utilising the test data provided.
9. Cross validation
10. Hypothesis Testing