Assignment 4

# Problem: Total of 70marks

Housing dataset which contains information about different houses. A model trained on this data that is seen as a good fit could be used to make certain predictions about a home — in particular, its monetary value. This model would prove to be invaluable for someone like a real estate agent who could make use of such information on a daily basis.

1. The objective is to predict the value of prices of the house using the given features using multiple regression
2. Perform time series analysis on the value of houses. Can you identify seasonal fluctuations or a trend?
3. Visualize different insights on a dashboard that could be important to a real estate agent using tableau
4. Write your observations in a document

# Here’s a summary of the attachment you need

* data\_description.txt - Has a full description of each column
* train.csv – contains the training set
* test.csv – contains the test set

# Here are the expected steps:

1. Variable definition (2marks): Explore the variables, their meanings
2. Data exploration and descriptive statistics (10marks): Familiarizing yourself with the data through an explorative process using visualization
   1. perform descriptive statistics and state important observations
   2. **Check if there’s is an outlier among the values within independent variables: If** there’s an outlier, you either drop the value, **normalize but create a flag mentioning that the particular variable was normalized or performed imputations.**
   3. Check for missing values:

**Here’s an example of an insight:** There are two class values '>50K' and '<=50K', meaning it is a binary classification task. The classes are imbalanced, with a skew toward the '<=50K' class label. This means I'll adjust the weights of each classes in my logistic regression model so it's not biased towards the 75% group

* + 1. '>50K': majority class, approximately 25%.
    2. '<=50K': minority class, approximately 75%.

1. Check for correlation and state correlated values (5marks)
   1. Check if there’s multicollinearity between variables and pick the right variables
   2. Ensuring that there is no significant correlation among the variables brought into the model and explain the model you dropped
2. Import the train and validate data to build and test the model (1marks)
3. Create a regression model and fit it with existing data and apply on test dataset (10marks)
4. Check the results of model fitting to know whether the model is satisfactory and explain why you choose the model (5marks)
5. Create a fictious data to test the model (5marks)
6. Use gradient descent to optimize the linear regression model (7marks)
7. Perform time series analysis on the prices of houses. Can you identify seasonal fluctuations or a trend? (10marks)
8. Visualize the data in tableau and showcase key insights (15marks)