Name: Shashwat Dhayade

Project Name: Predict Housing Prices

# Introduction

The housing dataset available from Kaggle is used for data mining and machine learning**,** which contains information about 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. This Kaggle project challenges you to predict the final price of each home. This white paper is intended to provide an overview of data mining on the Housing dataset, including **data preparation, exploratory data analysis, modeling, and evaluation**.

# Dataset:

The Housing dataset is available in several formats, including CSV, GZ, and TXT, and can be downloaded from **Kaggle.**

# Data Preparation:

The first step is to download the data. Next, the data will be cleaned and transformed into a format that can be easily evaluated. The missing values, outliers, and any other inconsistencies will be checked and handled properly. New features will be added to make a better dataset for evaluation.

# Exploratory Data Analysis:

After the data has been prepared, the next step is to perform exploratory data analysis (EDA). EDA involves visualizing and summarizing the data to gain insights into the relationships between the variables. Various features will be compared using histograms, and scatterplots to find a relationship between them. This will help in choosing the appropriate modeling techniques.

# Modeling

The next step is to select a modeling technique. I will use three algorithms for the Hosuing dataset, including **Random Forest, AdaBoost, and Catboost**. The goal of modeling here is to predict the final price of each home. There are 4 files which include –

**train.csv** - the training set

**test.csv** - the test set

**data\_description.txt** - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here. It contains 79 explanatory variables describing (almost) every aspect of residential homes

**sample\_submission.csv** - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

The modeling process involves splitting the data into training, validation, and testing sets, training the model on the training set, and then evaluating the model's performance on the validation set. Later submitting the results for the test dataset.

# Evaluation

Model evaluation is an important step in data mining on the Housing dataset. The metric used is [Root-Mean-Squared-Error (RMSE)](https://en.wikipedia.org/wiki/Root-mean-square_deviation) between the logarithm of the predicted value and the logarithm of the observed sales price to determine how well the model performs. The metrics will depend on the modeling technique used, but in general, a higher accuracy indicates a better-performing model.