# Ames Housing Price Prediction Project

This repository explains the Ames Housing Price Prediction Project's data analysis and modeling process in Python.

**Introduction**

Buying and/or investing in real estate market are one of the biggest decisions people make. It is important to make good real estate investments/purchase. We need to know whether a house is priced fairly or even underpriced.

I will use the Kaggle dataset to predict housing sale prices. The dataset contains 2580 records with 79 attributes for 2006-2010 years with detailed information about the house attributes, along with the sale prices. In my analysis, I predicted the price of Ames homes based on various predictors such as ‘OverallQual’, ‘GrLivArea’, ‘GarageCars’, ‘GarageArea’, ‘TotalBsmtSF’, ‘1stFlSF’, ‘YEarBuilt’, ‘FullBath’, etc.

**Project description**

This project aims to analyze and predict the price of Ames Housing market.

Source of the data: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

I was inspired to work on this project because of its relevance to almost everyone. People buy homes and invest in real estate. So, exploring how these home pricing models work is practical and exciting, which gives a better understanding of home prices. Furthermore, this real-world problem project greatly applies data science and machine learning techniques.

**Target audience**

The target audience are home buyers and real estate investors who need to evaluate home prices in Ames, IA.

**Dataset**

We are using the Kaggle dataset with 2580 records with 79 attributes. This dataset is a sample dataset for Housing market in Ames, Iowa. The dataset contains the corresponding features for Ames homes, like 'PID', 'GrLivArea', 'SalePrice', 'MSSubClass', 'MSZoning', 'LotFrontage', 'LotArea', 'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities', 'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType', 'HouseStyle', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'MasVnrArea', 'ExterQual', etc.

You can download this dataset from Kaggle. Link for the dataset: <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview>

**Research question**

My goal is to build a model using a [Kaggle dataset collected from Ames, Iowa](https://www.kaggle.com/code/leeclemmer/exploratory-data-analysis-of-housing-in-ames-iowa/notebook) that can help both realtors and prospective homeowners accurately price a house based on its available features**.**

**Steps completed**

Preprocess/ Clean/ Feature engineering

EDA of the Ames Housing data set, using Python

House Sale Price Predictive models – Linear Regression, KNN, Decision Tree, using Python

**Data Preprocessing and Exploratory data analysis**

The dataset contains missing values for 27 variables. I cleaned and preprocessed dataset, including removing duplicate rows, examining rows and columns with missing values, imputing some of those missing values, and engineering a few new variables. For example, I removed variables as 'Alley', 'PoolQC', 'Fence', 'MiscFeature’ with over 80% missing values. Also, I deleted all data with ‘MSZoning’ ‘C’ (commercial, agriculture and industrial) as these are not residential units. I deleted a variable “Condition2”, because 99% of values are “Norm”. I created a few new variables as ‘Age’, ‘Slope\_Gentle’, ‘RR\_prox’, ‘BsmtLivArea’.

The data was fairly right skewed with a few outliers with higher sale prices.  After a log transformation of the house sale price, the data became more normally distributed, what improved the regression models.

As a preparation to fit a machine model I removed outliers. Also, in order to perform machine learning on a dataset, I dummified nominal categorical variables, like 'MSSubClass', 'MSZoning', 'Street', 'LotShape', 'LandContour', 'LotConfig', 'Neighborhood', 'Condition1', 'BldgType', 'HouseStyle', 'RoofStyle', 'Exterior1st', 'Exterior2nd', etc.

**Modeling**

After completing the data preprocessing, exploratory data analysis and feature engineering, I built a few machine learning models. Models were selected based on which were most likely to have the highest accuracy. I selected three models, including Linear Regression, KNN, Decision Tree. I conducted a train and test split of 30% of the dataset. Each model was trained on the training split, and grid search and cross-validation, and to find optimal parameters for tree-based models.

During the model evaluation phase, linear regression model demonstrated highest R2 score.

**Conclusions**

My goal was to fit a model to predict the Sale Price of homes in Ames, Iowa. After trying to fit a few models, including Linear Regression, KNN, Decision Tree, I found out that

Linear Regression is the best ML model for Ames Housing price prediction with R2 91.47% .

**Next steps**

Future work could include:

Perform the model's refinement to improve accuracy. For example, consider the interactions between the variables and work on model stacking.

Incorporate additional features.

Multiple models could be assembled for greater machine learning power.

**Resources:**

The dataset in Kaggle: <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview>

https://towardsdatascience.com/

"Practical Statistics for Data Scientists 50 Essential Concepts", by Peter Bruce & Andrew Bruce, 2017, O'REILLY

"Machine Learning Mastery with R", by Jason Brownlee, 2016