US HOUSING PRICE PREDICTION

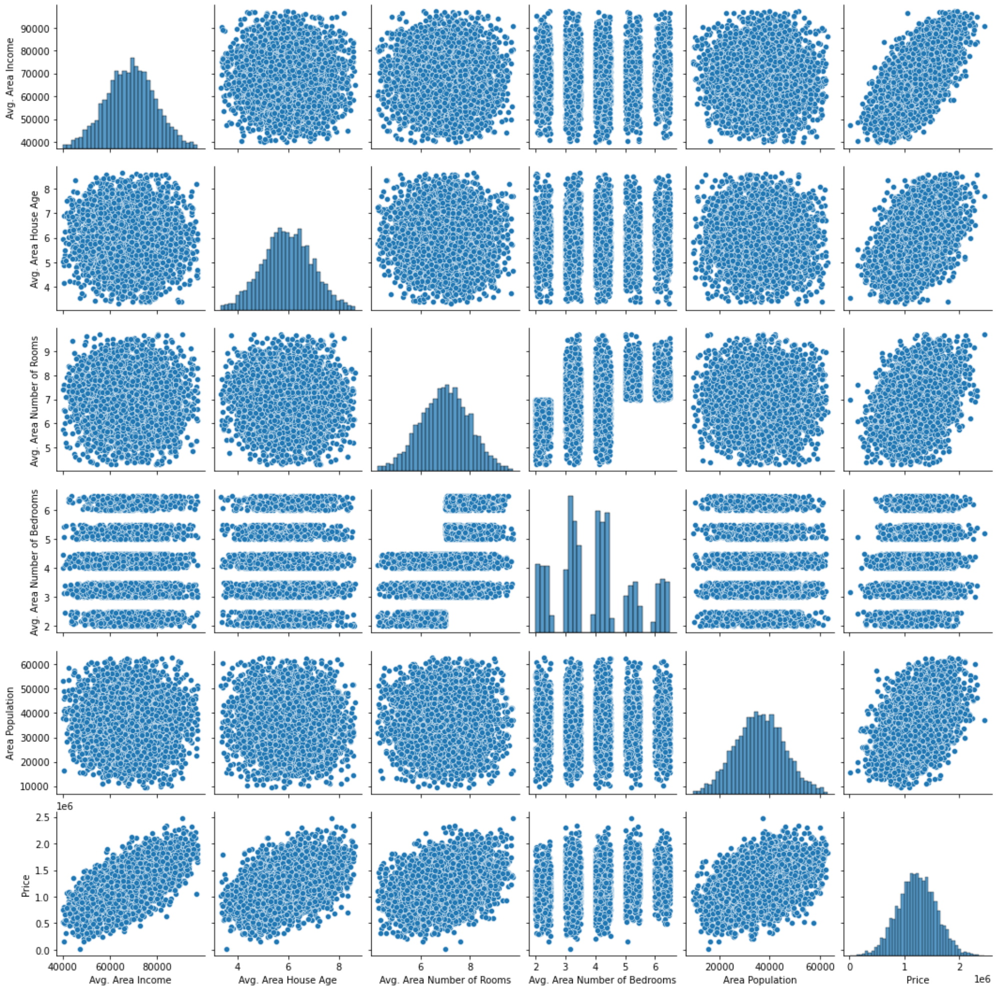
Business Introduction:

In the real estate market, a Comparative Market Analysis, or CMA, is utilized by the broker to present the seller with a proposed sale price and a comprehensive justification for this price. In this assignment we have given US Housing Market Data and Our analysis aims to develop a regression model to predict pricing of homes and could help the real estate agent in evaluating the selling price of the houses. The analysis intends to keep the pricing of homes closer to their “real value” (based on a concrete model) will result in lower resource use on the part of the broker/agency and in a lucrative sale.

Data Understanding:

Our Analysis will use the dataset consisting of 7 variables recorded for 5000 homes situated across the United States. This data will allow us to create a linear regression model to determine how different independent factors impact our dependent variable, Price. Knowing how each variable will affect the house's pricing will assist real estate brokers in determining an appropriate sales price for a home. As our model aims to predict the price of homes, which is a continuous variable, we began by identifying and cleaning the data, and removing all categorical data present. Based on our observations, we identified only one categorical value Address is present in our case, the rest is continuous data, and to train the machine learning model, we ensure our dataset does not contain any Missing (NAN/Null) values, which we found there aren't any.

A pair-plot shows both the distribution of single variables and the relationships between two variables. Pair plots are an excellent way to find trends for further analysis. So, we created a pair-plot to depict the association between dependent and independent variables. It’s showing a pair-wise relationship in a data set.

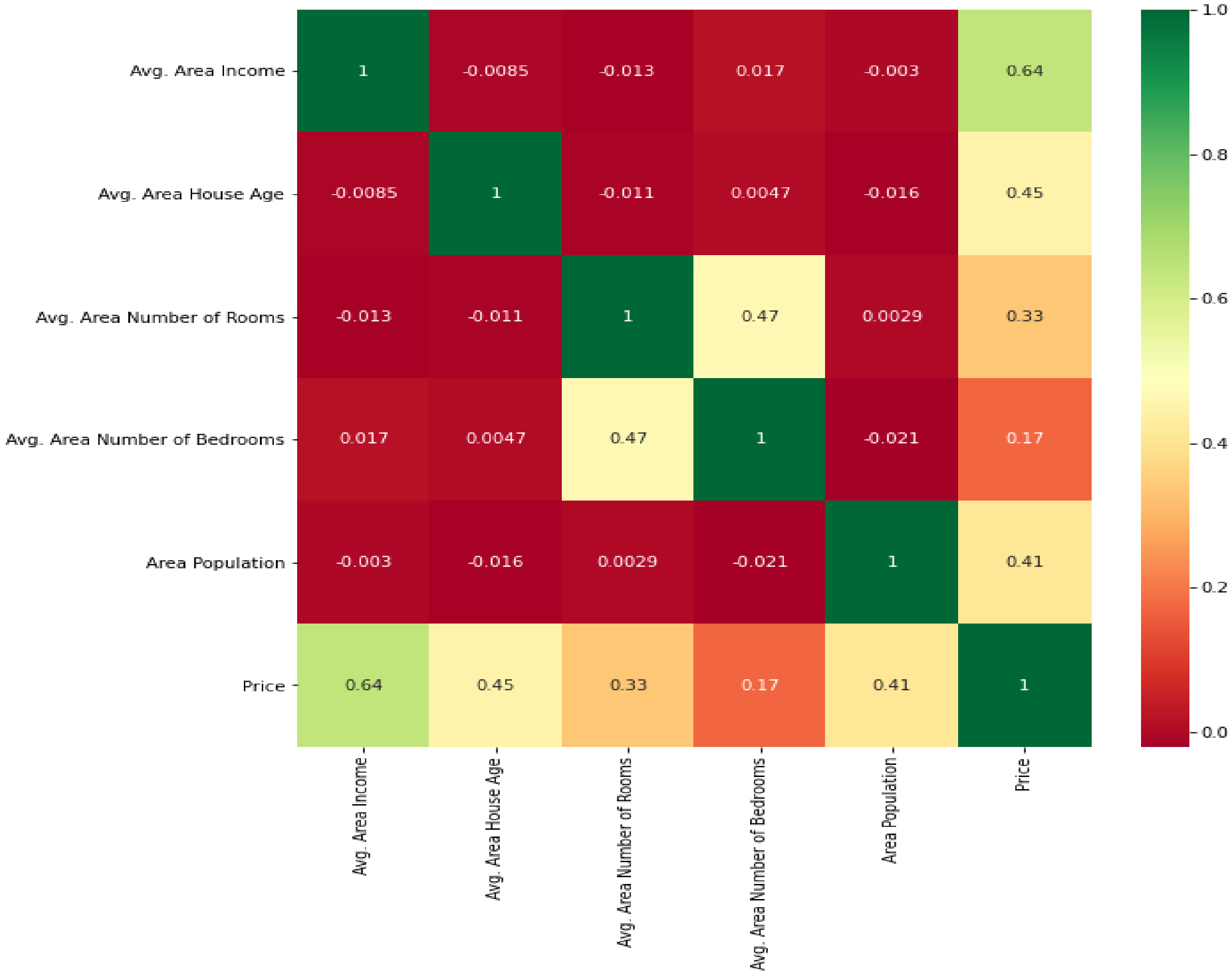


Data Exploration and Transformation:

In order to understand how and which of the given variables will impact the Price of homes, we implemented sklearn Feature Selection Model to identify the variables highly impactful with respect to Price. Based on the Univariate Method, we observed that the 'Area population' is the most influential variable in determining the price of a house, while the 'Avg. Area House of Area' has the least influence.

In addition, we did a correlation analysis of our independent variables against our dependent variable, price, and discovered obvious positive correlations between Area population and price.

Furthermore, before proceeding with our analysis, we ensure that the data fed into the model is normally distributed and does not contain any extreme input values/outliers influencing the model's performance, therefore the data has been processed by identifying and removing any outliers found. A total of 111 values were detected as outliers and were removed from the data pre-processing procedure.

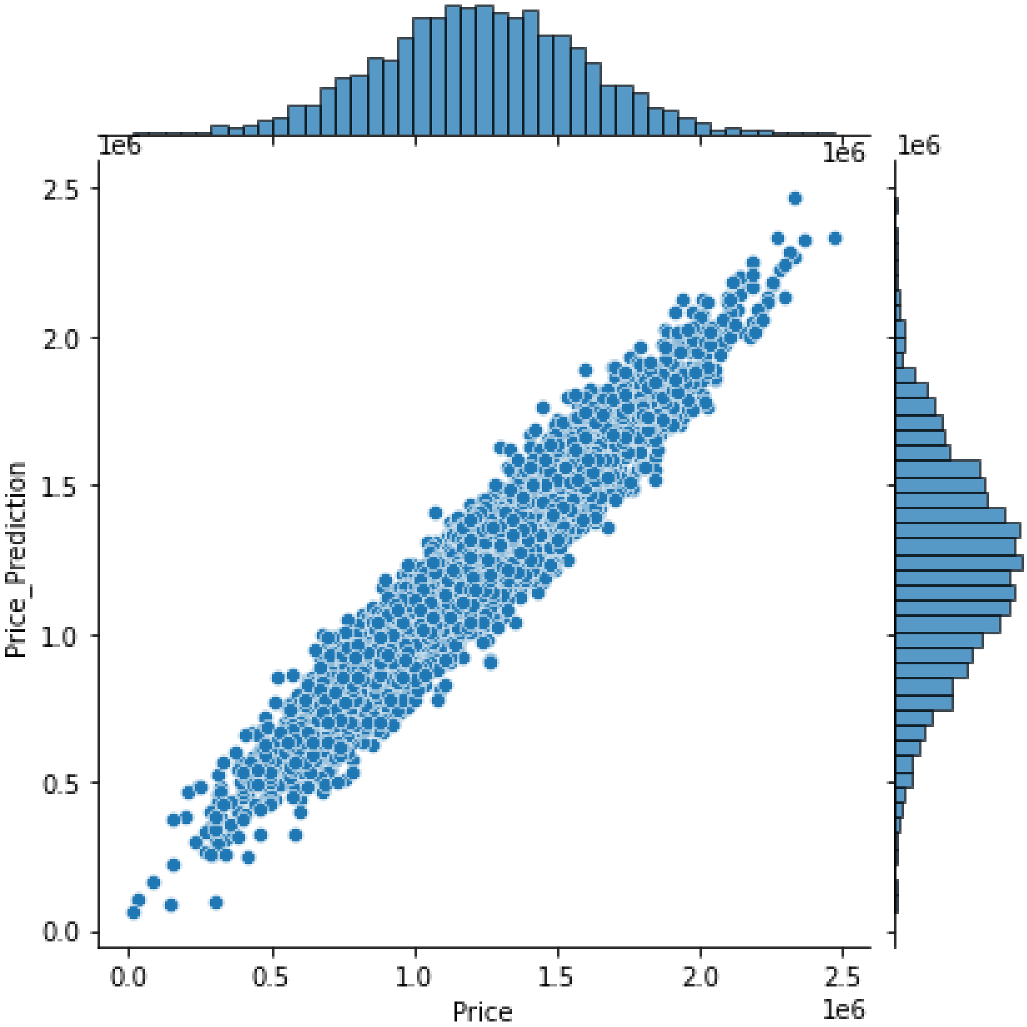


	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
Avg. Area Income	1	-0.0085	-0.013	0.017	-0.003	0.64
Avg. Area House Age	-0.0085	1	-0.011	0.0047	-0.016	0.45
Avg. Area Number of Rooms	-0.013	-0.011	1	0.47	0.0029	0.33
Avg. Area Number of Bedrooms	0.017	0.0047	0.47	1	-0.021	0.17
Area Population	-0.003	-0.016	0.0029	-0.021	1	0.41
Price	0.64	0.45	0.33	0.17	0.41	1

Heat Map depicting the correlation of Dependent Variable, price with other Independent variables, as we can see Avg Area Income is the highly correlated variable.

Modeling:

We know from the exploratory data analysis that selling price is closely connected with a number of variables, thus we used a Multiple Linear Regression Model to deploy an optimal pricing model for the US housing market. The model was trained on 80% of the data, and all explanatory factors were utilized to predict the response/dependent variable, Price. The model's accuracy was consistently more than 90.43 percent, and it was verified using the K-Fold cross validation approach. Aside from the Linear Regression Model, the accuracy of the model was also compared to other regression models such as Decision-Tree Regression and Lasso Regression. The accuracy of Linear Regression was identical to that of the Lasso Regression Model, which was 90.9 percent; however, the decision tree proved to be around



Evaluation:

To predict the Price of the given House Dataset, we have used three different supervised learning models, Linear Regression Model, Lasso Regression Model, Decision-Tree Regression. We employed a variety of Performance Metrics to analyze and compare model performance, including MAE, MSE, R-squared, RMSE, AIC, and Model. The output of the different metrics for the dataset stayed identical for the Linear Regression Model and Lasso Regression Model, and the model has an accuracy of 90.9%; also, the R-squared for both models is 0.9171, indicating that the line is close to the best fit line. Because the assessment measures for both models are quite similar, real estate agents may use either model to estimate the price of a property.

The OLS Regression Results depicting the R-squared, p-value, AIC and other key metrics.

```
OLS Regression Results
==============================================================================
Dep. Variable:              Price   R-squared (uncentered):              0.994
Model:                        OLS   Adj. R-squared (uncentered):         0.994
Method:             Least Squares   F-statistic:                     1.513e+05
Date:            Wed, 19 Jan 2022   Prob (F-statistic):                   0.00
Time:                    02:57:34   Log-Likelihood:                    -12671.
No. Observations:             978   AIC:                             2.534e+04
Df Residuals:                 977   BIC:                             2.535e+04
Df Model:                       1
Covariance Type:        nonrobust
==============================================================================
          coef    std err        t        P>|t|     [0.025     0.975]
------------------------------------------------------------------------------
x1      1.0038      0.003   389.014      0.000      0.999      1.009
==============================================================================
Omnibus:                    1.824   Durbin-Watson:                   2.002
Prob(Omnibus):              0.402   Jarque-Bera (JB):                1.833
Skew:                       0.066   Prob(JB):                        0.400
Kurtosis:                   2.834   Cond. No.                         1.00
==============================================================================
```

Interpretation of Result:

P-value:

P-Value is a statistical test that evaluates the probability of extreme results of a statistical hypothesis test when the Null Hypothesis is assumed to be valid. As a result, a predictor with a low p-value is likely to be a valuable addition to our model since changes in the predictor's value are associated to changes in the response variable. In our situation, the P-value for all models was the same, i.e., Zero, and a low p-value which is less than 0.05 indicates that we may reject the null hypothesis with a 95 percent confidence level and the test is statistically significant.

Coefficient:

The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant.

Following Interpretations can be made from the coefficients:

- A one-unit rise in Avg. Area Income results in a $21.542 increase.
- A one-unit rise in Avg. Area House Age results in a $164954.602 increase.
- A one-unit rise in Avg. Area Number of Rooms results in a $121022.512 increase.
- A one-unit rise in Avg. Area Number of Bedrooms results in a $1846.959 increase.
- A one-unit rise in Area Population results in a $15.045 increase.

Conclusion:

After exploring and analyzing the given US Housing dataset, we successfully created a supervised machine model that employs a Linear Regression Model, to predict the price of the US Housing. Also, the model identifies the relationship between the Price and other independent variable such as Avg. Area Income, Avg. Area House Age, Avg. Area Number of Bedrooms, and Area Population.

Although the intercept is negative, indicating that the model is estimating y values on average, a negative correlation in predicted values is required; yet, after correcting this issue, the model will aid real estate brokers in efficiently projecting US housing prices.