**Problem Statement**

Develop an estimated multiple linear regression model by April 30st 2024, to obtain a home price value based on individual zip code in Maricopa county. The predicted price of the is the target variable for the model. The model’s multiple coefficients will be sorted in descending order to determine the impact a specific house feature has on price. Some examples of house features are:

* Number of bedrooms
* Number of bathrooms
* Approx SQFT – living area square footage
* Dwelling Type – house, condo, mobile home, town house etc…

And so on.

There are 28 independent variables, one dependent variable and 3,117 observations.

The higher the coefficient, the greater the impact a particular feature has on the price of the house. A home developer can use the information to decide which combination of bedrooms, bathrooms, square footage, amenities etc.. are best candidates for maximizing profit on investments. The response variable, the house price, is calculated by a model created using machine learning methods. The calculated price capitalizes on each feature, using individual zip codes all in Maricopa county.

The zip code is a location feature which is useful to determine the type of housing adequate for the area while also considering the average price per square foot and the most sought-after mix of bedrooms, bathrooms, amenities etc… as predicted by the model. The house selling price can be calculated under select scenarios run by the model. The maximum profit for the developer is realized under the scenario where the difference between construction cost and predicted sale price is the greatest.

A minimum 70% coefficient of determination for the model, is considered acceptable. The p-value for the model must be less than 0.05 to be statistically significant. The model is a tool for the business executives to develop a data driven investment strategy to control costs, increase net profits and optimize the features offered in a newly built house.

**Context**

Maricopa county is the largest in Arizona. Has 4.5 million population and is the fastest-growing county in the United States, with more than 56,000 people added between July 2021-2022.

The housing market is dynamic and rapidly changing. Realtors use the MLS to find comps for recommending a price to their clients. A developer might be interested in finding the answer to questions like: For a particular zip code, what is the average market price for a specific housing type when features like square footage, number of bedrooms, bathrooms, stories, is considered. For example, knowing which amenities to offer in an HOA community, or how large the lots need to be to match buyers’ preferences is very valuable because the construction costs would vary accordingly and subsequently, the profit.

Limitations of the model: external factors are not considered in the model but it is assumed the sold price of houses on the market already reflect those. Examples are mortgage interest rates, unemployment rates, changes in property taxes policies and home insurance rates, county budget deficits that adversely affect infrastructure investments.

The data collected covers a time span of 30 days from February 1st 2024 to March 1st 2024. Manipulating the date range will change the predicted value. The rule of thumb is 6 months history of housing market is sufficient for running a comp analysis per realtors practice. This model looks at one month only in order to capture a snapshot in time and to minimize the influence of home sale price variations due to multiple interest rate cuts and rampant inflation. Model complexity can be increased to include homes sold in different time intervals. Current model only covers one month data.

**Criteria for success**

* Clean data by March 15th, 2024 and identify houses attributes that are fixed and variable, i.e.:
  + location is a fixed variable,
  + a house type is fixed and cannot be changed,
  + number of bedrooms, bathrooms, square footage is variable and can be adjusted.
* Obtain a model with relevant predictor variables by April 20th, 2024
* Use model’s coefficients to rank housings’ features based on effect on price from highest to lowest by March 25th, 2024
* Use the model to calculate price recommendations under scenarios like 1 bedroom, two bedrooms, three bedrooms, and so on and plot the results using confidence intervals to visualize price increase vs features by April 1st, 2024.

**Scope of solution space**

Use descriptive statistical analysis to describe data set. Perform a multiple linear regression analysis and an ANOVA statistic to find the relationship between multiple predictor variables, housing attributes, and the response variable, price. The data analysis to be completed by April 20th, 2024 and produce a model output, an estimated equation with multiple coefficients corresponding to attributes of housing. The reliability of the multiple linear regression to be checked by verifying assumptions of:

* linear relationship between predictor variables and response variable
* independence of residuals by performing a Durbin-Watson Test
* homoscedasticity of residuals by performing a Breusch-Pagan Test
* normality of residuals, by visually examining a quantile-quantile plot
* multicollinearity doesn’t exist among predictor variables, by calculating the VIF value (variance inflation factor) of each predictor variable

Deploy the model for continued use as new data becomes available.

**Constraints within solution space**

* An R-Square value close to minimum acceptable 0.70 indicates a lower reliability of the coefficients and thus the response variable - price - may be too high or too low.
* Demographics can influence the demand in a certain area, i.e. median income, unemployment rates, crime rate affecting the desirability of an area.
* Cost of living, infrastructure development, business investments are factors that can change the housing market dynamics from concept to realization of a housing development.
* Permits, water resources and local policies can promote or hinder housing developments despite model price predictions. The model is a tool in a tool box. Further research on local environment should be conducted to support the findings of model predictions.

**Stakeholders to provide key insight**

* D. R. Horton building developer
* John Burns, CEO of John Burns Research and Consulting
* Sebastian Ivan., the Database Manager
* Ernest K. S., Head of Data Science team

**Key Data sources**

Single CSV file Maricopa\_County\_Housing\_data.csv provided by database manager