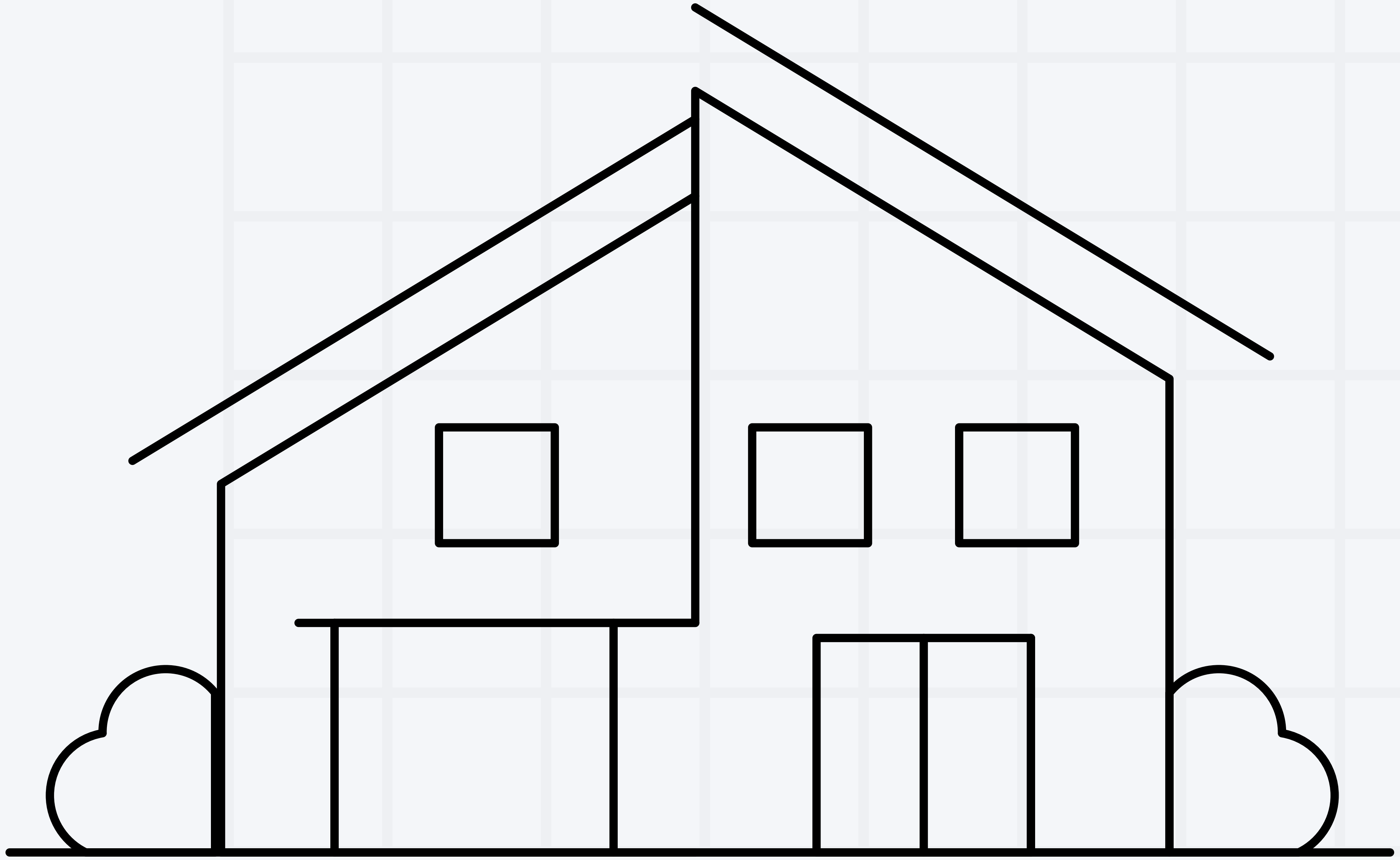
USA House Price Prediction

Linear Regression Method



Objective

This project will follow the Business Analysis (BA) workflow to address house price prediction using linear regression techniques. The business problem is creating a regression model that can accurately predict house prices based on the provided features. Therefore, real estate agents can utilize this model to evaluate the property.

Business Analysis workflow

01

Business Understanding

02

Data Understanding

03

Data Preparation

04

Modeling

05

Evaluation

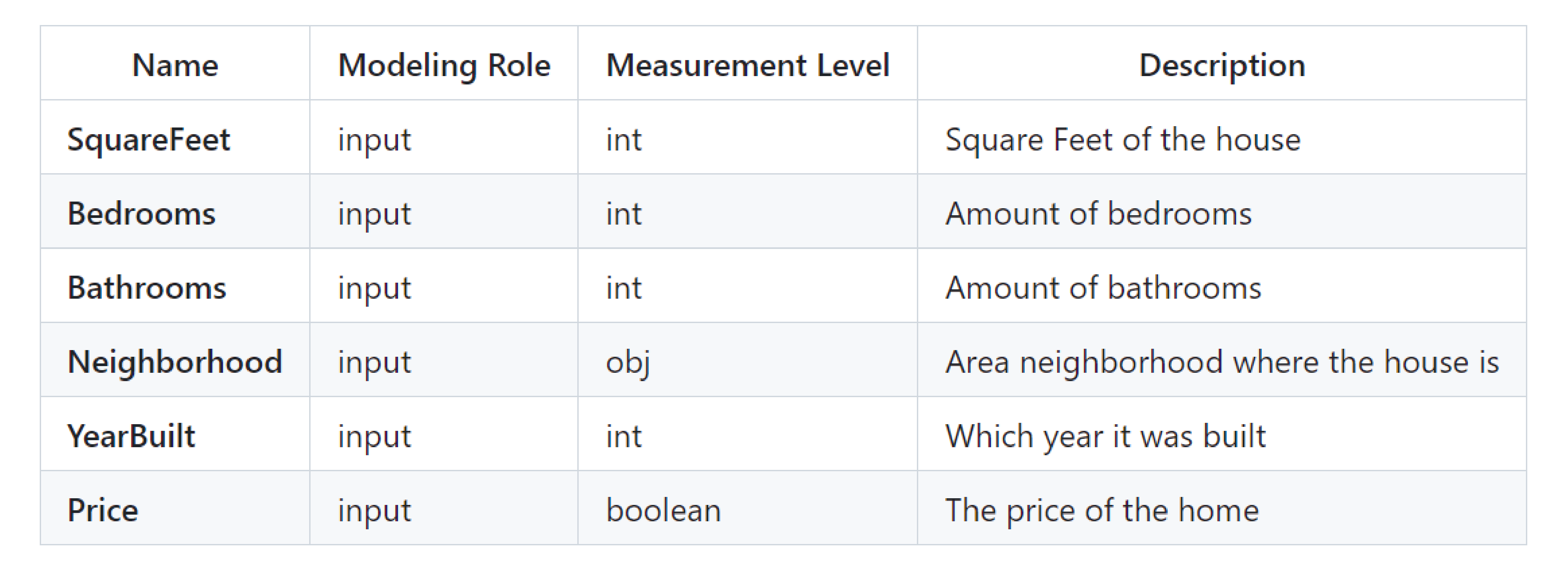
Business Understanding

The goal of this project is to develop a predictive model for housing prices based on various property attributes, including area, number of bedrooms and bathrooms, etc.

Data Understanding

The Diabetes dataset was loaded via Colab. The dataset is from Kaggle: https://www.kaggle.com/datasets/muhammadbinimran/housing-price-prediction-data (also please see housing_price_dataset.csv attached). Basic data analysis was performed to identify the shape of data, get column names, find missing values, and generate descriptive statistics.

- Data Dictionary



Name	Modeling Role	Measurement Level	Description
SquareFeet	input	int	Square Feet of the house
Bedrooms	input	int	Amount of bedrooms
Bathrooms	input	int	Amount of bathrooms
Neighborhood	input	obj	Area neighborhood where the house is
YearBuilt	input	int	Which year it was built
Price	input	boolean	The price of the home

Data Preparation

1 Remove Null
No missing data

2 Define variables

3 Scale X

4 Split train & test data

Modeling

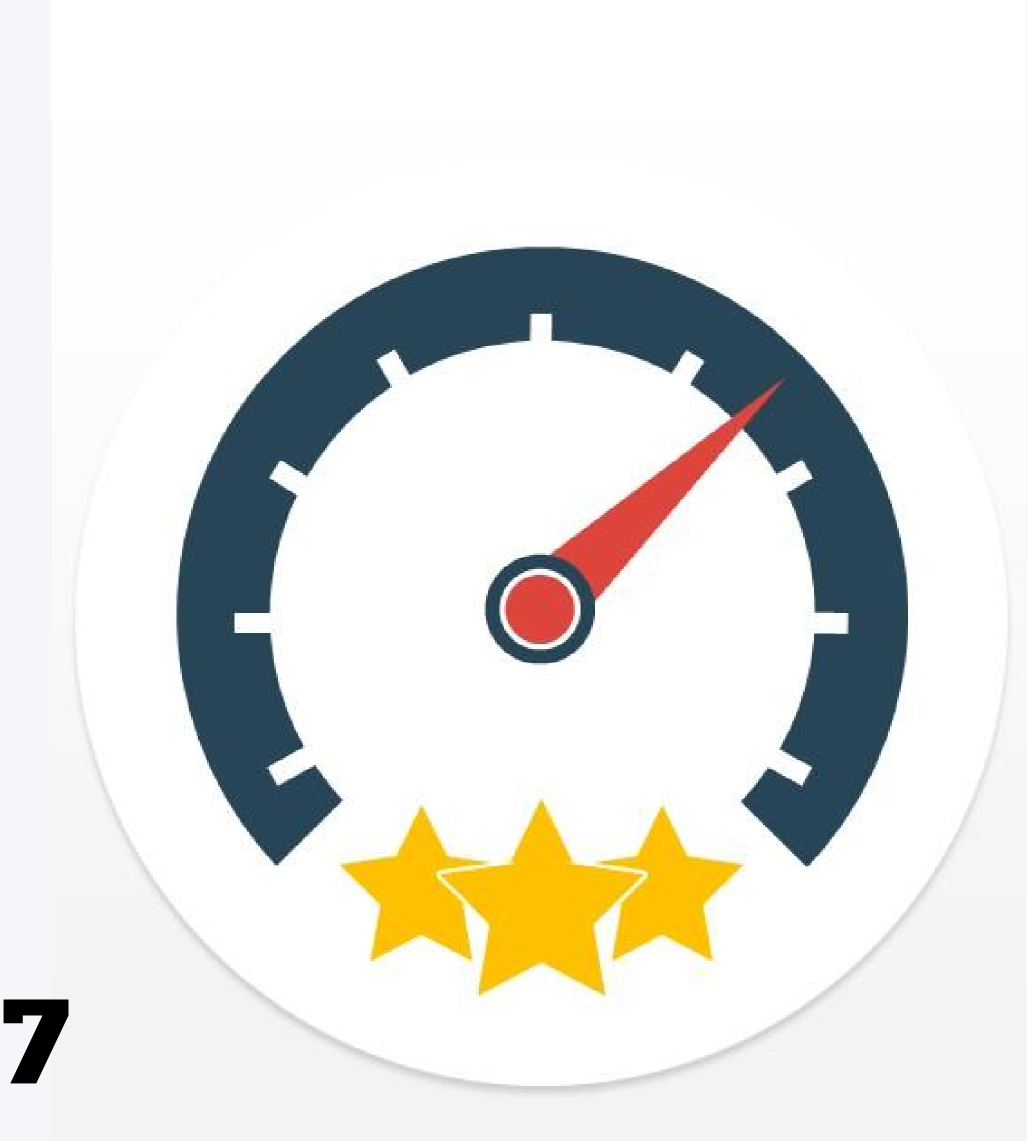
ANCOVA model:

$$
\begin{aligned}
\text{Price} = {} & 57328.725 * (\text{Square Feet}) \\
& + 5780.526 * (\text{Bedrooms}) \\
& + 2340.083 * (\text{Bathrooms}) \\
& + 230.036 * (\text{YearBuilt}) \\
& + (-209.535) * (\text{Neighborhood\_Suburb}) \\
& + 30.847 * (\text{Neighborhood\_Urban}) \\
& + 224727.762
\end{aligned}
$$

Reference group: Neighborhood_Rural

Evaluation

performance on test data



MAPE 25.14%

it indicates the average 25.14% difference between the predicted values and the actual values in the linear model.

R-square 0.57

the model explains 57% of the variance in the dependent variable based on the independent variables

Conclusion

In summary, while the model captures a moderate amount of variation in house prices, it has room for improvement in terms of reducing prediction errors. Further refinements, feature engineering, or considering additional factors could enhance its accuracy for better house price predictions.

Data collection

To perform a better model, data is pivotal. There are two ways to enrich the dataset:

1. include more independent variables
2. include more data points

Machine learning model

Linear regression is a clear and statistical learning technique for inferencing. However, exploring other regression models beyond linear regression, such as regulation or ensemble methods, be ideal. Different algorithms might capture complex relationships better than linear models.