STAT 4355.001- Project Proposal
Group 15

Elshad Jafarli, Pamela Kamdefwere , Natthiya Ngow , Daisy Rueda

Data : House Price Prediction Dataset

INTRODUCTION TO THE DATASET:

We found an interesting dataset from Kaggle about house price prediction in the real estate markets of Sydney and Melbourne. It contains historical housing prices along with various features that may influence price prediction. The dataset contains 2,000 observations and 9 variables, covering houses built between 1900 to 2023.

ANALYSIS GOAL:

Objective of our project is to find which variable(s) are the greatest predictors of house prices and since they clearly exhibit linear behavior we'll use the appropriate linear models for them to extract which are the most and least significant predictors.

DATA VARIABLES:

The dataset we have chosen consists of 18,000 data points with 9 variables(including price) and every variable having at least some kind of linear relationship with the dependent variable which is the house price. The prices of the houses vary greatly from $50,000 up to a $1M

- **Area**

  Area of a house is generally a great predictor for house prices since usually larger the area greater the price. Ranges from 500-5000 square feet

- **Bedrooms**

  Number of bedrooms a house has ranging from 1-5

- **Bathrooms**

  Number of bathrooms a house has ranging from 1-4

- **Floors**

  Number of floors a house has ranging from 1-3

- **YearBuilt**

  The year the house was built ranging from 1900 to 2023

- **Location**

  This variable indicates if the house is in a Downtown, Rural, Suburban or Urban

- **Condition**

  Ranges from Poor-Fair-Good-Excellent

- **Garage**

  This variable is binary classified as either Yes or No and doesn't include how many garages if it's more than 1.

- **Price**

ANALYSIS PLAN:

Objective:

Our goal for this project is to develop a linear regression model to predict house prices using features such as size, location, and quality. We aim to identify the factors that significantly influence house prices and evaluate the accuracy of our predictions.

Data Understanding:

We will use the Kaggle House Price Prediction dataset, which contains variables like living area, lot size, year built, neighborhood, and sale price. Our response variable will be the sale price, while the predictor variables will include both continuous and categorical features. We will thoroughly examine the dataset to understand its structure and identify potential issues, such as missing values and outliers.

Exploratory Data Analysis (EDA):

We will begin by cleaning the data, addressing any missing values and outliers. To understand relationships between variables, we will create visualizations such as scatter plots, histograms, and box plots, and calculate summary statistics. This will help us identify correlations and patterns that are important for building a reliable model.

Model Development:

We will build a multiple linear regression model and carefully check the assumptions, including linearity, normality of residuals, homoscedasticity, and multicollinearity. To select the most significant predictors, we will use methods like stepwise regression. If necessary, we will transform variables to improve model performance and accuracy.

Model Evaluation:

To evaluate the model, we will use metrics such as R-squared to assess the goodness of fit. We will perform residual analysis to check for violations of assumptions and use cross-validation to ensure model stability and generalizability.

Reporting:

Finally, we will interpret the model coefficients and discuss their significance in predicting house prices. We will also address any limitations of the model and propose improvements that could enhance accuracy or robustness in future analyses**.**

RESPONSIBILITIES:

We plan to work on the project together by assigning tasks and collaborating on various processes, including data collection, model building, evaluation, visualization, reporting, and presentation preparation.