Suggested Approach for the "American House Prices" Dataset

Contents

[Introduction 3](#_Toc165619420)

[Data Analysis including Seaborn Visualisations 3](#_Toc165619421)

[Read in the Data: Use pandas to load the CSV file 3](#_Toc165619422)

[Data Cleaning: Handle missing values, remove duplicates, and correct any data inconsistencies. 4](#_Toc165619423)

[Visualization with Seaborn 5](#_Toc165619424)

[Scatter Plot for 'Living Space' vs 'Price' 6](#_Toc165619425)

[Box Plot for 'Beds' vs 'Price' 7](#_Toc165619426)

[Correlation Matrix Heatmap 8](#_Toc165619427)

[Feature Selection 9](#_Toc165619428)

[Model Building and Training 9](#_Toc165619429)

[Conclusions 10](#_Toc165619430)

[Reference 11](#_Toc165619431)

# Introduction

We examine the "American House Prices" dataset in this analysis, which includes housing and demographic information for the top 50 American cities by population. This dataset offers a wealth of information, including demographic information on the population of the zip code and the median household income in addition to property details like the number of bedrooms and living space. Our objective is to apply linear regression to forecast home prices, providing insightful information to a range of stakeholders, including real estate brokers, investors, and legislators. The properties of the dataset will be described, the links between house features and prices will be examined, and a prediction model to estimate property values based on attributes will be developed. We hope to identify the key variables affecting housing values in these densely populated urban regions through this exercise.

# Data Analysis including Seaborn Visualisations

## Read in the Data: Use pandas to load the CSV file

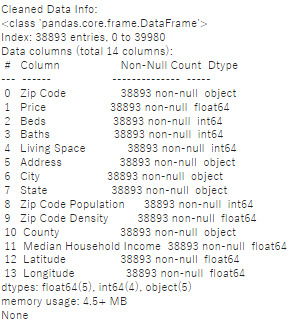
A screenshot of a computer

Description automatically generatedUsing a pandas DataFrame, the accompanying Python code effectively loads housing data from a CSV file, displaying the first five rows but only the first 10 columns to offer a focused view without excessive detail. In addition, it computes and outputs the total number of rows and columns in the DataFrame, providing a brief summary of the size and scope of the dataset. This method helps set up a more thorough examination by providing a basic understanding of the data structure.

## Data Cleaning: Handle missing values, remove duplicates, and correct any data inconsistencies.

A screenshot of a computer program

Description automatically generatedWe first loaded the data and conducted a preliminary analysis to find missing values and data type errors before beginning the data cleaning process for the American House Prices dataset. Specifically, to ensure numerical stability, we filled in any missing values in the "Median Household Income" by using the median. To make additional analysis easier, we additionally converted this column to a numeric form. We corrected any conversion issues by re-filling any gaps with the median value. To preserve the integrity of the dataset, duplicate entries were eliminated. Logical checks were also performed to make sure that there were no negative values in any places where they didn't make sense, such the 'Price' column. In order to keep leading zeros and prevent them from being mistaken for numerical data, zip codes were transformed into strings. Address entries were refined for proper formatting as part of additional cleaning. By ensuring that the dataset is solid and prepared for in-depth analysis and modeling, this thorough cleaning procedure offers a solid basis for precise insights and forecasts.

A white background with black text

Description automatically generatedA screenshot of a computer code

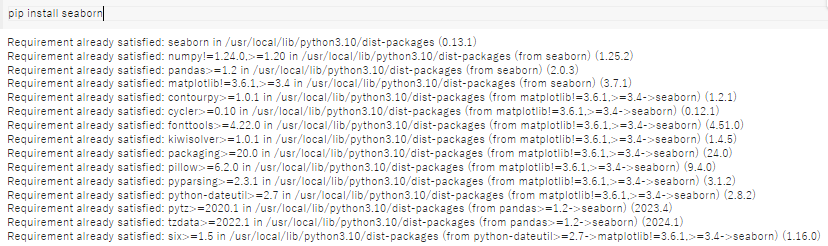
Description automatically generated

A screenshot of a computer

Description automatically generated

## Visualization with Seaborn

**pip install seaborn**

First, ensure you have Seaborn installed:

**Now, create the visualizations**

### Scatter Plot for 'Living Space' vs 'Price'

This scatter plot will help you visualize the relationship between the living space in square feet and the property price

A screenshot of a computer program

Description automatically generated

A graph showing the difference between living space and price

Description automatically generated

### Box Plot for 'Beds' vs 'Price'

The box plot can provide insights into the distribution of prices for different numbers of bedrooms in the properties.

A screenshot of a computer program

Description automatically generated

A graph of a number of beds

Description automatically generated

### Correlation Matrix Heatmap

This heatmap will display the strength of relationships between all numeric features, which is crucial for identifying potential predictors for the target variable 'Price'.

A screenshot of a computer code

Description automatically generated

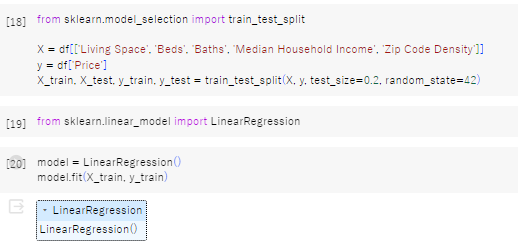
A screenshot of a graph

Description automatically generated

## Feature Selection

Because of their substantial correlations with {Price}, `Living Space}, `Baths}, and {Median Household Income} are the important features found in the correlation matrix analysis for house price prediction. Property size and luxury are indicated by `Living Space` and `Baths`, which are important pricing factors; on the other hand, the economic status of an area is reflected in `Median Household Income}, which has an impact on property values. Potential multicollinearity problems between {Living Space` and `Baths} are indicated by features like `Beds}, indicating that the model has limited additional value. Features with lower correlation, such as {Zip Code Density} and `Zip Code Population}, have less of an influence and are therefore less important to include in a predictive model that prioritizes simplicity and efficacy.

## Model Building and Training



A screenshot of a computer program

Description automatically generated

The above Python code snippet computes the R-squared and Root Mean Squared Error (RMSE) metrics in order to assess a regression model. With an average prediction error of 869509.63, the RMSE suggests that the model's predictions are not particularly reliable. With an R-squared value of roughly 0.286, the model has low predictive potential because it only accounts for about 28.6% of the variability in the objective variable (house prices). These metrics point to the necessity of improving the model's accuracy and explanatory power, perhaps by feature engineering, parameter tuning, or investigating different modeling strategies.

# Conclusions

'Living Space', 'Baths', and 'Median Household Income' were found to be important predictors of house prices in our linear regression analysis of the "American House Prices" dataset. These factors indicate property size, luxury, and neighborhood economic standing. Despite these realizations, the results of our model's performance—an RMSE of 869509.63 and an R-squared of 0.286—indicate that its accuracy and predictive capacity are restricted, indicating the need for additional work. To increase the model's accuracy and give real estate market participants more useful insights, possible enhancements include adding more detailed geographic data, utilizing sophisticated modeling approaches, and augmenting the dataset with new elements.

# Reference

He, Q. (2023, December 28). *Influence Factor Analysis and Forecast of US House Prices Based on Linear Regression and Time Series*. ResearchGate; EWA Publishing. <https://www.researchgate.net/publication/376887806_Influence_Factor_Analysis_and_Forecast_of_US_House_Prices_Based_on_Linear_Regression_and_Time_Series>

Wu, X., & Yang, B. (2022, April). *Ensemble Learning Based Models for House Price Prediction, Case Study: Miami, U.S*. ResearchGate; unknown. <https://www.researchgate.net/publication/365445388_Ensemble_Learning_Based_Models_for_House_Price_Prediction_Case_Study_Miami_US>

O’Farrell, S. (2018). *House Price Prediction*. Academia.edu. <https://www.academia.edu/38358601/House_Price_Prediction>