Module 7: Final Project Template

Title: Housing Price Prediction Model

Author: Nicole Pastrana

Introduction

This project aims to build a model to predict housing prices using linear regression. To achieve this we have tested multiple variable sets and their correlation to sale prices.

The sets created are:

- Variable Set 1: Year Built and/or Remodeled
- Variable Set 2: Bedrooms and Bathrooms
- Variable Set 3: Square Footage
- Variable Set 4: General Home Evaluation
- Variable Set 5: All Previous Variables Combined

The Data

An overview of the data includes:

Rows — Data includes 100 different houses.

Columns — 82 attributes/variables per house (‘SalePrice’ being one of them)

Nulls — We identified and removed attributes that didn’t include sufficient data to be able to use for predictions.

Data Exploration

We'll explore correlations between ***Sale Price*** and the 5 different sets of variables:

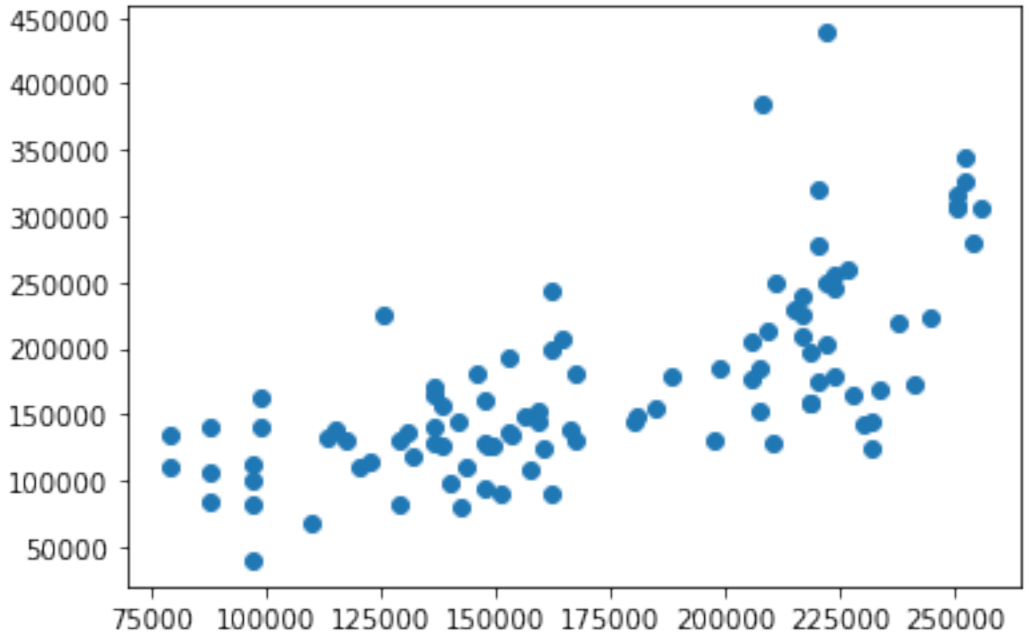
1. Year Built and/or Remodeled
2. Bedrooms and Bathrooms
3. Square Footage
4. General Home Evaluation
5. All Previous Variables Combined

Then, we'll decide what's the best combination of variables to achieve the highest square of the correlation (R^2).

Data Preparation and Correlation — Variable Set 1

Year Built and/or Remodeled

```
R^2 = 0.4753579080438711
```



Relationship between 'SalePrice' and a house's age and whether it has been remodeled or not.

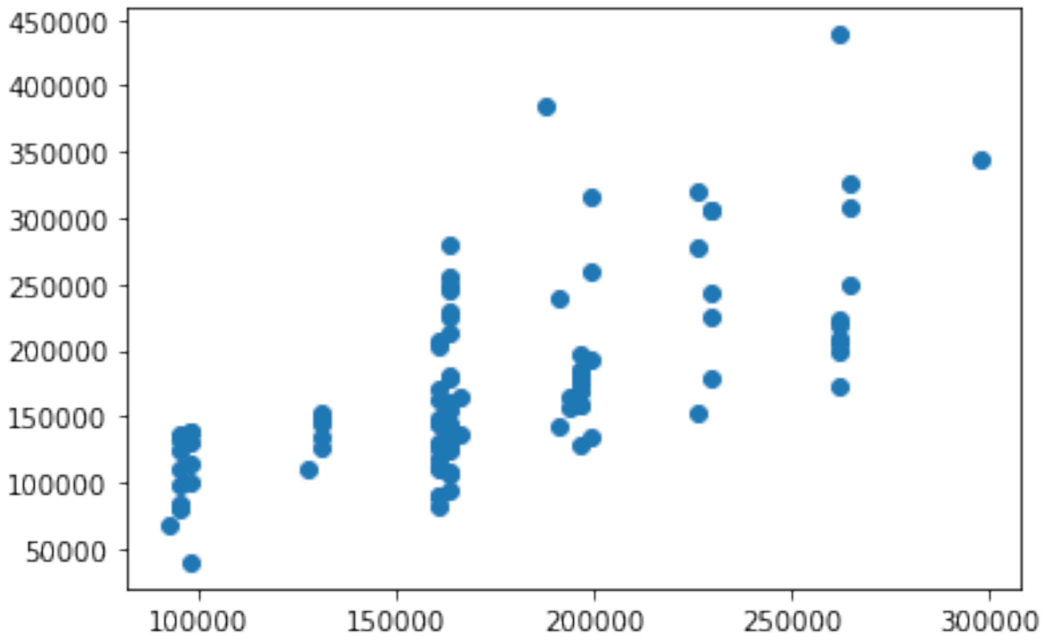
To explore this we:

- Add a 'Remodeled' column with values 1, if 'YearBuilt' and 'YearRemodAdd' are NOT equal, else 0.

Data Preparation and Correlation — Variable Set 2

Bedrooms and Bathrooms

```
R^2 = 0.4546432694700935
```



Relationship between 'SalePrice' and the number of bedrooms/bathrooms, and whether the property has been remodeled or not.

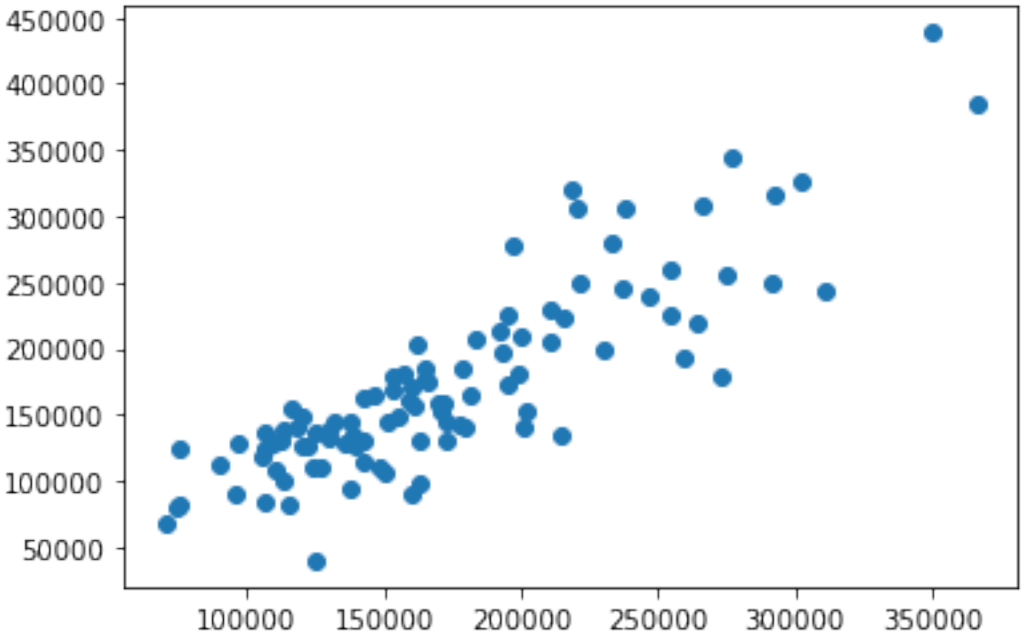
To explore this:

- Values were replaced on the half baths, columns from 1 to 0.5.
- All bathroom columns were added.

Data Preparation and Correlation — Variable Set 3

Square Footage

R^2 = 0.7405774384691417



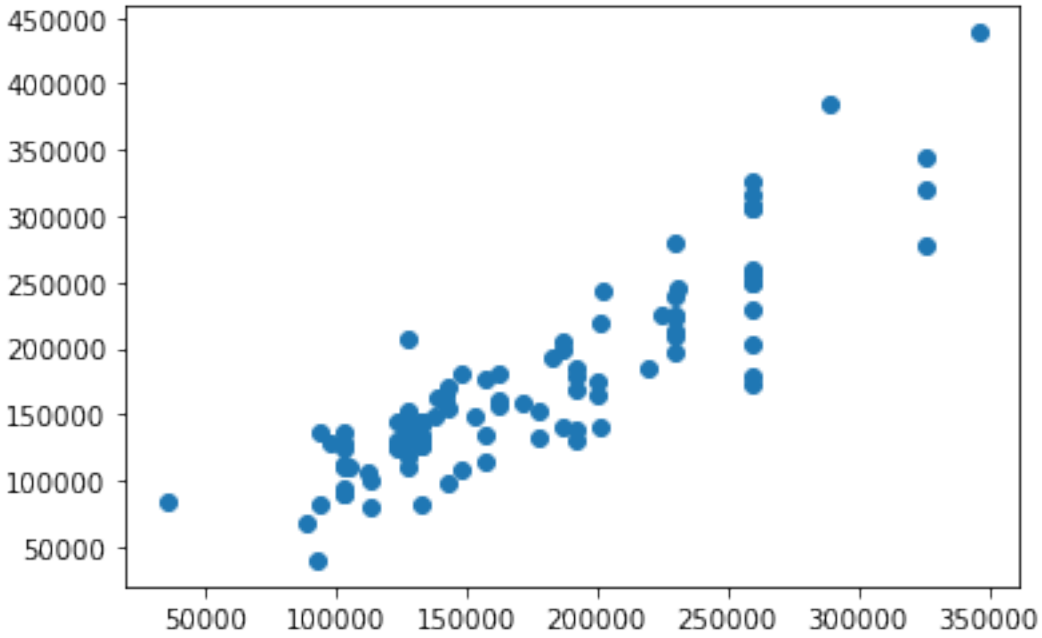
Explores if there's a correlation between a house's square footage and the sale price.
To achieve this:

- All deck and porch square footage was added into a new column 'SqFt_DeckPorch'.

Data Preparation and Correlation — Variable Set 4

General Home Evaluation

R^2 = 0.7791766006753547



Explores if there's a correlation between a homes quality/ condition with its sale price.

To achieve this we:

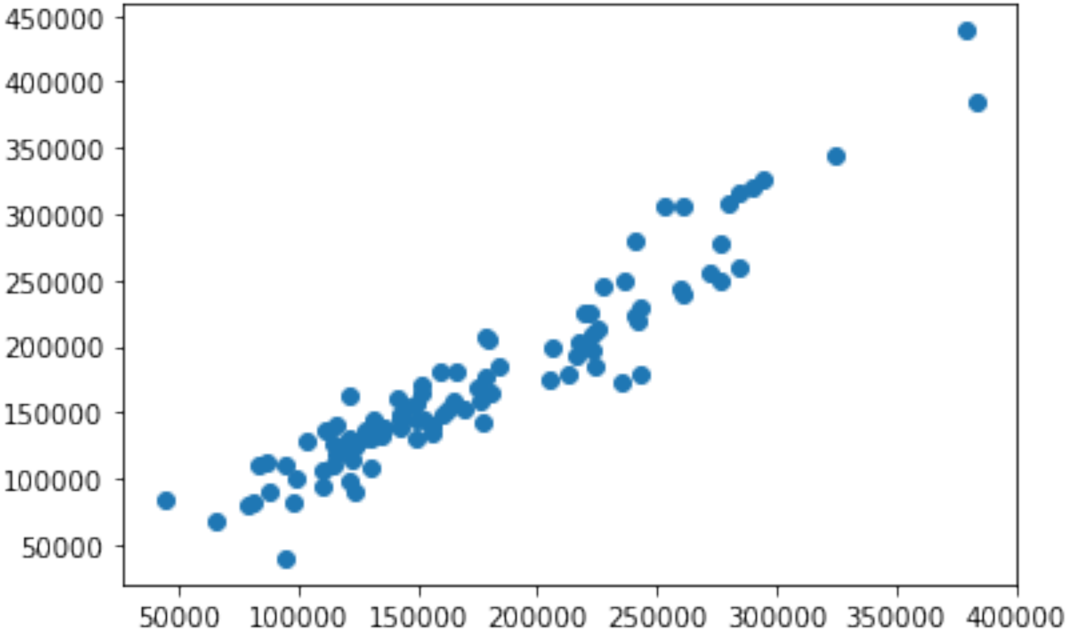
- Replaced string values with integers on some variables.

Data Preparation and Correlation — Variable Set 5

All Previous Variables Combined

```
R^2 = 0.9003659771675001
```

The more numeric data considered, the highest the correlation.



Project Description

Linear regression models the relationship between variables (dependent and independent). Its goal is to predict or forecast results and/or explain variations between them.

The **dependent** variable (Y) in this model is ‘SalePrice'.

The other variables selected are the **explanatory/independent** ones.

Analysis and Results

1. Year Built and/or Remodeled

   Includes 2 variables — Resulted in a ***low*** correlation

2. Bedrooms and Bathrooms

   Includes 5 variables — Resulted in a ***low*** correlation

3. Square Footage

   Includes 10 variables — Resulted in a ***good*** correlation

4. General Home Evaluation

   Includes 5 variables — Resulted in a ***good*** correlation

5. All Previous Variables Combined

   Includes 15 variables — Resulted in ***excellent*** correlation

Analysis and Results

The best correlation in proportion to the number of variables, was Set 4.

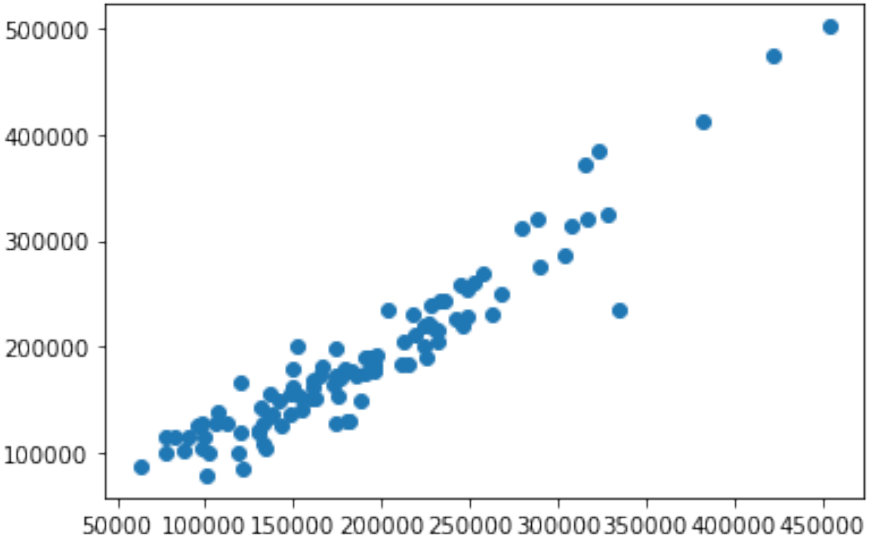
The best correlation overall was Set 5, having the most variables among them all.

We can conclude that transforming given data into a numeric datatypes to be able to have more variables, will further support the correlation.

Verification

Results for the `jtest.csv` dataset are very similar to the ones produced with `houseSmallData.csv`.

The model can be reproduced without glitches.



Scatterplot of all variables modified and combined, using `jtest.csv`.

Conclusion

Our goal was to build a model to predict housing prices using linear regression given certain data. To achieve this we have tested multiple variable sets and their correlation to sale prices, and then tested the combination of the variable sets to achieve the most accurate (highest correlation) results.

We studied the structure of the DataFrame and we cleaned and organized the data to be able to test different variable sets. Finally, we were able to achieve a R-squared result greater than 0.89 when we changed datatypes and combined multiple variables in Variable Set 5.

Through this project, I learned the importance of understanding the data to begin with. Understand how it can be formatted in a way that we can get more information out of it.

References

Department of Statistics and Data Science. “Linear Regression.” Yale University. 1998. http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm .