Model to Predict House Prices

Author: Melissa Belfer

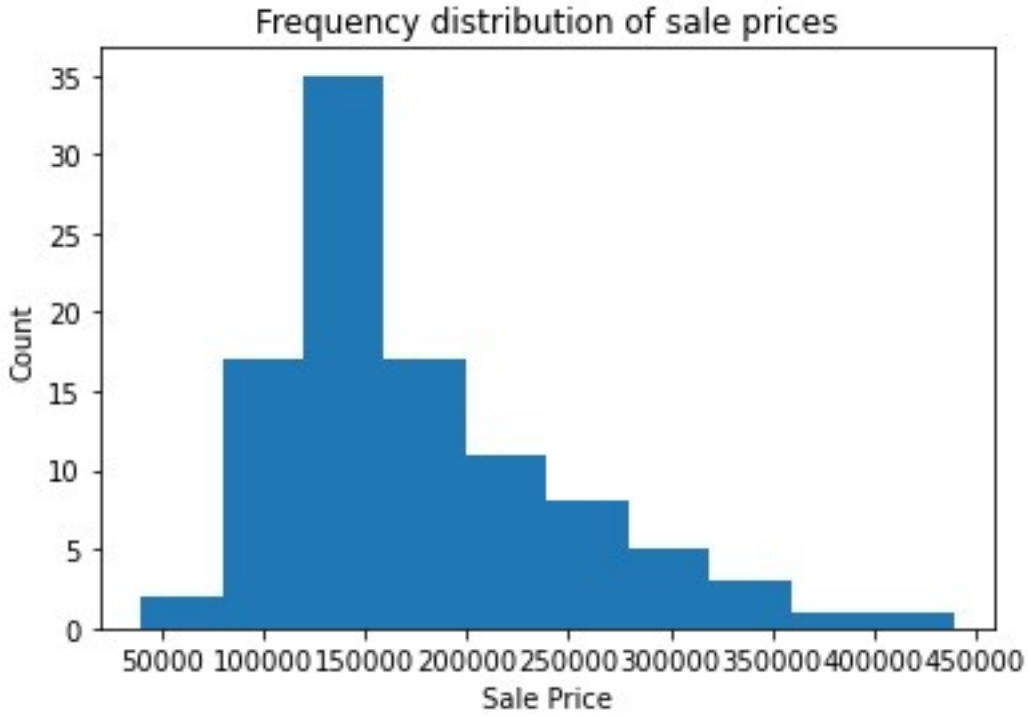
Introduction

The problem I am trying to solve is predicting the sale price of houses.

1. Analyze the dataset
2. Clean the data
3. Find the correlation of different variables against the sale price
4. Train a linear regression model with variables that have a high correlation to sale price
5. Test the different models
6. Find what model has the best score (the best chance of predicting the sale price)

The Data

- The dataset to explore has 82 columns
- The sale price column has a log normal distribution with a 1.17 skew
- The graph below shows the distribution of the sale price column from the training dataset

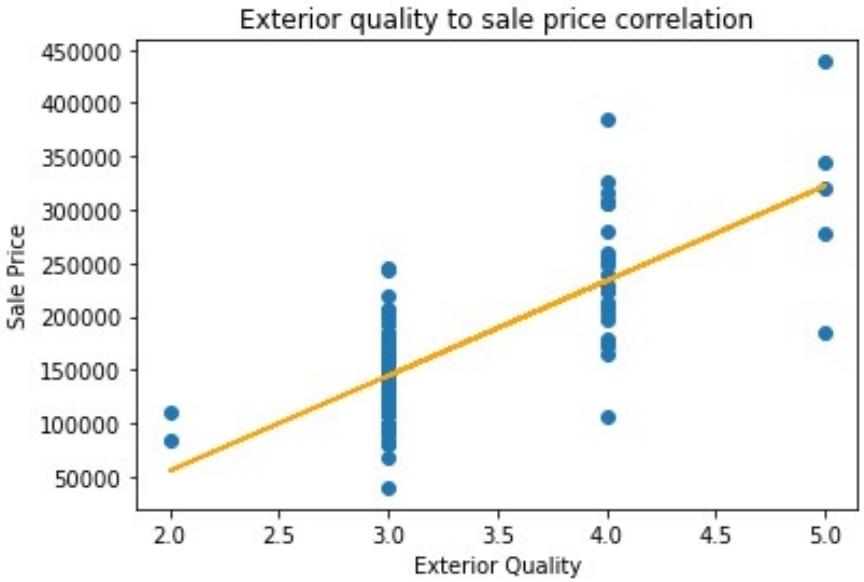


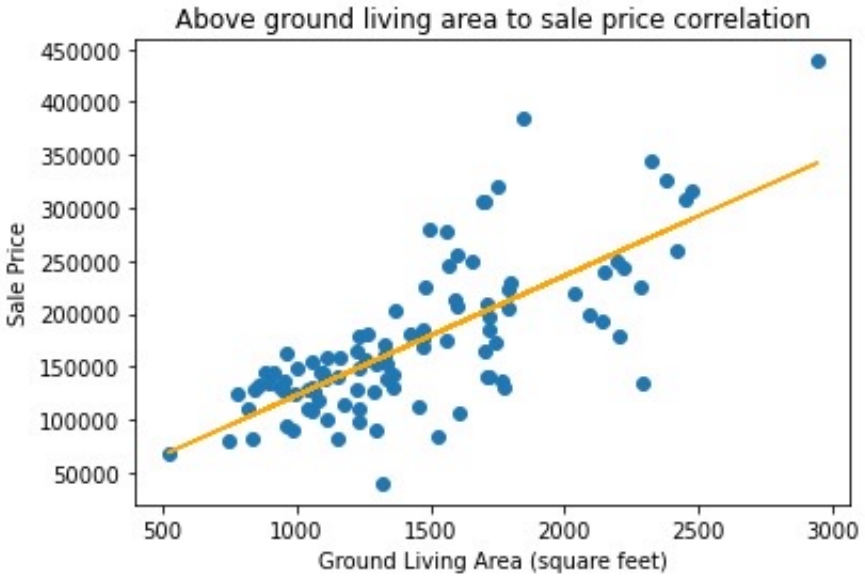
Data Exploration

- The overall quality of the house had the strongest correlation to the sale price
- The graph below shows the relationship between the two columns

Data Exploration

The above ground living area and exterior quality of the house had a strong positive correlation to the sale price



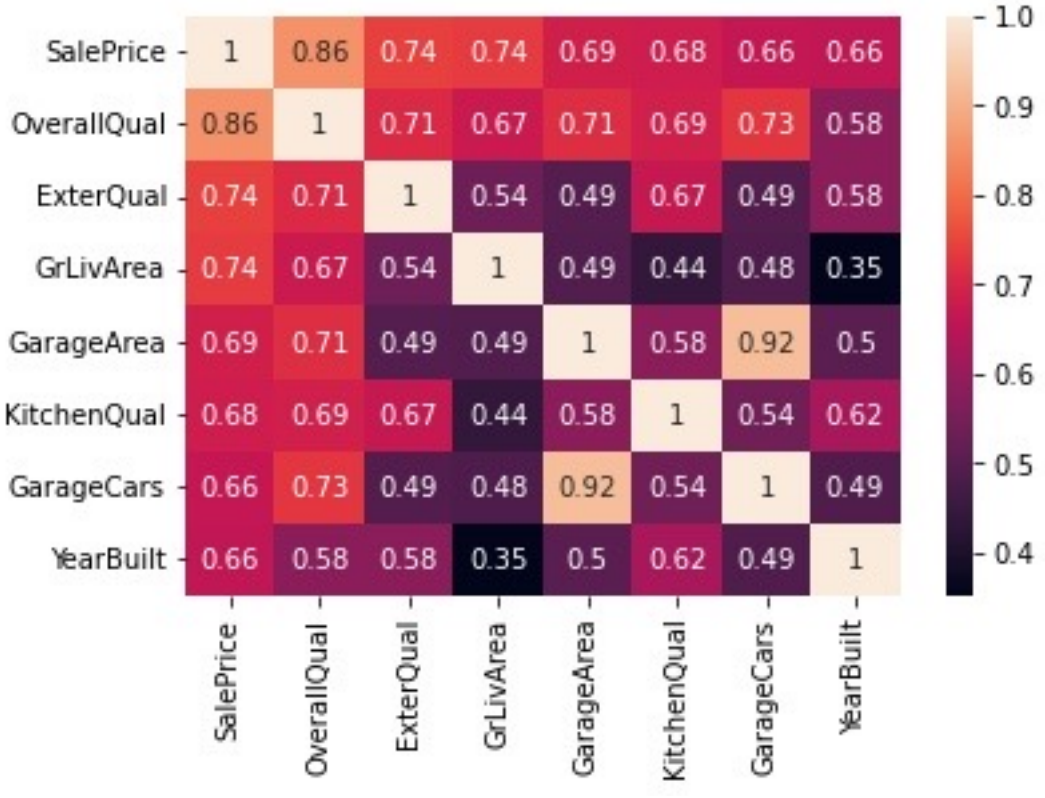


Data Preparation

1. Convert categorical data to a numeric value
    - Convert the exterior quality rating and kitchen quality rating to a 5-star rating system
    - Excellent is equivalent to 5 and poor is equivalent to 1
2. Extract only numeric columns
3. Replace null values with the average
4. Drop columns with only null values

Correlation

- Overall quality of the house has a very strong positive correlation to the sale price of 0.86
- Exterior quality of the house and ground living area in square footage has a strong positive correlation of 0.74 to the sale price.



Project Description

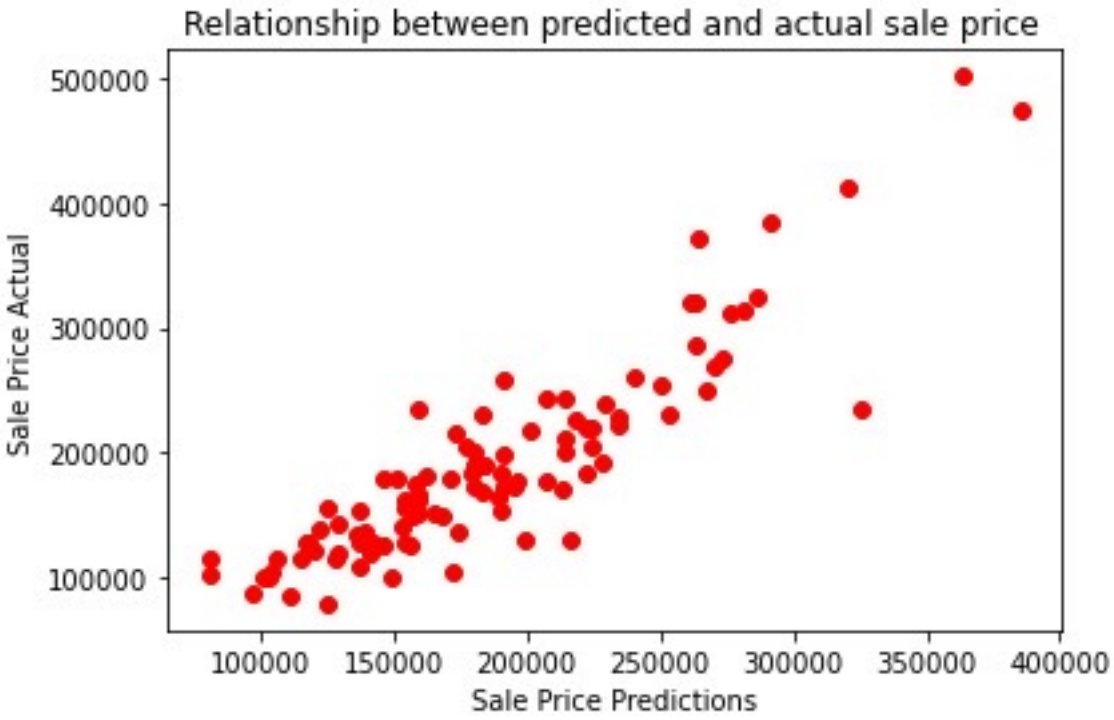
- The columns with the strongest correlation to the sale price are OverallQual, ExterQual, GrLivArea, GarageArea, KitchenQual, GarageCars, YearBuilt, and HeatingQC
- All columns except HeatingQC had a correlation value above 0.65
- Used the linear regression algorithm to predict the sale price column
- Used a variation of the following columns OverallQual, ExterQual, GrLivArea, GarageArea, KitchenQual, GarageCars, YearBuilt, and HeatingQC to create different linear regression models
- Compared each model against the test dataset
- Found the set of independent variables with the highest R squared value

- I analyzed 6 different models with different variations for the independent variables
  - Set 1 = 'OverallQual', 'ExterQual'
  - Set 2 = 'OverallQual', 'ExterQual', 'GrLivArea', 'GarageArea', 'KitchenQual'
  - Set 3 = 'OverallQual', 'ExterQual', 'GrLivArea', 'GarageArea', 'KitchenQual', 'GarageCars'
  - Set 4 = 'OverallQual', 'ExterQual', 'GrLivArea', 'GarageArea', 'KitchenQual', 'GarageCars', 'YearBuilt'
  - Set 5 = 'OverallQual', 'ExterQual', 'GrLivArea', 'GarageArea', 'KitchenQual', 'GarageCars', 'YearBuilt', 'HeatingQC'
  - Set 6 = 'OverallQual', 'ExterQual', 'GrLivArea', 'GarageArea', 'KitchenQual','YearBuilt'

- I discovered increasing the number of independent variables did not mean the models score was going to increase
- The models score on the *training* dataset made a huge jump from set 1 to set 2
  - Set 1 = 0.769
  - Set 2 = 0.823
- The models score went down on the *test* dataset for column set 4 through 6
  - Set 3 = 0.78
  - Set 4-6 = ~0.75
- The models score from column set 4 through 6 stayed consistent
  - Score in the *training* dataset = 0.84
  - Score in the *test* dataset = .75

Verification

- The final columns used for my model were 'OverallQual', 'ExterQual', 'GrLivArea', 'GarageArea', 'KitchenQual', and 'GarageCars'
- The score of the model on the training dataset was ~0.824
- The score on the test dataset was ~0.781



Conclusion

- The final equation for my linear regression model to predict the sale price of a house is the following:

SalePrice = (17665.26545183 * OverallQual) + (25624.99501943 * ExterQual) + (41.07489011 * GrLivArea) + (79.59400665 * GarageArea) + (10179.51947658 * KitchenQual) + (-8875.89573741 * GarageCars) + -132883.48735087246

- I learned that adding more independent variables to the model did not always increase the score of the model
- It sometimes decreased the models score

References

- Dhingra, Deepanshi. "All you need to know about your first Machine Learning model – Linear Regression." Analytics Vidhya. Data Science Blogathon, May 25, 2021. https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/#:~:text=In%20the%20most%20simple%20words,the%20dependent%20and%20independent%20variable.

- Moffitt, Chris. "Guide to Encoding Categorical Values in Python." Practical Business Python. February 06, 2017. https://pbpython.com/categorical-encoding.html.

- Zach. "How to Plot Line of Best Fit in Python (With Examples)." Statology. October 05, 2021. https://www.statology.org/line-of-best-fit-python/.