Module 7: Final Project Template

House Price Estimation
(Improvement vs. Dr. Williams' model)

Author: Stephen Barnes

Introduction

My project improved upon Dr. Williams' model by taking the following steps:

1.) Incorporating a broader range of independent x-variables;

2.) Cleaning the csv file to 'drop' and/or 'impute' data as necessary;

3.) Addressing the issue of multicollinearity;

4.) Calculating correlation and linear regression, and;

5.) Plotting the results in a scatter graph.

The Data

Data Overview

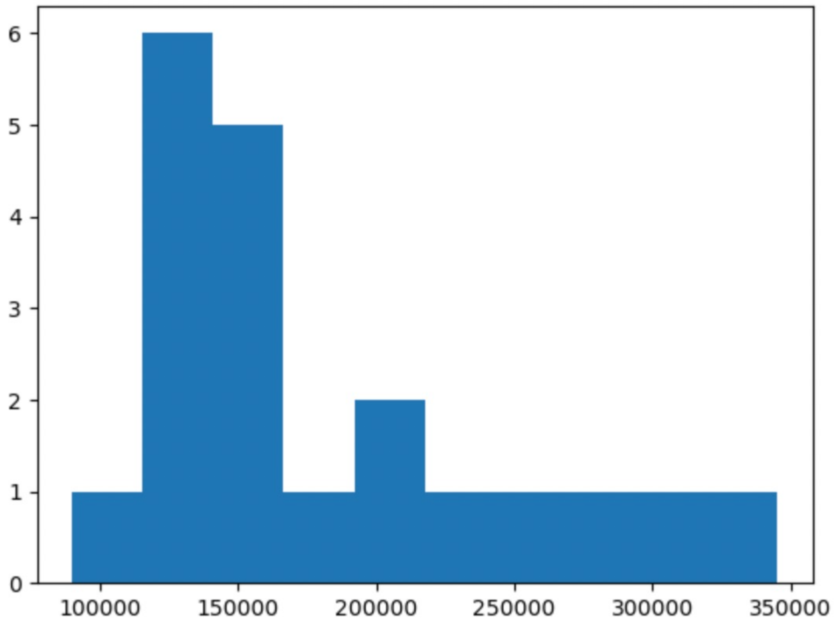
- The project’s data related to home sales, including sale price.
- The project’s goal was to determine the relationship between “SalePrice” and the other house characteristics (independent, x-variables).
- Using multiple linear regression, I predicted house prices based on 14 “cleaned” house characteristics.
- The independent x-variables in my model provided > 96% predictability to “SalePrice”, thereby improving on Dr. Williams’ model.

Data Exploration

The most significant graphs I produced to display the relationships between variables was:

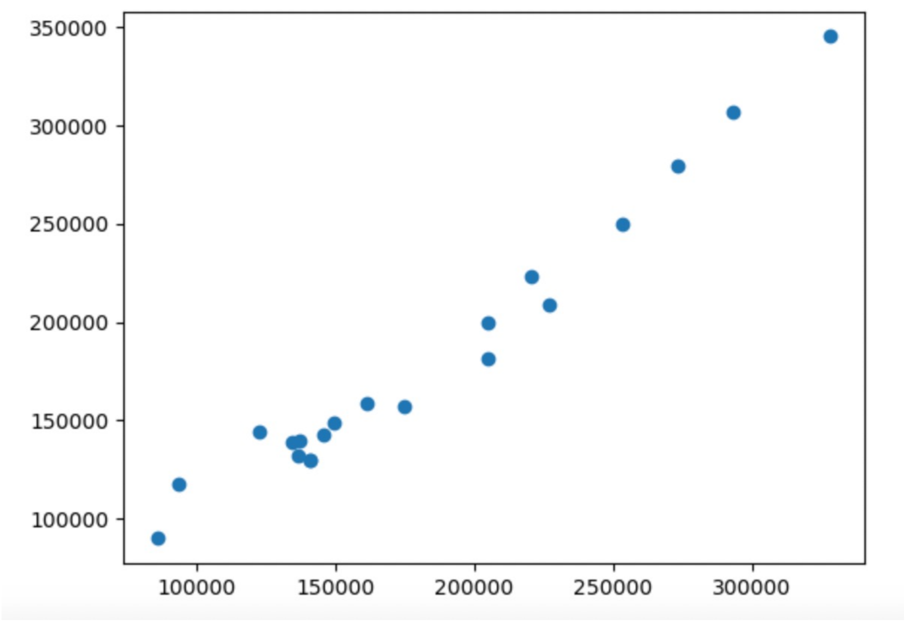
Histogram

Shows where most houses are priced. Starts process to identify variables that place houses in their respective bins.



Scatter Graph

Shows linear relationship between the house characteristics in my model and their impact on sale price.



Cleaning the data:

1.) Identify columns with missing data and impute them;

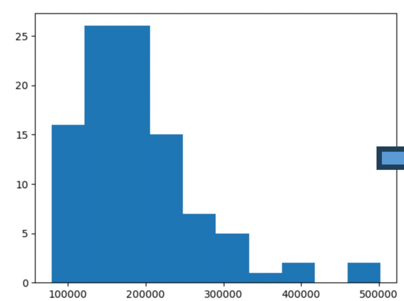
```
Columns with missing values: ['LotFrontage', 'Alley', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFi
nType2', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature']
```

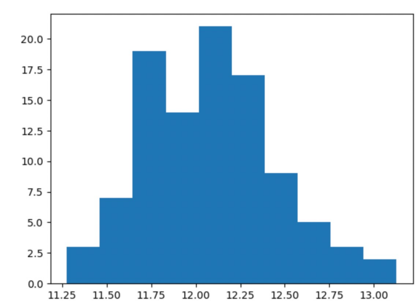
2.) Address multicollinearity

```
SalePrice       1.000000
OverallQual     0.807380
MasVnrArea      0.788274
FullBath        0.721954
TotRmsAbvGrd    0.699634
YearBuilt       0.699627
YearRemodAdd    0.698731
GarageArea      0.696998
BedroomAbvGr    0.681291
GrLivArea       0.676909
TotalBsmtSF     0.651318
GarageYrBlt     0.649557
LotFrontage     0.593996
WoodDeckSF      0.575730
GarageCars      0.571377
```

3). Deal with skewed data

Create a histogram and, if the data is skewed, re-run the histogram using the log of your variable (e.g. SalePrice in our case).





Correlation

Summary of correlation results:

1. The table below shows the house characteristics most correlated to “SalePrice”.

SalePrice	1.000000
OverallQual	0.807380
MasVnrArea	0.788274
FullBath	0.721954
TotRmsAbvGrd	0.699634
YearBuilt	0.699627
YearRemodAdd	0.698731
GarageArea	0.696998
BedroomAbvGr	0.681291
GrLivArea	0.676909
TotalBsmtSF	0.651318
GarageYrBlt	0.649557
LotFrontage	0.593996
WoodDeckSF	0.575730
GarageCars	0.571377

2. I predicted house prices based on “cleaned” variables with more than 50% correlation to “SalePrice”.

3. My project highlights the importance of:

- including multiple independent ‘x’ variables to predict a dependent ‘y’ variable
- including strongly correlated variables
- excluding variables that have multicollinearity

Project Description

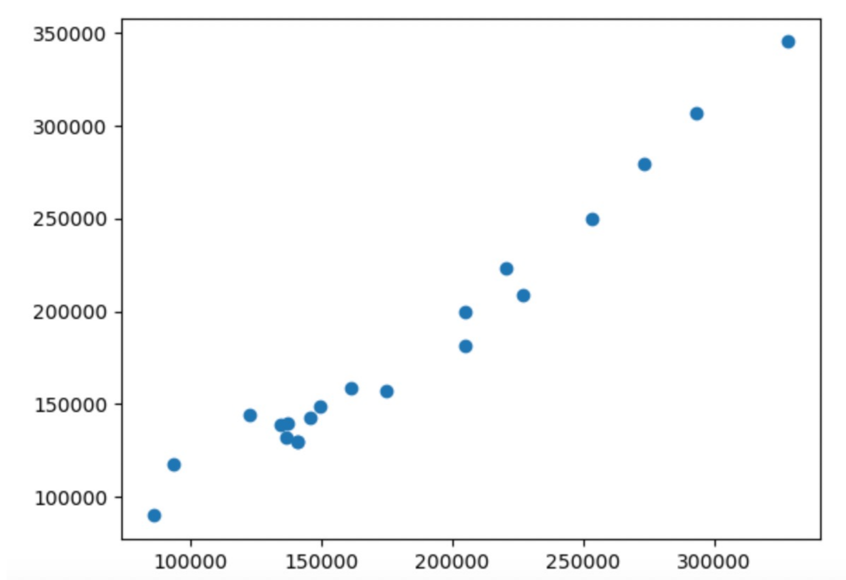
- The most important algorithm used in this project was Multiple Linear Regression.

- LMR is a predictive modeling approach for estimating the relationship between several independent variables and one dependent variable.

<u>Once we determine correlation</u>

```
SalePrice       1.000000
OverallQual     0.807380
MasVnrArea      0.788274
FullBath        0.721954
TotRmsAbvGrd    0.699634
YearBuilt       0.699627
YearRemodAdd    0.698731
GarageArea      0.696998
BedroomAbvGr    0.681291
GrLivArea       0.676909
TotalBsmtSF     0.651318
GarageYrBlt     0.649557
LotFrontage     0.593996
WoodDeckSF      0.575730
GarageCars      0.571377
```

<u>Use a scatter graph to visualize relationships</u>



- The variables I used for my model were those with > 50% correlation to ”SalePrice”

SalePrice	1.000000
OverallQual	0.807380
MasVnrArea	0.788274
FullBath	0.721954
TotRmsAbvGrd	0.699634
YearBuilt	0.699627
YearRemodAdd	0.698731
GarageArea	0.696998
BedroomAbvGr	0.681291
GrLivArea	0.676909
TotalBsmtSF	0.651318
GarageYrBlt	0.649557
LotFrontage	0.593996
WoodDeckSF	0.575730
GarageCars	0.571377

- However, I cleaned the data by imputing missing values and removed variables with possible multicollinearity before calculating final correlation and linear regression

Verification

- My model performed well with test data.
- The data itself needed a lot of cleaning and was skewed, which I corrected using the log of SalePrice.
- I ran into several challenges using test data but, in the end, I got my python code to execute the prediction model.

(* Please note that I used GPT4 extensively to figure out why I kept getting errors)

Conclusion

- My project successfully implemented a step-by-step analysis from data preparation to feature selection, and model training to evaluation.

- The model achieved an impressive score of slightly over 96%, confirming its reliability and robustness in predicting house prices – and outperforming Dr. Williams' model from Module 7.

<u>Key Take-Aways</u>:

1.) It's very important to include multiple independent 'x' variables to predict the dependent 'y' variable.

2.) A good model should always include clean, strongly correlated variables (e.g. for our analysis: OverallQual, MasVnrArea, and FullBath to SalePrice).

3.) Make sure the data is clean (e.g. remove variables that have multicollinearity)

References

The only reference material I used was provided by:

- OpenAI’s GPT4 (https://openai.com) – used to help find code to clean the data, regardless of data type.

*Please add slides as required.