Predicting Home Prices

By: Corina Lentz

Table Of Contents

- Problem Statement
- Cleaning & Processing
- Exploratory Data Analysis
- The Model
- The Results
- Conclusion & Recommendations
- Resources

Problem Statement

The real estate market climbed to $36.2 trillion in 2020 [1]. In 2021 the demand for homes has only increased [2]. With this booming real estate market, predicting how much a home will sell for with a reasonable level of accuracy is more important than ever. That is where Lentz Data Analysis comes in. We will use the Ames Iowa Housing dataset to build a multiple linear regression model that can predict the sale price of homes. We believe the most accurate model will include the overall quality of the home, the home type, and the number of bedrooms/bathrooms, along with measurements of the size and quality of any basement or garage features the home might have.

Cleaning & Processing

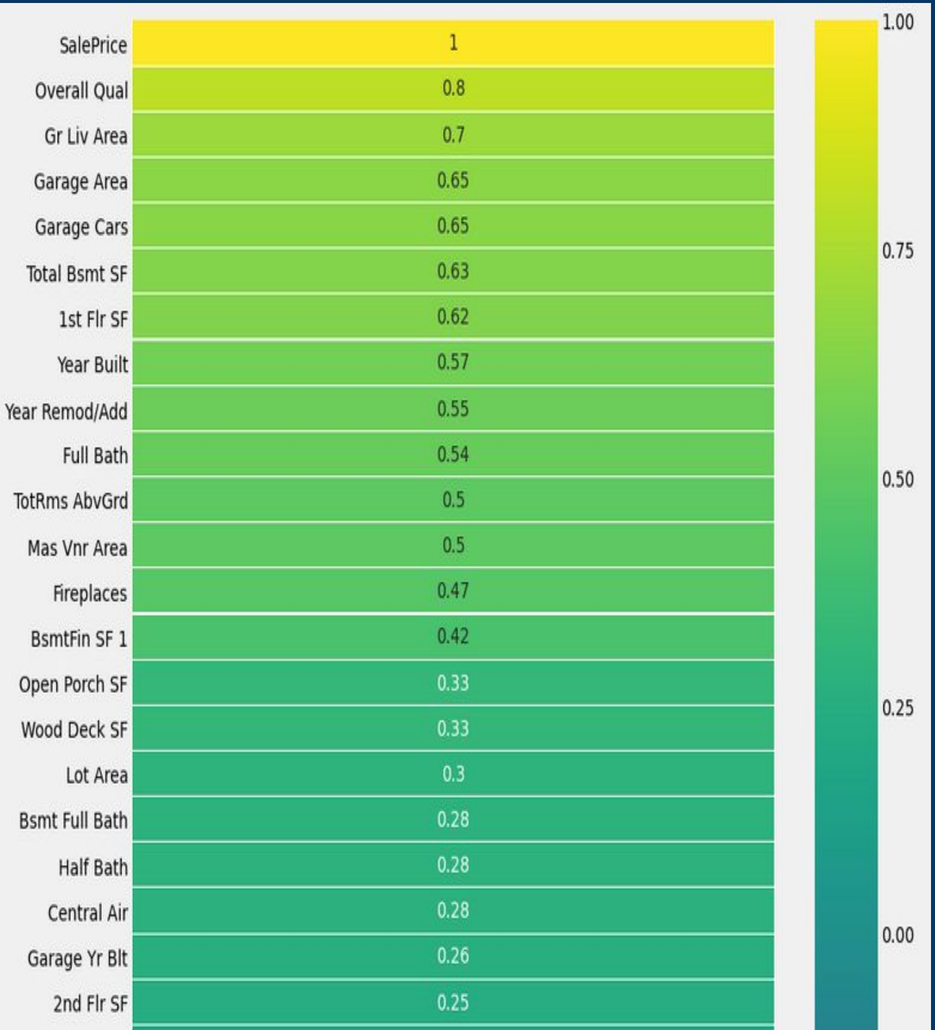
- Updated missing values to indicate that the home doesn’t have the given feature (where appropriate).

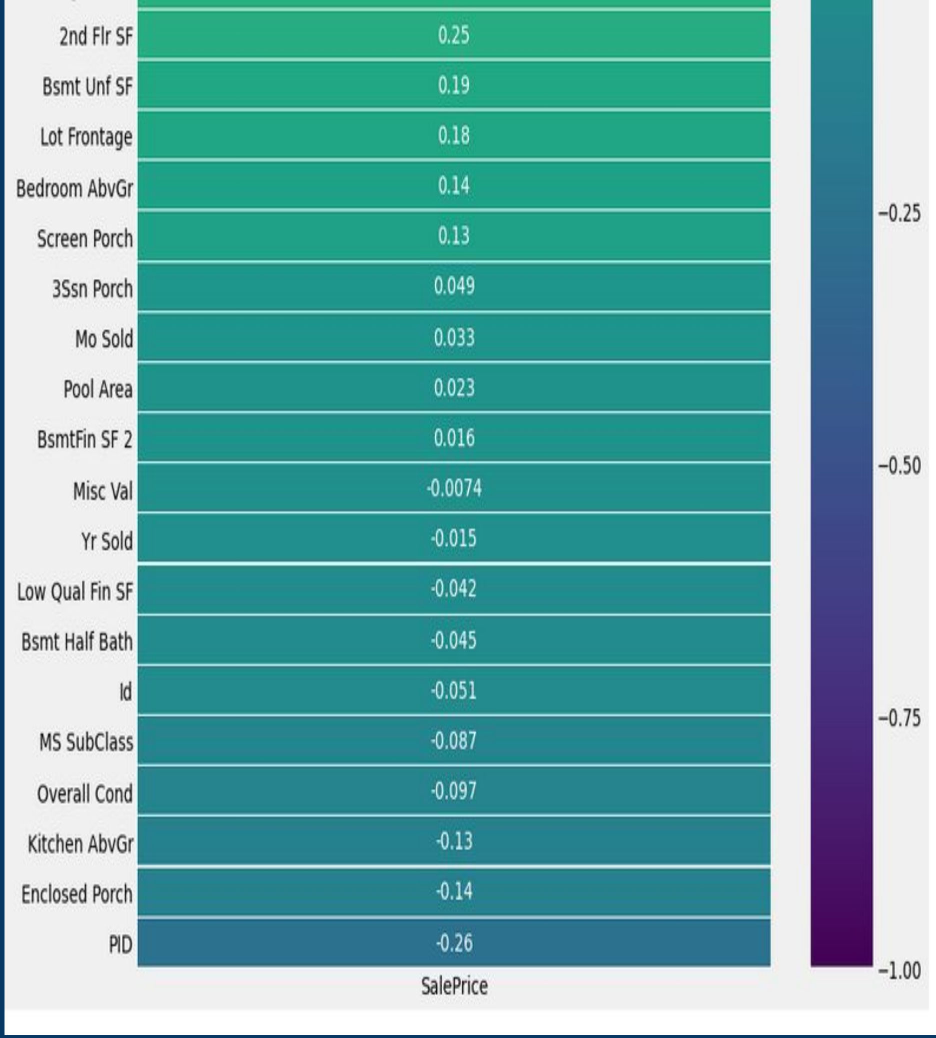
- Dropped columns that had less than 20% of non-null values (‘Alley’, ‘Pool QC’, ‘Fence’, and ‘Misc Feature’).

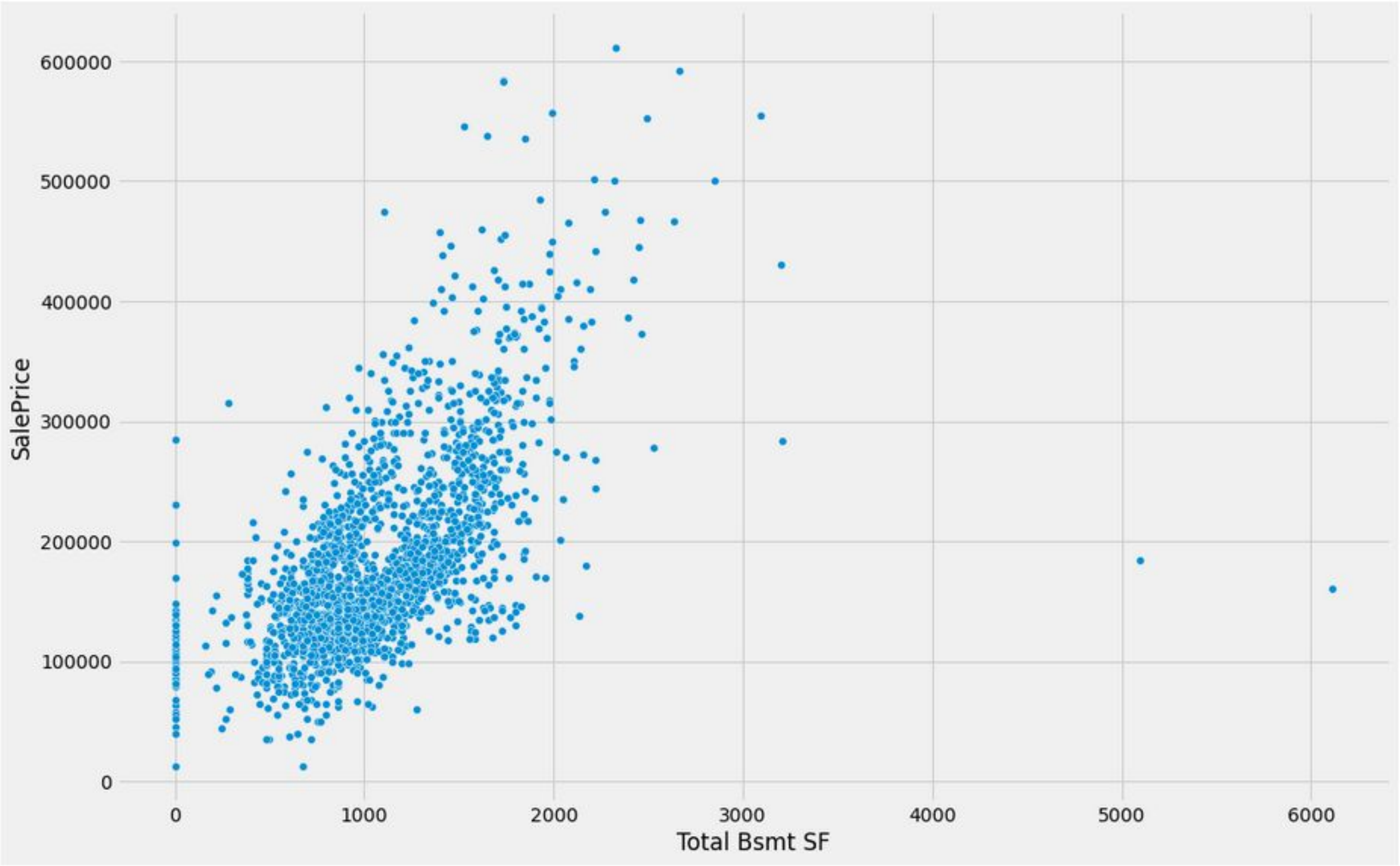
- Changed the ‘Central Air’ column to binary.

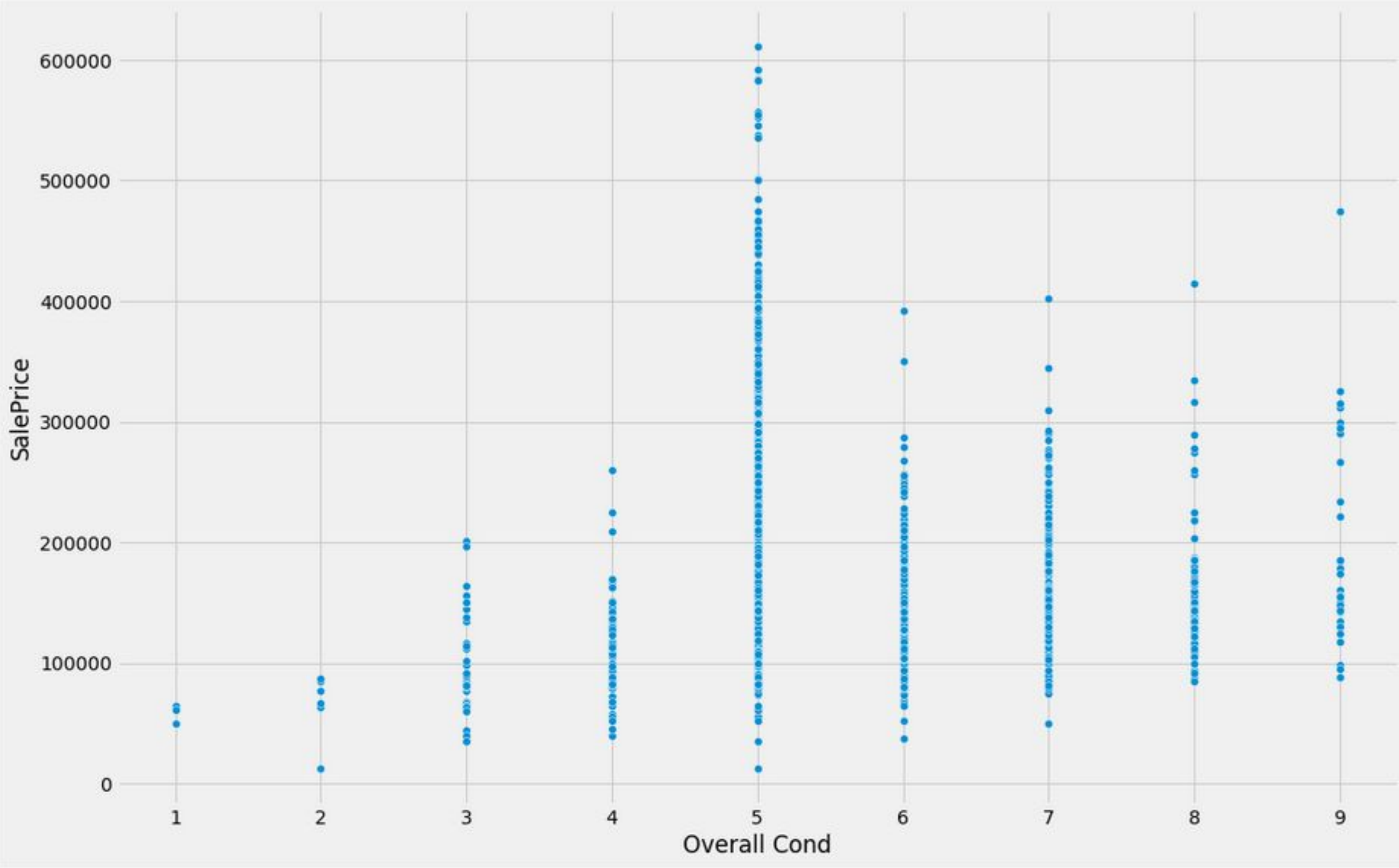
- One-hot encoded categorical columns for the model ('Overall Qual', 'Bsmt Cond', 'BsmtFin Type 1', 'BsmtFin Type 2', 'Fireplace Qu', 'Garage Type', 'Garage Finish', 'Garage Cond', and 'Garage Qual')

Exploratory Data Analysis


SalePrice 1
Overall Qual 0.8
Gr Liv Area 0.7
Garage Area 0.65
Garage Cars 0.65
Total Bsmt SF 0.63
1st Flr SF 0.62
Year Built 0.57
Year Remod/Add 0.55
Full Bath 0.54
TotRms AbvGrd 0.5
Mas Vnr Area 0.5
Fireplaces 0.47
BsmtFin SF 1 0.42
Open Porch SF 0.33
Wood Deck SF 0.33
Lot Area 0.3
Bsmt Full Bath 0.28
Half Bath 0.28
Central Air 0.28
Garage Yr Blt 0.26
2nd Flr SF 0.25
1.00
0.75
0.50
0.25
0.00


2nd Flr SF 0.25
Bsmt Unf SF 0.19
Lot Frontage 0.18
Bedroom AbvGr 0.14
Screen Porch 0.13
3Ssn Porch 0.049
Mo Sold 0.033
Pool Area 0.023
BsmtFin SF 2 0.016
Misc Val -0.0074
Yr Sold -0.015
Low Qual Fin SF -0.042
Bsmt Half Bath -0.045
Id -0.051
MS SubClass -0.087
Overall Cond -0.097
Kitchen AbvGr -0.13
Enclosed Porch -0.14
PID -0.26
SalePrice
−0.25
−0.50
−0.75
−1.00


600000
500000
400000
300000
200000
100000
0
SalePrice
0
1000
2000
3000
4000
5000
6000
Total Bsmt SF


SalePrice
600000
500000
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
Overall Cond

In choosing the variables for our model, we used the relationships we discovered during EDA (particularly through the heatmap) to select features that would have the most impact; while also focusing on features that aligned with our initial hypothesis. We selected features that:

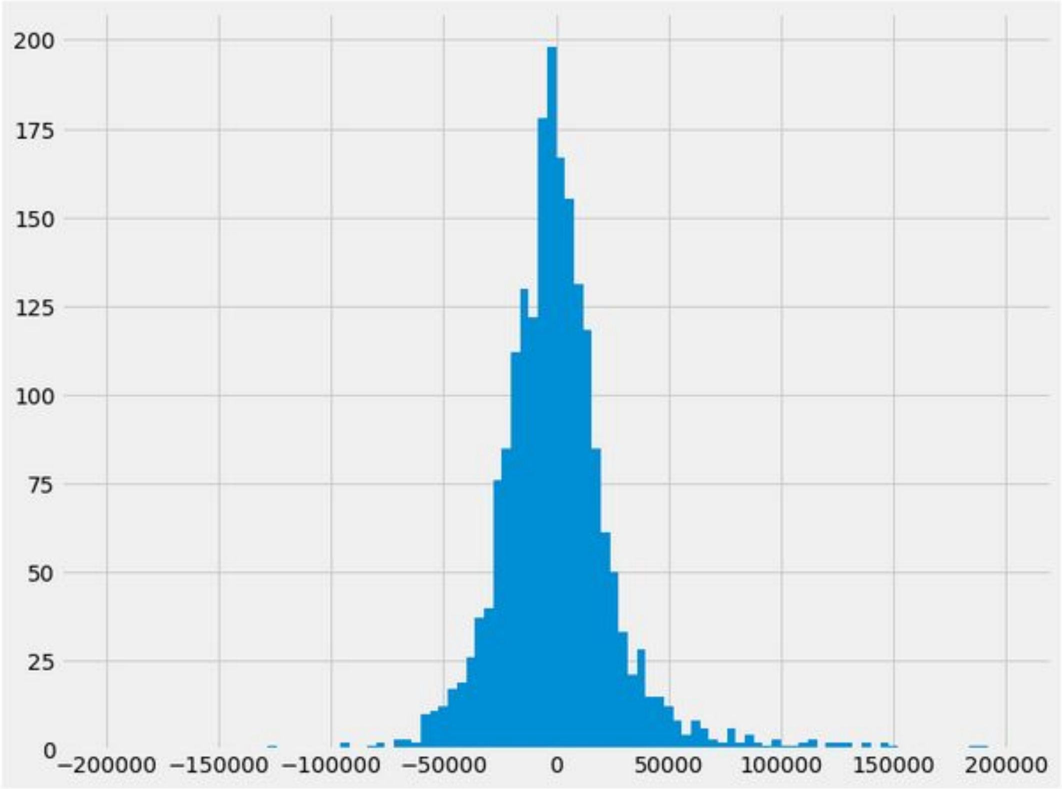
- Measure the overall quality of the home
- Indicate the home type
- Measure the number of bathrooms/bedrooms
- Measure the overall size of the home
- Measure any additional home features (garage, basement, etc)
- Measure the quality of any additional home features

In the end we had 18 numerical features and 9 categorical features (that we processed through one-hot encoding), for a total of 68 feature columns that were used in our multiple linear regression model.

The Model

The Results

- Train Score- 0.8492
- Test Score- 0.8734
- R^2- 0.85501
- RMSE- 30172.34



Conclusion & Recommendations

- The features that we included in our model do have measurable impact on the sale price of the home, however they do not provide the most complete or accurate calculations of the home sale price.

- We recommend using this model to develop an even more accurate model by dropping outliers from the data, possibly removing some of the redundant features (such as ‘Total Bsmt SF’ when we have ‘Bsmt Fin SF’) and adding in other features that may offer better insight such as measurements of the masonry and decks included in the home.

Resources

1. https://www.forbes.com/sites/brendarichardson/2021/01/26/housing-market-gains-more-value-in-2020-than-in-any-year-since-2005/

2. https://www.cnn.com/2021/06/16/homes/us-housing-market-offers/index.html

Data Dictionary: https://www.kaggle.com/c/dsir-524-project-2-regression-challenge/data