Determining Home Prices

Case of Ames, IA

Ruchika Sah

Introduction

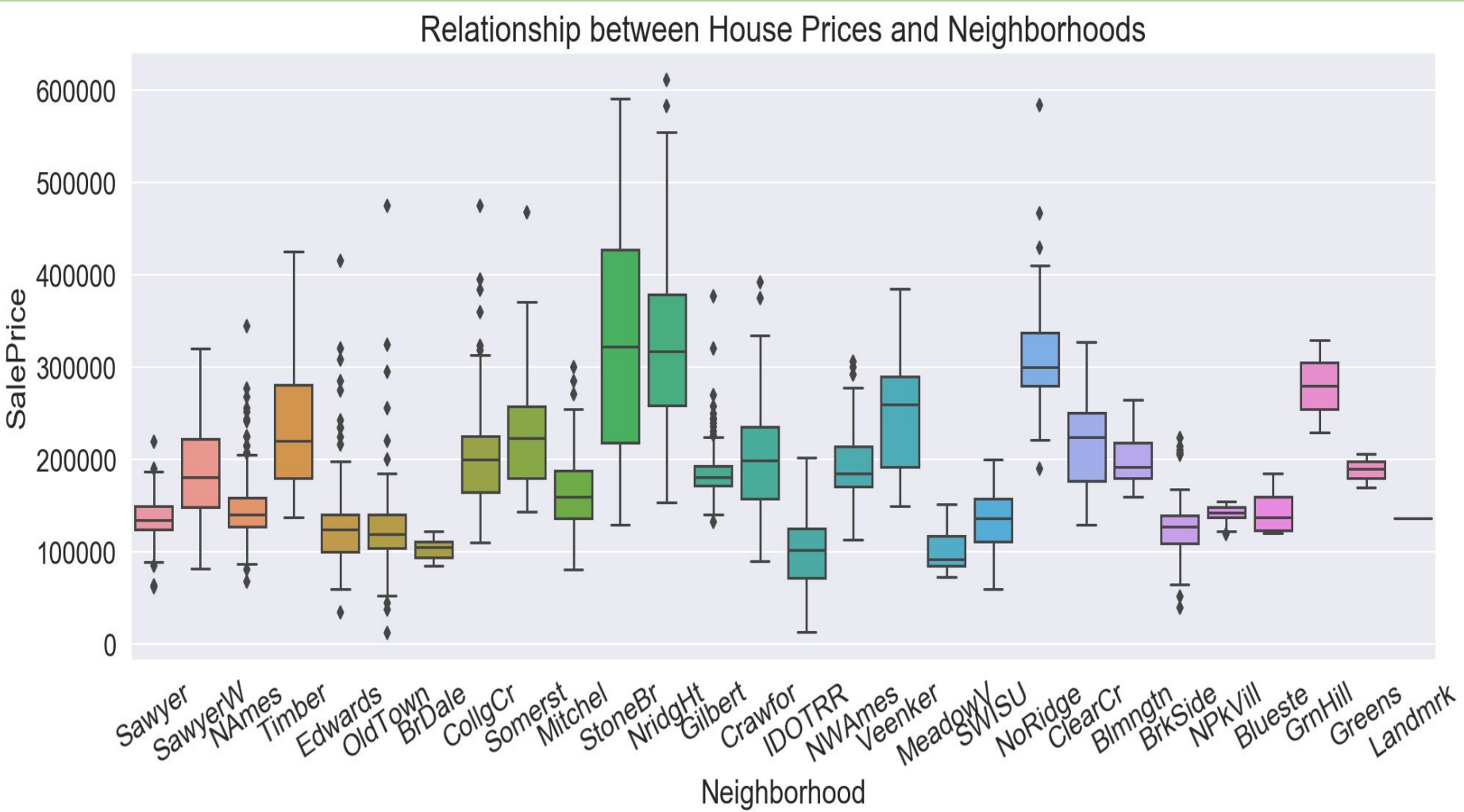
Big, Better, Best!

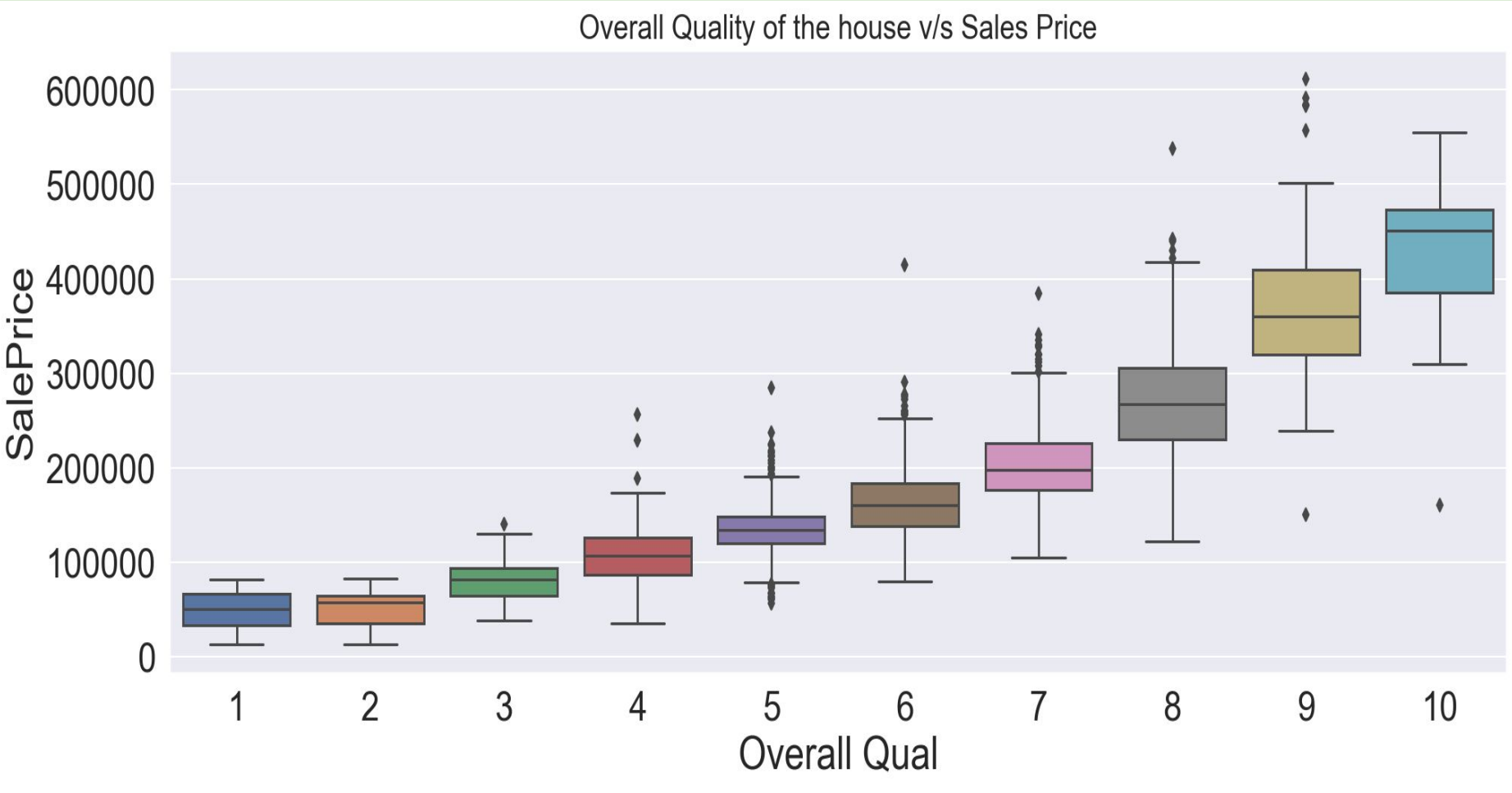
- Location, Location, Location!
- First Impressions
- Space
- What is the age?
- Upgrades and Updates

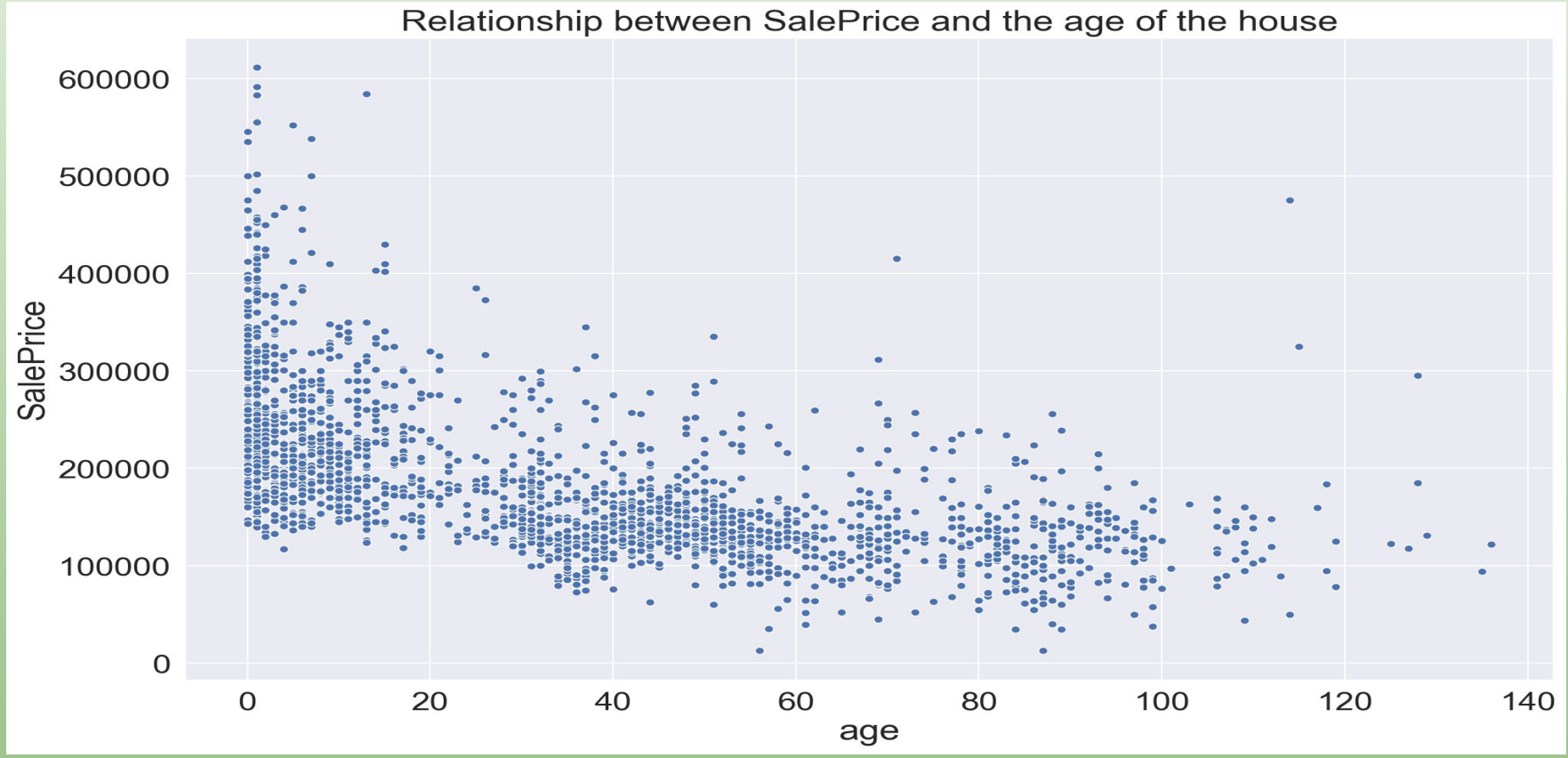
The Data: Initial Observations

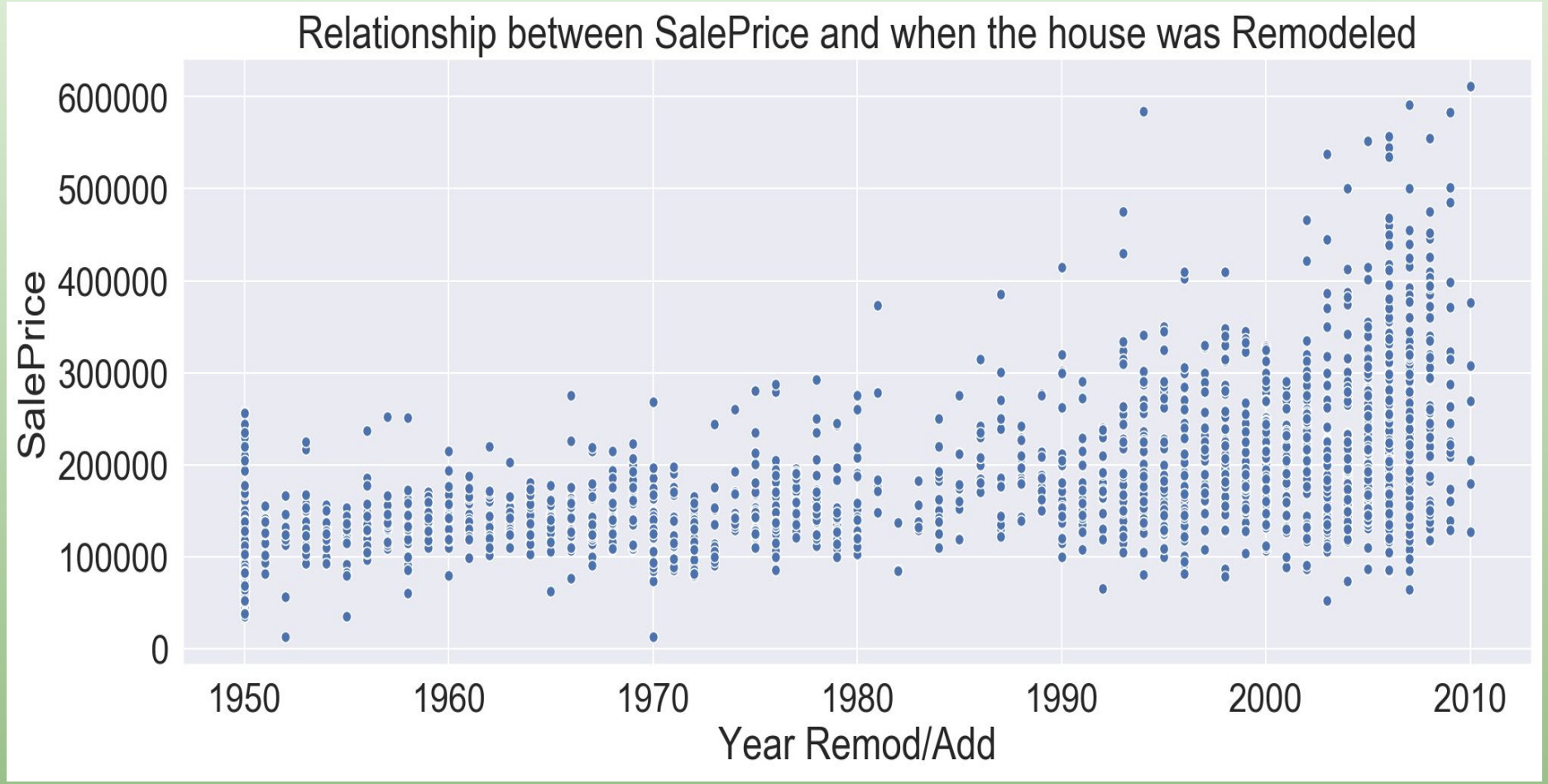
- The Dataset has 81 variables. Of these there are 20 continuous, (excluding our target, the Sales price) 23 Nominal, 23 Ordinal and 14 discrete variables.
- Some of the variables have ‘NA’ which is actually a category and does not imply missing data.
- The majority of the data is not numerical.
- The data is divided into two sets - The training set and the testing set

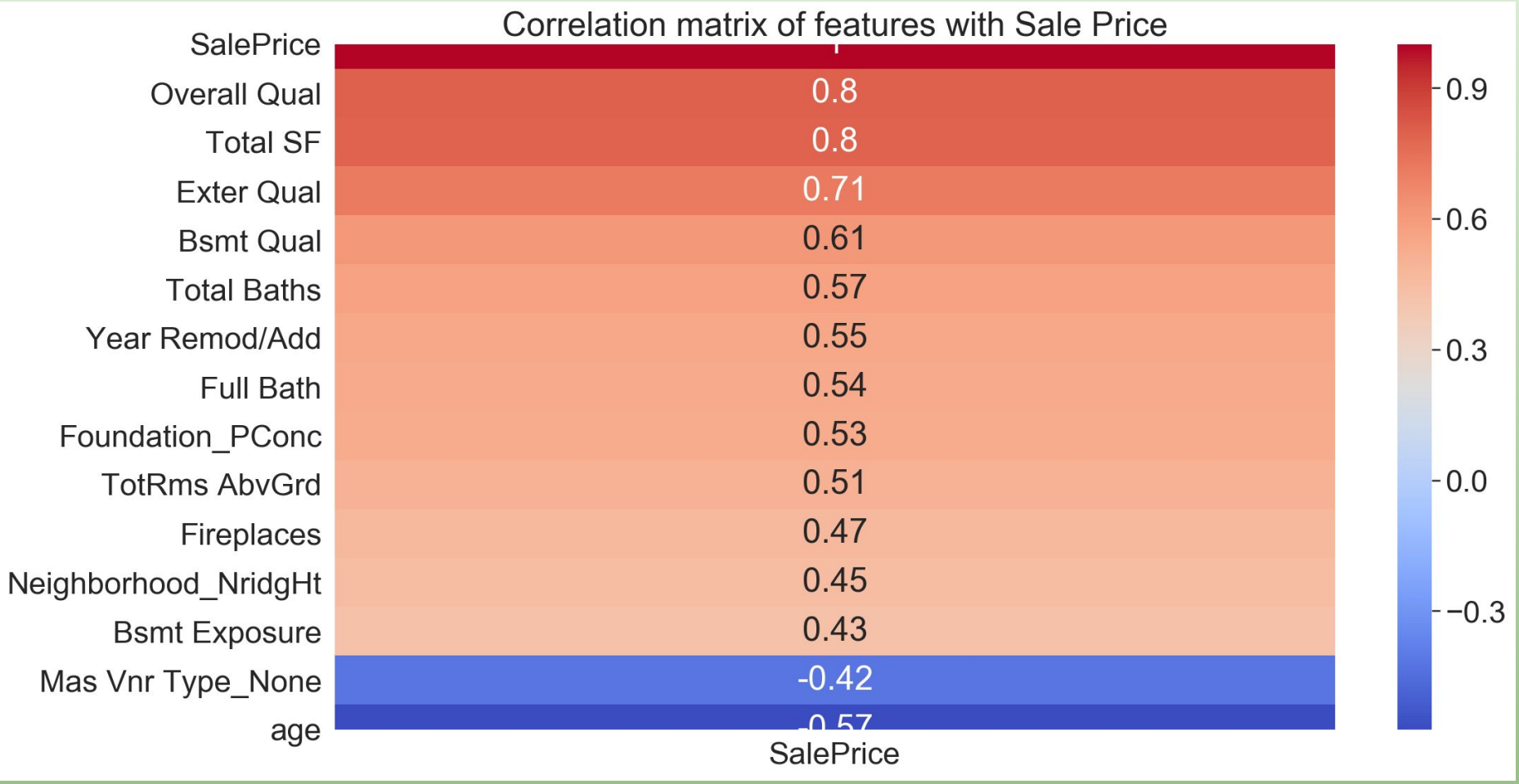

Living Area V/s the Sales Price
SalePrice
600000
500000
400000
300000
200000
100000
0
0
2000
4000
6000
8000
10000
12000
Total SF


Relationship between House Prices and Neighborhoods
SalePrice
600000
500000
400000
300000
200000
100000
0
Sawyer
SawyerW
NAmes
Timber
Edwards
OldTown
BrDale
CollgCr
Somerst
Mitchel
StoneBr
NridgHt
Gilbert
Crawfor
IDOTRR
NWAmes
Veenker
MeadowV
SWISU
NoRidge
ClearCr
Blmngtn
BrkSide
NPkVill
Blueste
GrnHill
Greens
Landmrk
Neighborhood


Overall Quality of the house v/s Sales Price
SalePrice
600000
500000
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
10
Overall Qual


Relationship between SalePrice and the age of the house
SalePrice
600000
500000
400000
300000
200000
100000
0
0
20
40
60
80
100
120
140
age


Relationship between SalePrice and when the house was Remodeled
SalePrice
600000
500000
400000
300000
200000
100000
0
1950
1960
1970
1980
1990
2000
2010
Year Remod/Add


Correlation matrix of features with Sale Price
SalePrice
Overall Qual
Total SF
Exter Qual
Bsmt Qual
Total Baths
Year Remod/Add
Full Bath
Foundation_PConc
TotRms AbvGrd
Fireplaces
Neighborhood_NridgHt
Bsmt Exposure
Mas Vnr Type_None
age
0.8
0.8
0.71
0.61
0.57
0.55
0.54
0.53
0.51
0.47
0.45
0.43
-0.42
SalePrice
0.9
0.6
0.3
0.0
−0.3

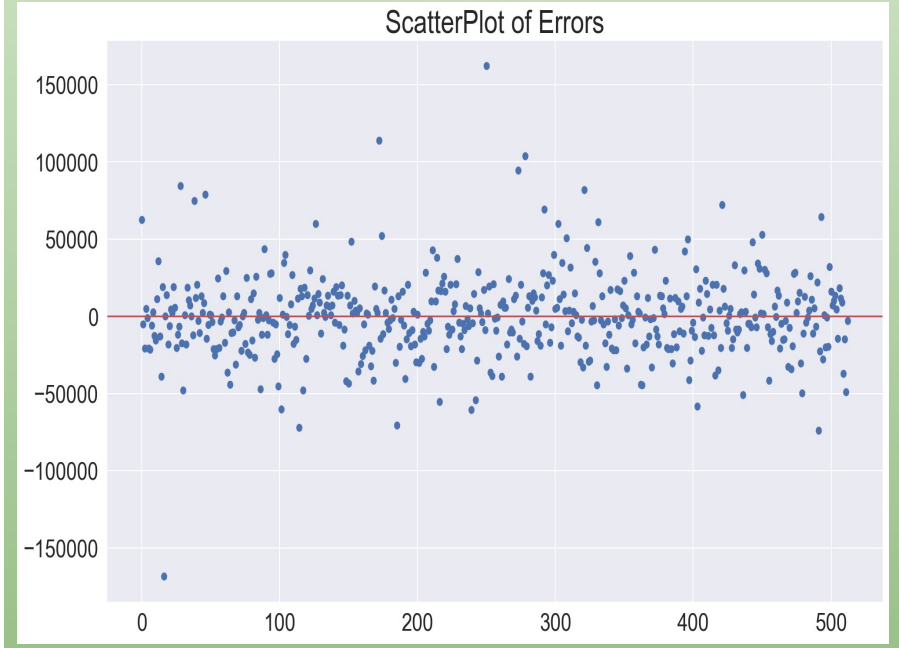
Results of the Linear Regression Model

The model was tried with different combinations of variables and also with extra features that were engineered. I also dropped some variables and created dummies and weighted variables for the non numeric data.

The model that was the best fit had an R-Square of 0.91 for my training split and 0.88 for the testing split with a cross validation score of 0.86.

Variable	Coefficient	P-Value
Total Square Footage	36.225	0.00
Overall Quality	9459.42	0.00
Overall Condition	5775.57	0.00
Neighborhood - Northridge Heights	4.40E+04	0.00
Neighborhood - Brookside	3.43E+04	0.04
Neighborhood - Crawford	4.65E+04	0.00
Basement Exposure	5749.4	0.00
Basement Finished Area	2925.93	0.00
Year Remodeled /Updated	74.66	0.44
Age of the house	-595.02	0.00

Linear Regression Results - The line of best fit for both the training and the testing data seems pretty good. The errors seem homoscedastic but there are outliers.



Conclusions:

- As expected, we found that the overall size of the house and the overall quality of the house were the most significant predictors of house prices - accounted for 77% of the variability in the sales price.
- The location of the house was pretty important as well.
- One surprising conclusion that I did not expect was that the basement exposure and how much of the basement was ‘finished’ was also important in predicting the prices of houses
- The neighborhood of Crawford had a very large and significant Beta. Although my boxplot showed a low mean in the price of houses. I think that would mean that it is an upcoming neighborhood.

Next Steps

Any relationship between the age of the house and the size and condition of the house. Are newer houses in better locations and are they bigger?

Get more data over the past to see if factors that affect prices have changed through the decades.

Dig deeper into the data to find why there are outliers.