Image source: https://www.diversity.iastate.edu/what-we-do/ames

Predicting House Sale Prices in Ames, Iowa

Shruthi Sarode
Date: 08/24/2021

Table Of Contents

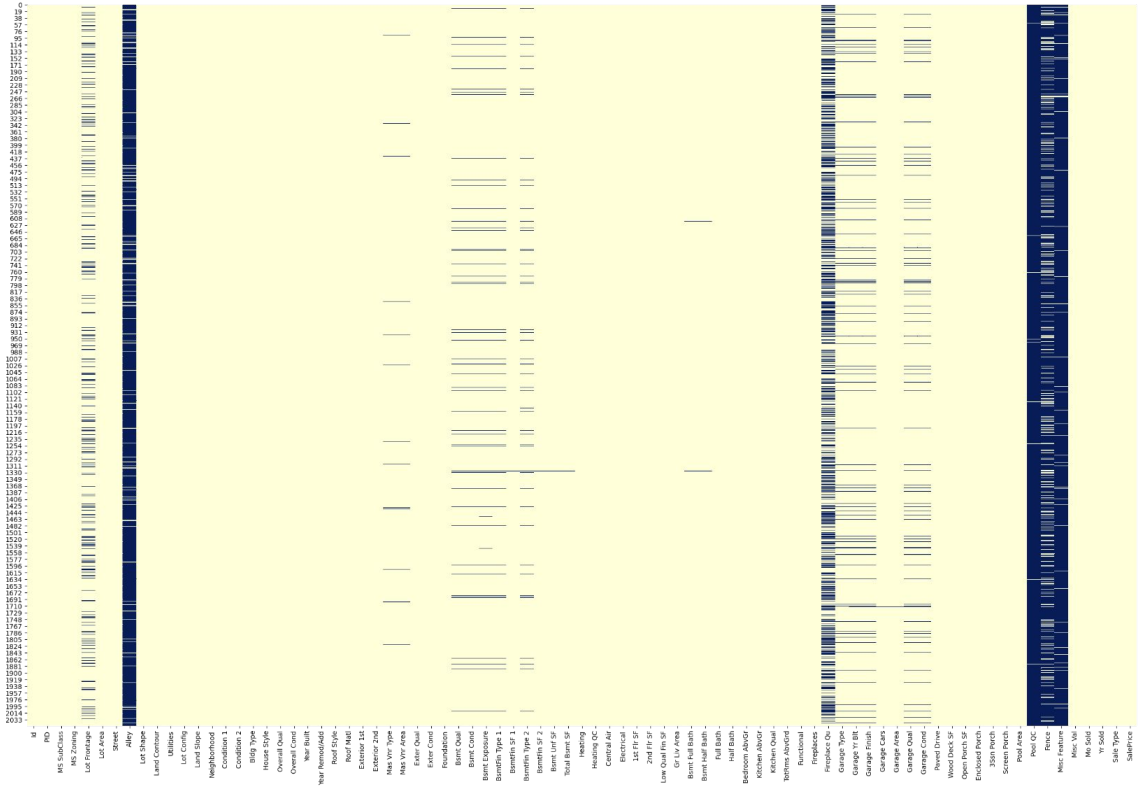
Executive Summary

- The Ames Housing dataset contains all residential home sales in Ames, Iowa between 2006 and 2010.
- The data set contains many explanatory variables on the quality and quantity of physical attributes of residential homes
- Most of the variables describe information a typical home buyer would like to know about a property (square footage, number of bedrooms and bathrooms, size of lot, etc.).
- Giving train and test dataset- We model on the train and fit on the test

This project aims to find out what drives the price of a house- Is it the neighborhood? The size of the house? The amenities? Or something else?

EDA for the Train Dataset

- The train dataset was checked for Missing values. Required substantial amount of cleaning.
- Dropped those variables with more than 90% null values(Alley, Pool QC, Misc Feature)



Visualizing the missing data with heatmap

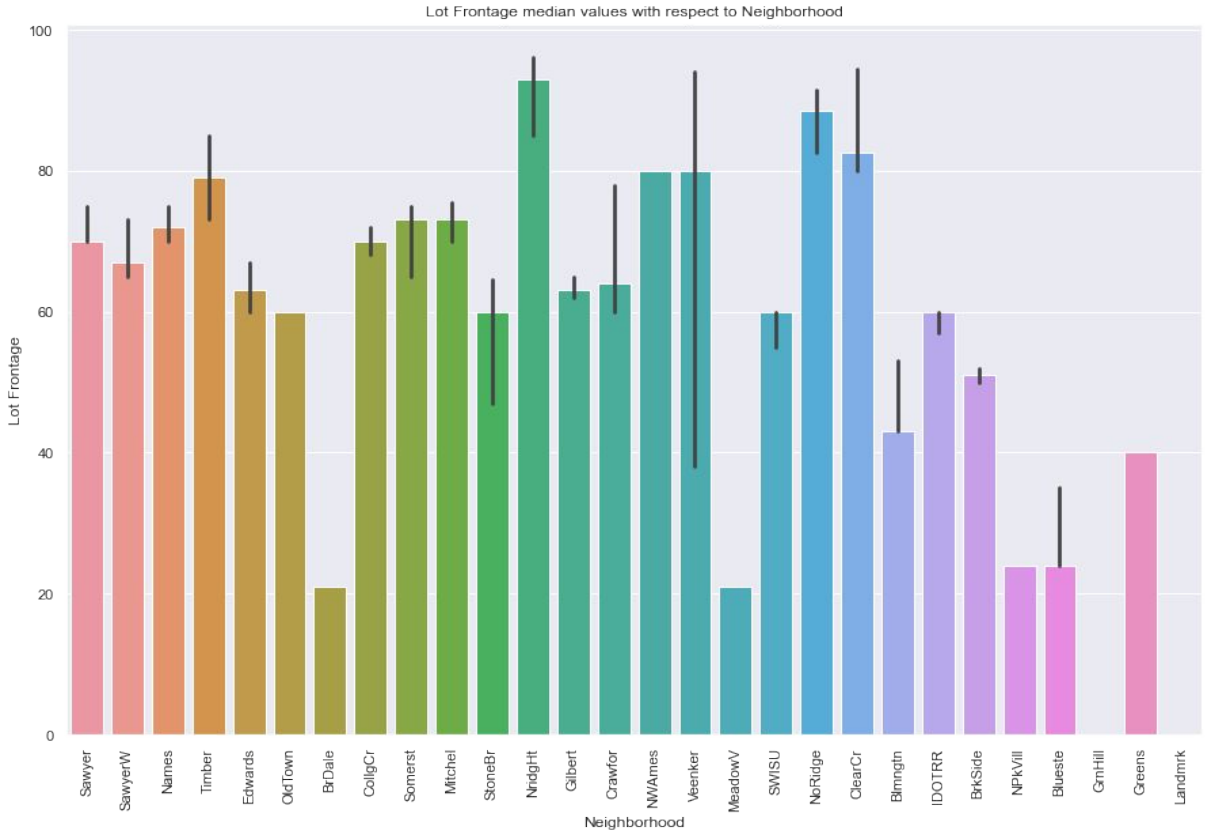
Missing Values

- Many of the categorical features used NaN as a value
- NaN was mapped to NA (not applicable) or NONE where a feature was not part of the house.
- Houses with NO garages, houses with NO Basement, NO Fireplaces, etc.
- Checking if: Total Bsmt SF ==0 correspond to the 51 Null values in the Bsmt Qual and Bsmt Cond
- Appropriate values were filled.

```
# We can fill NA to the Garage columns corresponding the the Garage Area = 0
train['Garage Yr Blt'] = train['Garage Yr Blt'].fillna(0)
train['Garage Finish'] = train['Garage Finish'].fillna('NA')
train['Garage Qual'] = train['Garage Qual'].fillna('NA')
train['Garage Cond'] = train['Garage Cond'].fillna('NA')
```

Lot Frontage Missing Values

- Lot Frontage-330 missing values were imputed based on the median values for the houses grouped by neighborhood.



Feature Engineering

- Changed ordinal categorical columns to numerics (such as quality scales for kitchen, fireplaces, garages, etc.).
- Mapped - {'Ex':5, 'Gd':4, 'TA':3, 'Fa':2, 'Po':1, 'NA':0}
- Binary columns for Street(Paved=1 or Gravel=0), Central Air(Y=1, N=0)
- Looked at any outliers. Eg: Garage built year= 2207 changed to 2007
- Lot frontage and SalePrice plot- some outliers with high Lot Frontage and low Sale price
- Generally if a lot frontage is high then the Saleprice should be high as well
- Dummy Columns for - Land Contour, MS Zoning, Foundation, Exterior, Condition, etc

Feature Engineering

- Feature Creation- To reduce the number of features and make more general property features.
- Age= Year sold - Year Built
- The 'Gr Liv Area' is the sum of '1st Flr SF' and '2nd Flr SF', we can delete the 'Gr Liv Area' and keep only 1st Flr SF and 2nd Flr SF
- 'Total Bsmt SF' is the sum of 'BsmtFin SF 1' and 'Bsmt Unf SF', we can drop 'Total Bsmt SF '
- 'Garage Cars' and 'Garage area': creating a column for Garage_area per car = Garage_area/Garage_Cars, and drop Garage_Area column
- Total rooms above grade includes all the rooms. We can separate this information into 'total bedrooms' and 'other rooms', and drop the Total Rooms column.
    - Other Rooms = Total room ABove grade - Bedroom above grade

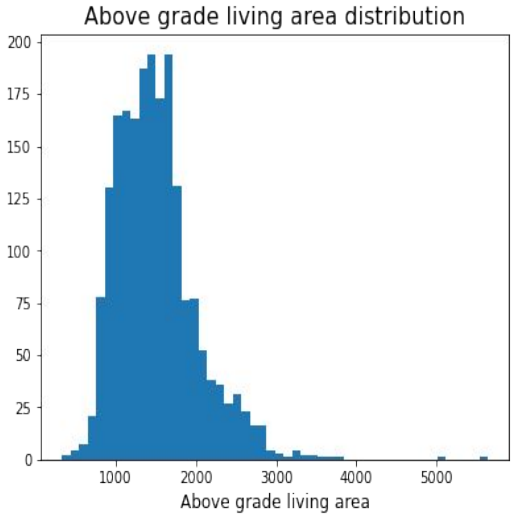
Feature Engineering

- Ordinal column for the Neighborhood- the neighborhood column was divided into 4 categories based on the overall score of the house
- ['overall_score'] = ames_train['Overall_Qual'] + ames_train['Overall_Cond'] +ames_train['Exter_Qual'] +ames_train['Exter_Cond']
- ames_train.groupby('Neighborhood')['overall_score'].mean().sort_values()

```
neigh_dict = {1 : ['BrDale', 'Landmrk', 'MeadowV', 'Edwards', 'SWISU', 'IDOTRR', 'Sawyer'],
              2 : ['Names', 'BrkSide', 'ClearCr', 'OldTown', 'Gilbert', 'NPkVill', 'Mitchel'],
              3 : ['SawyerW', 'CollgCr', 'NWAmes', 'GrnHill', 'Blueste', 'Blmngtn', 'Crawfor'],
              4 : ['Timber', 'Somerst', 'NoRidge', 'Greens', 'Veenker', 'NridgHt', 'StoneBr']}
```

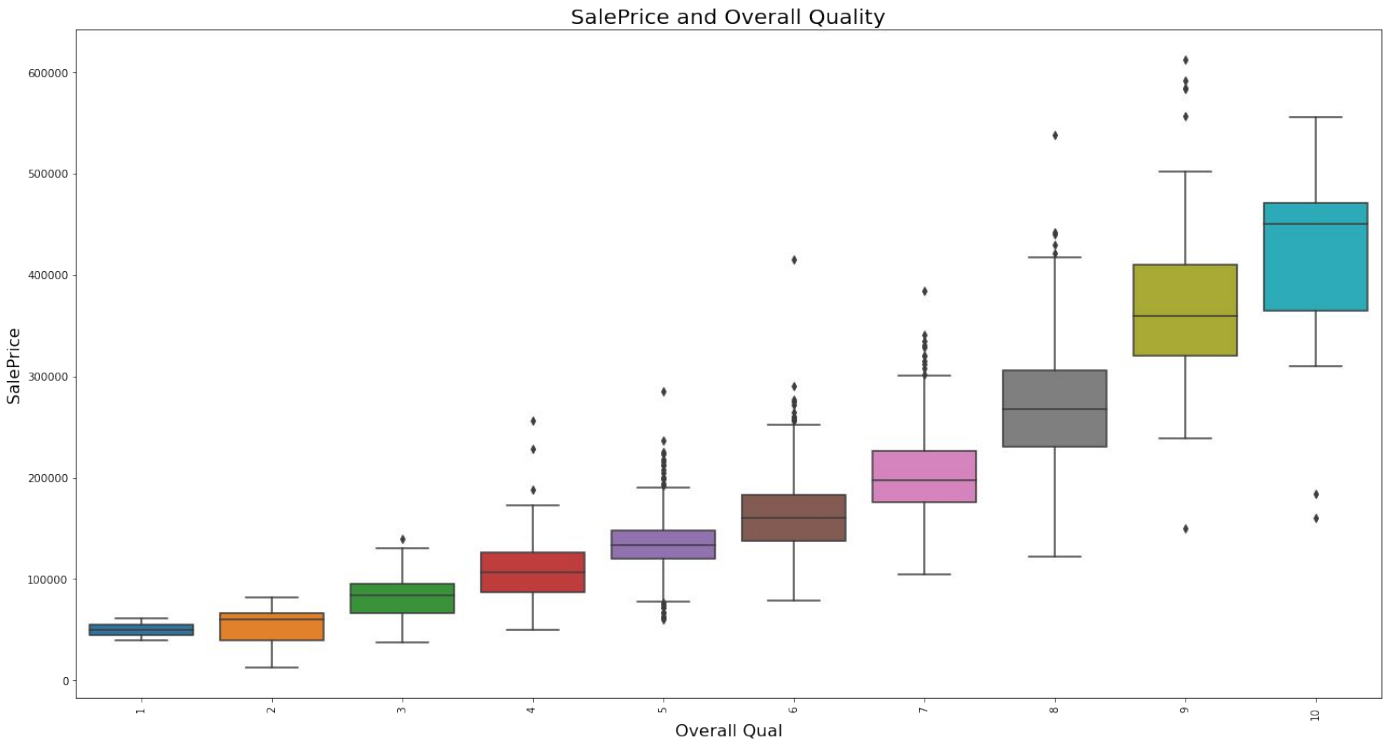
Some Visualizations

- The relationship between sale price and Above Grade living area
- There are indeed clear linear relationships.
- The histogram for the grade living area is right skewed.
- Plots for SalePrice and Garage area, Basement area, 1st Floor SF, etc had similar distributions



Some Visualizations

- The boxplot for over all quality and Sale price has a linear relation as seen.
- As Expected we can see that higher the Overall Quality higher is the Sale price for a house.
- There are some outliers as well



Preprocessing for Test Data

- The test data had some missing values as well.
- the missing values were imputed accordingly and
- Same dummy columns were created
- The columns were matched to the train dataset.

Modeling

- We started by selecting all the features we had except the ID and PID columns (99 features)
- Then scaled the data since we have different units represented in the dataset.
- Did k-fold cross validation to get a baseline model score.
- With these features, several model options were looked at, and then ran these models across our testing data.
  - Linear Regression
  - LASSO
  - Ridge
- Each model was scored using the same metrics
  - R2
  - RMSE

Modeling

Model 1- Includes all the 99 features to start with.

Model 4- After filtering the outliers in the Lot area and the Home SF and selected features based on correlation with the target variable.

Model 5- After filtering the outliers in the Lot area and the Home SF and including all the features that I initially had (99 features) .

Model 1	RMSE	Model Score	Prediction Score
Linear	28602.404	0.87199	0.871016
Ridge	28776.46	0.868472	0.869441
LASSO	28805.138	0.86511	0.86918

Model 4	RMSE	Model Score	Prediction Score
Linear	25833.043	0.87049	0.8755
Ridge	25743.79	0.87038	0.8763
LASSO	25740.29	0.8703	0.8763

Model 5	Model Score	Prediction Score
Linear	0.9281	0.9162
LASSO	0.9053	0.9036

RMSE- 22730.701

Modeling- Top Attributes for Model 6

Model 6: LASSO

- Features- 99
- Log transformed target variable
- Outliers removed from Lot area and Home SF
- Lasso regression cross validation mean score: 0.9174
- Lasso regression model score: 0.930
- Lasso regression prediction score: 0.94

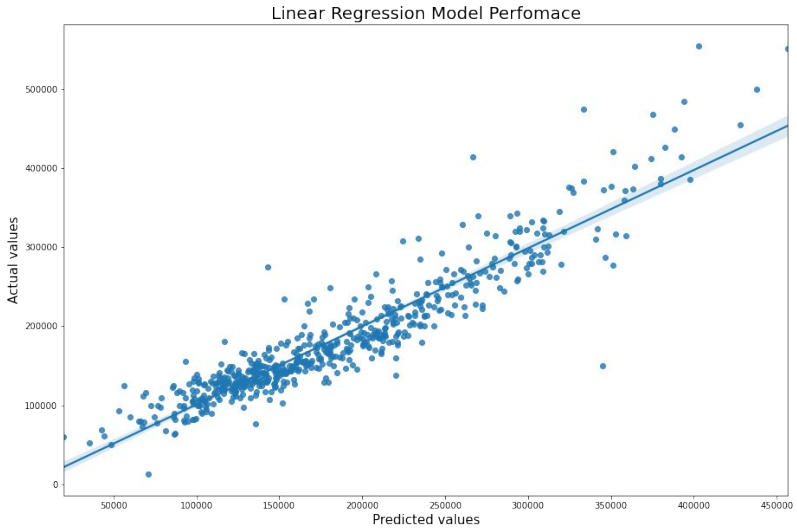
For Every Unit increase in Xi, (Variable), we expect to increase the Sale Price by Coef %, holding all else constant.

Eg: For 1 year increase in the age of a home, we can expect a decrease of about 6.5% in the price.

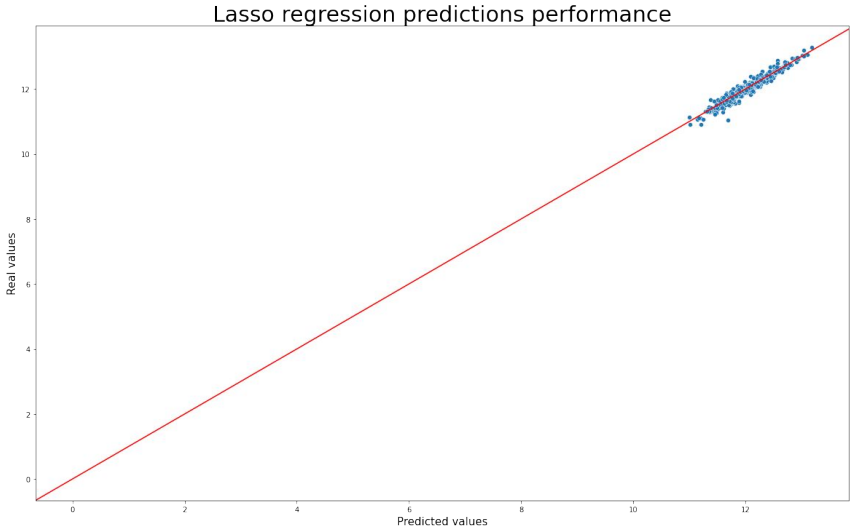
For 1 point increase in the overall quality, holding all other features constant, we can expect, on average, a 7% increase in the price of a home

variable	coef
1st_Flr_SF	0.103266
2nd_Flr_SF	0.096965
Overall_Qual	0.07207
Age	-0.065894
BsmtFin_SF_1	0.055685
Overall_Cond	0.045172
Lot_Area	0.031402
Functional	0.024816
Neighb_Qual	0.021992
Garage_Cars	0.021592
Bsmt_Unf_SF	0.01967
Kitchen_Qual	0.019572
Exter_Qual	0.017018
MS_Zoning_RM	-0.016313
Screen_Porch	0.014468

Modeling: Visually comparing the models



Model 1



Model 6 with outliers removed and normalizing Sale price

Conclusions

Based on this model we can conclude that the top features :

- 1st FLoor and 2nd Floor area(total above grade living area),
- Age,
- Basement SF,
- Overall quality, Overall Cond
- Lot area
- Functional,
- Neighborhood ,
- Garage Car capacity
- Kitchen quality and Basement Finish

Have the greatest impact on the potential Sale Price of a home in Iowa.

Conclusions and Next steps:

- For data cleaning, one of the most important thing was to identify the categorical and numeric- MS subclass
- Normality of the target variable helped to improve the accuracy of the model.
- This model generalizes well for Ames Iowa but cannot be used for other regions of the country.
- Would like to see how adding other features can affect the home price- School rating for a neighborhood, transportation, commercial information.
- Using a cross-validation grid search to find best alpha
- Explore other algorithms such as
    - Elasticnet Model
    - XGBoost model

Thank You!