# House Prices in Ames, Iowa: Initial findings

Chirag Limbachia, M.S. in Biomedical Engineering, Wright State University.

Capstone Project 1: Milestone Presentation

# Problem Statement

- Several features influence price of a house such as, the building type, total number of rooms, lot area, garage size, masonary work, location, utilities, and many more.

- Can house price be estimated (predicted) using such features? Can we identify features that influence house price the most? How important is the building type in determining the price? What influences the price more: location or size (sq.ft.)? How much value does remodeling add to the house?

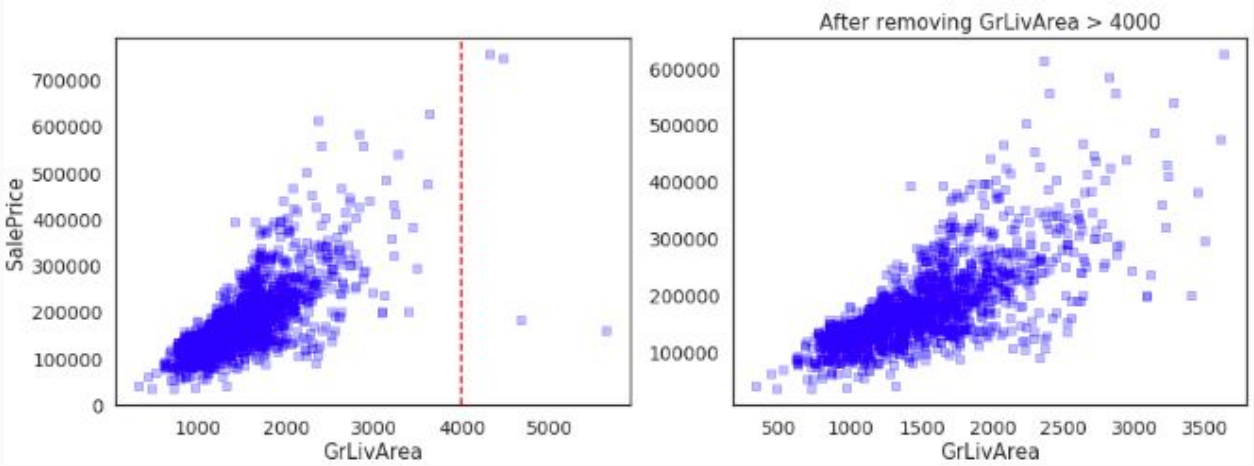
- Being able to estimate a price based on the available information about the house can be useful for a house buyer to negotiate the right price. It can also be useful for a real estate investor in terms of assessing the realistic value of the housing property.

# Data: Source and Description

- Ames Housing dataset, compiled by Dean De Cock is used in this project. The dataset is available on kaggle. It consists of 1460 instances and 79 explanatory variables that describe almost every aspect of residential homes in Ames, Iowa.

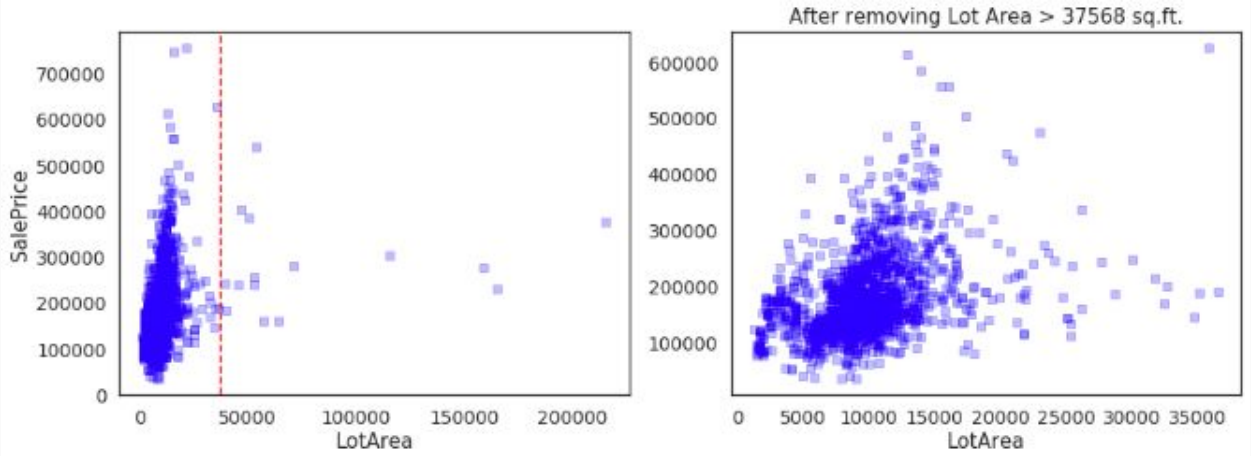
# Data Wrangling: Outlier Removal

- There were outliers in the dataset. Grade Living Area (GrLivArea) plotted against the Sale Price showed some outliers. These outliers were removed from the data.



# Data Wrangling: Outlier Removal

- Lot Area plotted against the Sale Price also showed some outliers. These data points were also removed.
- After outlier removal, total number of instances came down to 1443 from 1460.

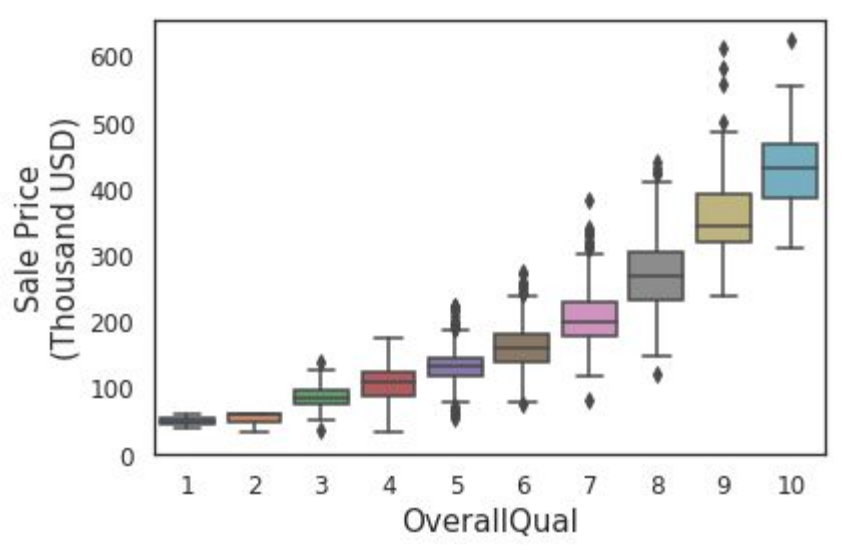


# Exploratory Data Analysis (EDA)

# EDA: Initial Findings

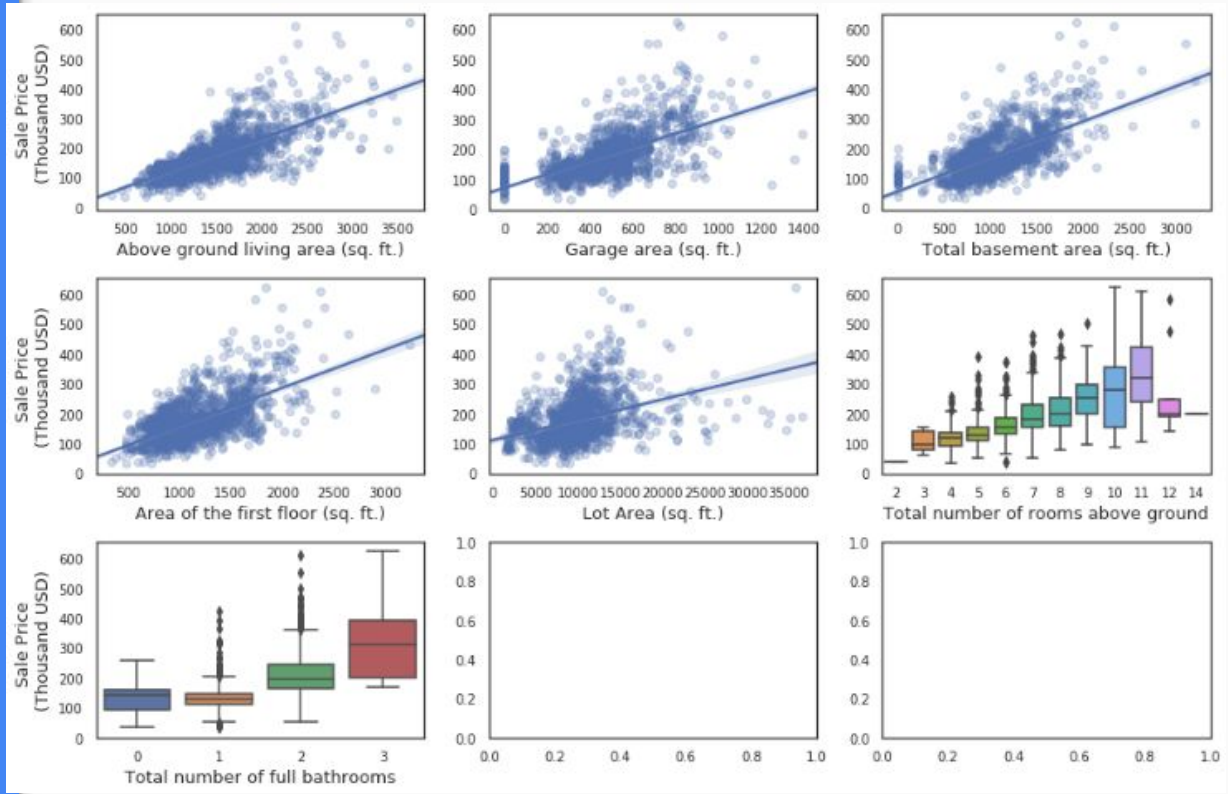
Several features seem to influence the sale price.

Sale price increases as the overall quality (OverallQual) of the house increases.



# EDA: Initial Findings

Sale price is also correlated with variables that reflect size of the housing property such as, grade living area above ground, garage area, total basement area, area of the first floor, lot area, total number of rooms, number of full bathrooms, etc.

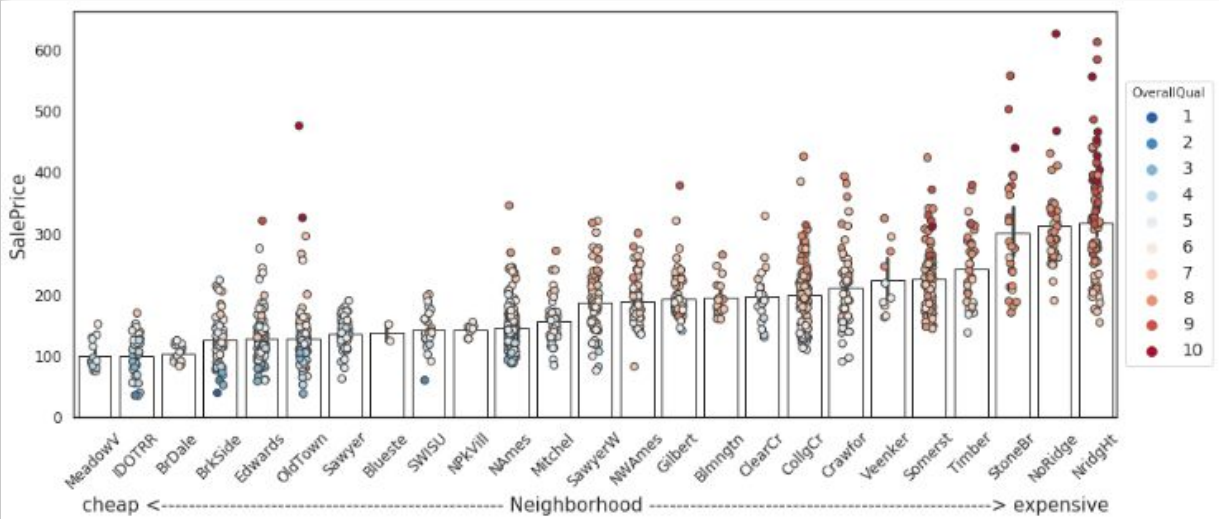


# EDA: Initial Findings

Neighborhood has an influence on the sale price.

Houses in Northridge Heights, Northridge, and Stone Brook have the highest average sale prices.

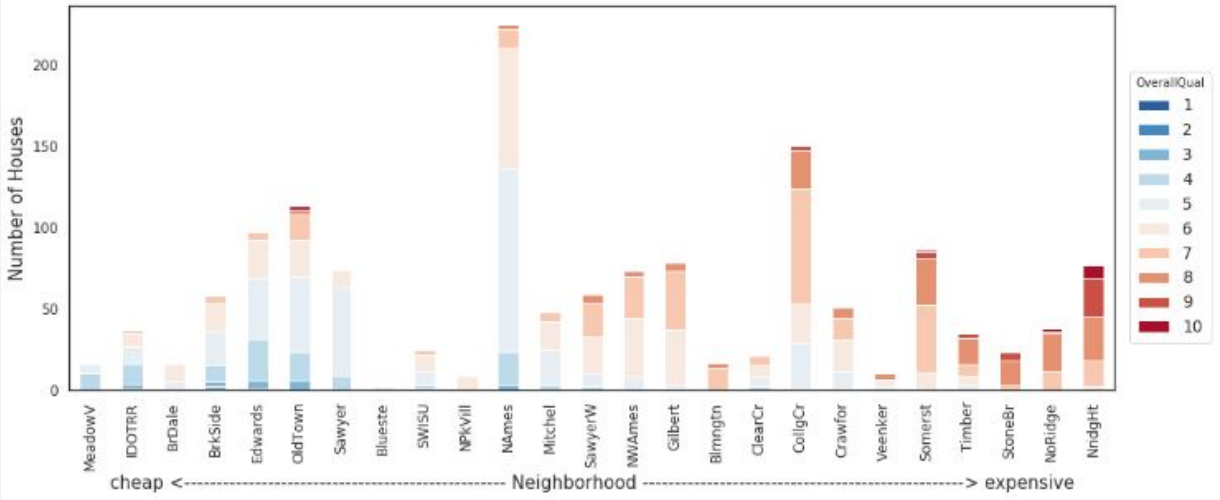
Houses in Meadow Village, Iowa DOT and Rail Road, and Briardale have the lowest average sale prices.



# EDA: Initial Findings

Northridge Heights, Northridge, and Stone Brook have higher concentration of better quality houses.

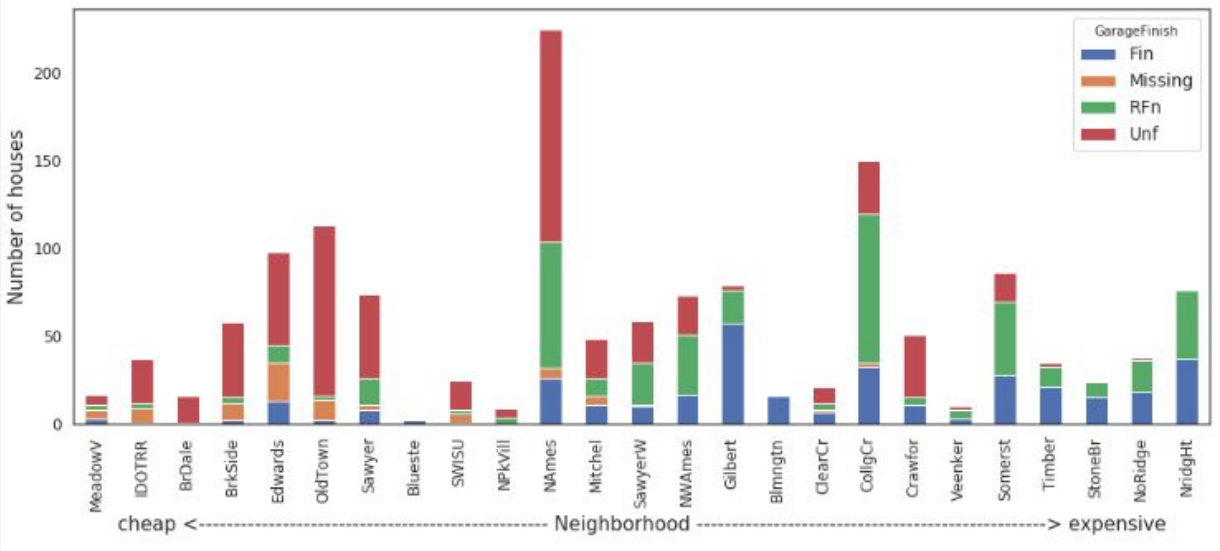
This explains why houses in these neighborhoods have a higher average sale price.



# EDA: Initial Findings

Expensive neighborhoods like Northridge Heights, Northridge, and Stone Brooks tend to have houses that either have a finished or a roughly finished garage. Rarely do they have houses with no garage.

Cheaper neighborhoods like Meadow Village, Iowa DOT and Rail Road, and Briardale mostly have houses with unfinished or no garage.

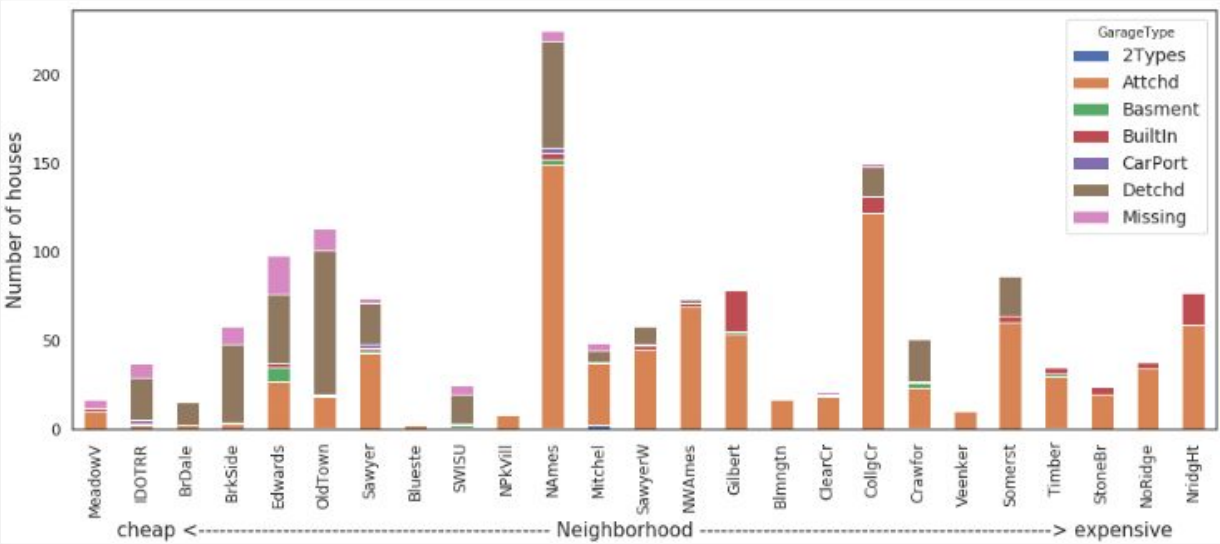


# EDA: Initial Findings

Attached garage are common across all neighborhoods except a few cheaper neighborhoods.

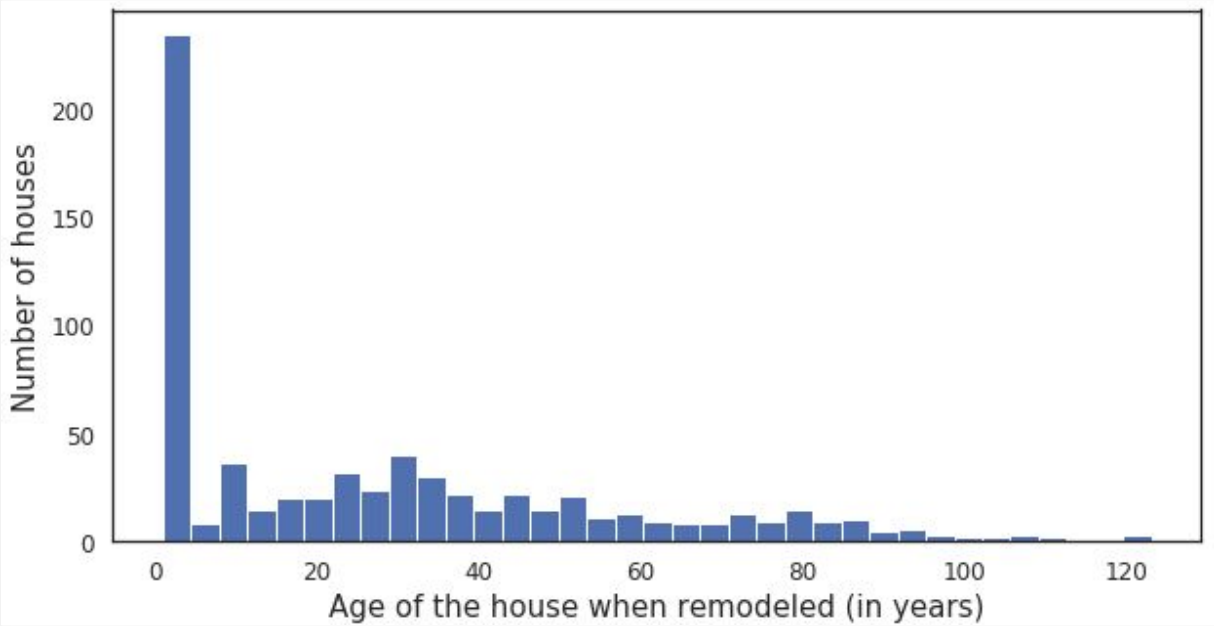
Built-in garage are more common among houses in expensive neighborhoods.

Detached or no garage are most common in cheaper neighborhoods.



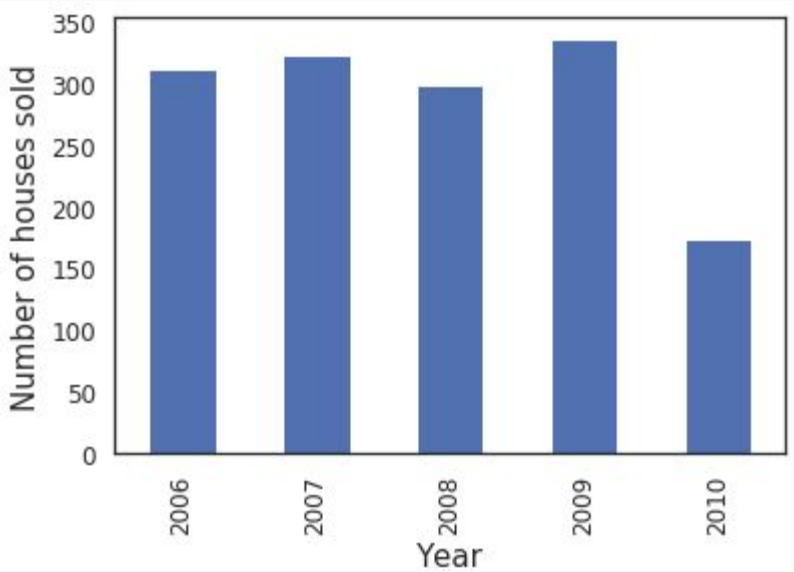
# EDA: Initial Findings

Most houses are remodeled within a couple years after they are built.



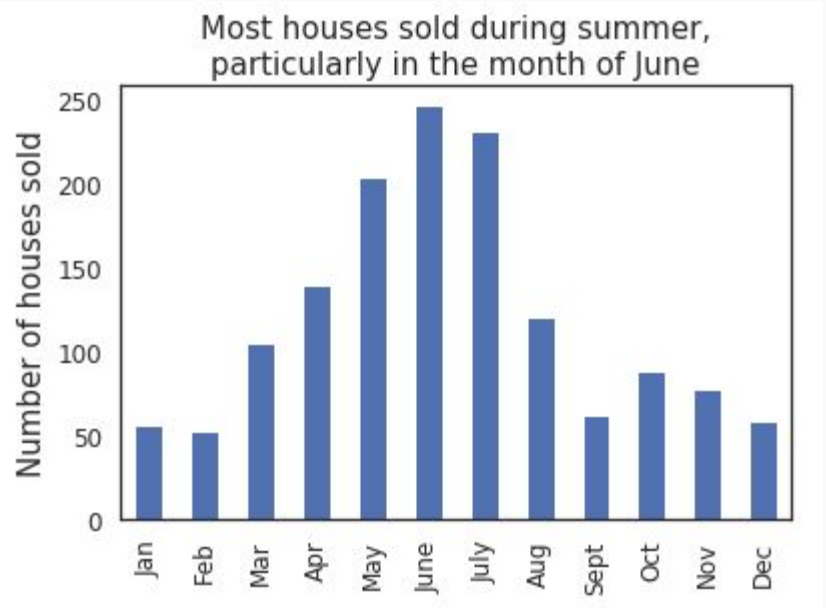
# EDA: Initial Findings

All houses were sold between 2006-2010.



# EDA: Initial Findings

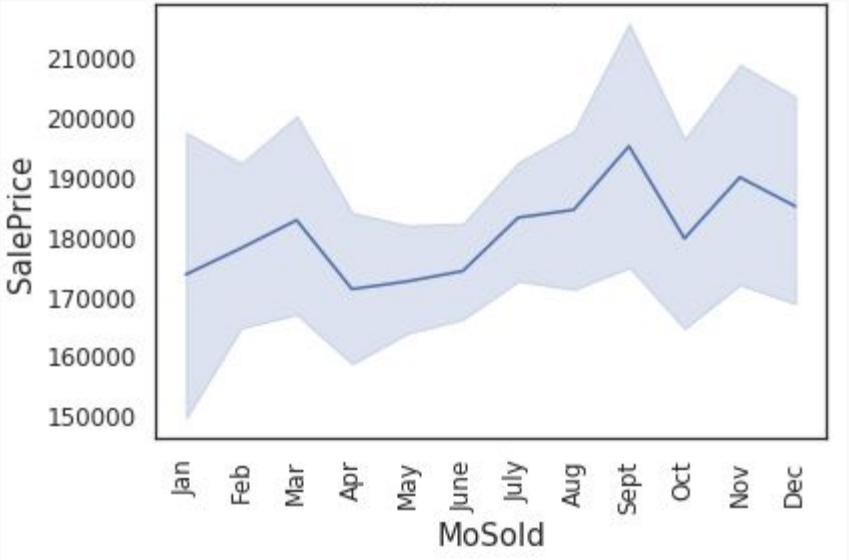
Most houses are sold during summer; highest in the month of June.



# EDA: Initial Findings

Sale price fluctuates across the year. Falls by a few thousand dollars ($10,000) by the end of winter (in April) and start rising again by May-June, peaking in September.

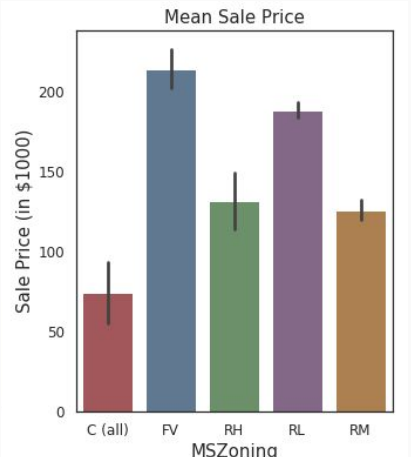
This explains why most houses get sold during summer.

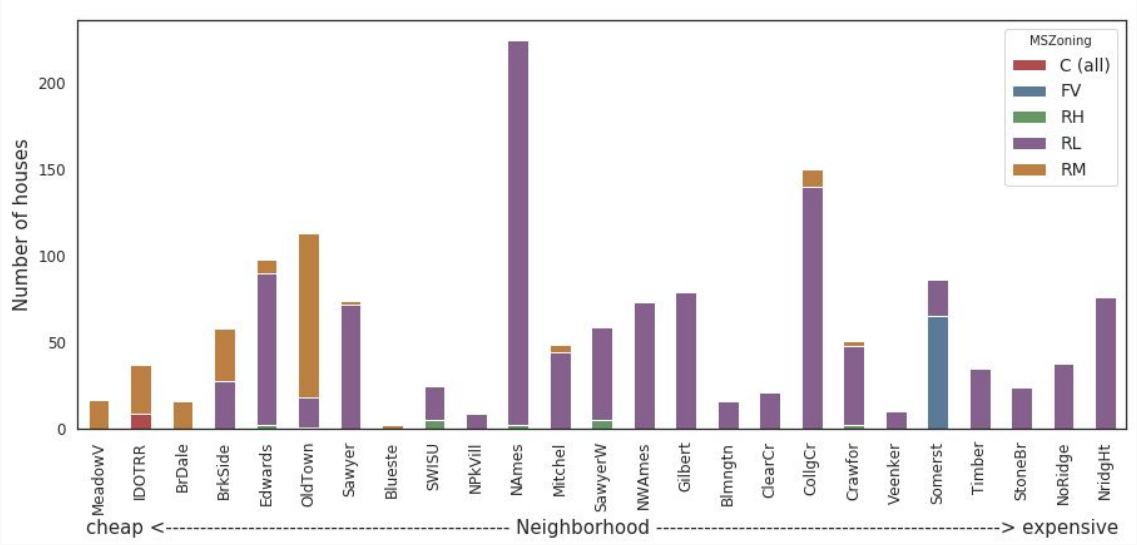


# EDA: Initial Findings

Houses located in floating village residential (FV) zones are most expensive, whereas those located in the commercial (C (all)) zones are the cheapest.

Most houses are located in low density residential (RL) zones. In Somerset, most houses are located in the floating village (FV) residential zone (Note: houses in this zone also have the highest average sale price). Only in the Iowa DOT and Rail Road (IDOTRR) neighborhood do we find houses located in commercial (C all) zones (Note: houses in this zone have the lowest average sale price)



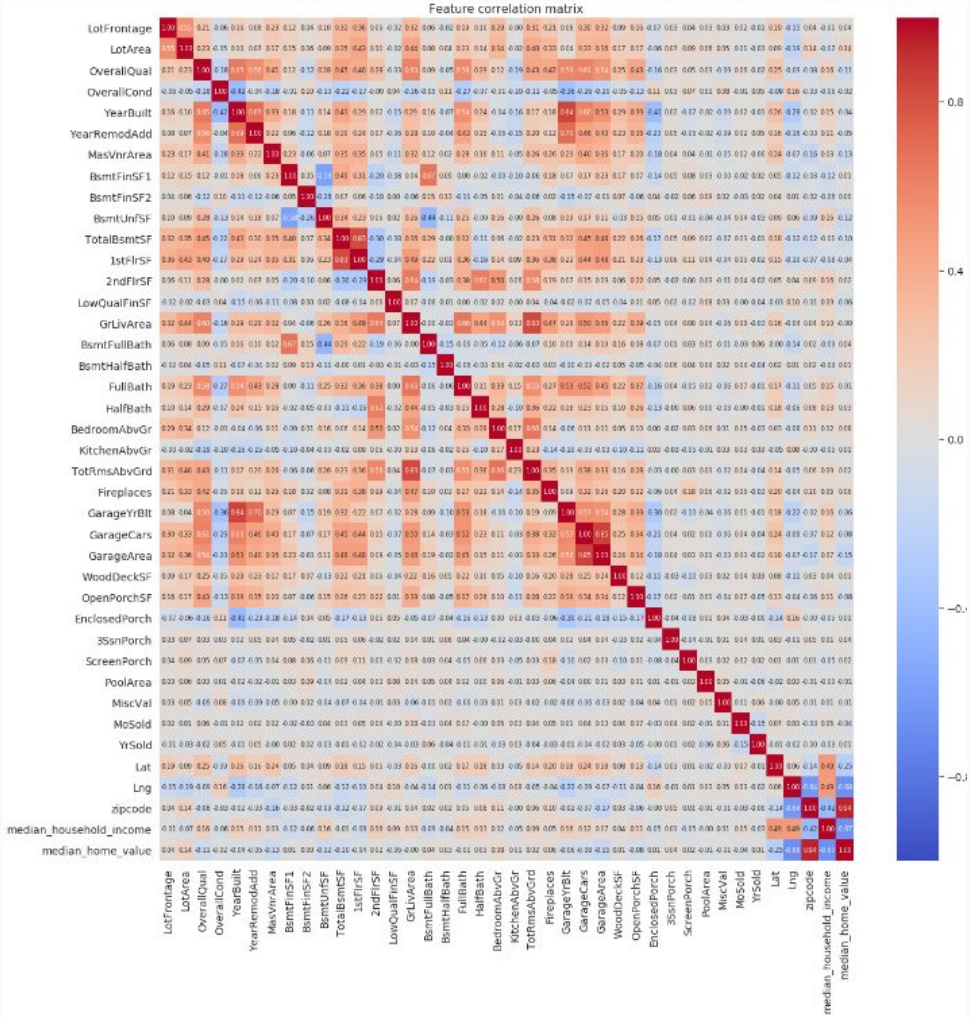



# Feature Engineering

# Numeric Feature Correlations (spearman)

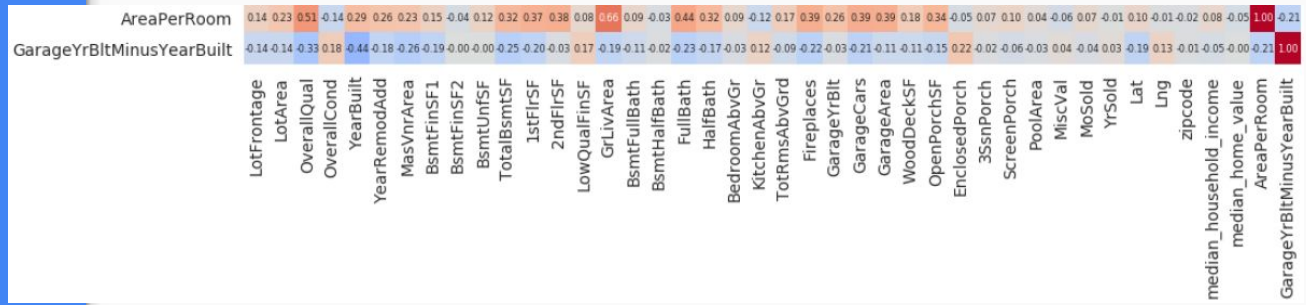
Several features are highly correlated with one another. Highest correlation between house and garage year built (YearBuilt and GarageYrBlt). New feature, GarageYrBltMinusYearBuilt, is created which models the difference between the year in which the garage and the house was built.

Second highest correlation is between the total number of rooms (TotRmsAbvGrd) and total square feet area above ground (GrLivArea). New feature, AreaPerRoom, is included which models the average area per room by dividing the TotRmsAbvGrd by GrLivArea.



# Numeric Feature Correlations
## (spearman)

Unlike the parent features, the newly added features are not that highly correlated with the remaining features. Neither are they that highly correlated with the parent features.



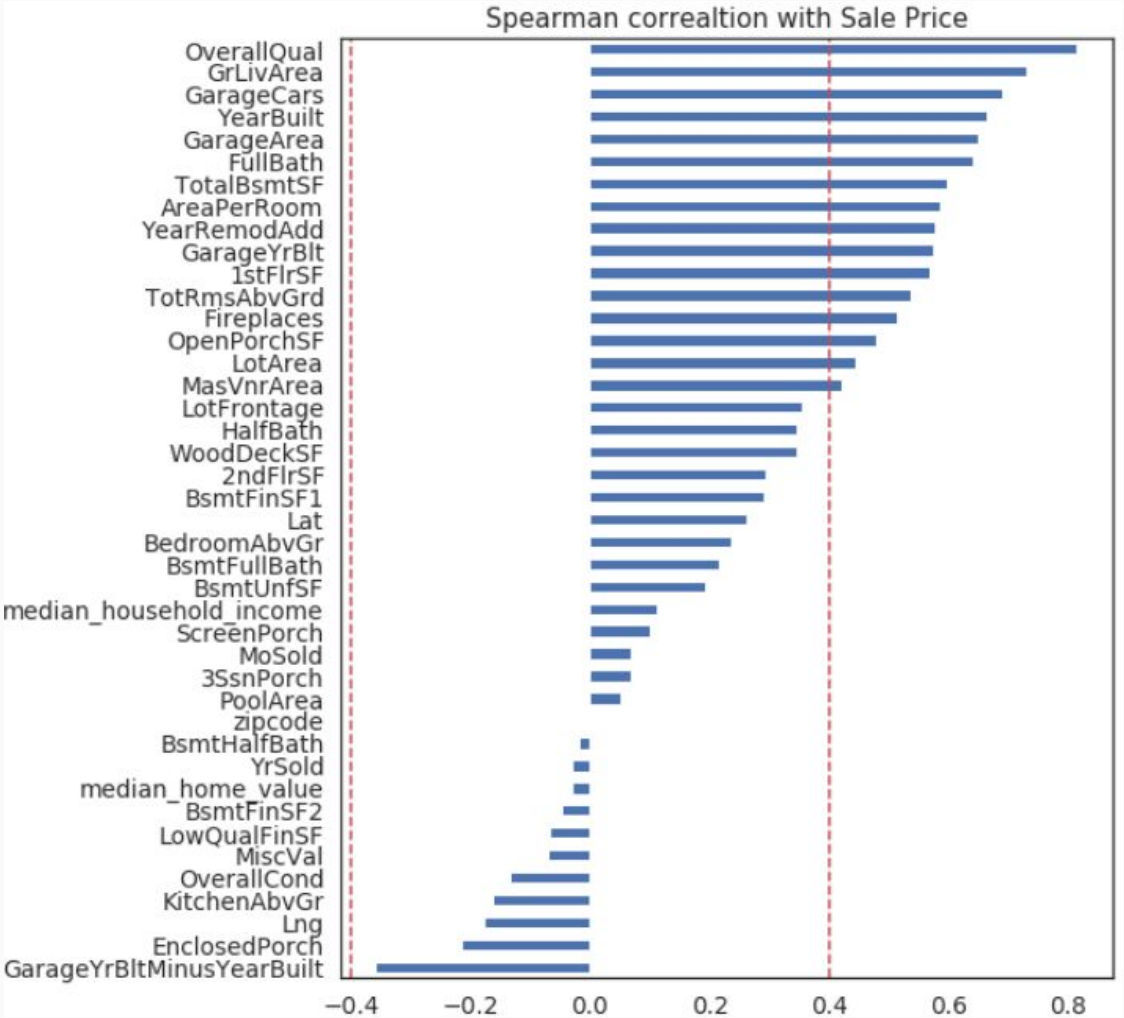
# Correlation with Sale Price

Several features are highly correlated (> 0.4) with the sale price.

One of the new features, AreaPerRoom, is also highly correlated (~0.6) with the sale price. In fact, it is more correlated than one of its parent features from which it is derived (TotRmsAbvGrd).

GarageYrBlitMinusYearBuilt is another new feature. It has the highest negative correlation with the sale price.

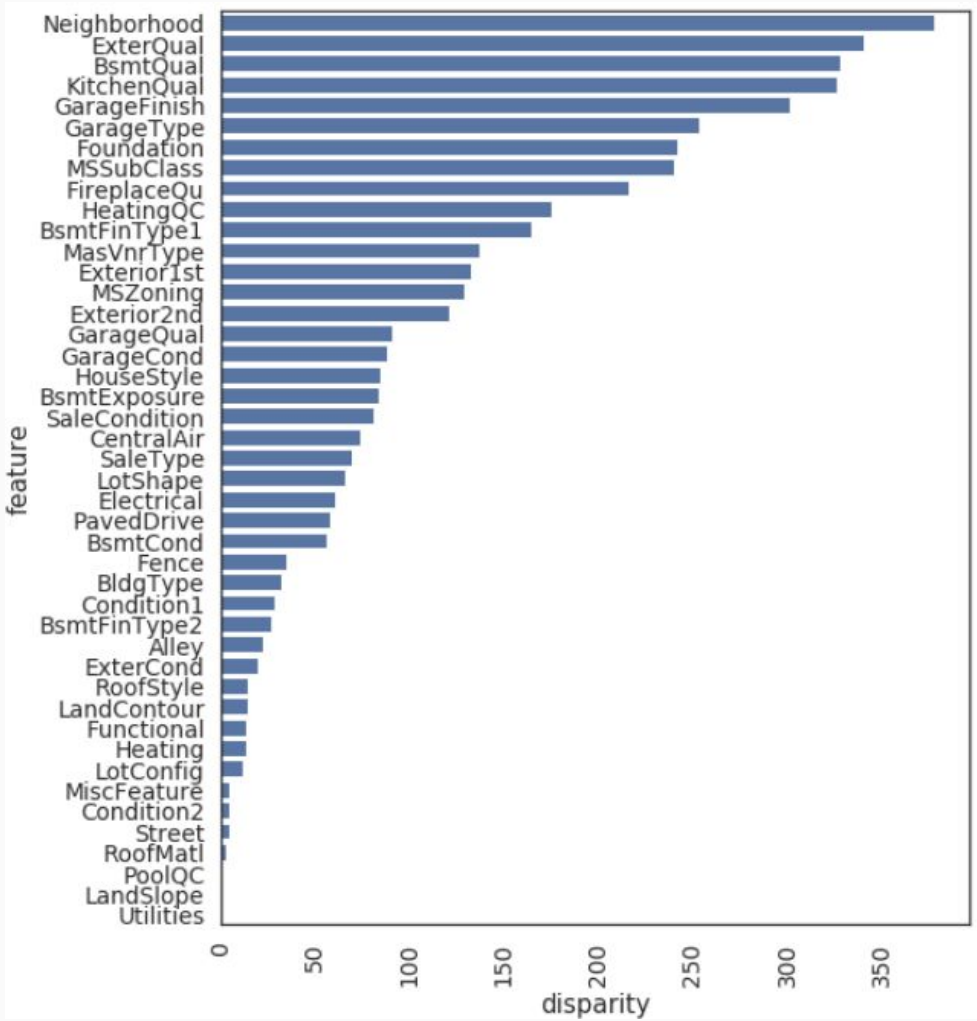
Both new features are goods candidates for predicting the sale price.



# Categorical Features: Kruskal ANOVA

The above bar plot shows the influence of categorical features on the sale price. Features with higher disparity have greater influence on the sale price. Disparity is derived from the p-value obtained from ANOVA analysis of a particular feature.

Features like Neighborhood, exterior quality (ExterQual), basement quality (BsmtQual), kitchen quality (KitchenQual), garage finish (GarageFinish) seem to have high influence on the sale price.



# Feature Summary

- Sixteen out of the 44 categorical features are actually ordinal features. Based on the provided feature descriptions, appropriate ordinal labels were assigned to the levels of the ordinal features.
- Remaining 28 categorical features were One-Hot encoded  (for details see the jupyter notebook). Now all features in the dataset are numeric (continuous, ordinal and One-Hot encoded).
- Features that showed a correlation of 0.4 or greater with sale price, were multiplied with each other and included as new features in the dataset. Interaction was kept limited only up to between two features.
- The final dataset consists of 1443 examples and 433 features.

# Thanks!

For questions/comments, reach me at:

chirag.limbachia1@gmail.com