Ames House Predictions

Group 1

Problem Statement

Our Client is a Property Agents. They are looking to expand into the Iowa property market. We are tasked to predict the prices of houses based on their characteristics.

Our client is also interested in understanding what are the important variables that affects house prices in order to understand what are the most important data to be collected to predict the prices of future house listings on their site.

Some Key Questions

- Do Neighbourhood matter?
- Do house size matter?
- Is a certain area of the house more important?
- How important is the finishing of the houses?
- Does the age of the house matter?

Data Set

We are provided with Ames Housing data set which contains information of more than 2800 properties sold in Ames, Iowa between 2006 and 2010. The data set has over 70 columns of different features relating to houses.

Method of Approach

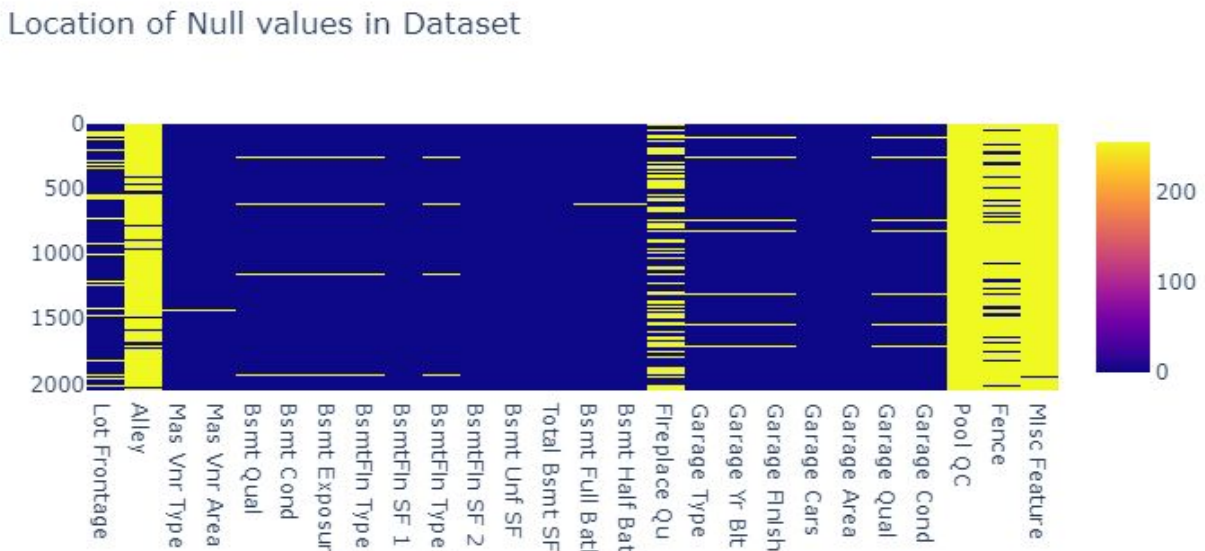
Model use:

- Linear Regression
- Lasso Regression
- Ridge Regression

Exploratory Data Analysis:

- Recursive Feature Elimination
- Correlation
- Intuition - reading the data dictionary

Null Values Handling

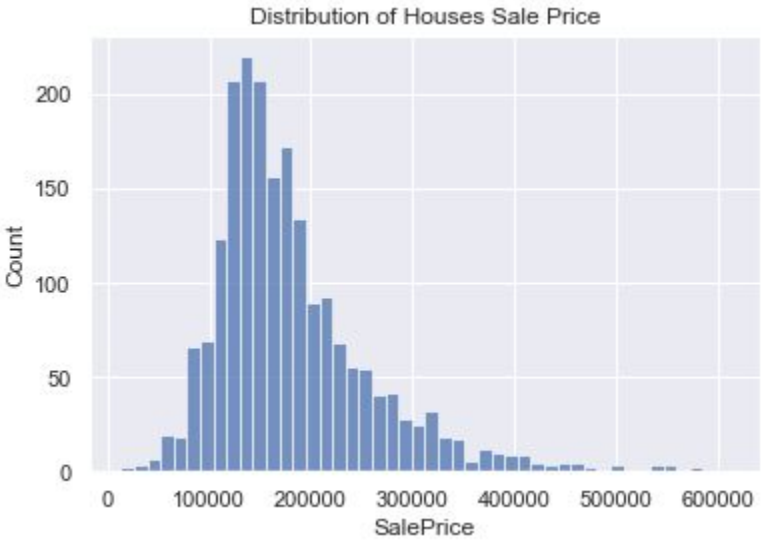


Feature	Percentage of Null Values
Pool QC	99.5%
Misc Feature	96.8%
Alley	93.2%
Fence	80.5%
Fireplace Qu	48.8%
Lot Frontage	16.1%

- Many of the Null values occur on the same rows for multiple columns suggesting that they are not mutually exclusive.

Exploratory Data Analysis

Saleprice

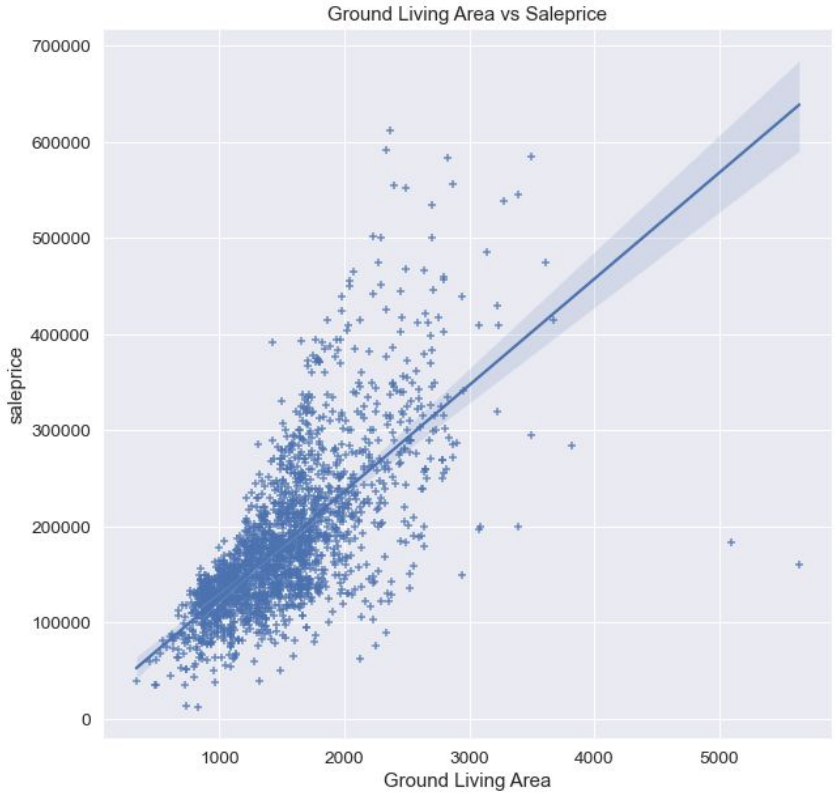


The average Sale price of houses is $181,470

50% of the houses cost between $130,000 and $214,000

The Houses Sale Price have a right skewed distribution

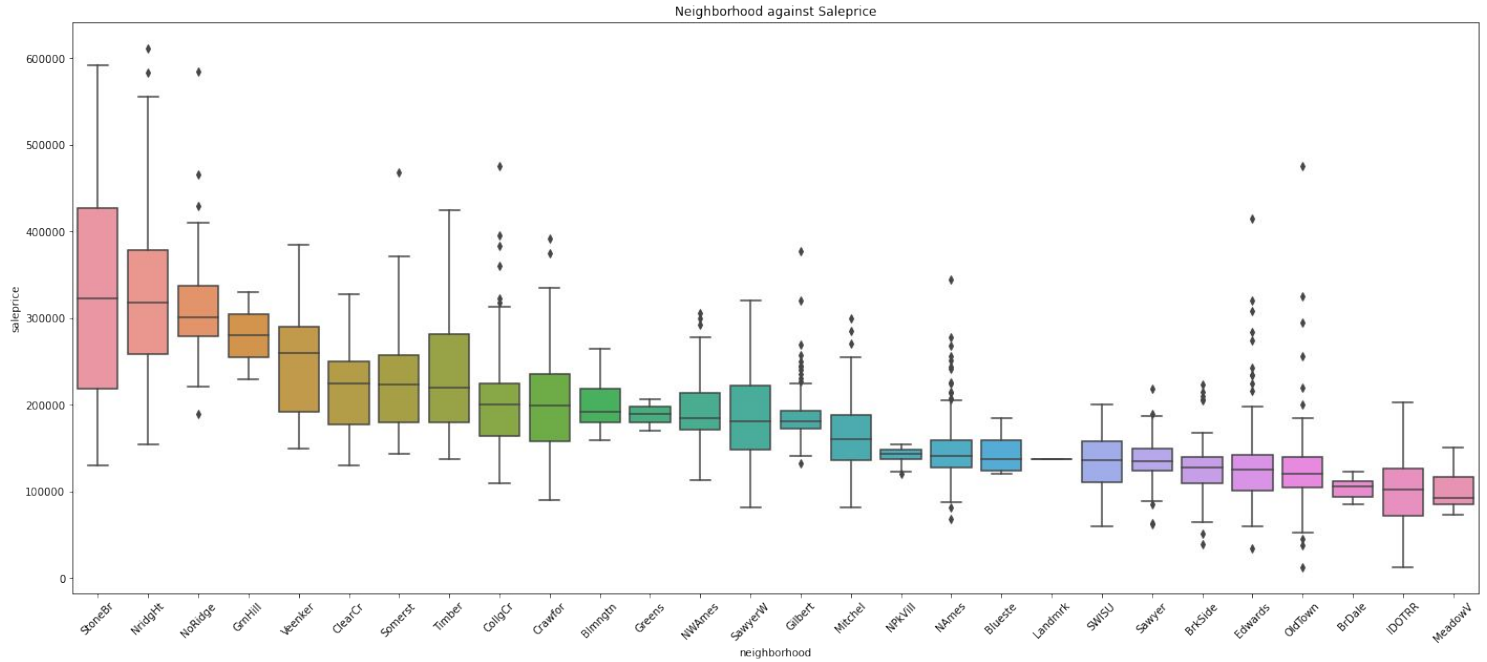
Living Area VS Sale Price



Living area have positive relationship against saleprice

- Two outlier
  - A relatively low sale price for a large house

Neighborhood VS Saleprice



The property location impacts the sale price

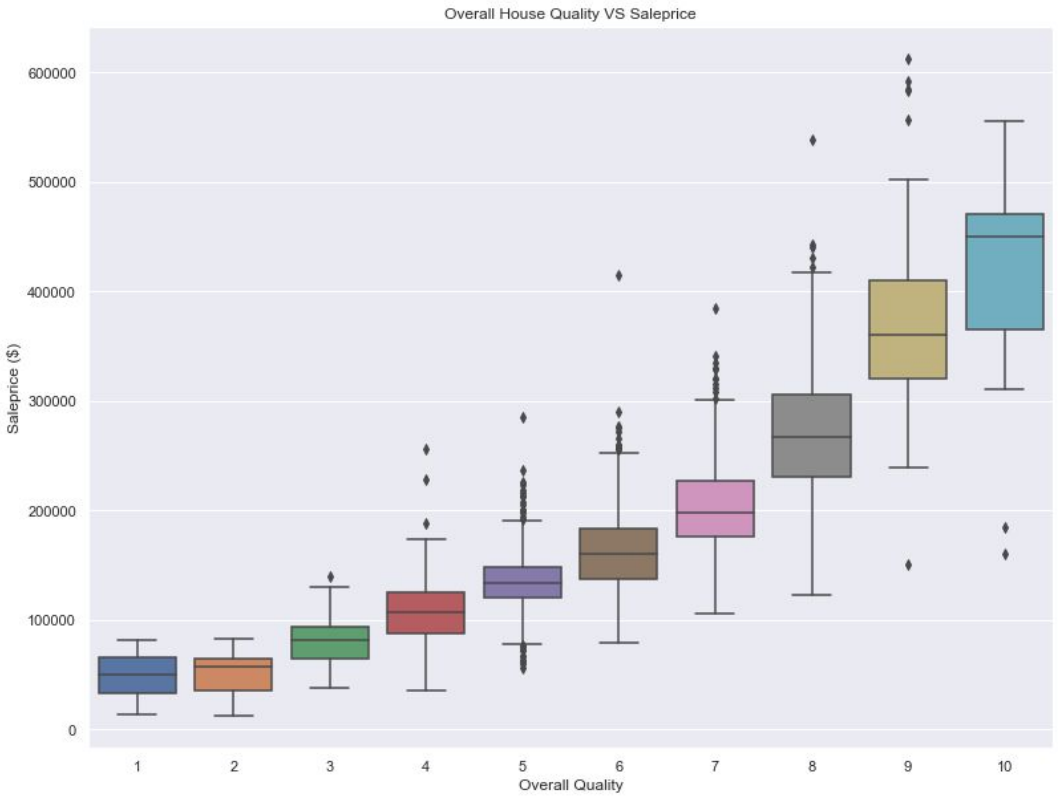
Recursive Feature Elimination

Recursive Feature Elimination

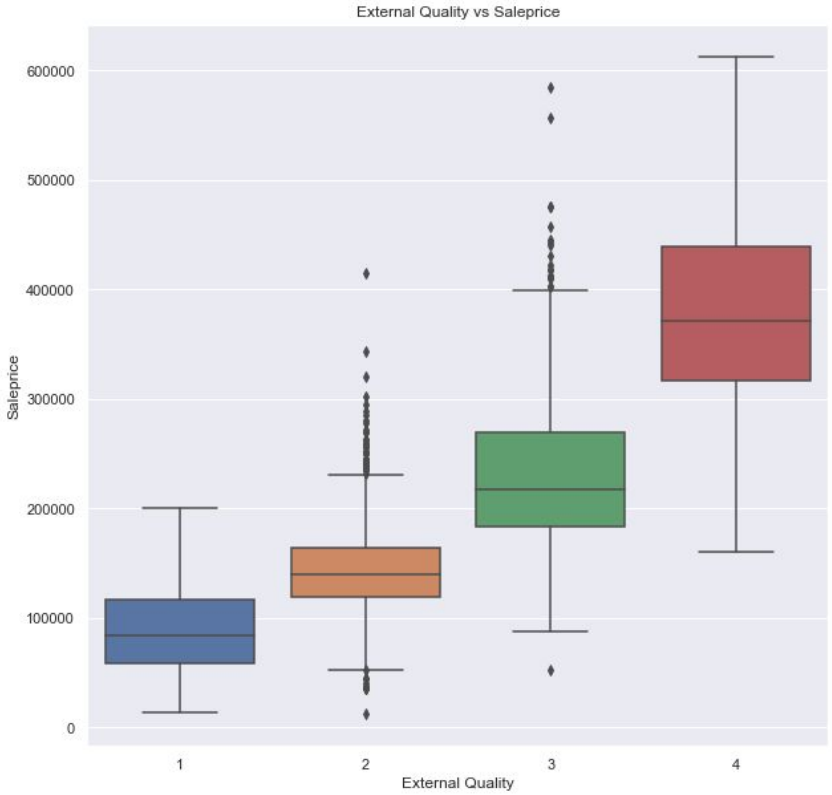
Overall Quality
Basement Full/Half Bath
Full/Half Bath
Total Bedroom/ Kitchen/ Rooms
Fireplace
Garage capacity

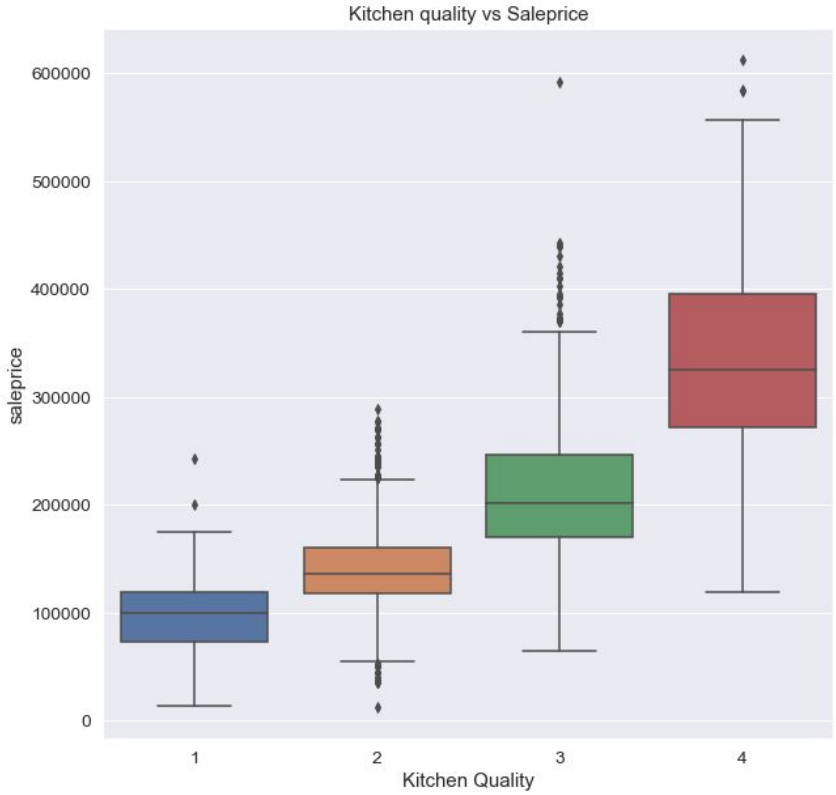
- As indicated by RFE, these variables might be important factors of sale price
- The RFE ranked the most basic but important features as the top priority
- This gives us a good indication of what features affects the sale price

RFE - Overall Quality

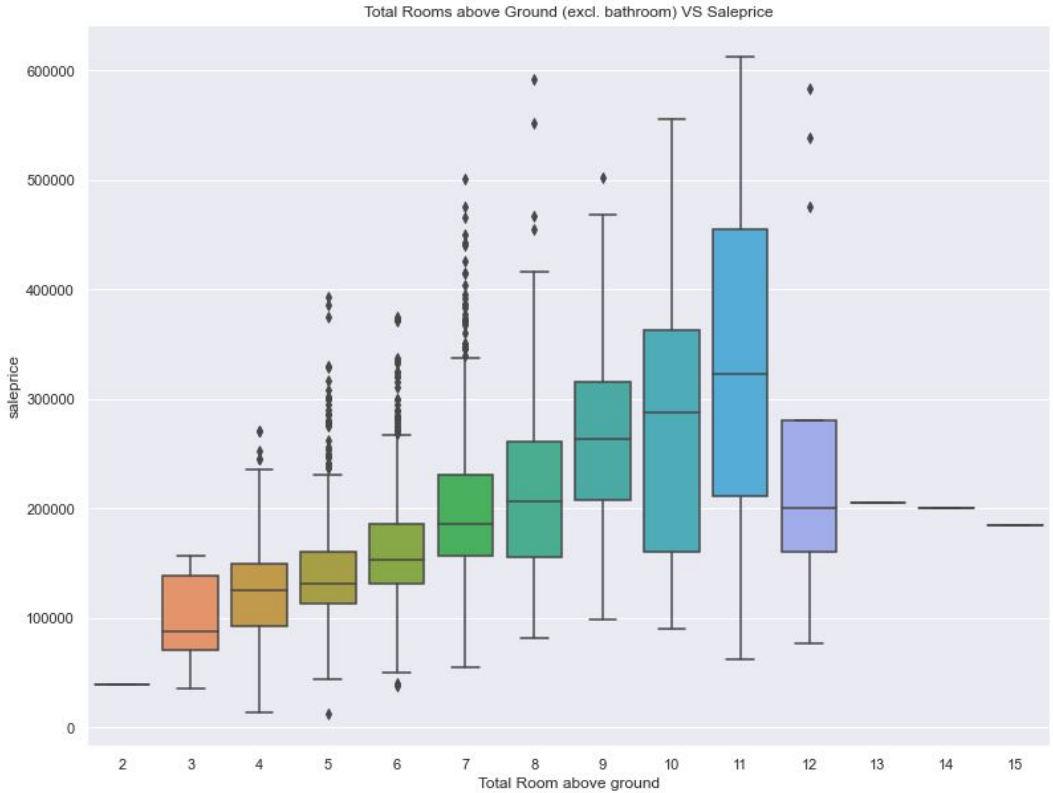


Additional findings - “Quality” affects the price



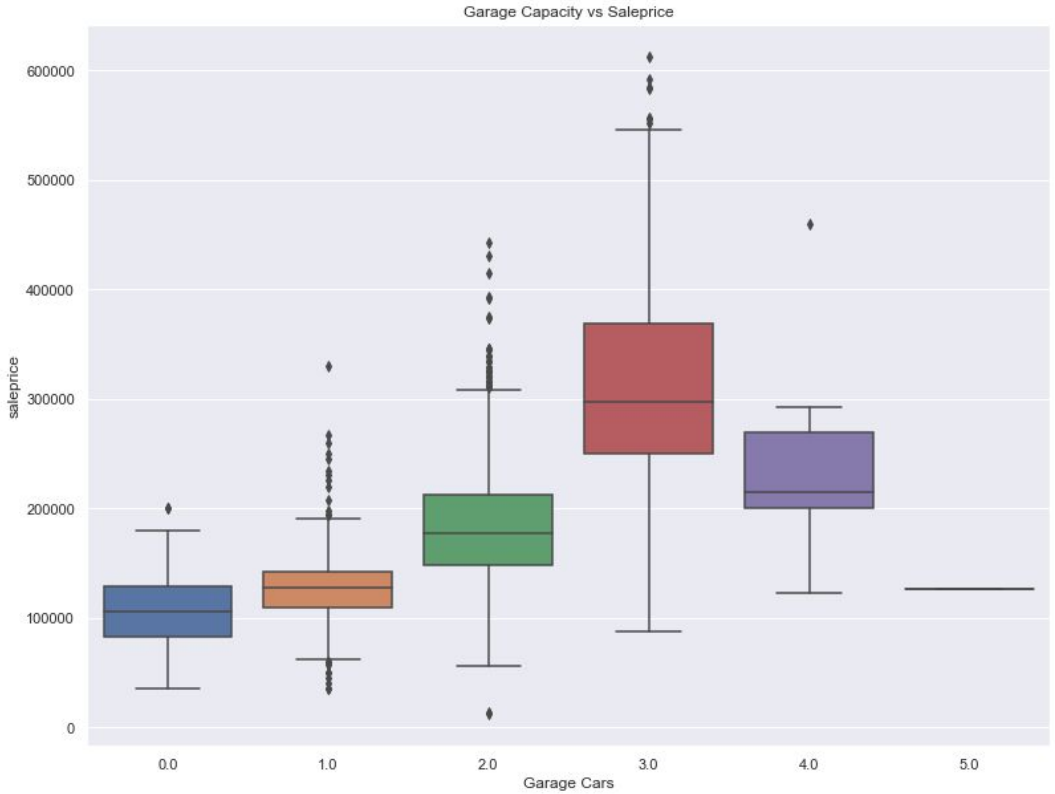


RFE - Total Rooms Above Ground



- Total Rooms above ground positively affects the sale price, which reflect the total livable space of the land

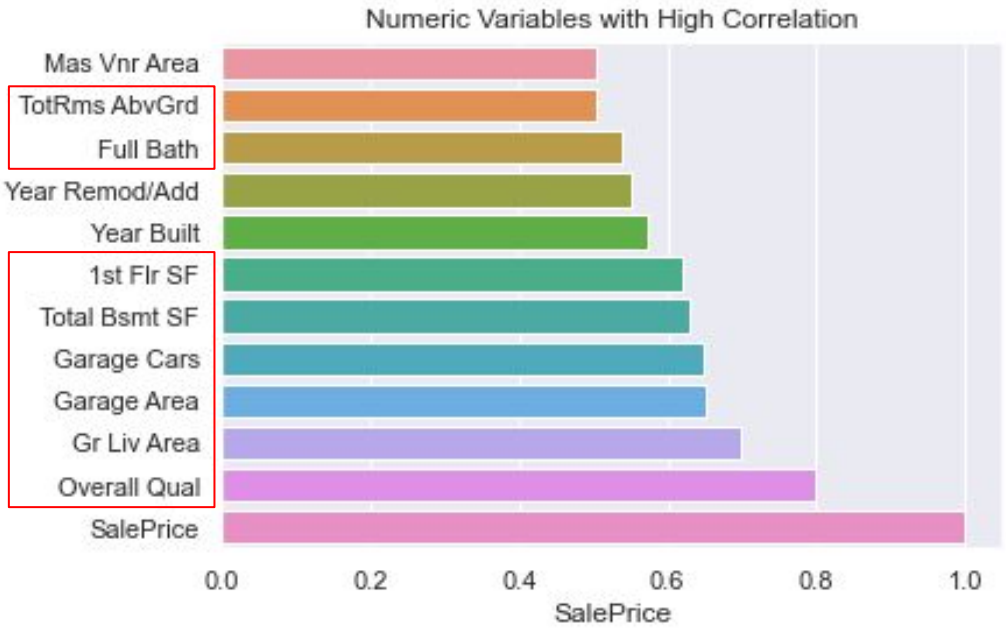
Garage Cars (Capacity)



- Assumed that most people in US owns car. With that assumption, it make sense that the saleprice increases with the garage car capacity as it is demanded.

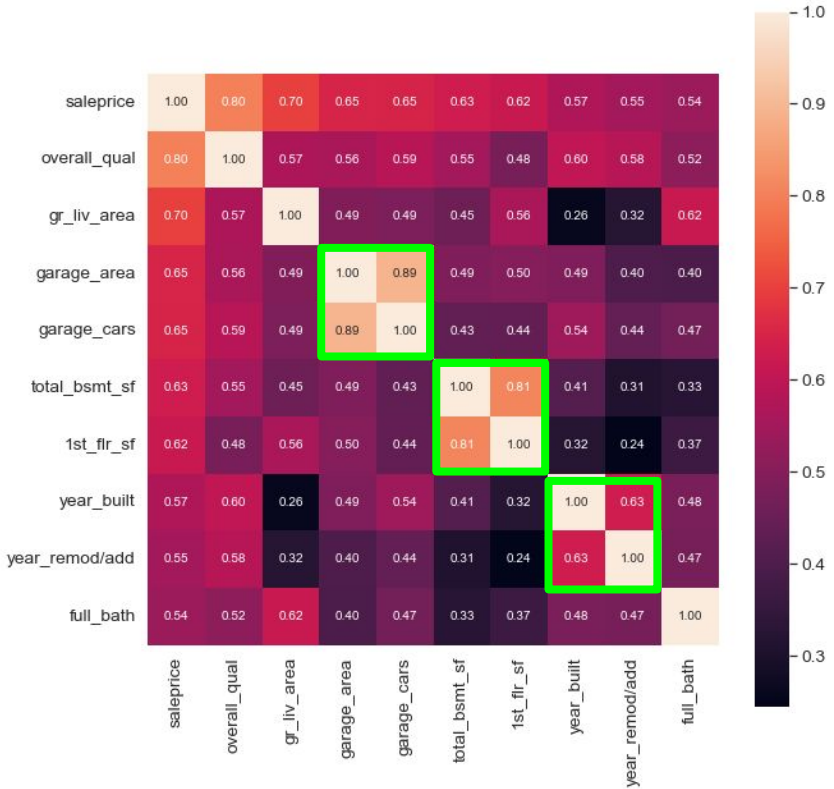
Correlation

Correlation



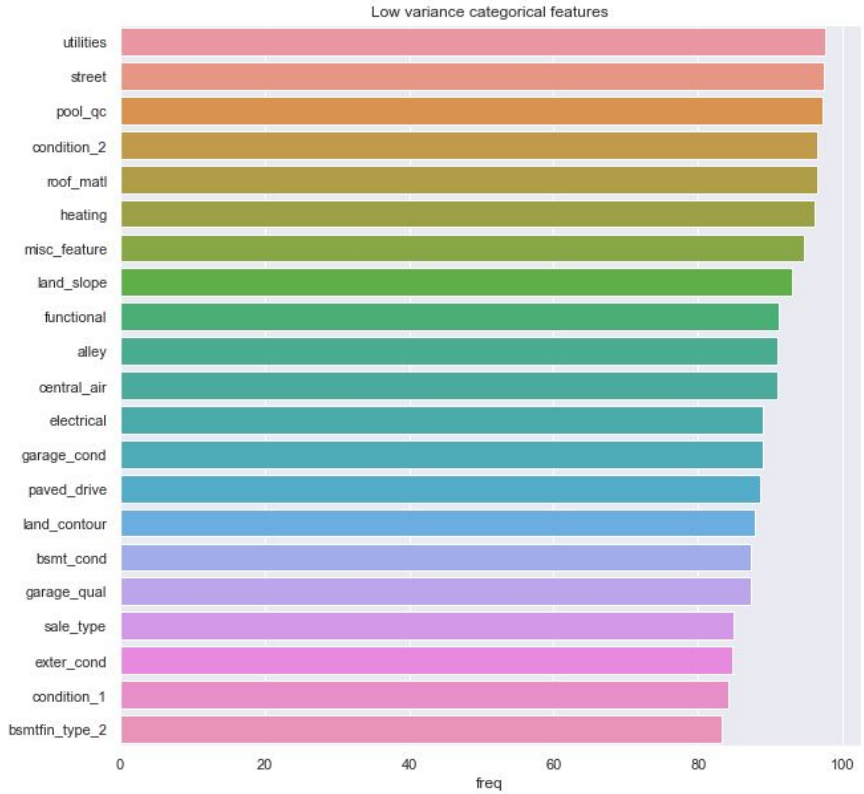
- Correlation shows similar result
  - Overall quality, ground living area and 1st floor area (could be equivalent to the total number of rooms), garage cars and number of bathrooms are similarly shown in the recursive feature elimination.

Correlation - Multicollinearity



- Garage cars and garage area show multicollinearity, garage cars feature will be drop in favour of garage area as it is a continuous feature
- 1st floor area and total basement area shows multicollinearity also.
- Year built and year remod/add also shows multicollinearity.

Categorical Features with low variance



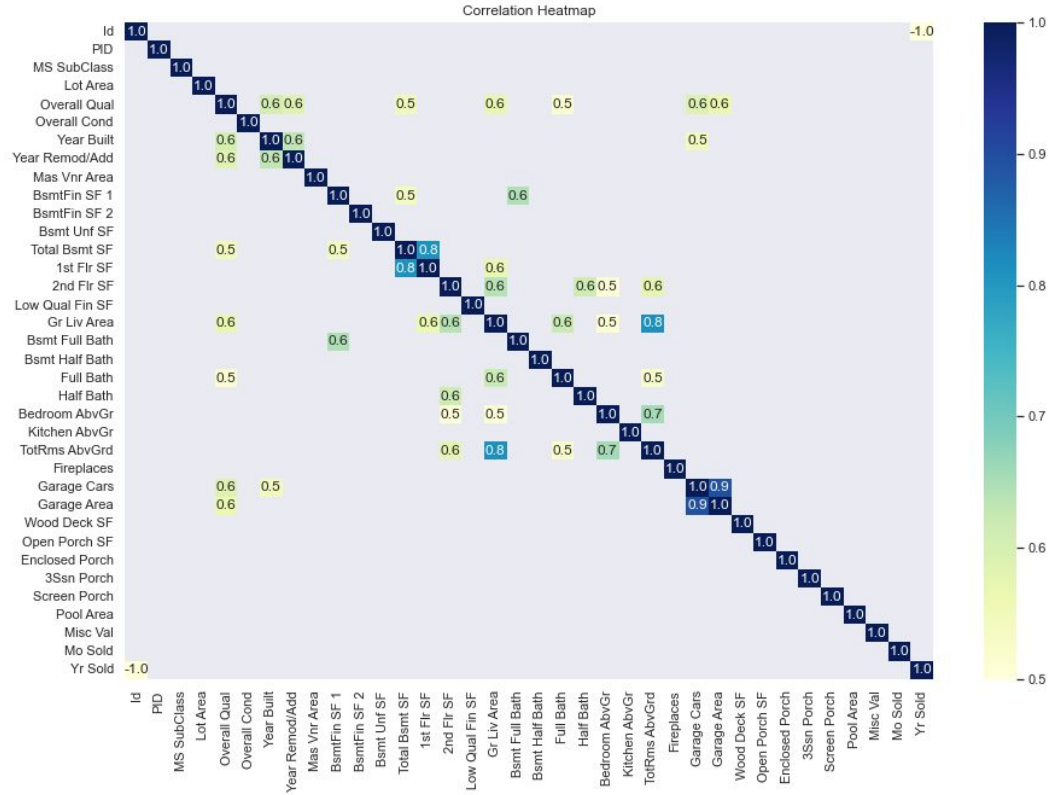
- Categorical Features where >80% of the data is the same value
- These features are eliminated

Feature Engineering

Feature Engineering - encoding

- After dropping features, categorical variables remaining:
  - 13 Ordinal Categorical variables
  - 10 Nominal Categorical variables
- Ordinal Encoding by rank for Ordinal Categorical features
- One Hot Encoding for Nominal Categorical Feature

High Correlation



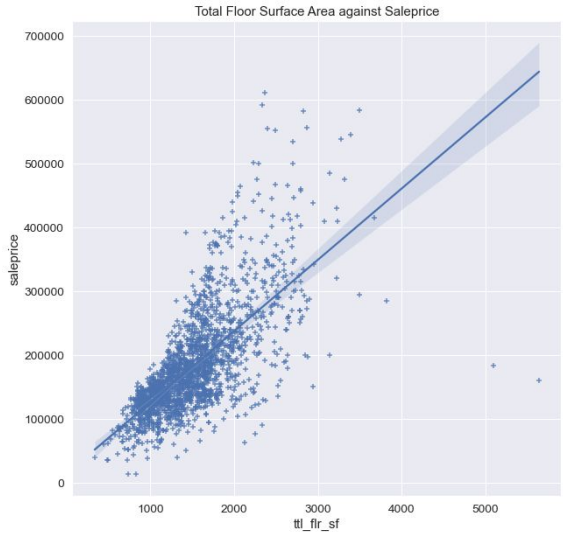
- Some features measuring square feet & number of rooms are correlated with one another

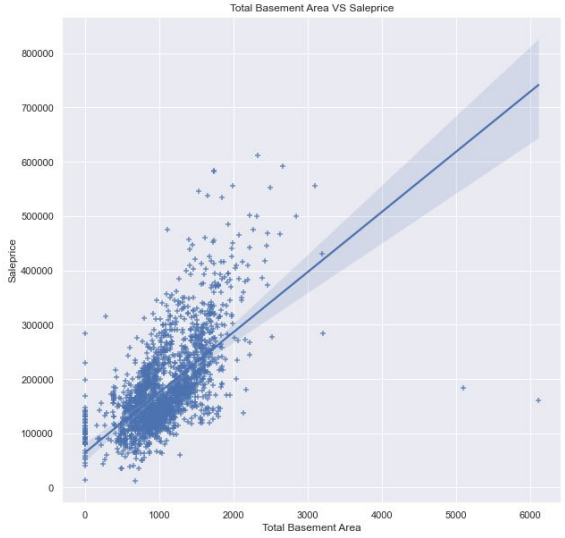
House Finishing

- Quality x Condition
  - Overall
  - Exterior
  - Basement
  - Garage
- New features generally shows a positive relationship with Sale Price

Feature Engineering - Total Area

- Basement
  - Basement Finishing * Basement SF
  - Combined 5 features
- Above Ground
  - Sum of SF (Low Quality SF discounted by 50%)
- Outdoor Area
  - Sum of Wood Deck, Open Porch, Enclosed Porch, 3Ssn Porch. Screen Porch SF
- Two outliers





Feature Engineering - Total Bathroom

- Full Bath + Half Bath * 0.5
- Consolidated amount for basement and above ground
- Total number of bathroom has positive relationship with Sale Price

Feature Engineering - Kitchen & Fireplace Score

- Multiply count of features and their quality
  - Kitchen
  - Fireplace
- Both new features shows a slight positive relationship with

Feature Engineering - House Age

- Year Sold - Year Built
- Negative relationship between house age and house age

Model Evaluation and Selection

Model	CV R-Squared	Validation R-Squared	RMSE	Alpha
Linear Reg	0.89	0.91	23022	NA
Ridge	0.91	0.91	22887	1.12
Lasso	0.91	0.91	22961	1.0

- Features are scaled to values between 0-1
- Ridge Model
- Best RMSE score (Lowest)
- Alpha of 1.12
- RMSE on test dataset - 21648

Important Features

- Features extracted from our final model (high coefficient)
  - Above Ground SF (most important)
  - Overall Finishing Score (2nd most important)
  - Basement Finishing + Square Feet
  - Neighbourhood (Greenhill & Stone Bridge)
  - House Age

Summary

- Ridge Regression model is a useful model to help predict future prices
- Key questions:
  - Neighbourhood is a strong influencer on Sale Price
  - House Size is an important feature from our model
  - The inner house area above the ground floor seems to be the most valuable area
  - The overall finishing (Quality and Condition) also influences Sale Price quite significantly
  - House age is also a good predictor of houses
- Limitations of our findings
  - There are other social, economic and political factors that are likely to heavily influence house prices. We are not able to control for these variables with the information provided in our dataset.

Recommendation

1. We can build a Ridge Regression model to help predict prices of houses that will be added to the listing from the property agency
2. The most important data to collect are above ground living area and overall finishing (Quality and Condition), Lot Area, Basement SF & Finishing, neighbourhood and house age.
3. More external data, such as Demographics and Interest rate should be collected and added to our prediction model