Capstone Project 1: Milestone Report

Chirag Limbachia

1. Problem Statement

Buying a house can be a very challenging process. It takes time, patience, and a lot of research to find a house you like and negotiate the right price for it. There are several features that influence the house price, 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 these 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? These are a few questions among many more that I intend to explore and answer.

This analysis can be useful for a house buyer in terms of negotiating the right price for the house. It can also be useful for a real estate investor in terms of assessing the realistic value of the housing property.

2. Data Description and Wrangling

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. Although the dataset is well structured, it did require a fair amount of data wrangling/cleaning. There were outliers and a lot of missing values in the dataset. There was also an opportunity to gather more information such as latitude-longitude location, zip code, and median household income from other sources (geocoding, and uszipcode libraries) to augment the dataset. The final wrangled data is made available as a pickle file on GitHub.

Following were the data wrangling steps performed:

- Outlier identification and removal
- Imputation of missing values
- Data augmentation

2.1. Outlier identification and removal

Distributions of some of the numerical features like Grade Living Area (GrLivArea), and Lot Area were skewed. The skewness was mainly due to presence of outliers. Outliers in GrLivArea are identified visually by plotting it against the target variable (Sale Price) (Figure 1). Points beyond GrLivArea of 4000 seemed like outliers and hence removed. Similarly for the Lot Area, points beyond the 99th percentile (marked with a red dashed line in Figure 2) seemed like outliers and hence removed. After outlier removal, the total number of instances came down to 1382 from 1460.

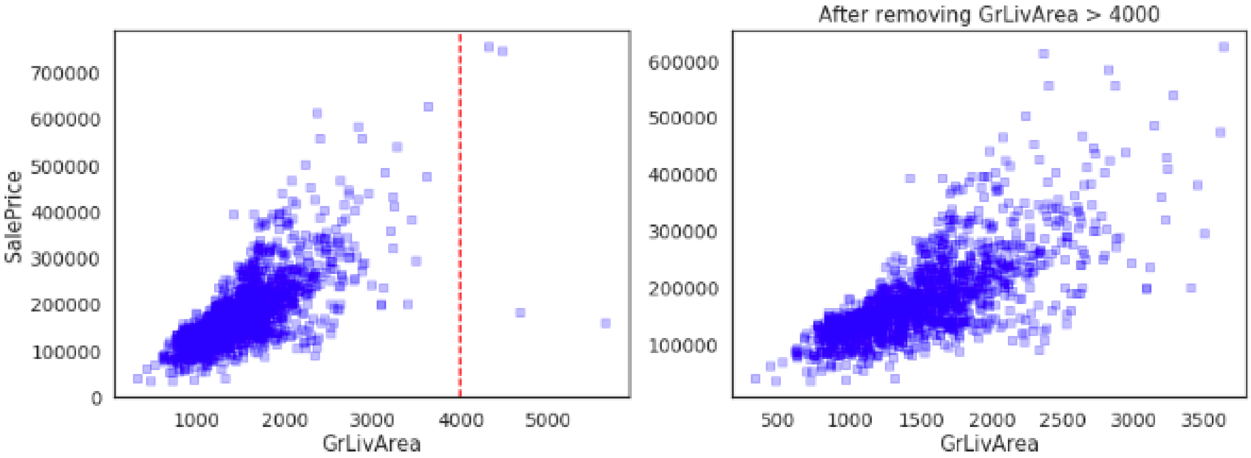


Figure 1. GrLivArea vs. Sale Price. Before (left) and after (right) outlier removal.

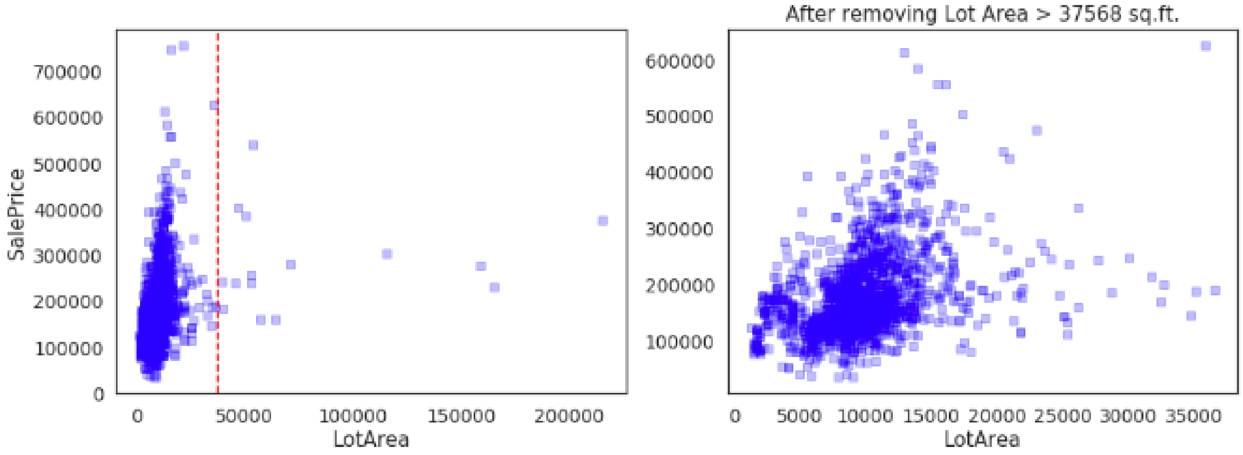


Figure 2. LotArea vs. Sale Price. Before (left) and after (right) outlier removal.

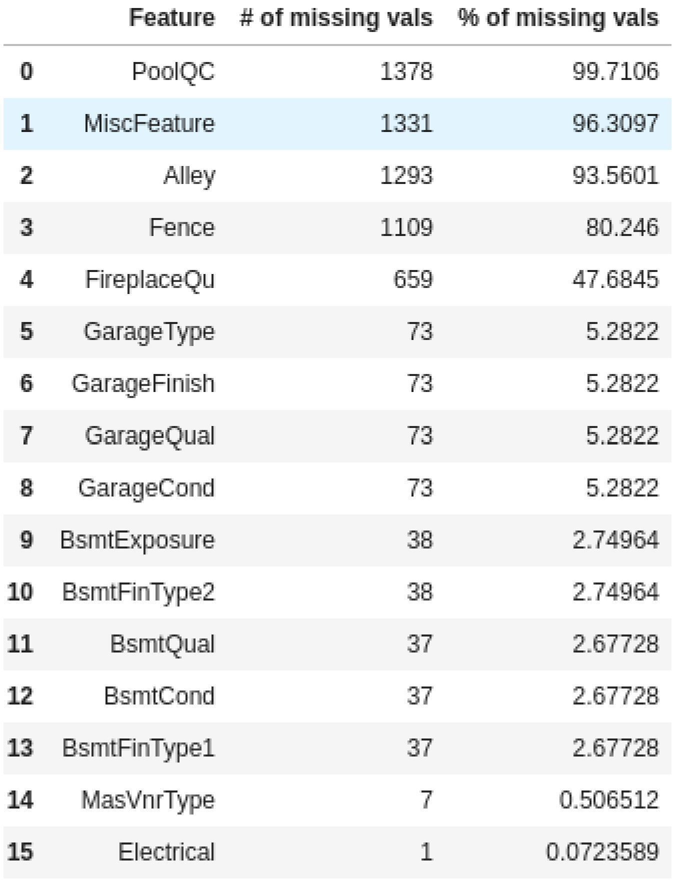
2.2. Imputation of missing values

Categorical features

There were several categorical features with missing values (see Table 1). All except two features had structurally missing values due to non-existence of the said (or related) feature. For example, if a house does not have a Fireplace, it has a missing value for the FireplaceQu (fireplace quality). Therefore, a new category called “Missing” was imputed for missing instances of the categorical features.

Other two features (Electrical & MasVnrType) had random missing values. These instances were imputed with the most frequently occurring (mode) category.

Table 1. Categorical features with total number and percentage of missing values.



	Feature	# of missing vals	% of missing vals
0	PoolQC	1378	99.7106
1	MiscFeature	1331	96.3097
2	Alley	1293	93.5601
3	Fence	1109	80.246
4	FireplaceQu	659	47.6845
5	GarageType	73	5.2822
6	GarageFinish	73	5.2822
7	GarageQual	73	5.2822
8	GarageCond	73	5.2822
9	BsmtExposure	38	2.74964
10	BsmtFinType2	38	2.74964
11	BsmtQual	37	2.67728
12	BsmtCond	37	2.67728
13	BsmtFinType1	37	2.67728
14	MasVnrType	7	0.506512
15	Electrical	1	0.0723589

Numerical Features

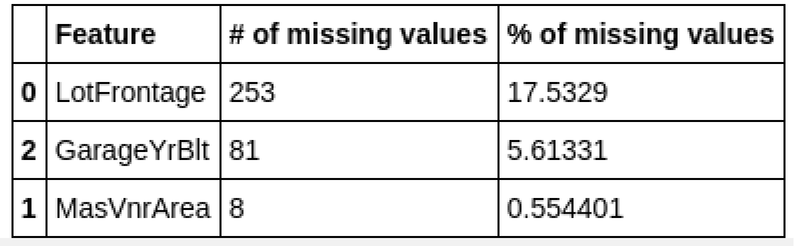
There were three features with missing values: Lot Frontage, Garage Year Built (GarageYrBlt), and Masonry Veneer Area (MasVnrArea).

GarageYrBlt: Values were missing for those houses that did not have a garage. The median of Garage Year Built was imputed for missing instances.

MasVnrArea: Related to the categorical feature Masonry Veneer Type (MasVnrType). Group specific MasVnrArea medians were computed and imputed for missing instances in MasVnrArea. Grouping was done by the type of masonry veneer used in that house (MasVnrType).

LotFrontage: Related to two categorical features, Lot Configuration (LotConfig) and LotShape. Again, group specific medians were computed and imputed for missing instances in LotFrantage.

Table 2. Numerical features with total number and percentage of missing values.



	Feature	# of missing values	% of missing values
0	LotFrontage	253	17.5329
2	GarageYrBlt	81	5.61331
1	MasVnrArea	8	0.554401

2.3. Data Augmentation

The Neighborhood feature of the dataset consisted of names of the neighborhood in which a particular house was located. Using the neighborhood name, city and state information, geo coordinates and zip codes were extracted from Google’s Geocoding API. Furthermore using the zip codes, demographic information such as median household income, and median home value were also extracted using the *uszipcode* library.

3. Initial Findings

This section highlights some interesting trends and correlations in the data.

Several features seem to influence the sale price. This section will talk about a few important ones. It will also ask and address some interesting questions that are not centric to the problem at hand (i.e., prediction of house price). Knowing answers to these questions can help in making informed business decisions.

House prices are highly correlated with the overall quality of the house (OverallQual); better quality houses have a higher sale price (Figure 3).

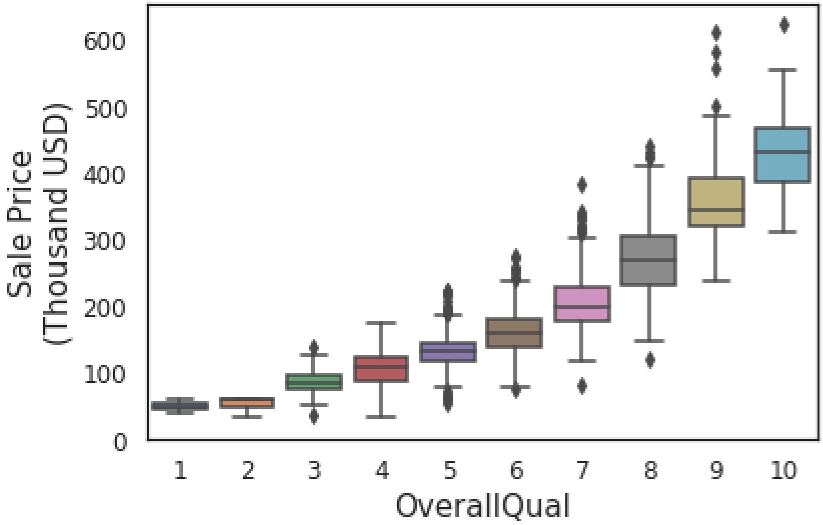


Figure 3. Influence of house’s overall quality on sale price. Better quality houses cost more.

Sale price is also correlated with variables that reflect size of the housing property such as, size of the 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 (Figure 4).

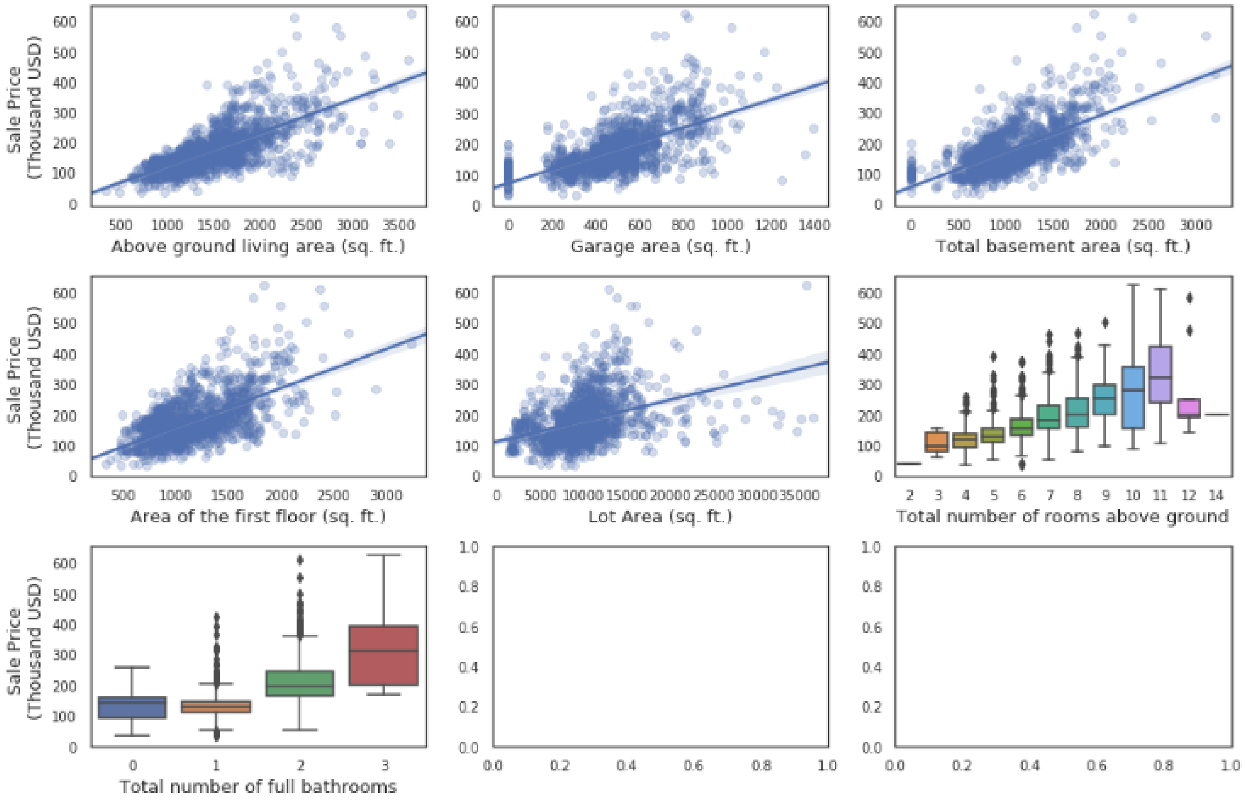


Figure 4. Influence of variables that reflect size of the housing property on sale price.

Neighborhood also has an influence on the sale price. Houses in Northridge Heights, Northridge, and Stone Brook have the highest average sale prices. Whereas, houses in Meadow Village, Iowa DOT and Rail Road, and Briardale have the lowest average sale prices (Figure 5).

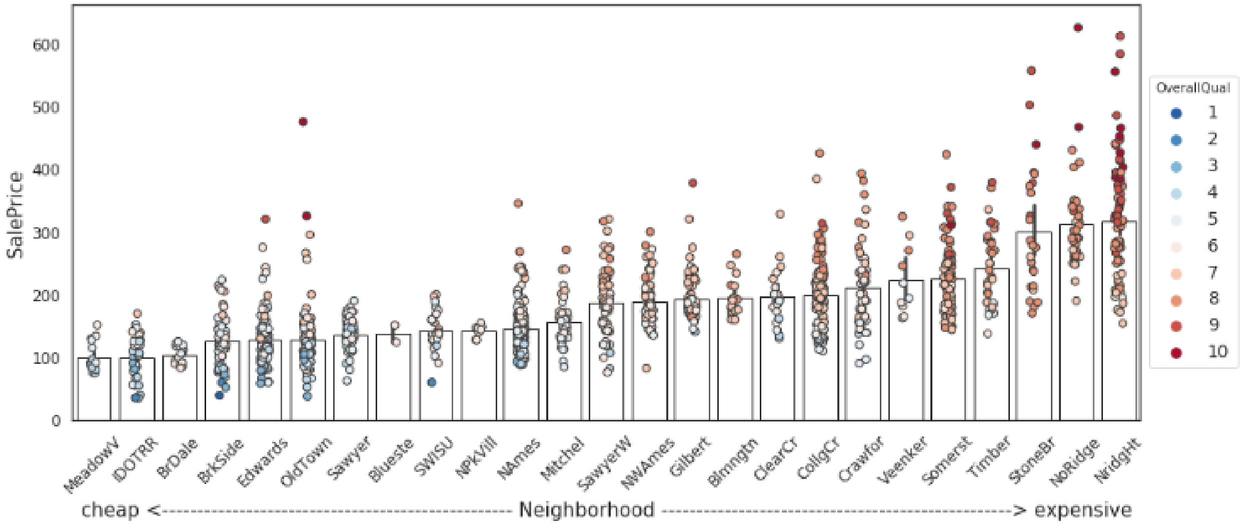


Figure 5. Average sale price of houses in each neighborhood.

Northridge Heights, Northridge, and Stone Brook also have higher concentration of better quality houses. This explains why houses in these neighborhoods have a higher average sale price (Figure 6).

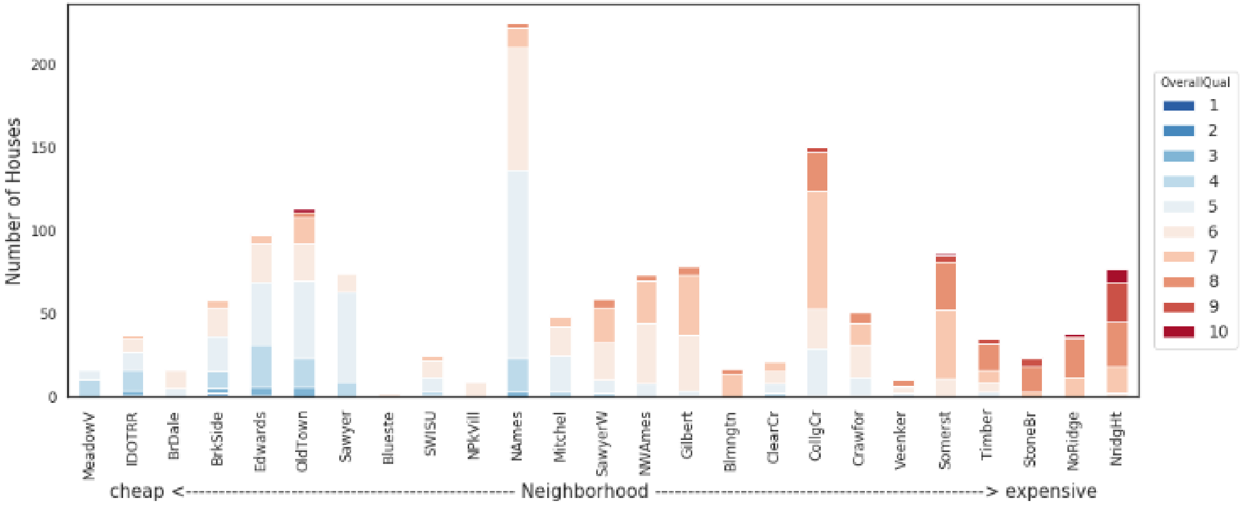


Figure 6. Total number of houses in each neighborhood. Colormap reflects the total number of houses of a particular quality..

Expensive neighborhoods like Northridge Heights, Northridge, and Stone Brooks also tend to have houses that either have a finished or a roughly finished garage. Rarely do they have houses with no garage. Whereas, cheaper neighborhoods like Meadow Village, Iowa DOT and Rail Road, and Briardale mostly have houses with unfinished or no garage (Figure 7).

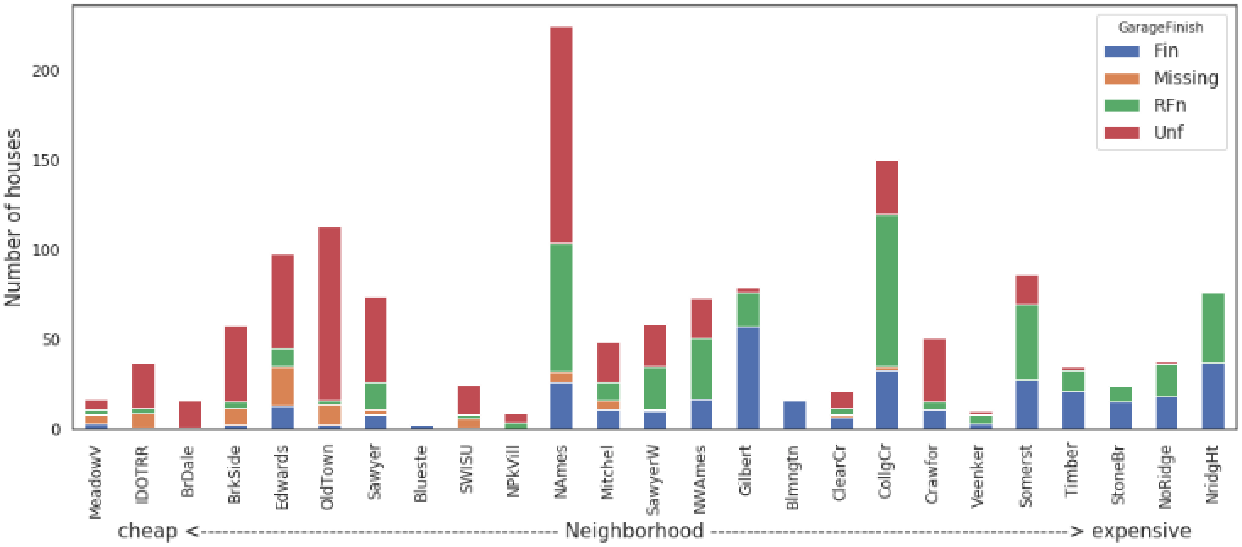


Figure 7. Total number of houses in each neighborhood. Colormap reflects the total number of houses of a particular kind of garage finish.

Attached garage are common across all neighborhoods except a few cheaper neighborhoods. Built-in garages are more common amongst houses in expensive neighborhoods. Whereas, detached or no garage are most common in cheaper neighborhoods (Figure 8).

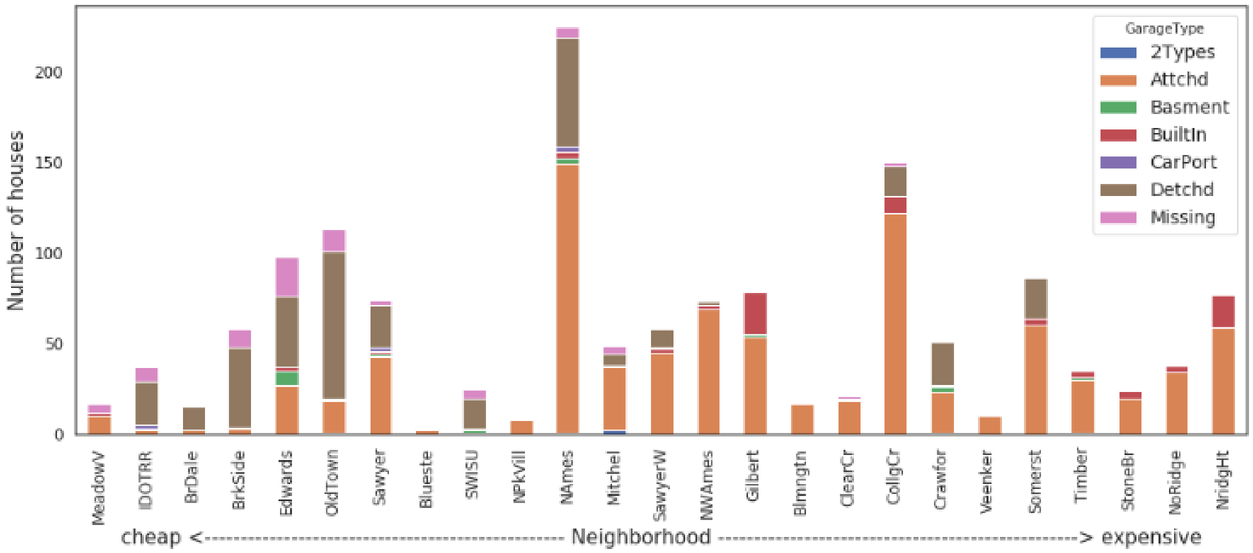


Figure 8. Total number of houses in each neighborhood. Colormap reflects the total number of houses of a particular type of garage.

Most houses are remodeled within a couple years after they are built (Figure 9).

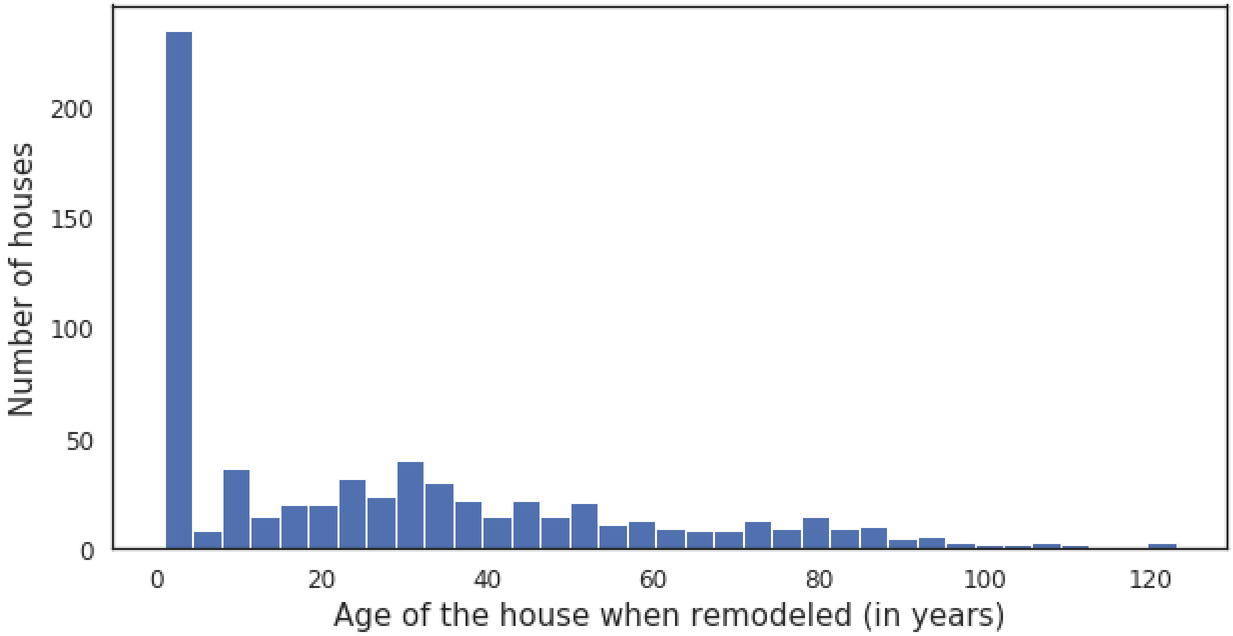


Figure 9. Number of houses remodeled at a particular age.

All houses were sold between 2006-2010 (Figure 10).

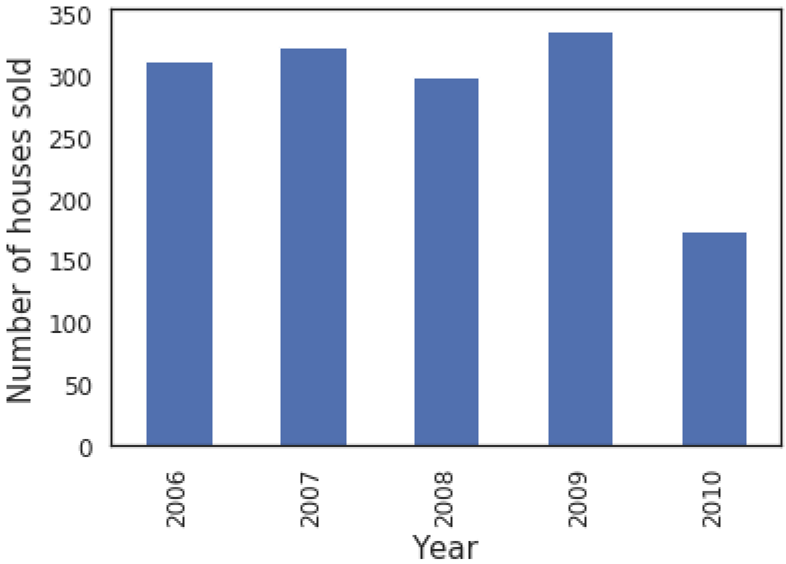


Figure 10. Number of houses sold in a particular year.

Most houses got sold during summer; highest in the month of June (Figure 11).

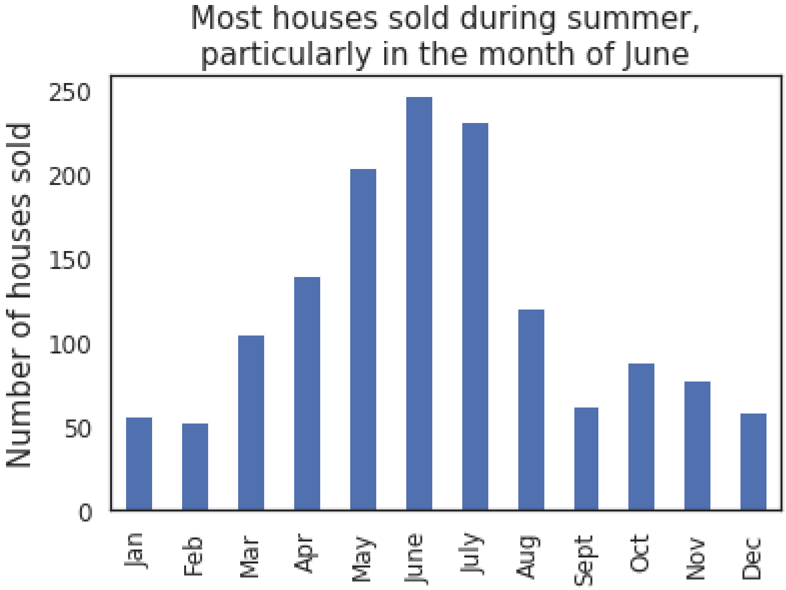


Figure 11. Number of houses sold in a particular month.

Sale prices seem to fluctuate across the year. They fall 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 (Figure 12).

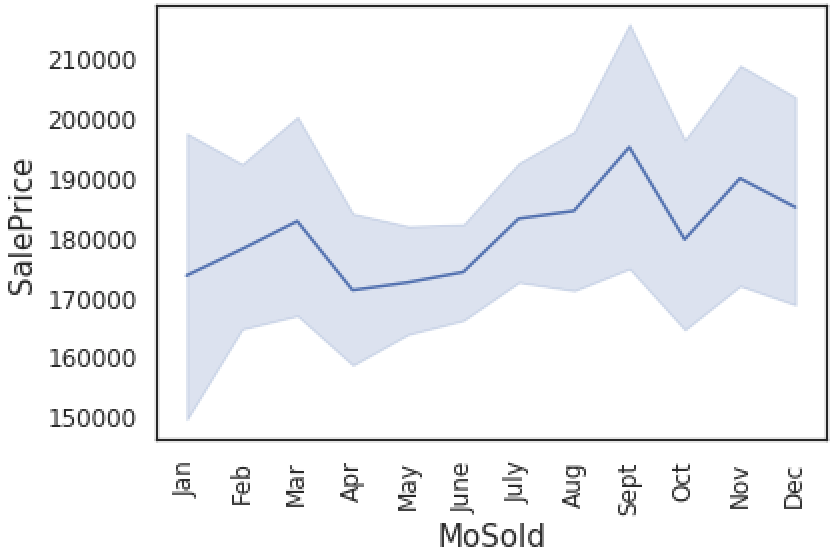


Figure 12. House price fluctuation across a year.

Houses located in floating village residential (FV) zones are most expensive, whereas those located in the commercial (C (all)) zones are the cheapest (Figure 13).

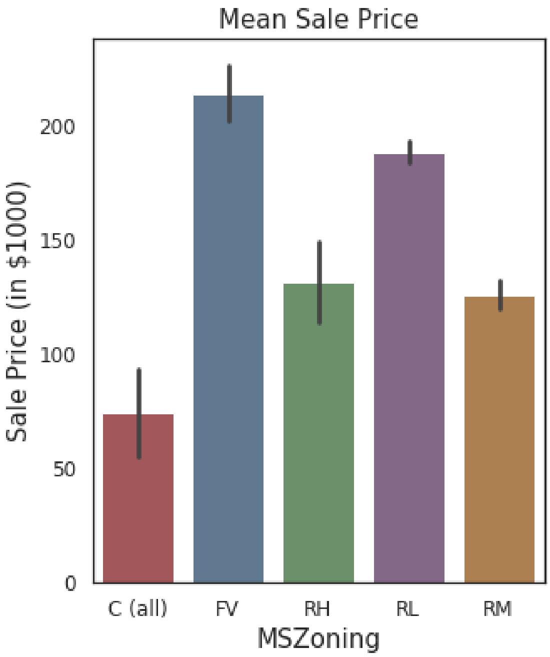


Figure 13. Average Sale Price by Zoning.

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).

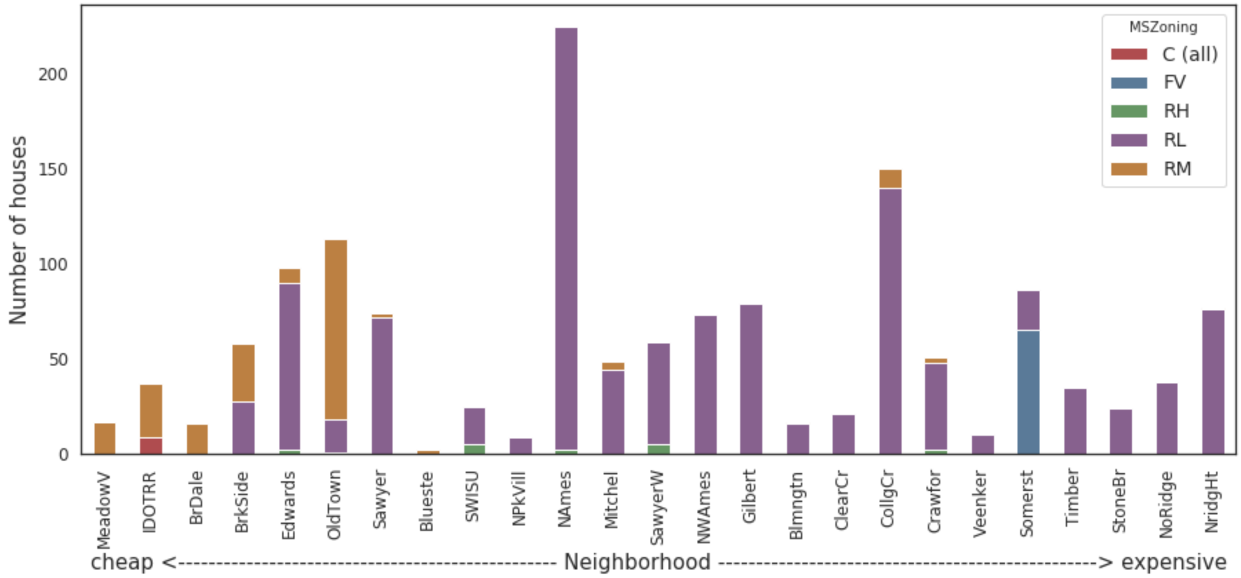


Figure 14. Average Sale Price by Neighborhood and Zoning.

4. Feature Engineering

4.1. Numeric Features

Feature correlation matrix

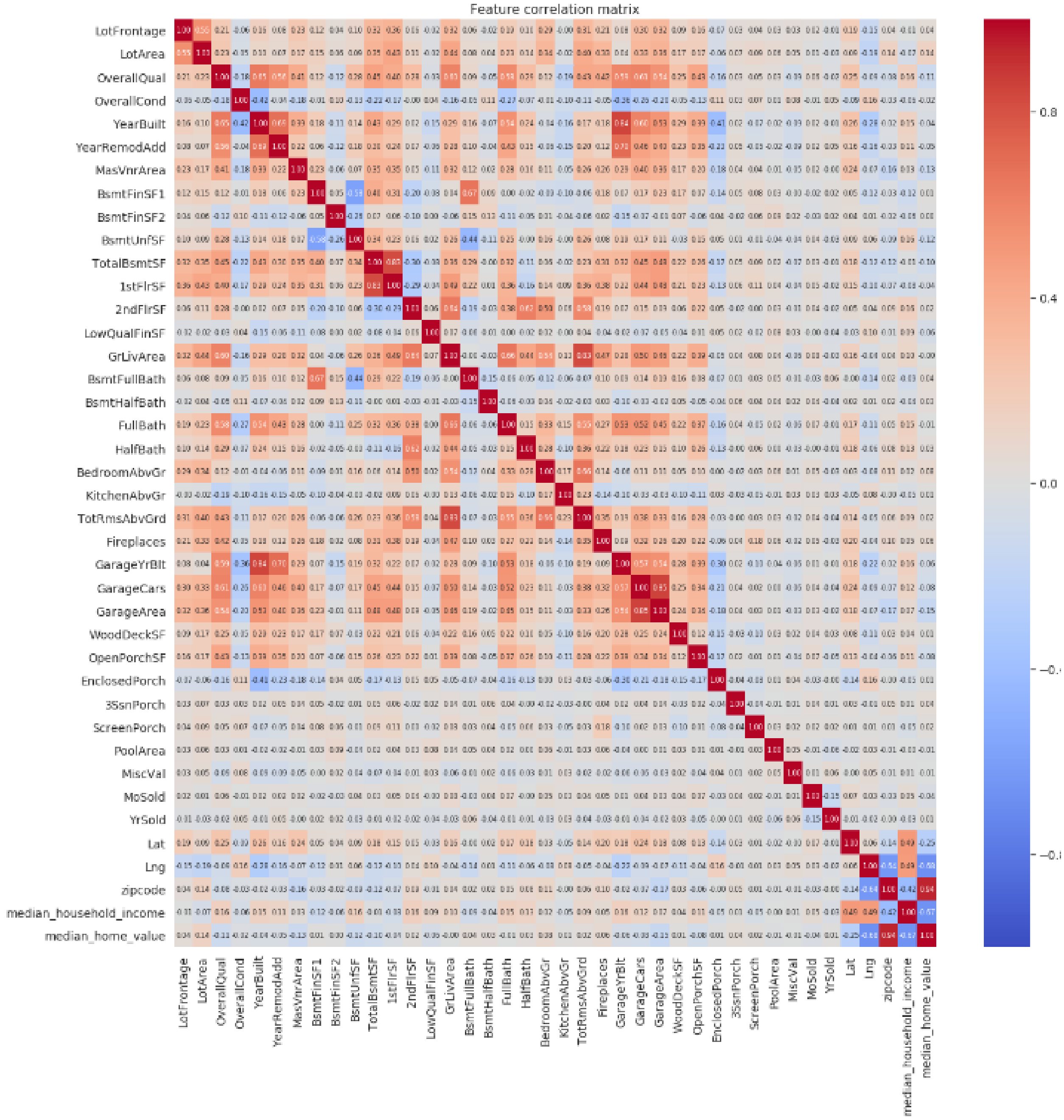


Figure 15. Pairwise correlations (spearman) of all numerical features.

Several features are highly correlated with one another. Highest positive correlation is between house and garage year built (YearBuilt and GarageYrBlt). This is not unexpected, as most house garages are built along with the house itself. A new feature, GarageYrBltMinusYearBuilt, is added which models the difference between the year in which the garage and the house was built.

Second highest positive correlation is between the total number of rooms (TotRmsAbvGrd) and total square feet area above ground (GrLivArea). This is also expected as houses with greater area tend to have more number of rooms. Another new feature, AreaPerRoom, is included which models the average area per room by dividing the TotRmsAbvGrd by GrLivArea.

Correlation of the newly added features with the rest of the features.

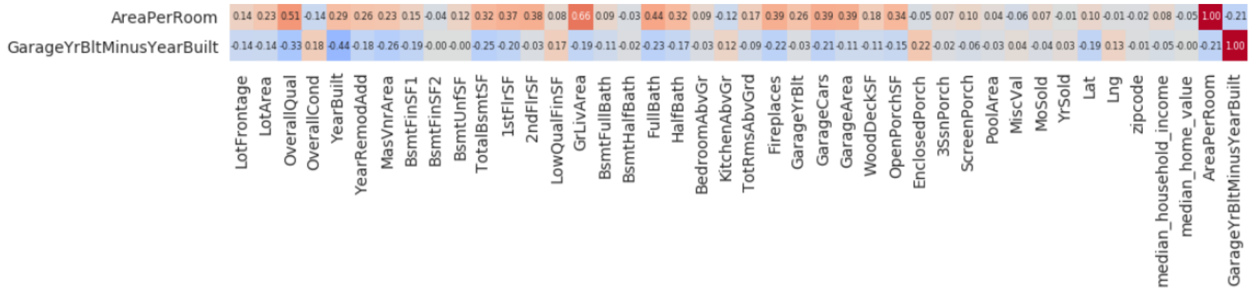


Figure 16. Spearman correlation of newly added features with the rest of the features.

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

Correlation with the target (sale price)

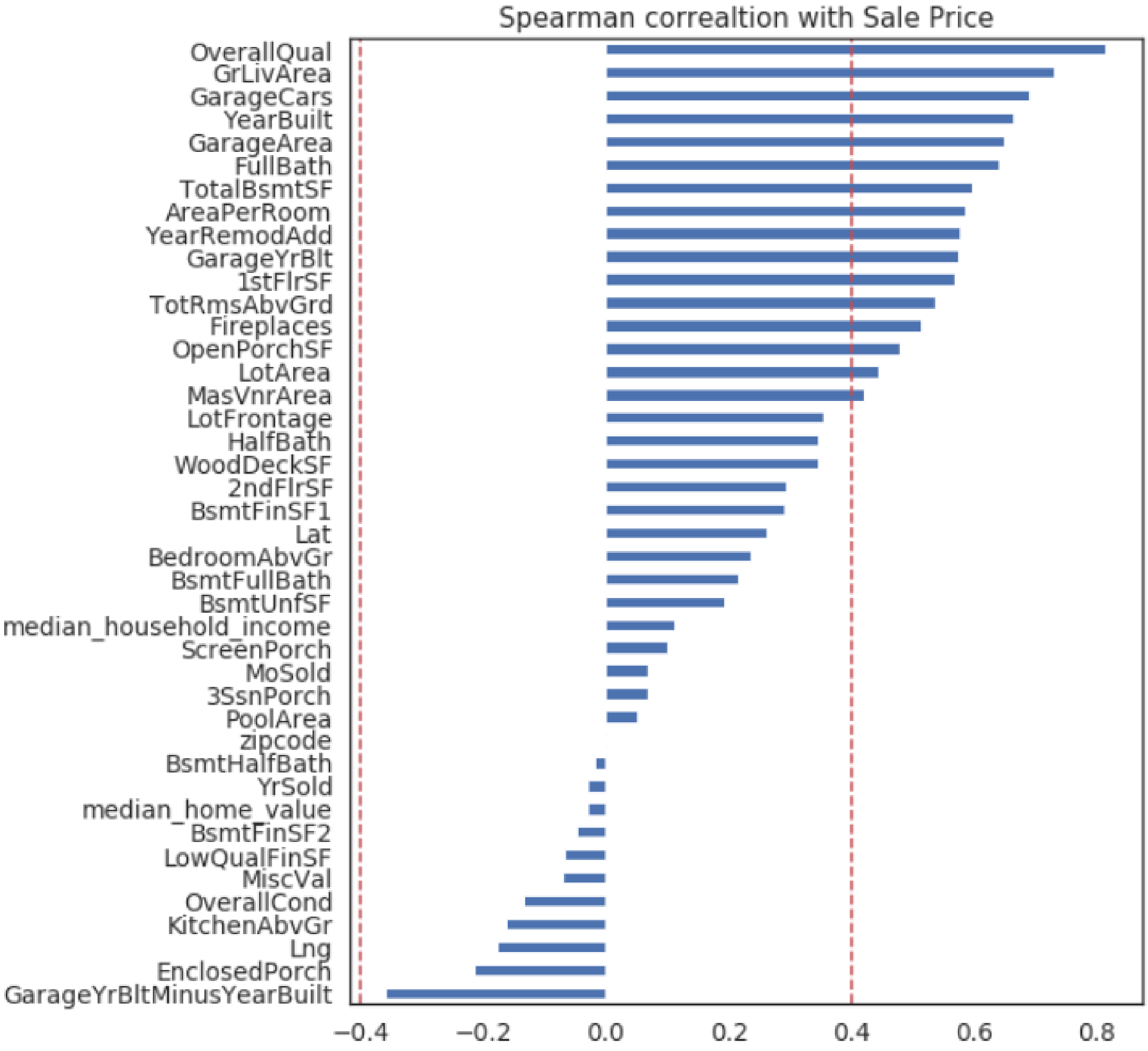


Figure 17. Feature correlations with the sale price.

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.

4.2. Categorical Features

Kruskal (non-parametric) ANOVA

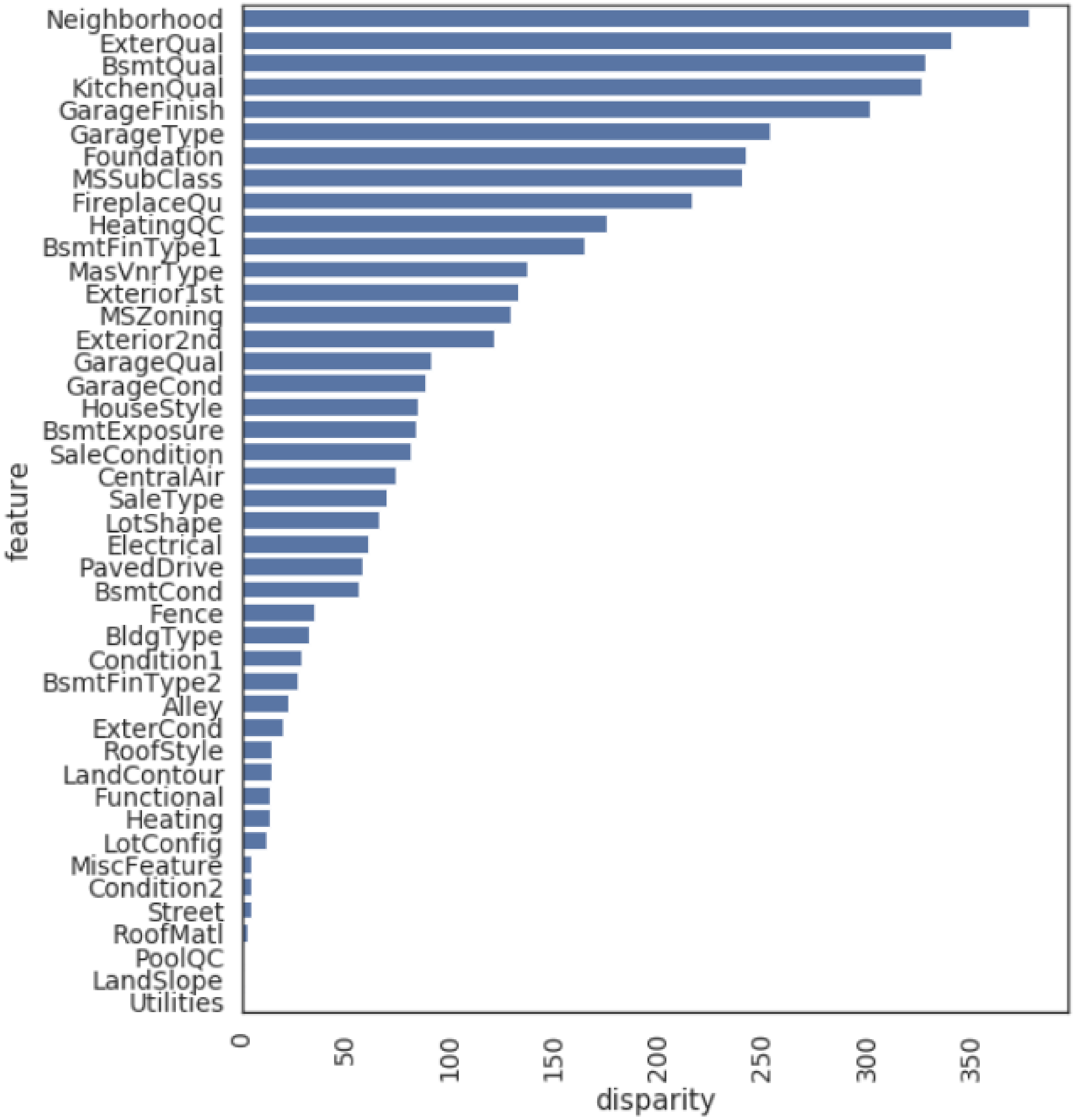


Figure 18. Non-parametric ANOVA disparity measure for all categorical features with respect to sale price.

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.

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.