u3224942 EDA Report An Extensive analysis into the Ames Housing Dataset

Somtochukwu Nnajide

2024-04-16

Contents

1. Abstract

Rigorous pre-processing has been performed while maintaining dataset consistency across the train and test sets. Transitioning to an extensive exploratory data analysis where problems of interest are explored and insights are derived. Further pre-processing is performed with the help of random forests to feature engineer new variables and select significant features.

Finally an iterative and thorough modelling process is done with modelling algorithms; k-nearest neighbour, random forest and linear regression across different transformations to the target variable with the aim of selecting the best model with the lowest RMSE to predict sale price.

For the purposes of keeping this report readable, no code blocks will be shown in the pdf document. Please refer to the markdown file for code analysis if necessary.

2. Problem Identification

The Ames Housing Dataset is a renowned resource in machine learning and statistics, containing information on residential properties in Ames, Iowa, USA, along with their sale prices. With 2930 observations and 82 variables, it offers a comprehensive view of housing market dynamics (Kaggle, 2024). Key attributes include sale price, lot area, overall quality and neighborhood and these variables cover various factors influencing housing prices, such as size, age, quality, location, and amenities.

The dataset is often used to build predictive models on sale price and for advanced research in data science and machine learning.

Variables of Interest I will focus on are overall quality, neighbourhood and month sold, along with other feature engineered results in the report. Some of the questions I will attempt to address are:

- Which neighbourhoods are more expensive than the rest ?
- Which neighbourhoods have a higher build quality than the rest ?
- What was the trend of median and mean price over the years ?
- What was the trend of number of sales across the years ?
- Was there a seasonal effect on the number of sales ?
- Does home quality affect its sale price ?
- What is the relatioship between a large square feet and sale price ?
- Are houses with more bathrooms more expensive ?

etc.., amongst many othe questions and problem areas I will be addressing in this report.

3. Data Pre-processing

Data preprocessing is a crucial step in the data analysis pipeline that involves transforming raw data into a clean and structured format suitable for analysis (Mesevage, 2021). It encompasses a variety of techniques and procedures aimed at preparing the data for further analysis and modelling (Mesevage, 2021). The importance of data pre-processing is highlighted below:

- Quality Assurance: Raw data often contain errors, inconsistencies, missing values, or outliers. Data cleaning helps identify and rectify such issues, ensuring data quality and reliability.

- Accuracy: Clean data leads to more accurate analysis and modeling outcomes. By removing errors and inconsistencies, preprocessing ensures that the insights derived from the data are trustworthy and reflect the true characteristics of the underlying data.

- Better Insights: Data preprocessing encourages the extraction of meaningful patterns, trends, and insights from the data. By removing noise and irrelevant information, preprocessing enhances the signal-to-noise ratio, making it easier to identify relevant patterns and relationships.
- Improved Model Performance: Clean and well-preprocessed data are essential for building accurate and robust predictive models. Models trained on dirty or unprocessed data are likely to produce unreliable predictions and perform poorly in real-world applications.

In my approach, I made sure to reflect all pre-processing performed on the train set unto the test set. This ensures consistency between both datasets and avoids modelling mismatches. Due to size of the dataset, I decided to clean every variable by a variety of methods depending on what is appropriate, methods are inclusive of replacing NA values, imputing medians, imputing modes, re-factoring orders etc.

Missing values identification and handling

Columns with missing values were identified in each dataset. The table below shows an overview of the columns and the amount of NAs in descending order.

```
## Columns with missing values in train set
```

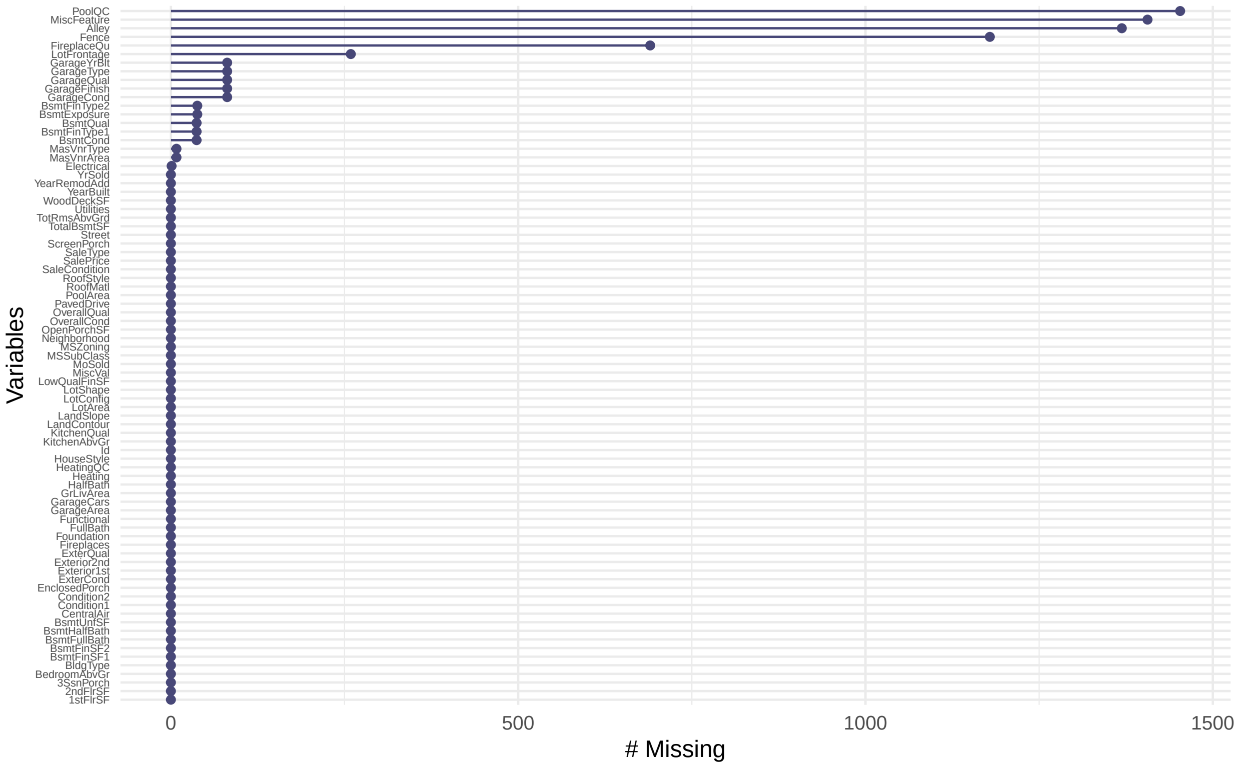
```
##       PoolQC  MiscFeature        Alley        Fence  FireplaceQu  LotFrontage
##         1453         1406         1369         1179          690          259
##   GarageType  GarageYrBlt GarageFinish   GarageQual   GarageCond BsmtExposure
##           81           81           81           81           81           38
## BsmtFinType2     BsmtQual     BsmtCond BsmtFinType1   MasVnrType   MasVnrArea
##           38           37           37           37            8            8
##   Electrical
##            1
```

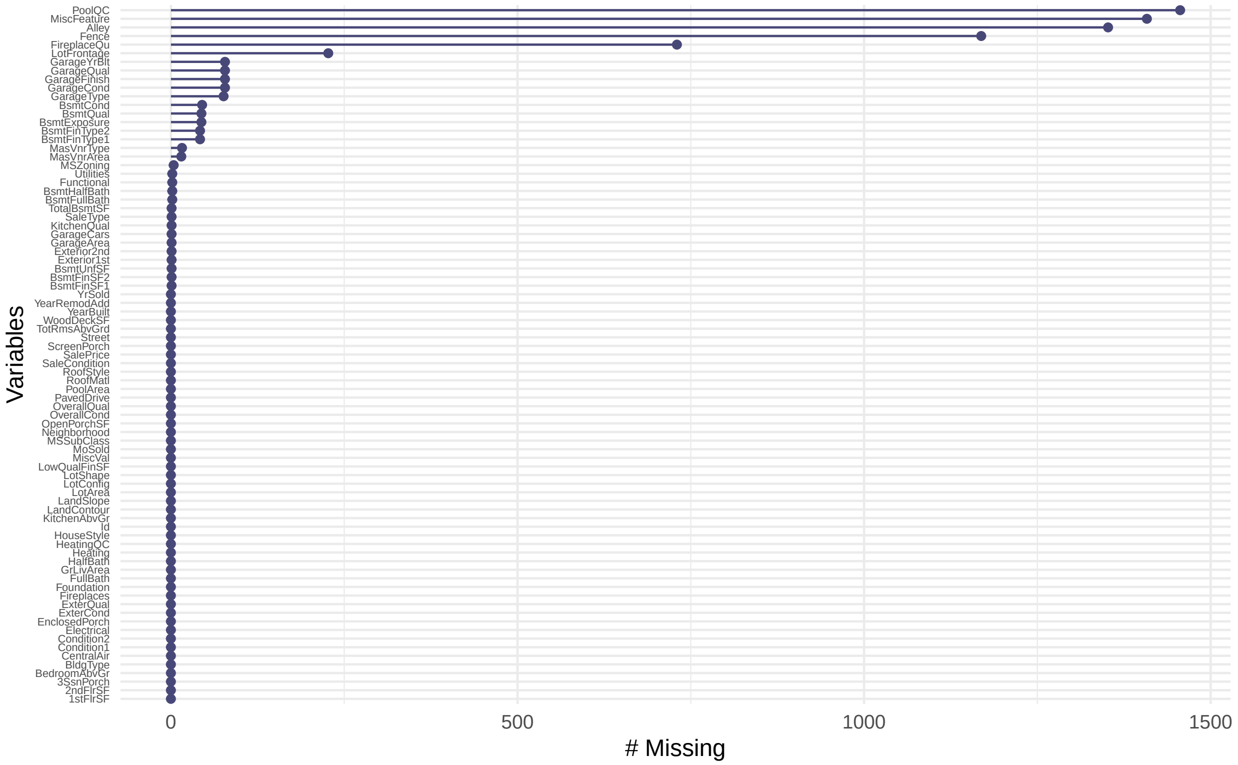
```
##
## Columns with missing values in test set
```

```
##       PoolQC  MiscFeature        Alley        Fence  FireplaceQu  LotFrontage
##         1456         1408         1352         1169          730          227
##  GarageYrBlt GarageFinish   GarageQual   GarageCond   GarageType     BsmtCond
##           78           78           78           78           76           45
##     BsmtQual BsmtExposure BsmtFinType1 BsmtFinType2   MasVnrType   MasVnrArea
##           44           44           42           42           16           15
##     MSZoning    Utilities BsmtFullBath BsmtHalfBath   Functional  Exterior1st
##            4            2            2            2            2            1
##  Exterior2nd   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF  KitchenQual
##            1            1            1            1            1            1
##   GarageCars   GarageArea     SaleType
##            1            1            1
```

The tables above indicate there are differences in the total number of missing values in both datasets. We also observe more than 50% of values in "PoolQC", "MiscFeature", "Alley" and "Fence" are marked as "NA". Further investigation will be done to determine if these are actually missing values or are labels marked as "NA" for the variable.

The plots below are a visual representation of the missing values.

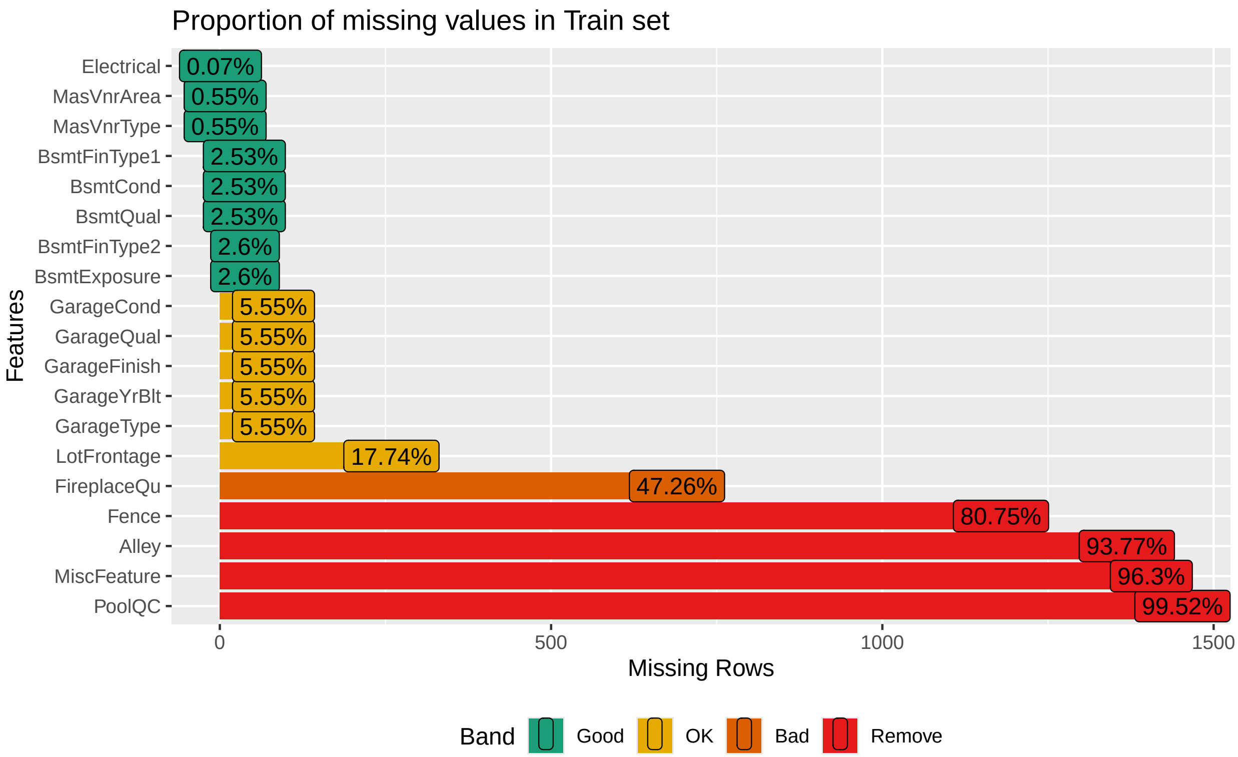


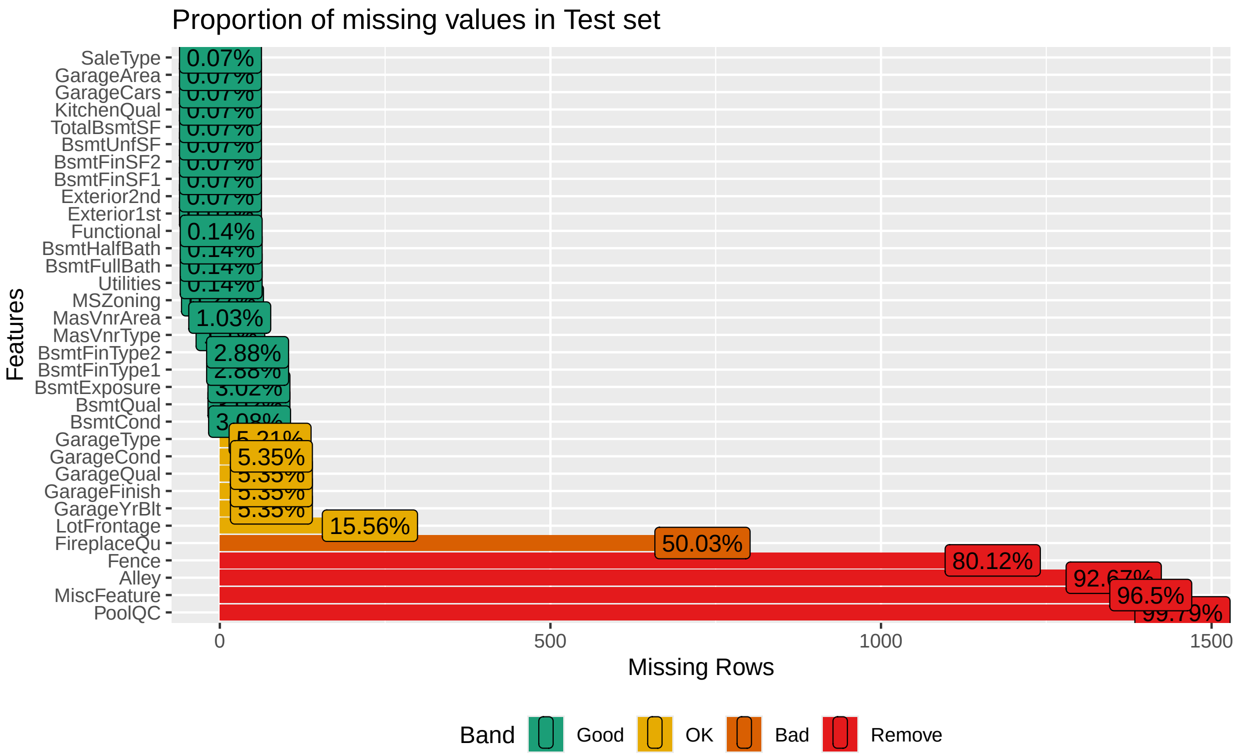


We also notice there are more columns with missing values in the test set than training set, most probably has to do with how the data was split.

```
## There are 19 columns with missing values in the train set
## There are 33 columns with missing values in the test set
```

Determine the proportion of missing values in both dataframes.

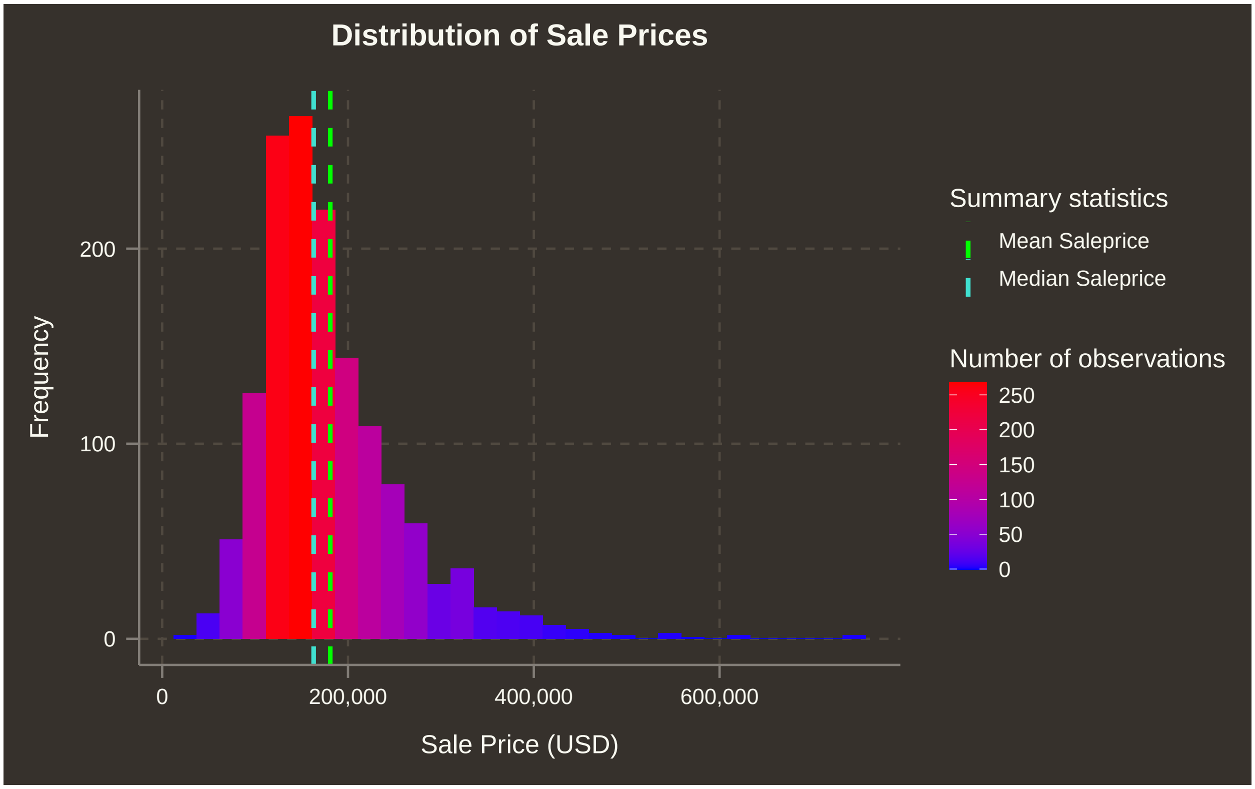

Proportion of missing values in Train set


Proportion of missing values in Test set

Distribution of target variable

Knowing the proportions of the missing values, let's go further a bit and explore the distribution of the SalePrice in train_data using ggplot to understand its skewness and identify potential outliers.

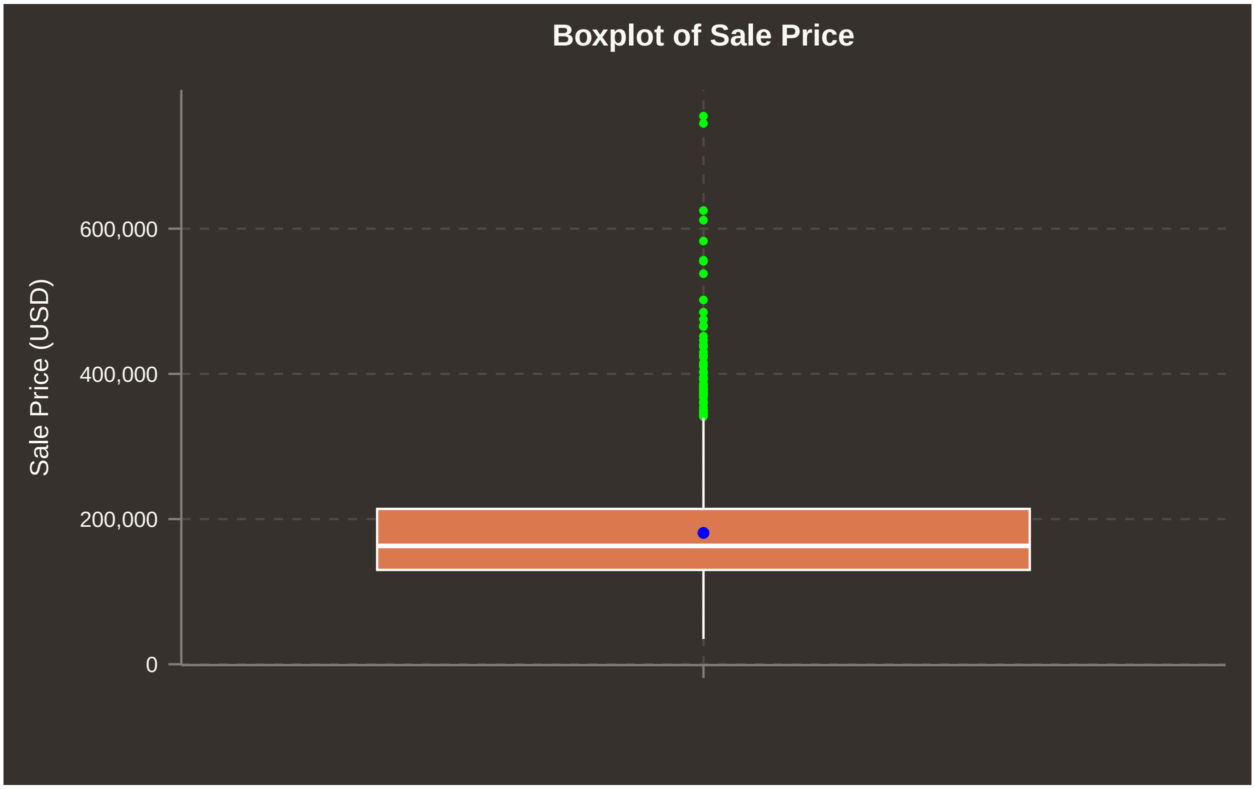
Histogram



A right skewed histogram makes sense in this context as more expensive sales are generally expected to be less frequent. Good to note the mean and median saleprices are in the range of $150k to $200k, which may indicate homes in Ames are on average affordable.

Some extreme outliers observed with sales price above $600,000. Further investigation will be done to determine which variables strongly influence a high sale price. This might also broadly indicate homes in commercial aread or homes built on high land value.

Boxplot



Boxplot plotted to place emphasis on the histogram's skew. A better representation of outliers is also observed, noting that they are significantly more than what the histogram potrays.

Further processing will be done to reduce the number of extreme outliers and bring the target variable closer to a normal distribution.

Exploring Numeric and Categorical variables

In this sub-section, we explore the numeric and categorical variables so we can identify the distributions and frequencies of the variables in train_data. This will also help us develop further insight into the "NA" values we identified earlier.

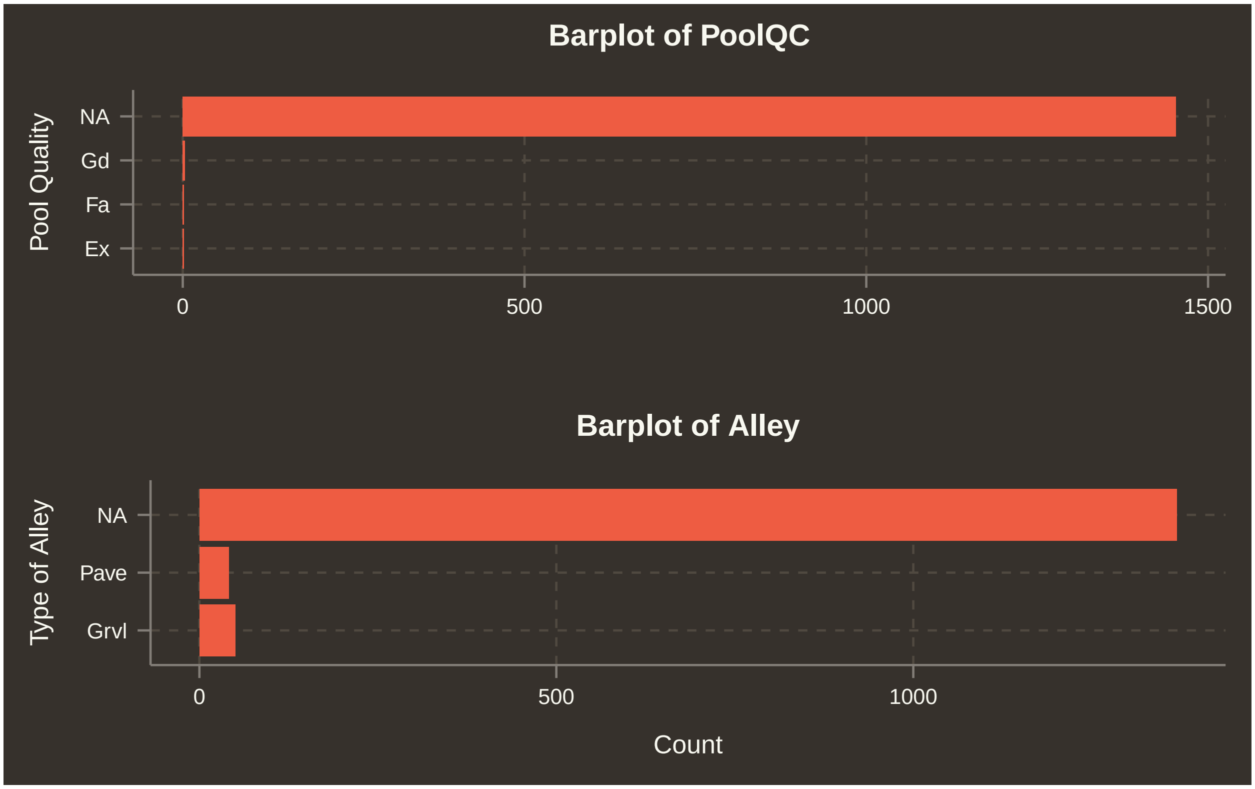
Please note, the plots are too many to be shown in this report so please refer to the code in the markdown file.

From the plots, we identified that most NAs are not really missing values, but in most cases it means a "None". By example, we have a lot of NAs in "PoolQC" and Alley, but investigating the metadata further, we found out that you cannot have "PoolQC" if you do not have a pool and cannot have an "Alley" if there is no alley access.

The same reasoning goes for some other features too. The final decision to handle these is not delete the variables but rather replace the "NA" values with "None" for better interpretation

Bar plots of "PoolQc" and "Alley" with "NA" values are shown below to demonstrate my reasoning.

Barplots of PoolQC and Alley



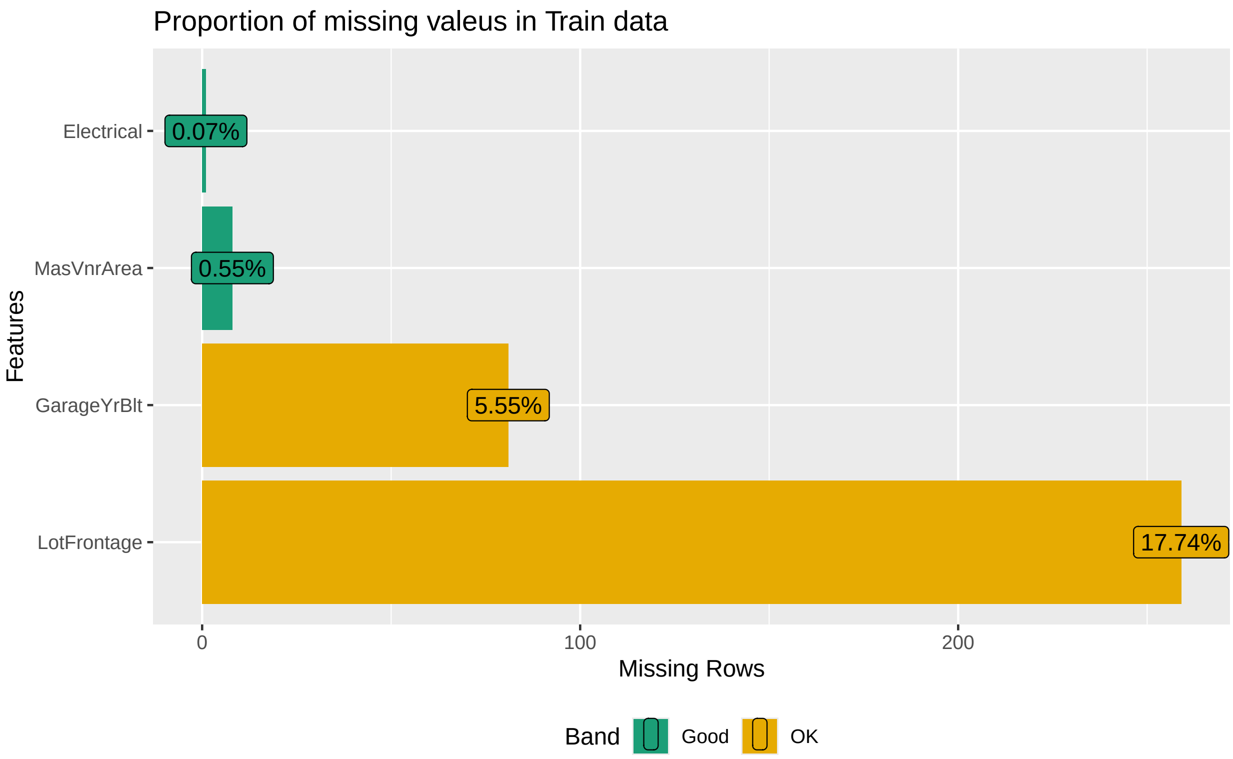
From the metadata, we understand that "NA" in PoolQC means "No Pool" and "NA" in Alley means "No alley access". Therefore these are not missing values in the sense that they do not exist in the dataset.

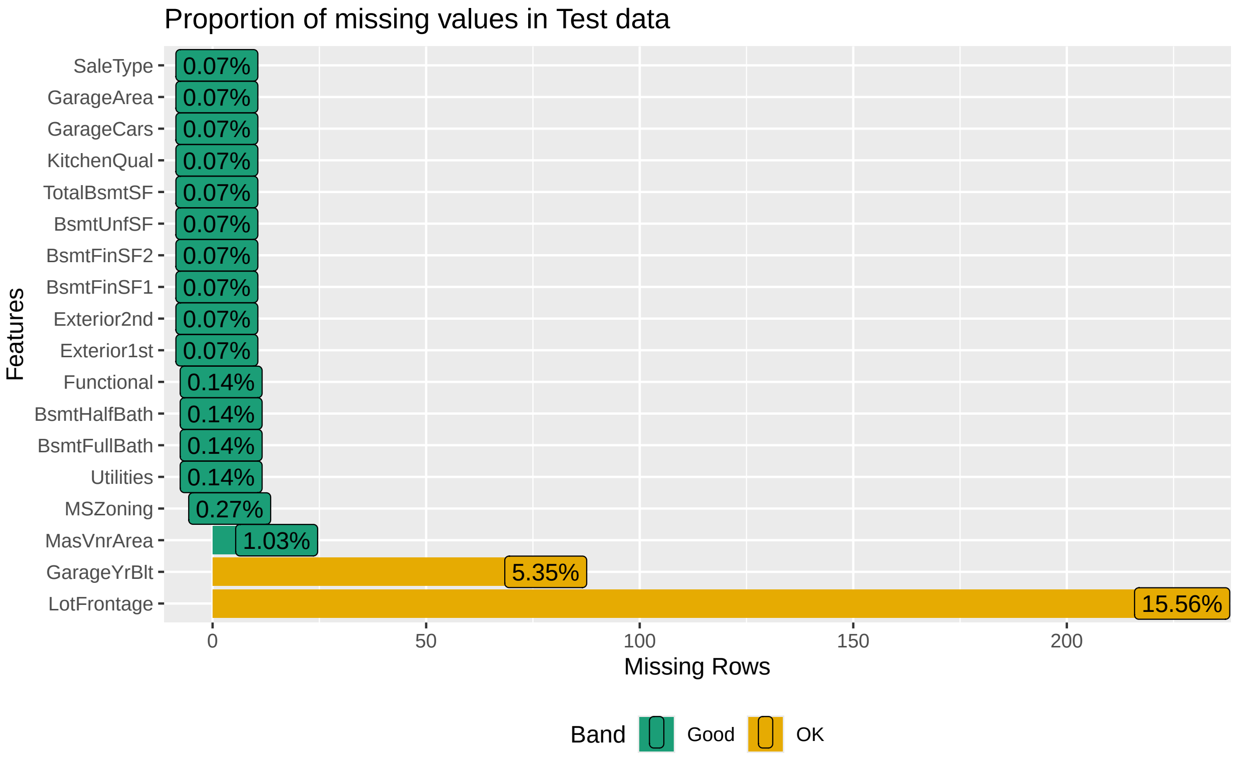
Also interesting to note is that a number of homes do not own pools. Could it be because pool maintenance is expensive in Ames? Recall how the mean and median sale price of homes were in the $150k - $200k range, perhaps pools are not common in homes with these price tags in America? On the assumption that these price tags represent the lower middle class.

Replacing NA values with "None"

NA values in categorical vaues where appropriate replaced with "None". Please refer to markdown file to view code.

Proportion of missing values after replacement.





```
## Columns with missing values in train set after NA replacement

## LotFrontage GarageYrBlt  MasVnrArea  Electrical
##         259          81           8           1
```

```
##
## Columns with missing values in test set after NA replacement

##  LotFrontage  GarageYrBlt   MasVnrArea     MSZoning    Utilities BsmtFullBath
##          227           78           15            4            2            2
## BsmtHalfBath   Functional  Exterior1st  Exterior2nd   BsmtFinSF1   BsmtFinSF2
##            2            2            1            1            1            1
##    BsmtUnfSF  TotalBsmtSF  KitchenQual   GarageCars   GarageArea     SaleType
##            1            1            1            1            1            1
```

Based on these findings, the number of missing values in “LotFrontage”, “GarageYrBlt” and “MasVnrArea” across both sets are still quite high.

More investigation will be done to determine how to handle them by checking the structure and summary of the dataset.

Data Formatting and Categorization

Further investigation into the structure and summary of train_data points out that multiple categorical values can be transformed into factors or ordered factors to ensure categorical data are correctly represented in the dataset. The variables fall into one this two categories:

Nominal Variables (No Order): Variables like MSZoning, Street, and Neighborhood represent categories without any inherent ordering. We will convert these into factors to ensure that modeling algorithms treat them as distinct categories rather than numbers with magnitudes.

Ordinal Variables (Order Matters): Variables like ExterQual, BsmtQual, and HeatingQC possess a meaningful order in their categories. We will convert these into ordered factors to preserve this order.

Factors

Conversion of nominal variables. Please refer to markdown to view code.

Ordered Factors

Conversion of ordinal variables. Please refer to markdown to view code.

MSSubClass

This variable identifies the type of dwelling involved in the sale. The classes are encoded as numbers but are inherently categorical, as seen below:

- 20 1-STORY 1946 & NEWER ALL STYLES
- 30 1-STORY 1945 & OLDER
- 40 1-STORY W/FINISHED ATTIC ALL AGES
- 45 1-1/2 STORY - UNFINISHED ALL AGES
- 50 1-1/2 STORY FINISHED ALL AGES
- 60 2-STORY 1946 & NEWER
- 70 2-STORY 1945 & OLDER
- 75 2-1/2 STORY ALL AGES
- 80 SPLIT OR MULTI-LEVEL
- 85 SPLIT FOYER

- 90 DUPLEX - ALL STYLES AND AGES
- 120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER
- 150 1-1/2 STORY PUD - ALL AGES (Not in the dataset we are working with)
- 160 2-STORY PUD - 1946 & NEWER
- 180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
- 190 2 FAMILY CONVERSION - ALL STYLES AND AGES

Therefore, we will re-value the codes to their labels to improve readability. Please refer to markdown to view code.

LotFrontage

Since "LotFrontage" might vary by neighborhood, a common strategy is to impute missing values based on the median LotFrontage of the neighborhood. Please refer to markdown to view code.

MasVnrArea

Since "MasVnrArea" is numeric and has a few missing values, these could be imputed by a central tendency measure. Please refer to markdown to view code.

Electrical

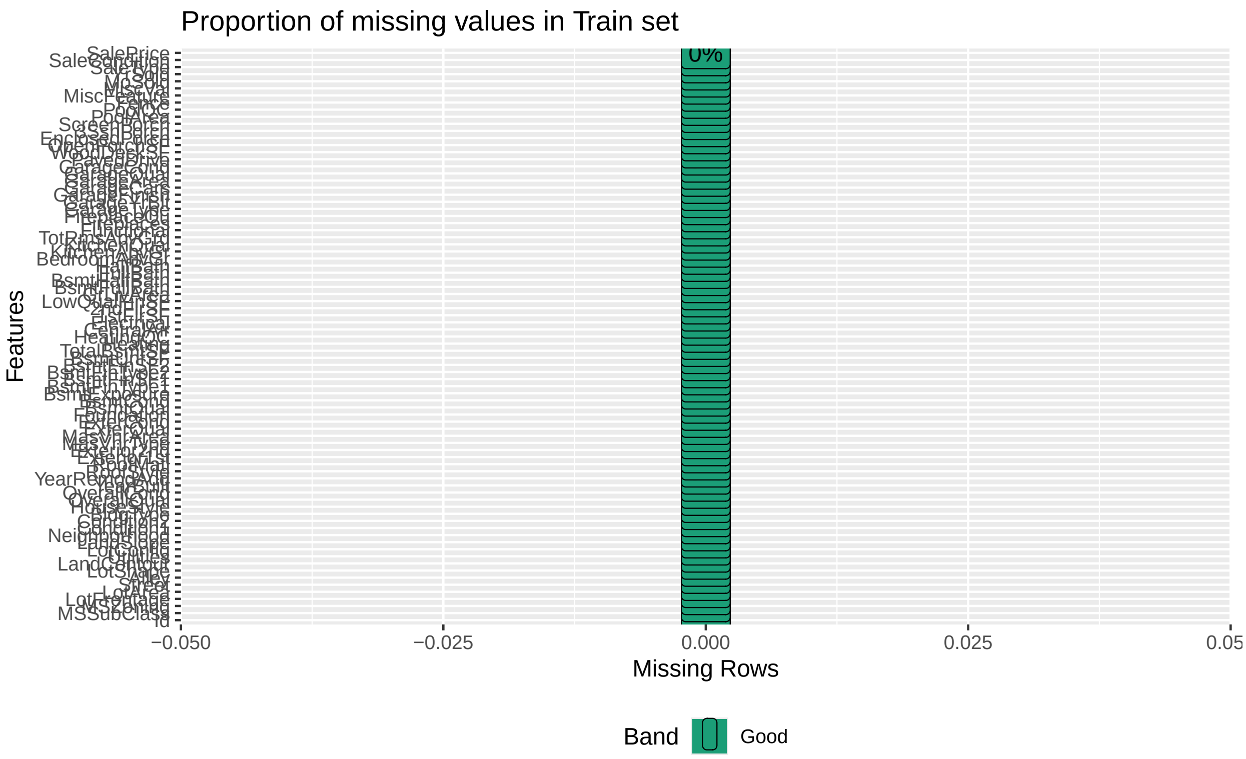
Since there is only one missing value in Electrical, and this is a categorical variable, the simplest approach would be to impute this with the mode. Please refer to markdown to view code.

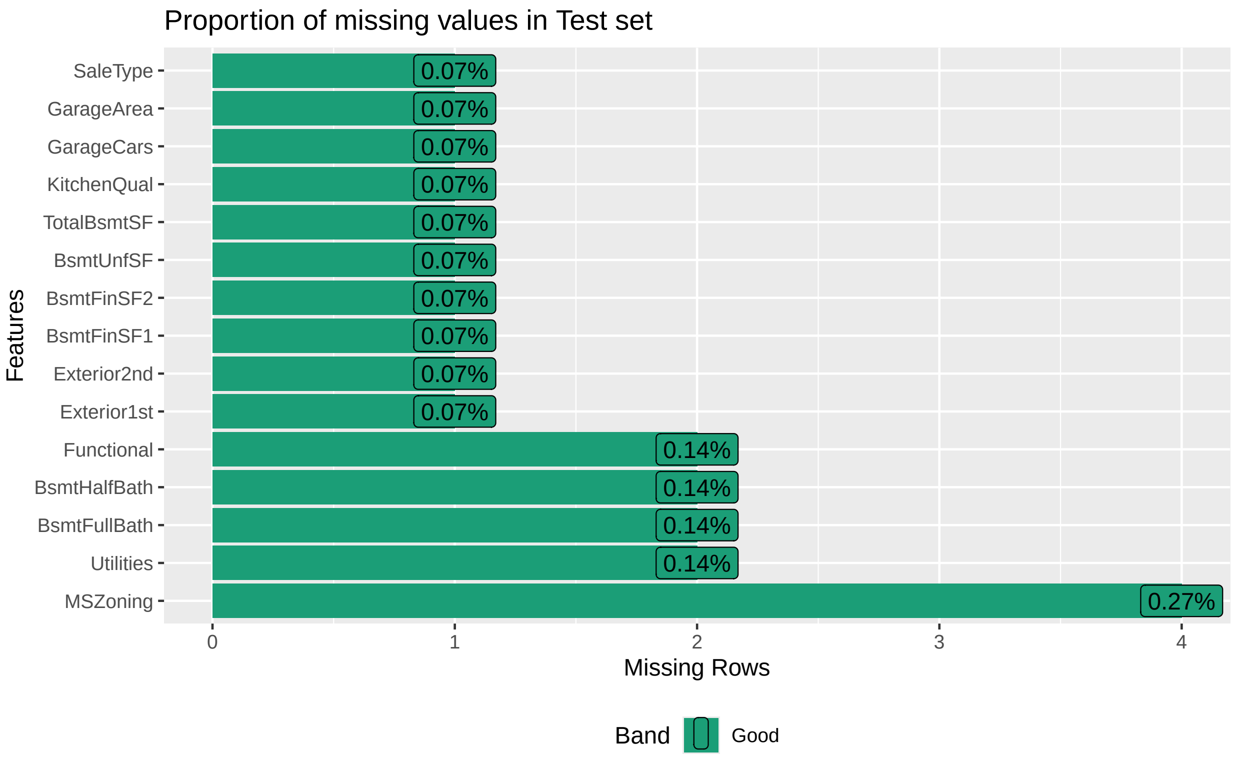
Impute GarageYrBlt with YearBuilt

For homes where GarageYrBlt is missing, we'll use the year the house was built (YearBuilt). This assumes that the garage was constructed concurrently with the house, since the datasets were inspected and we notice that, frequently, the YearBuilt is similar to the GarageYrBlt.

Please refer to markdown to view code.

Verifying imputations





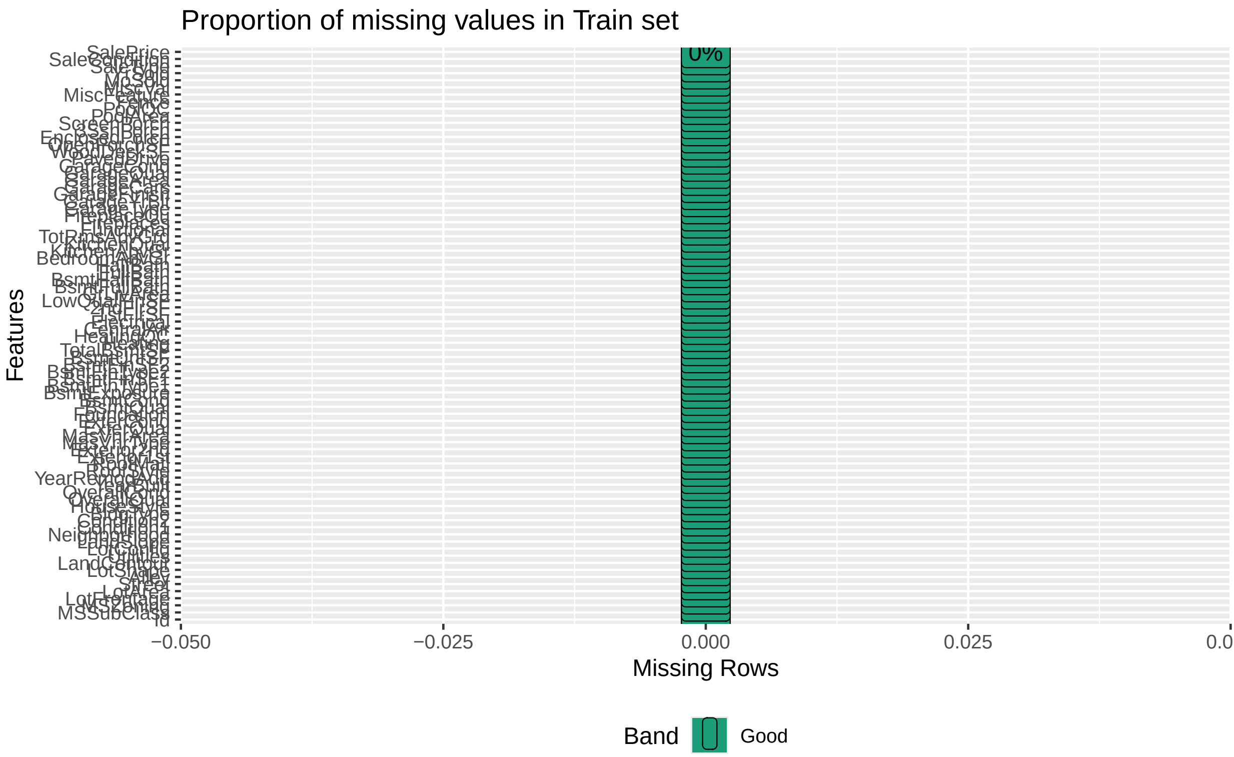
We observe there are still some missing values in the test set. Manual imputation will be used to handle them.

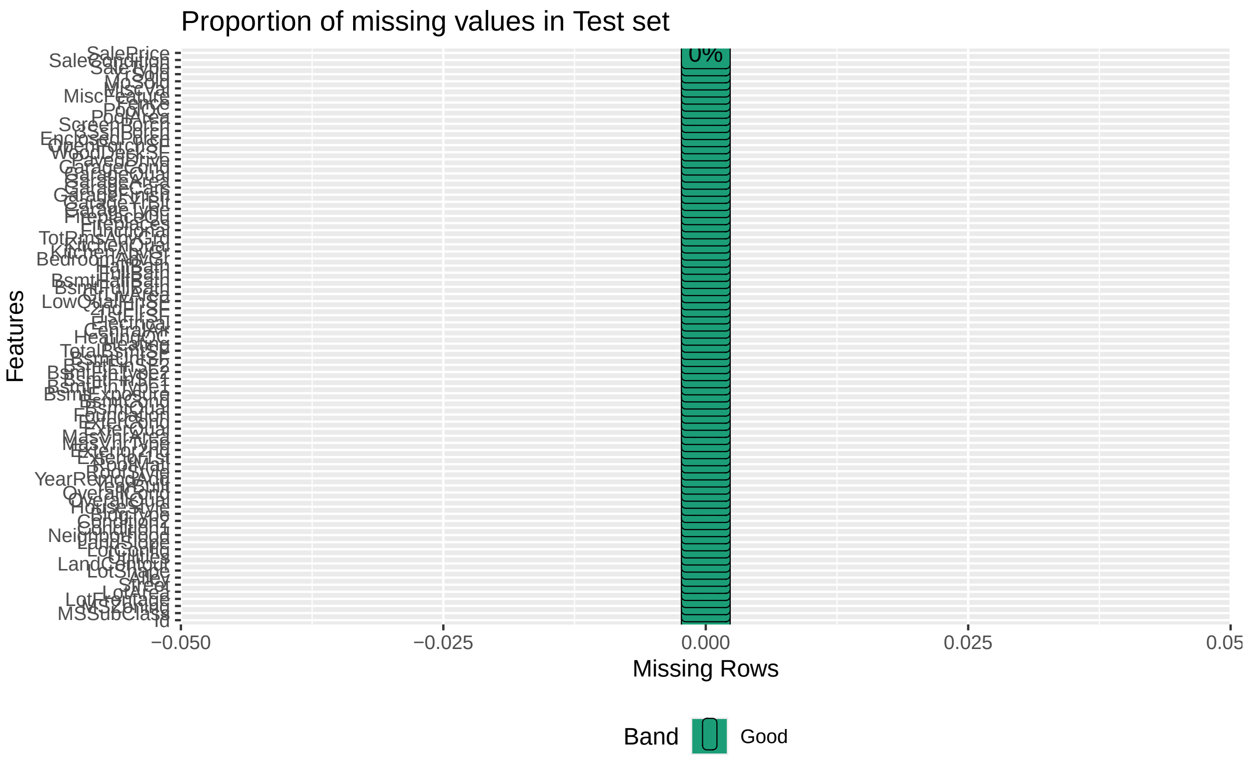
- MSZoning - Imputation with the mode of this categorical variable.
- Utilities - As this is typically ‘AllPub’, imputation with the mode.
- Exterior1st and Exterior2nd - Imputation with the mode of these categorical variables.
- BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF - Missing, likely no basement; set to 0.
- BsmtFullBath and BsmtHalfBath- Missing, likely no basement; set to 0.
- KitchenQual - Imputation with the mode as it’s a categorical variable.
- Functional - Imputation with ‘Typ’ as it’s the most common category indicating typical functionality.
- GarageCars and GarageArea - Missing, likely no garage; set to 0.
- SaleType - Imputation with the mode.

Please refer to markdown to view code.

```
## Number of NAs left in train set: 0
```

```
## Number of NAs left in test set: 0
```





Up to this point, all NA values have been handled, variables have been factorised, categorised and re-encoded and the datasets are ready for exploratory analysis and modelling

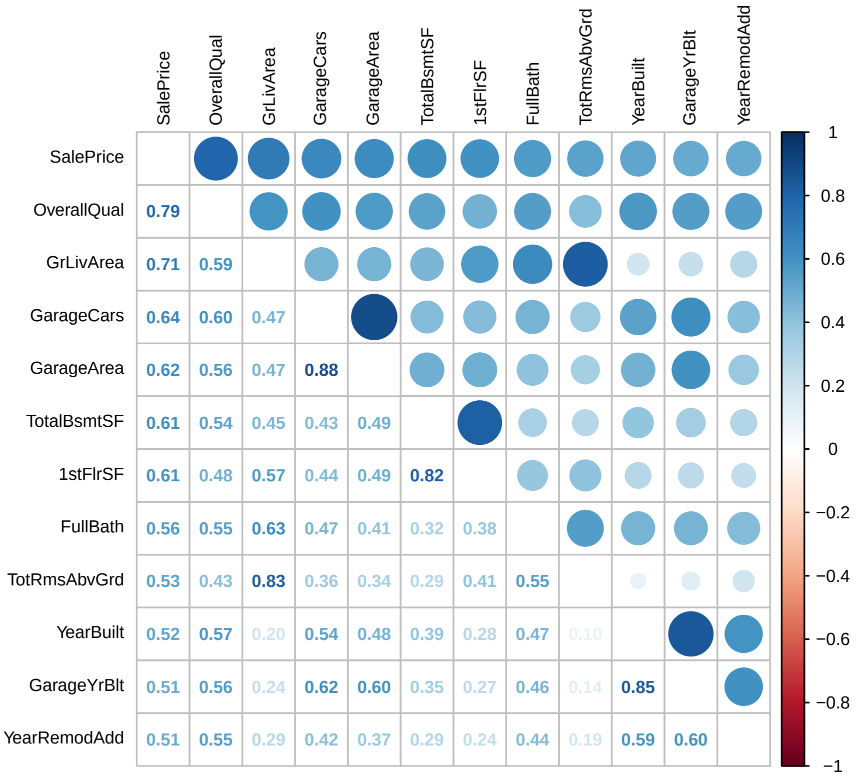
4. Exploratory Data Analysis

In this section, I explore relationships between variables of interest and their relationship with SalePrice. Some questions explored include;

- Neighbourhood - which neighbourhoods are more expensive than the rest ?,
- Yrsold - what was the trend of median price over the years ?
- MoSold - was there a seasonal effect on the number of sales ?
- Overall Quality - does home quality affect its sale price ?

Further exploration will be conducted after new variables have been feature engineered in the next section.

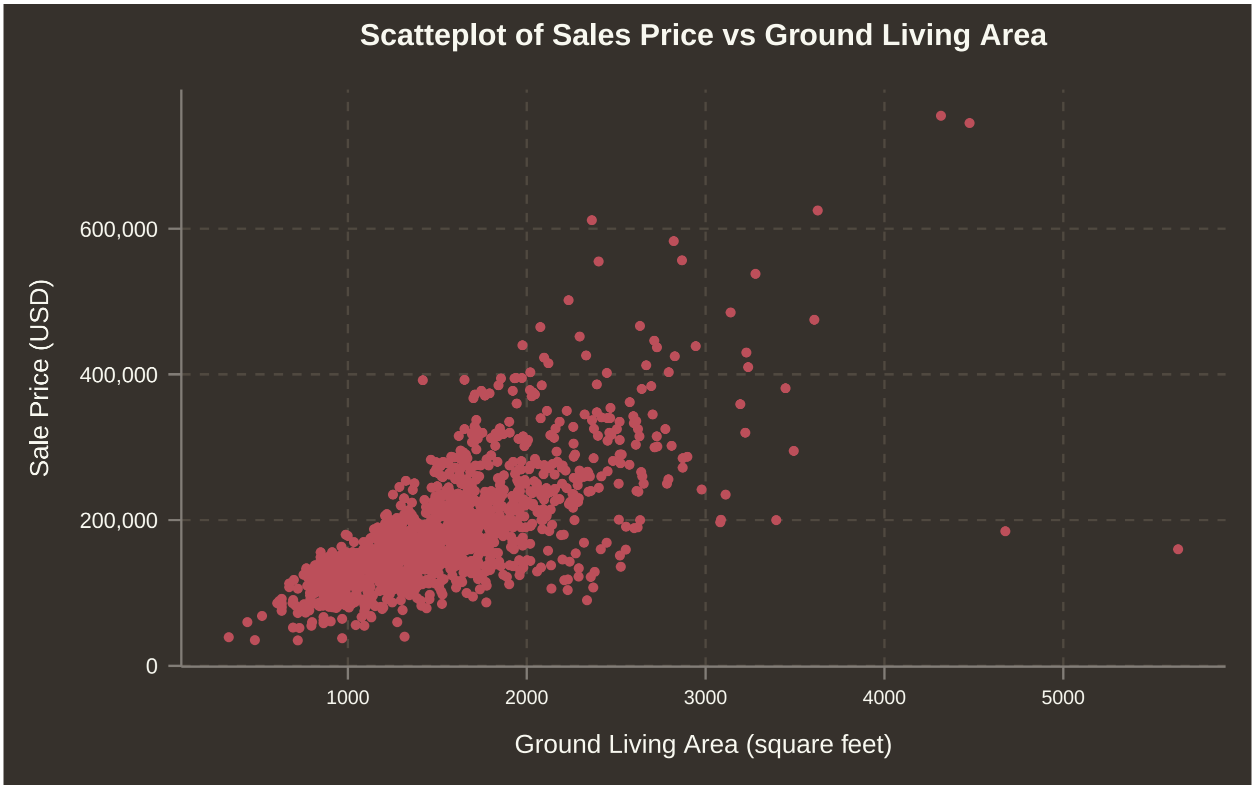
Pre-Feature Engineering Correlations



Correlations which were above 0.5 with SalePrice was plotted develop a bit of insight into some of the questions listed above. Overall Quality has the highest correlation with SalePrice hence we expect that more expensive homes should tend to have a higher quality. We will investigate this trend.

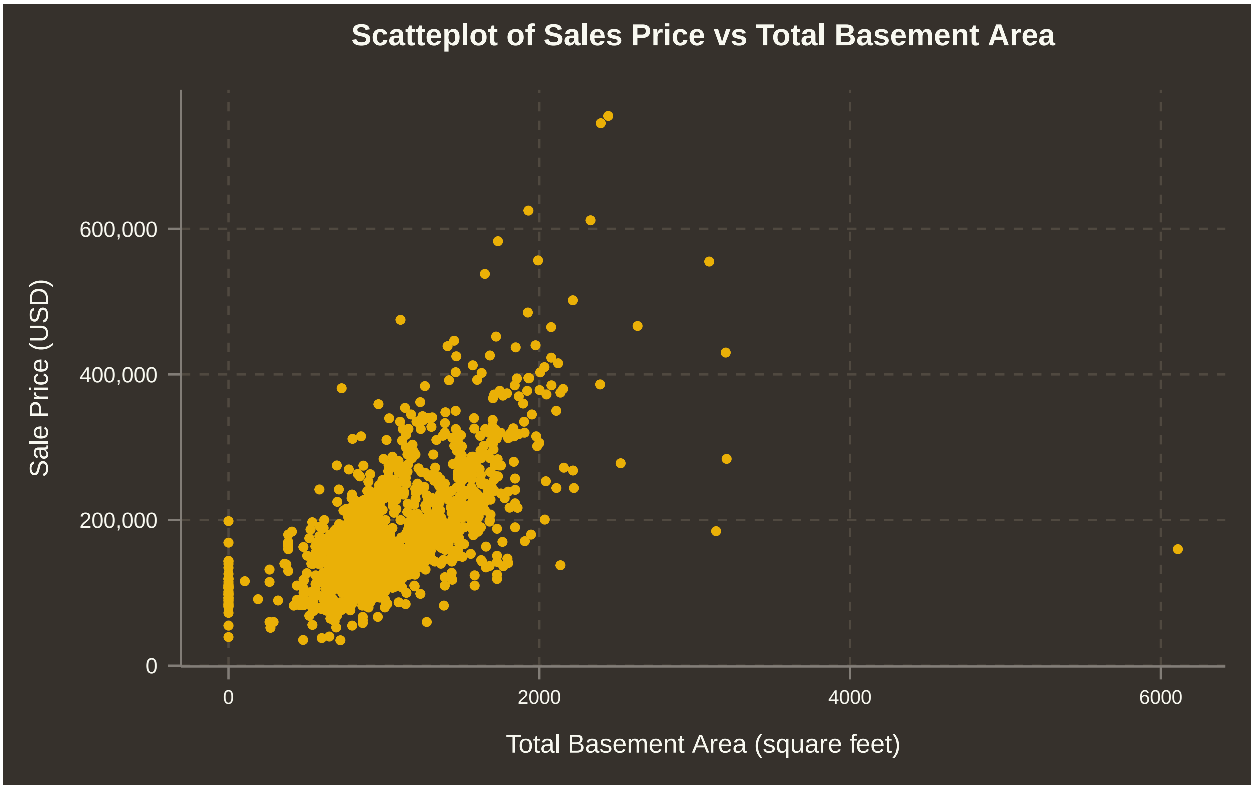
Please note, this correlation was done before feauture engineering therefore rankings of variables might change. Multicollinearity will also be handled later in this report.

Scatterplot between SalePrice and Ground Living Area



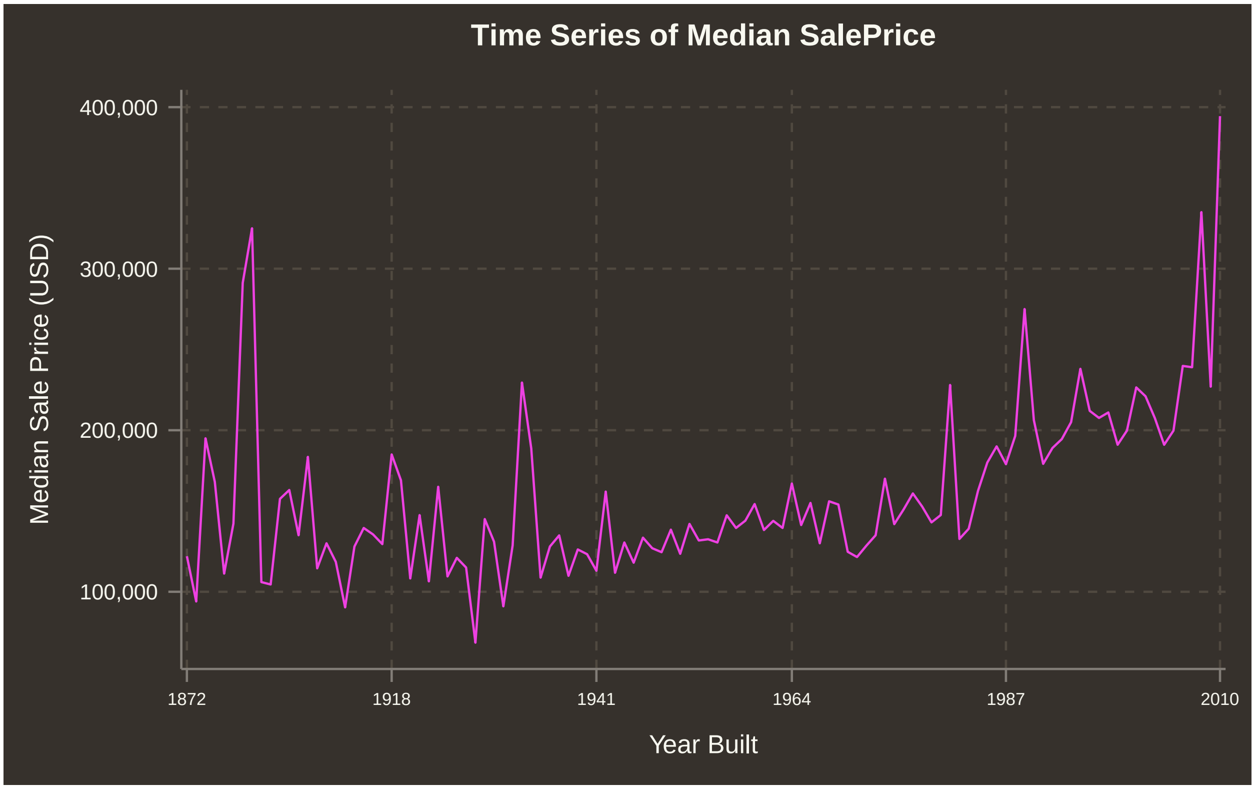
Positive strong correlation between SalePrice and Ground Living Area. Indicating that bigger homes tend to have a higher fee. Some outliers spotted where large homes have relatively low prices. This might be due to the age of the house, or other factors at play.

Scatterplot between SalePrice and Total Basement Area



Positive strong correlation between SalePrice and Total Basement Area. Indicating that homes with large basements tend to have a higher fee. Some outliers spotted where large basements have relatively low prices. Again, this might be due to the age of the house, basement quality or other factors at play.

Time Series of Median Sales Price

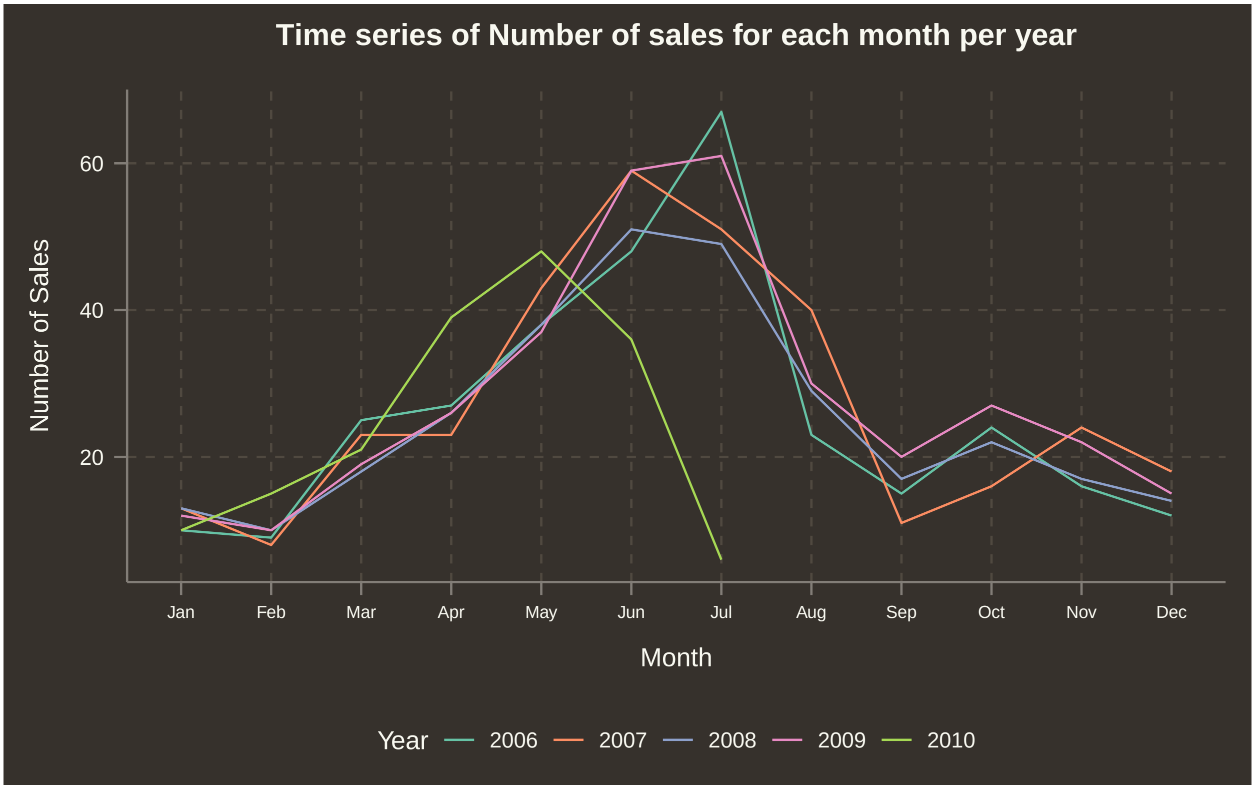


Time series shows trend of median prices gradually increasing over the years. Interesting spikes to note are the sharp increase in price shortly after 1872, a time period where there was an economic surge in North America and the 2007 recession.

However, this is immediately followed by a sharp downturn, representing The Panic of 1873 (Library of Congress, 2024). This period lasted between 1873 up until 1878. This economic downturn later became known as the Long Depression after the stock market crash of 1929.

Also note the effect of The ’07 Recession, before the spike up shortly before 2010.

Time series of number of sales for each year

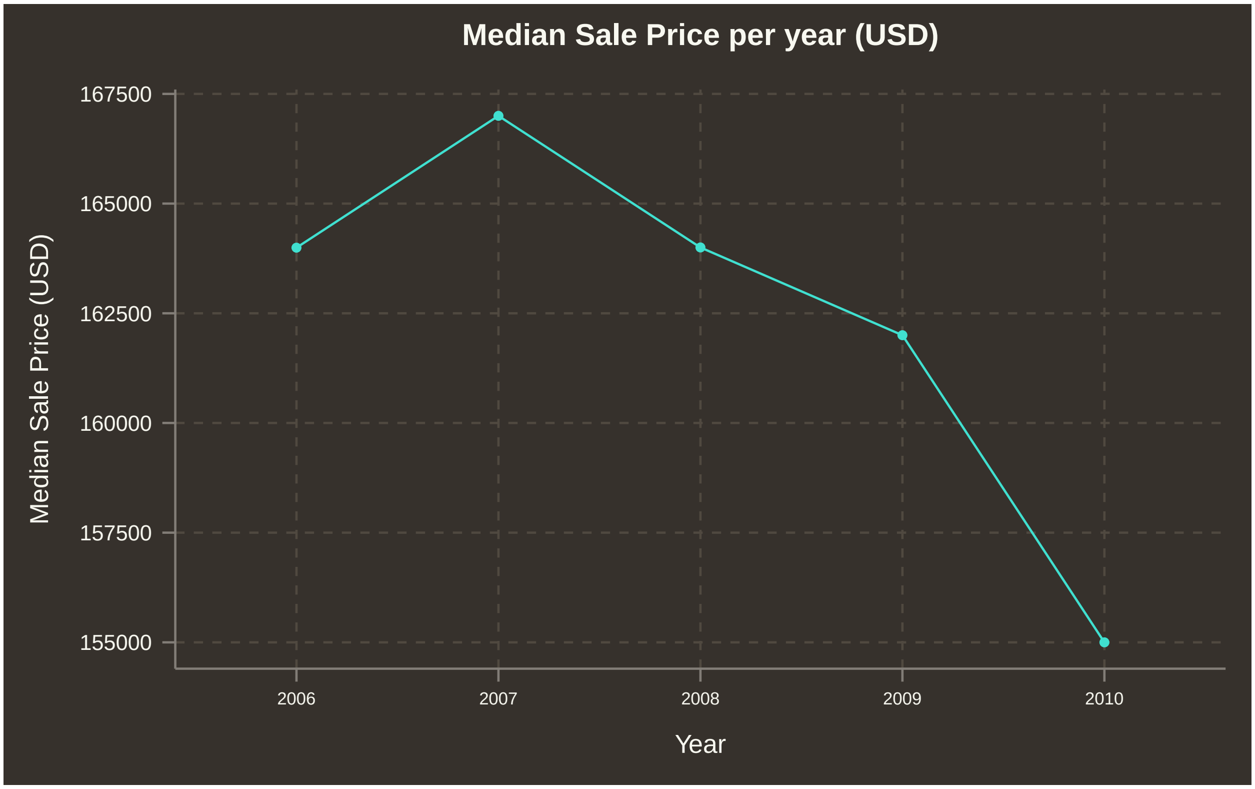


We see a trend where sales tend to increase across all years between the months of February to July, spanning across end of winter to the middle of summer. Sales begin to decline after August, most probably due to school terms resuming and a transition into the holiday period.

Seasonality is definitely at play here. Home buyers are less aggressive during the holiday months of November to January and the peak home buying season seems to occur during June, July and August.

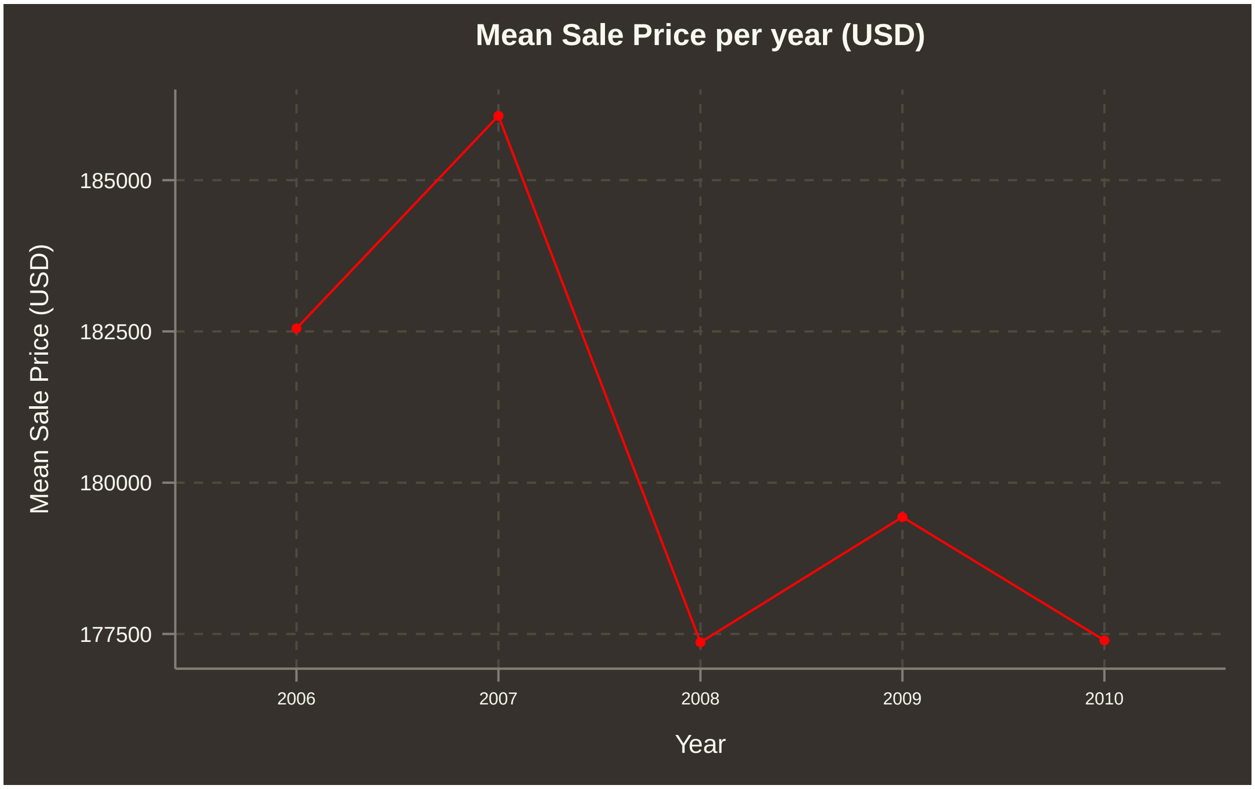
Also note, trend line for 2010 ends in July, indicating the dataset does not have observations for months after July in 2010. These observations could be in the test data or they were simply not recorded.

Trend of median sale price of each year



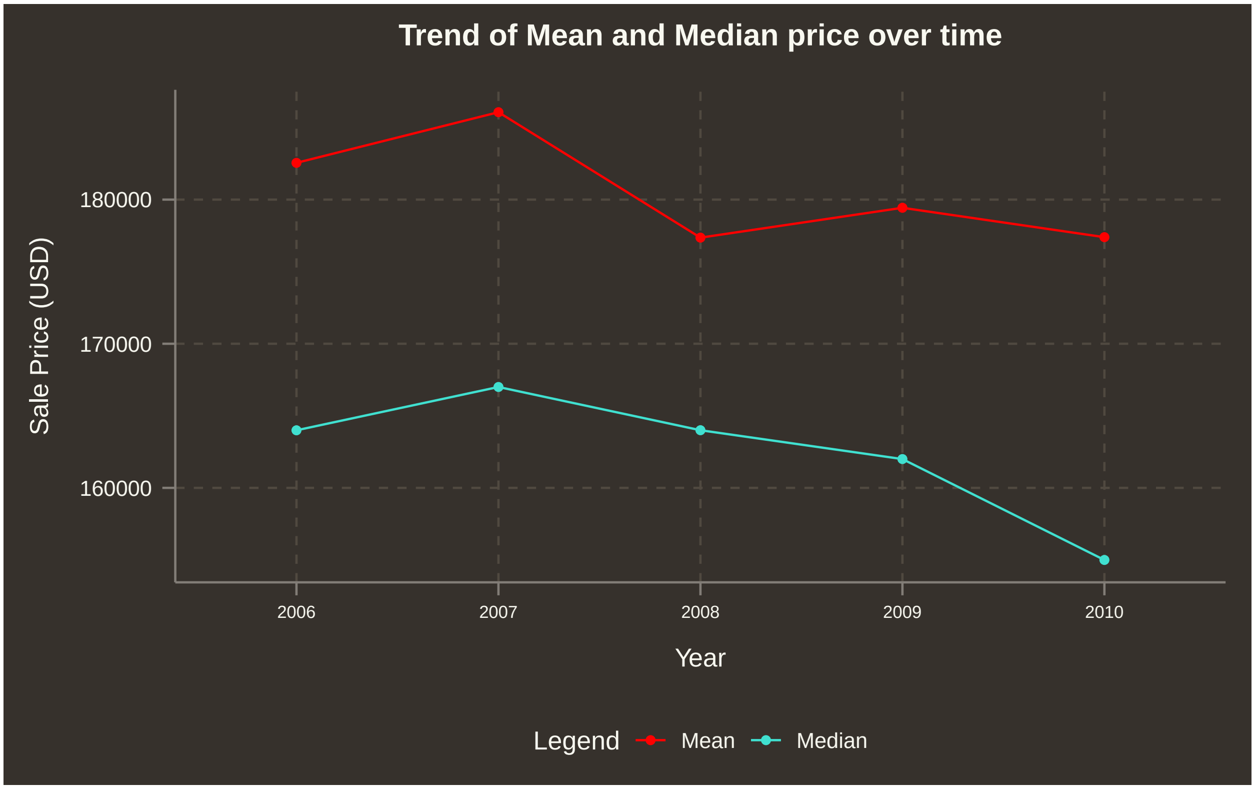
Trend shows a sharp decrease in median sale price after 2007, highlighting the strong effect of the 2007 recession. An economic downturn caused by the burst of the US housing market and the global financial crisis (Investopedia, 2023).

Trend of mean sale price of each year



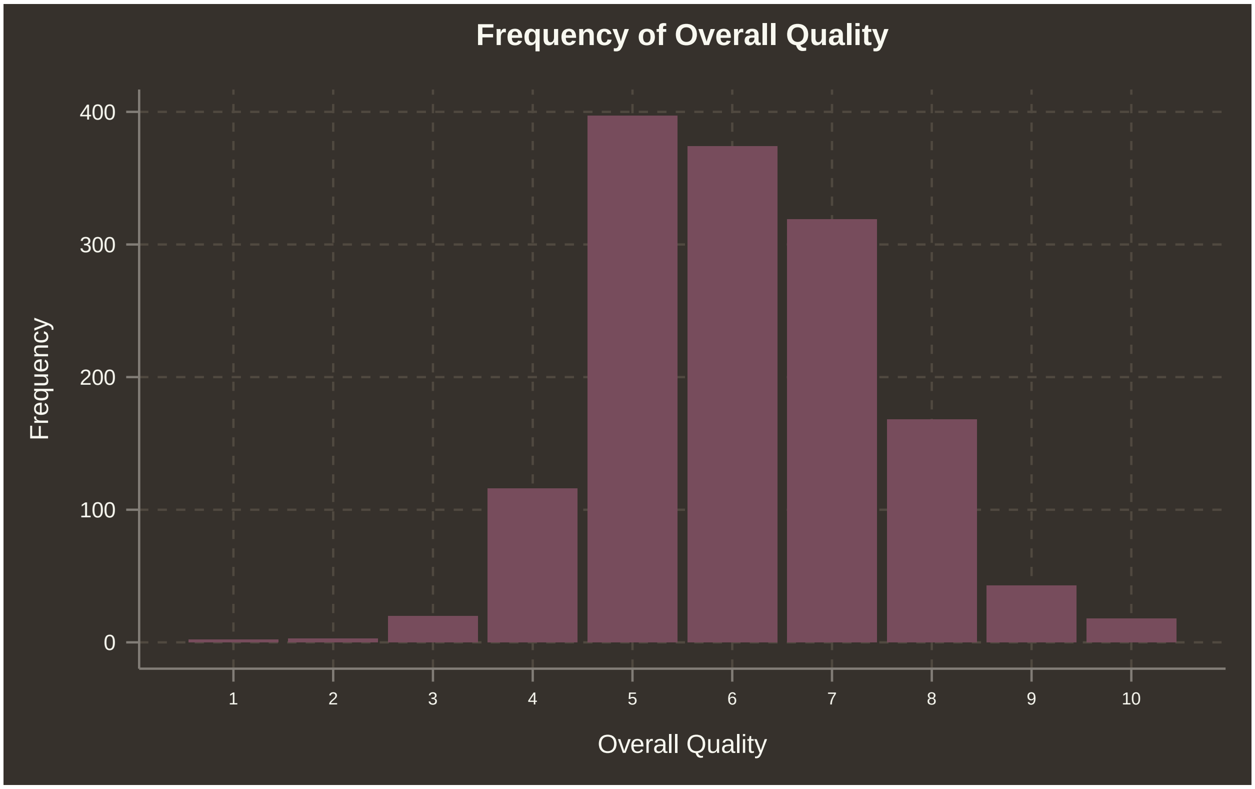
Similar decreasing trend in the average sale price after 2007.

Combined trends for a holistic view



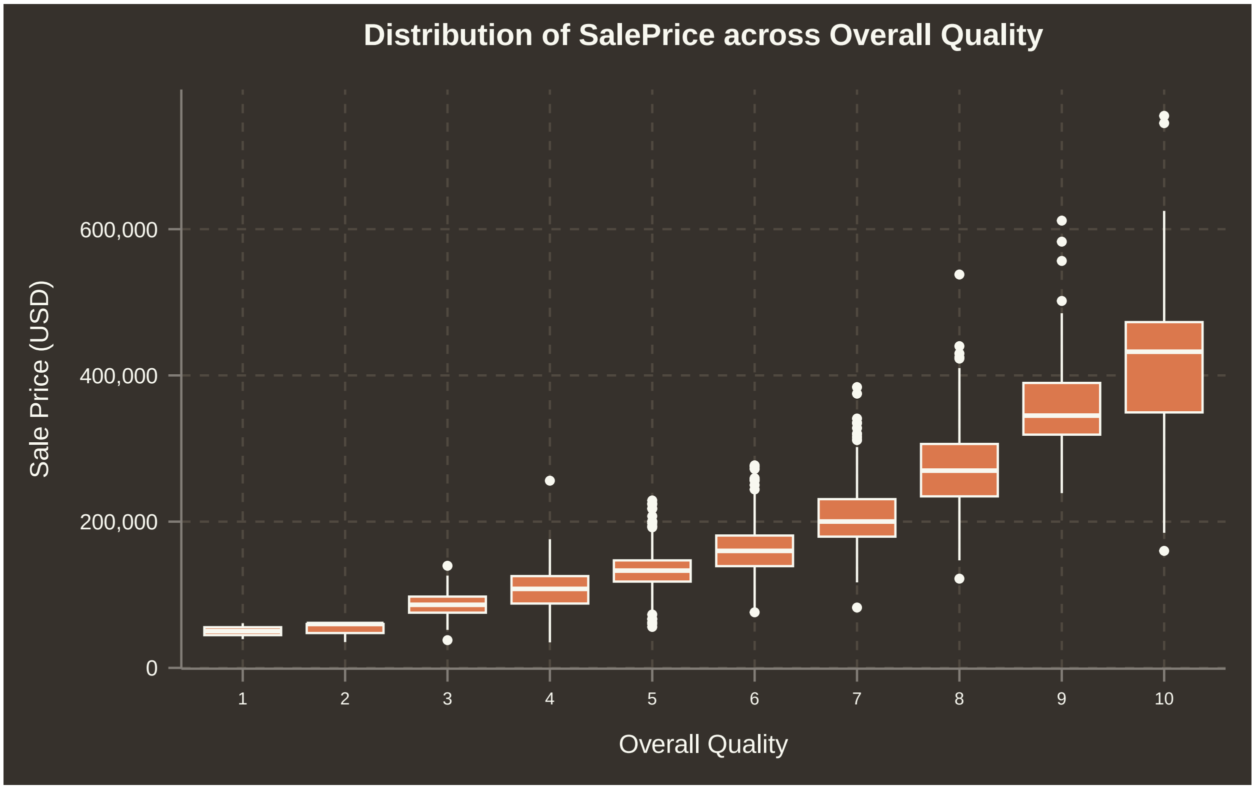
Average sale price generally higher than the median sale price. This was also highlighted in the histogram plot earlier in this report.

Frequency of Overall Quality



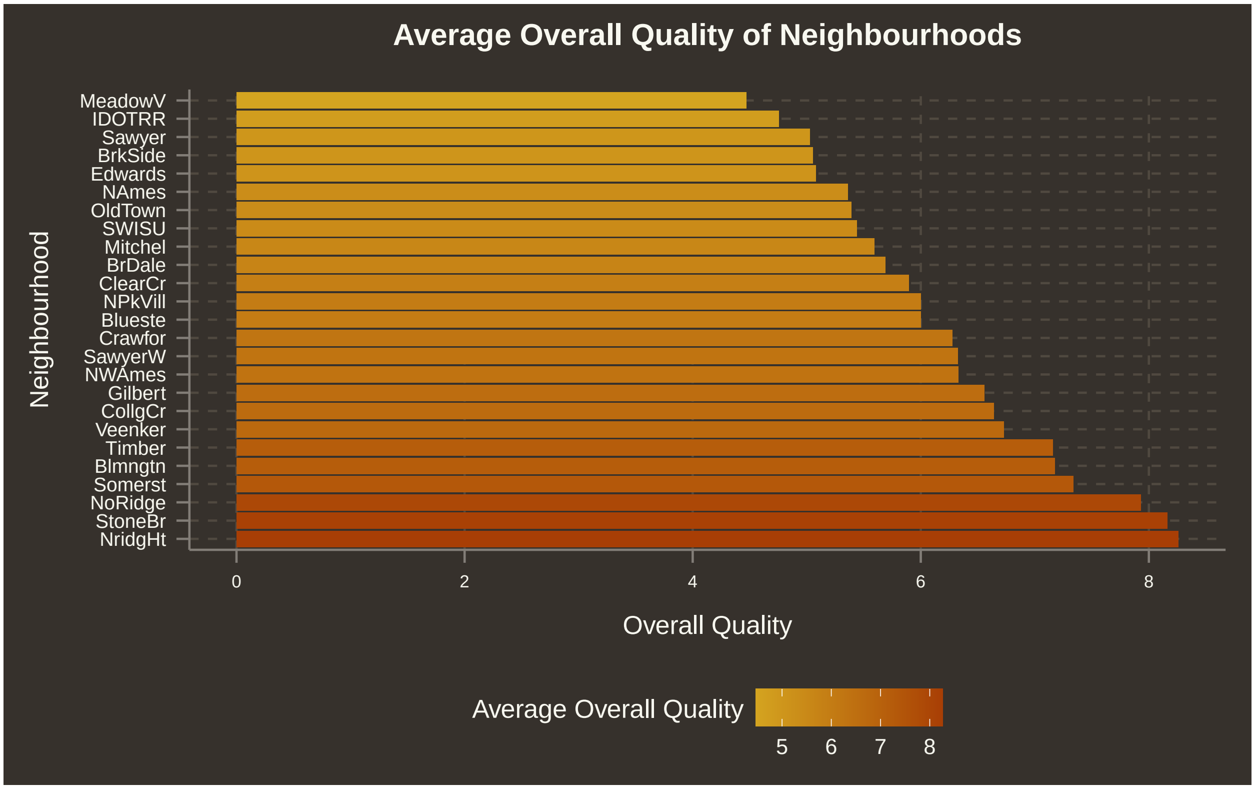
We observe that more houses have an Overall Quality between 5 and 7, which might correlate to the median sale price being affordable. Fewer houses have a quality between 8 and 10 and we would expect these homes to have a higher median SalePrice than most.

Distribution of SalesPrice across OverallQuality



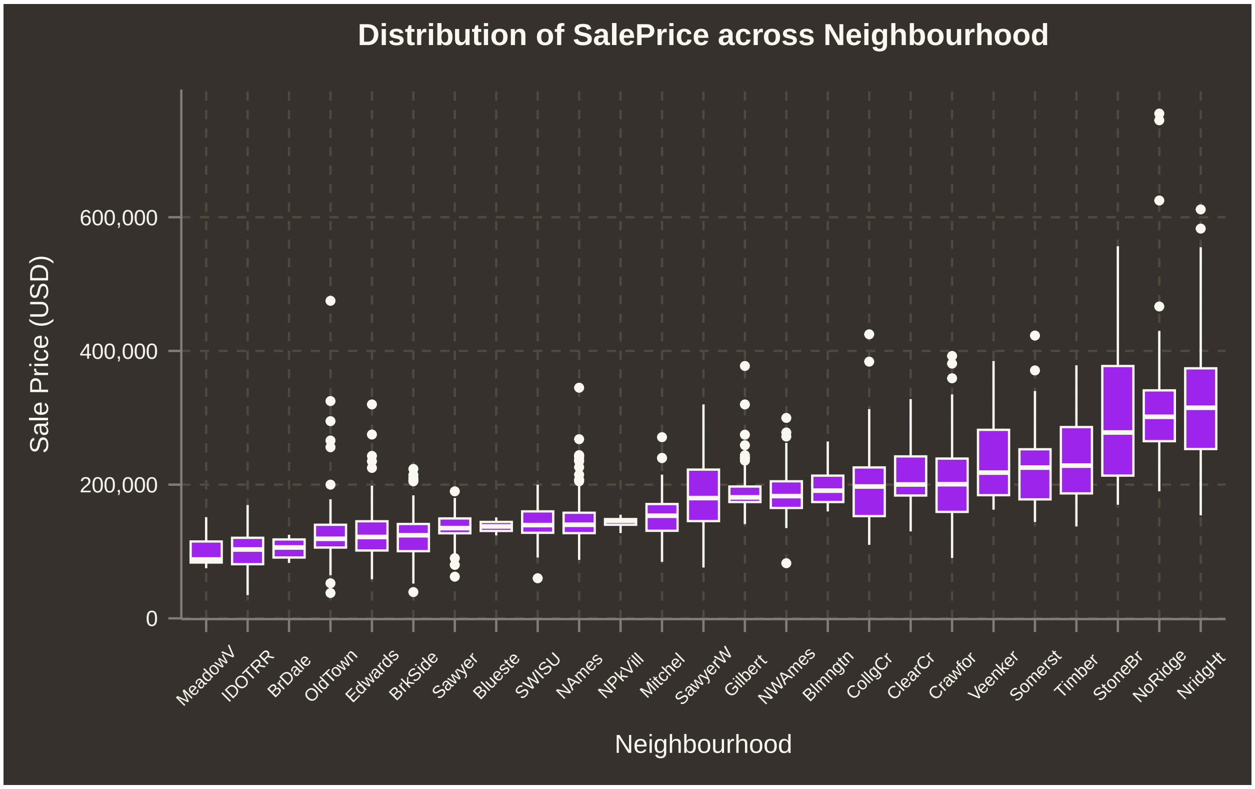
Distribution confirms that homes with higher Overall Quality have a higher Sale Price, satisfying the expectations we established earlier.

Average Overall Quality of neighbourhoods



Distribution of average quality across neighbourhoods gives us an idea of which neighbourhoods we expect to have a higher median sale price than most. From this plot, the top three neighbourhoods are NridgHt, StoneBr, and NoRidge.

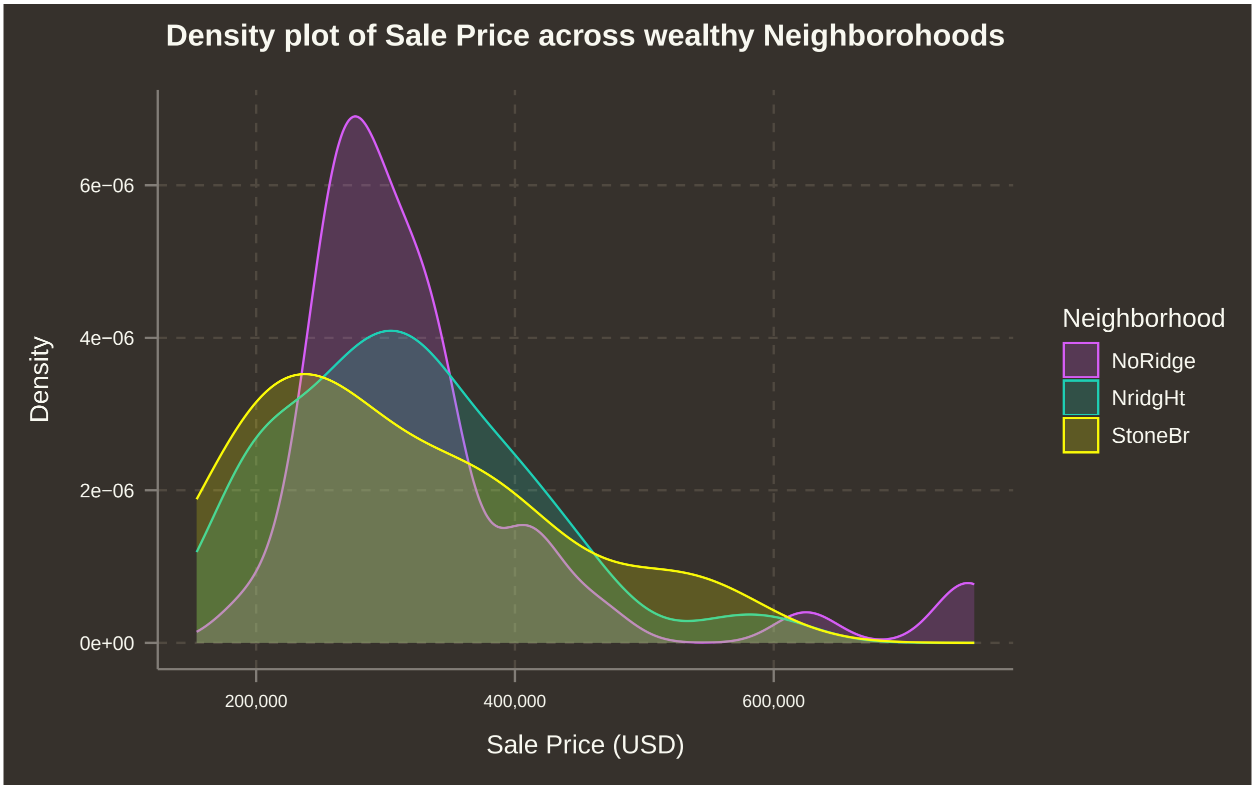
Distribution of SalePrice across Neighbourhood



Distribution of sale price across neighbourhood supports our assumptions of neighbourhoods with high overall quality will have a higher median price than most. NridgHt, StoneBr, and NoRidge are again the top three neighbourhoods in this plot, a direct correlation with the plot above.

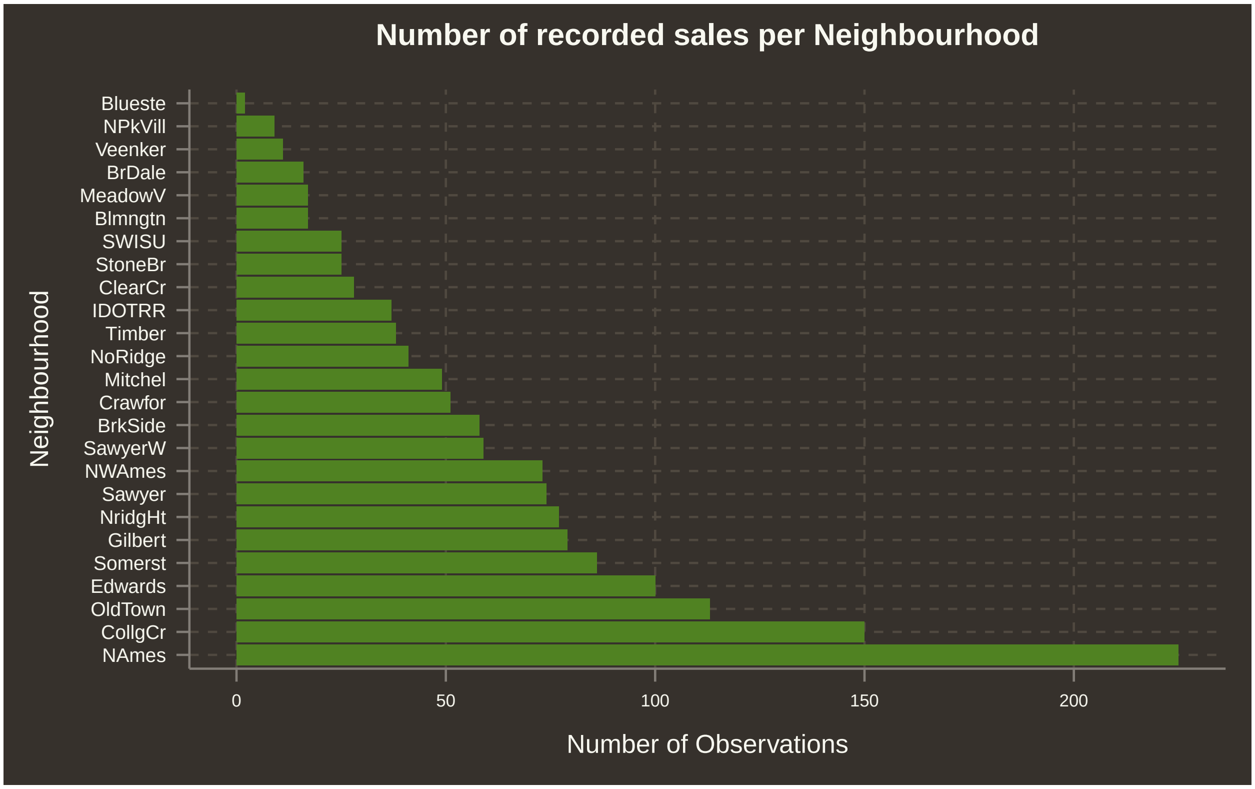
Distribution of SalesPrice in wealthy neighbourhoods

A neighbourhood will be considered wealthy if the median saleprice is $100,000 above the median SalePrice, ie, greater than $263,000



Again, we see the neighbourhoods selected for this plot are NridgHt, StoneBr, and NoRidge. Most of their prices are in the range $200,000 to $400,000 with the NoRidge neighbourhood in particular being the suburb with extreme house prices. Maybe this is the newest neighbourhood in Ames? or the neighbourhood with the highest commercial/land value?

Number of recorded sales per neighborhood



Most sales were recorded in NAmes, a neighbourhood with a median sale price of $140,000. Expectedly, the neighbourhoods with the highest median sale prices; NridgHt, StoneBr, and NoRidge, recorded fewer sales.

From these EDAs, we expect Neighbourhood and Overall Quality to be significant variables in predicting SalePrice.

5. Further preprocessing

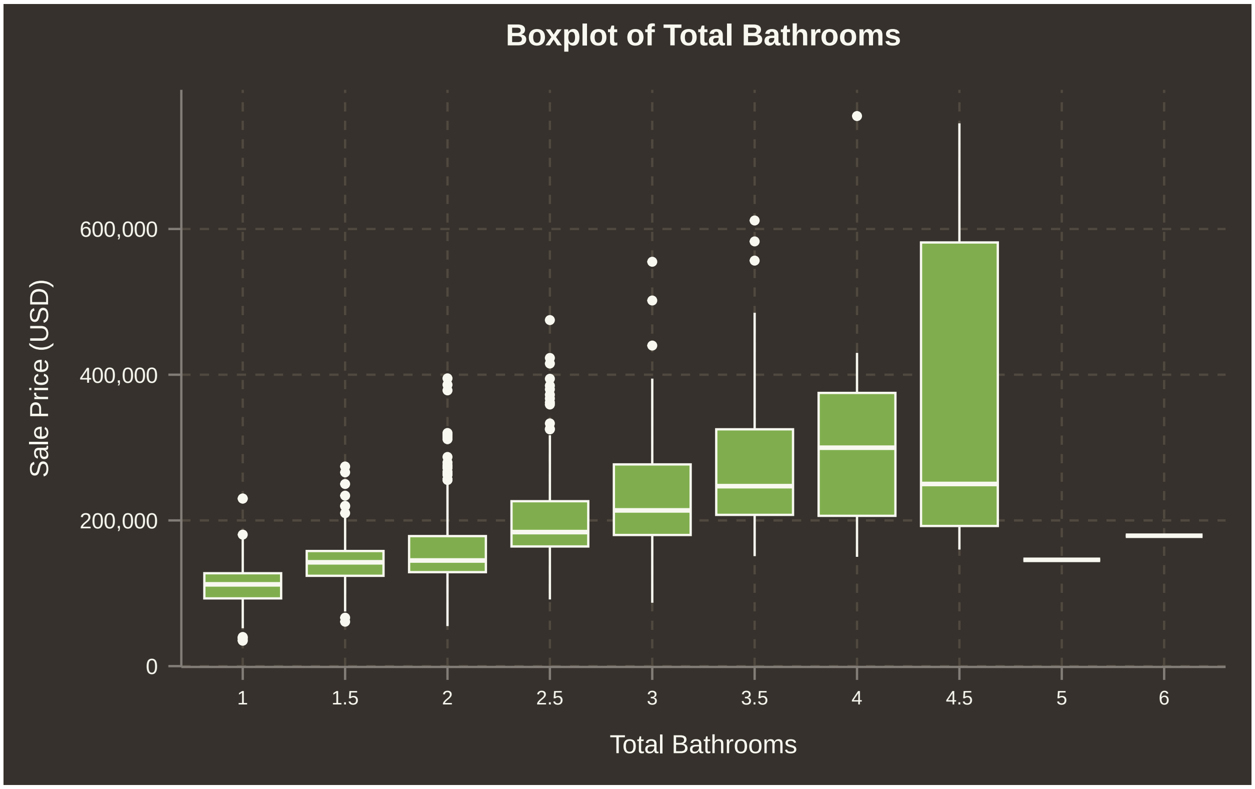
In this section, I perform further analysis into variables, aggregate variables into new ones which I believe will have a strong predictive relationship with SalePrice, handle multicollinearity where necessary, and select significant variables for predictive modelling based off correlation plots and variable importance plots from random forests.

Feature Engineering

Creating Total Bathrooms

The TotalBathrooms feature aggregates all bathroom data into a single feature by summing up the counts of full and half bathrooms in both the basement and above-grade (non-basement) areas of the house. Full baths count as one, while half baths count as 0.5, acknowledging that half baths have less utility than full baths.

Boxplot of Total Bathrooms



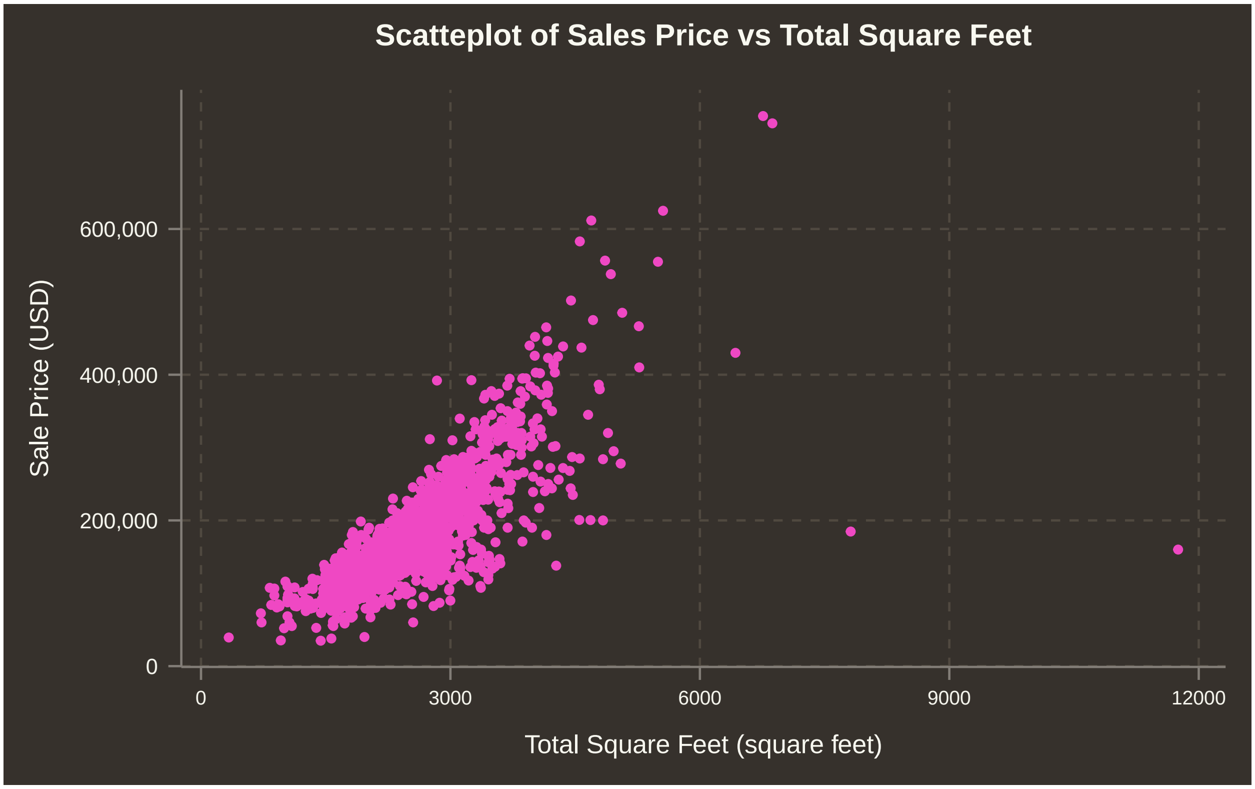
The distribution follows the assumption that homes with more bathrooms will fetch a higher price. The solid dashed lines at Total Bathrooms 5 and 6 is due to the fact that there is only one observation for each of the levels, so their median sale price will simply be their sale price, as shown in the table below:

Id	SalePrice	TotalBathrooms
739	179000	6
922	145900	5

Creating Total Square Feet

The TotalSquareFeet feature combines the living area above ground (GrLivArea) and the total basement area (TotalBsmtSF) to provide a better measure of the total usable area of the house. This aggregation might provide a more impactful predictor than considering these areas separately because combined space often translates more directly to consumer perceptions of size and value.

Scatterplot of Sale Price vs Total Square Feet

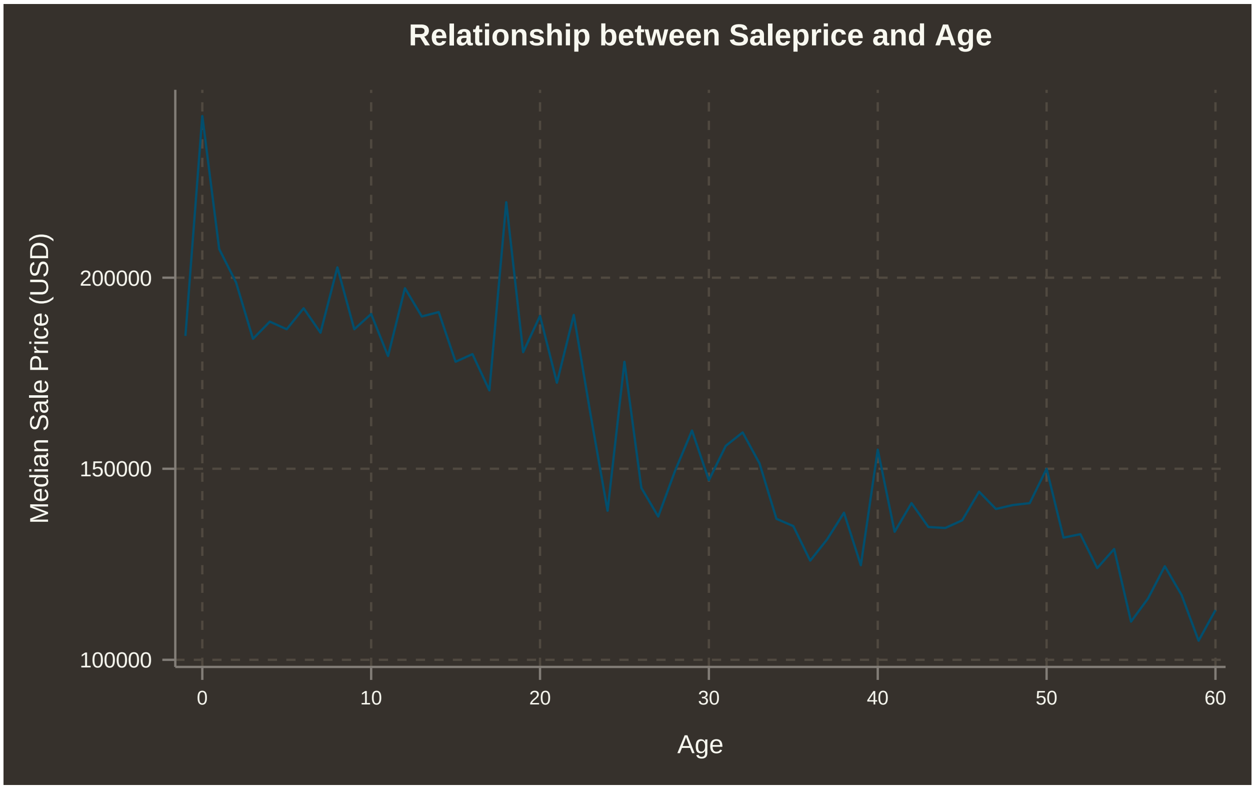


Scatterplot shows a strong postive relationship between sale price and total square feet, indicating that larger homes are more expensive. Some outliers observeed where large homes have a low sale price.

Creating House Age, Remodeled and IsNew

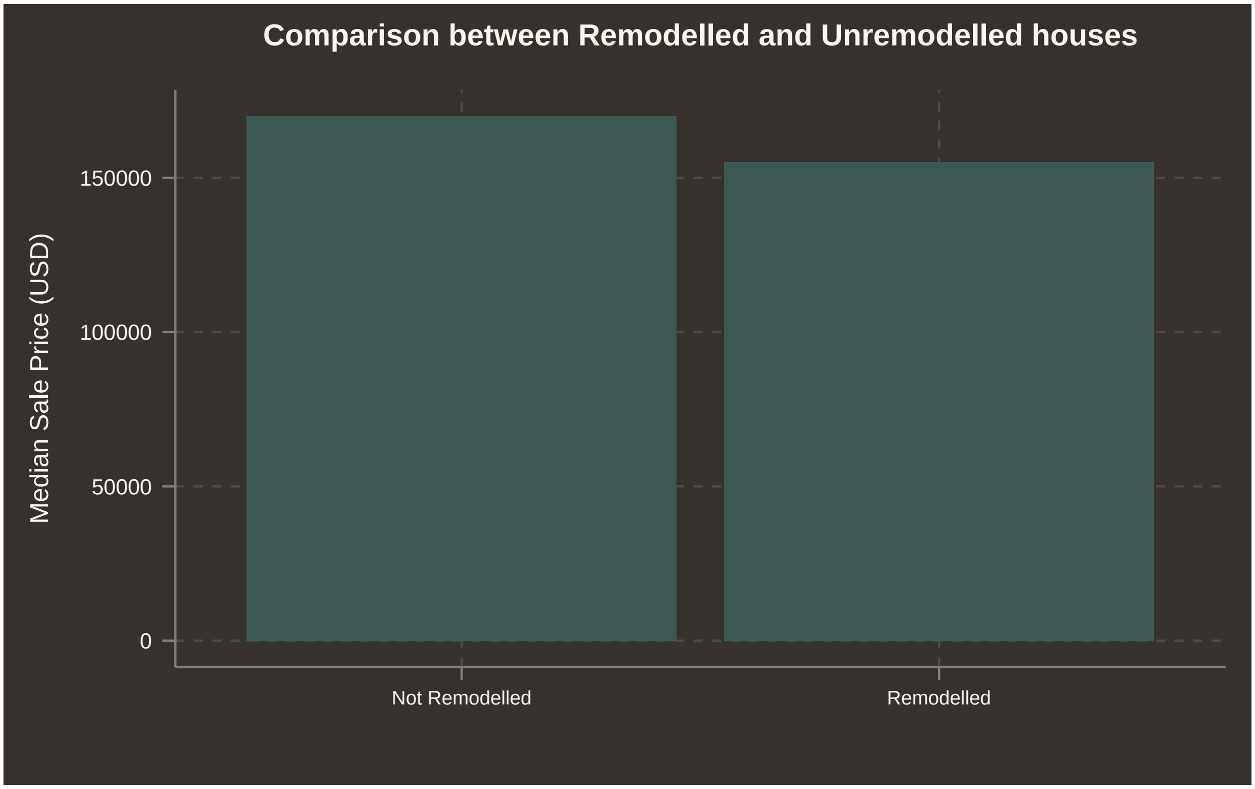
There are 3 variables that are relevant with regards to the Age of a house; YearBlt, YearRemodAdd, and YearSold. YearRemodAdd defaults to YearBuilt if there has been no Remodeling/Addition. I will use YearRemodeled and YearSold to determine the Age. However, as parts of old constructions will always remain and only parts of the house might have been renovated, I will also introduce a Remodeled Yes/No variable. This should be seen as some sort of penalty parameter that indicates that if the Age is based on a remodeling date, it is probably worth less than houses that were built from scratch in that same year.

Relationship between Age and Saleprice



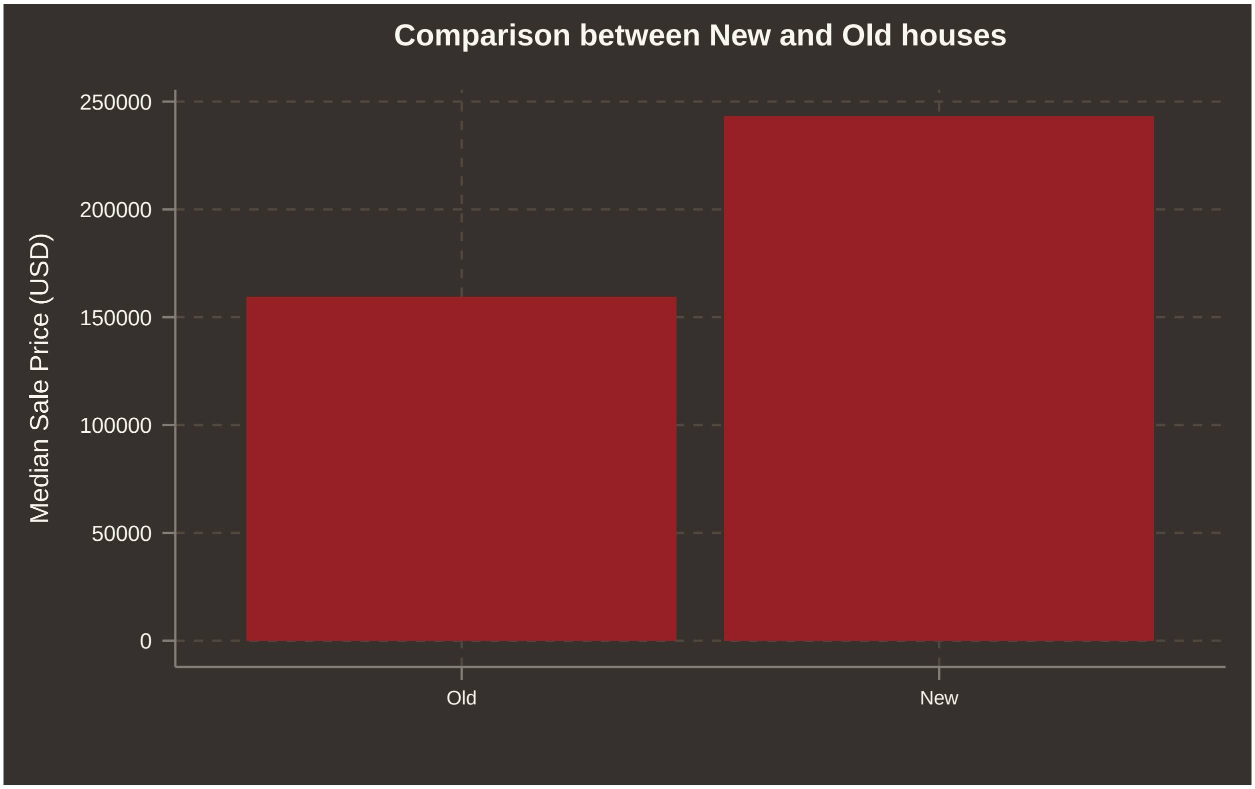
Trend shows expected relationship where older homes will have a lower sale price.

Comparison between Remodelled and Unremodelled houses



Remodelled homes have a lower median price than homes not remodelled.

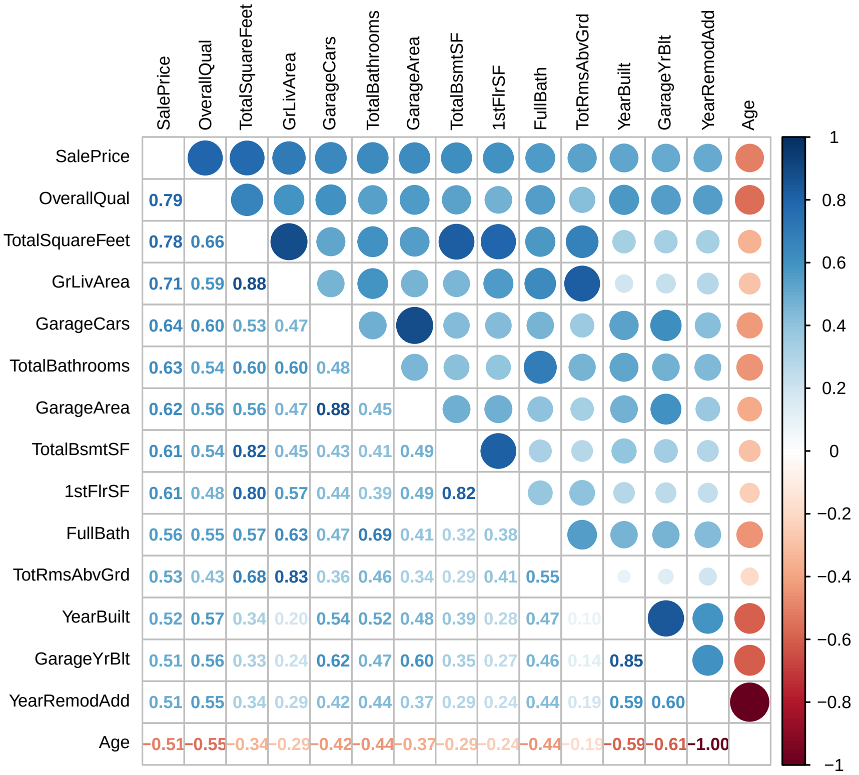
Comparison between New and Old houses



As expected, newer homes are more expensive than old homes.

Post Feature Engineering Correlations

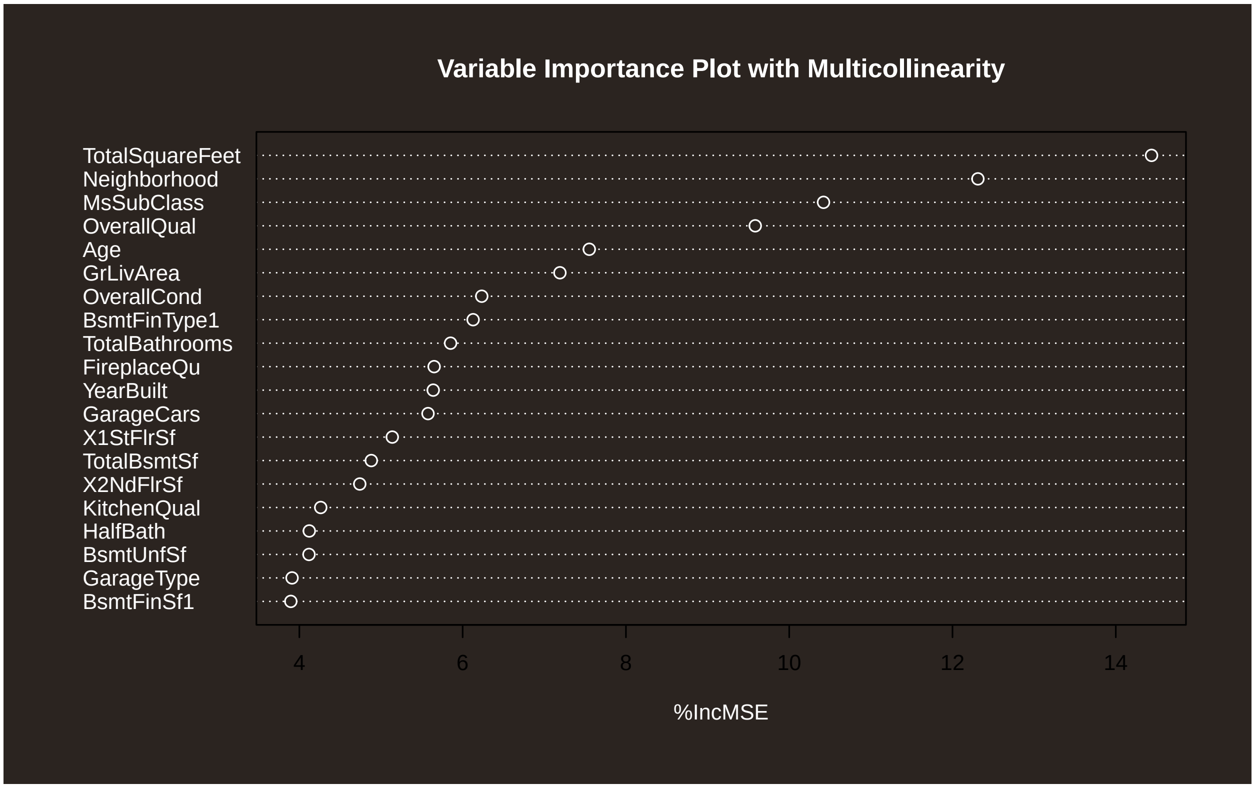
Correlation plot with new variables.



Clean variable names

Ensuring variables are in proper format for random forests. Please refer to markdown to view code.

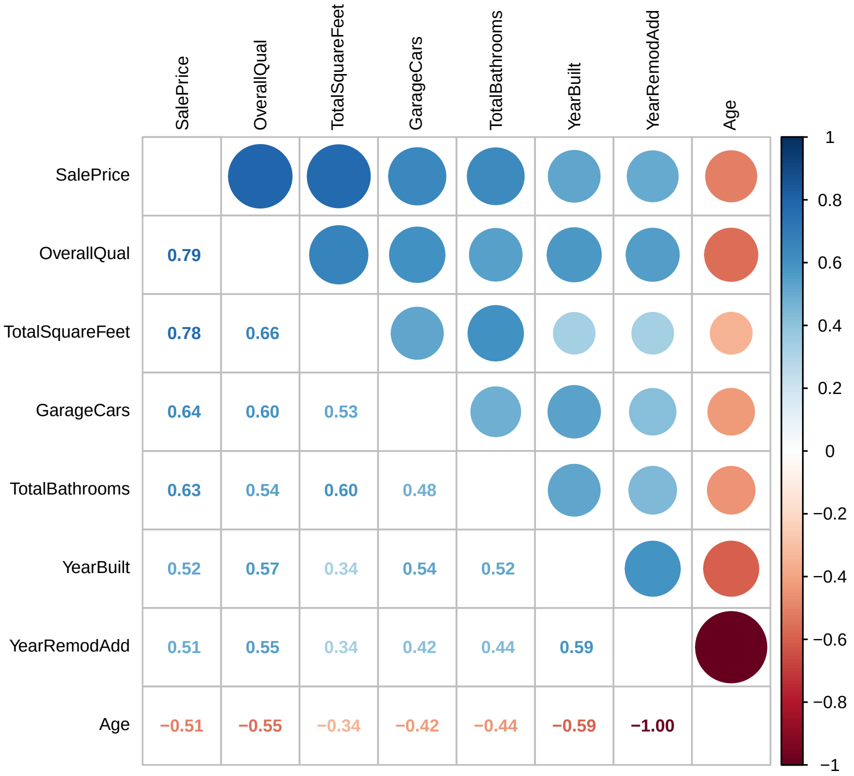
Random Forest with multicollinearity



Handling multicoliinearity

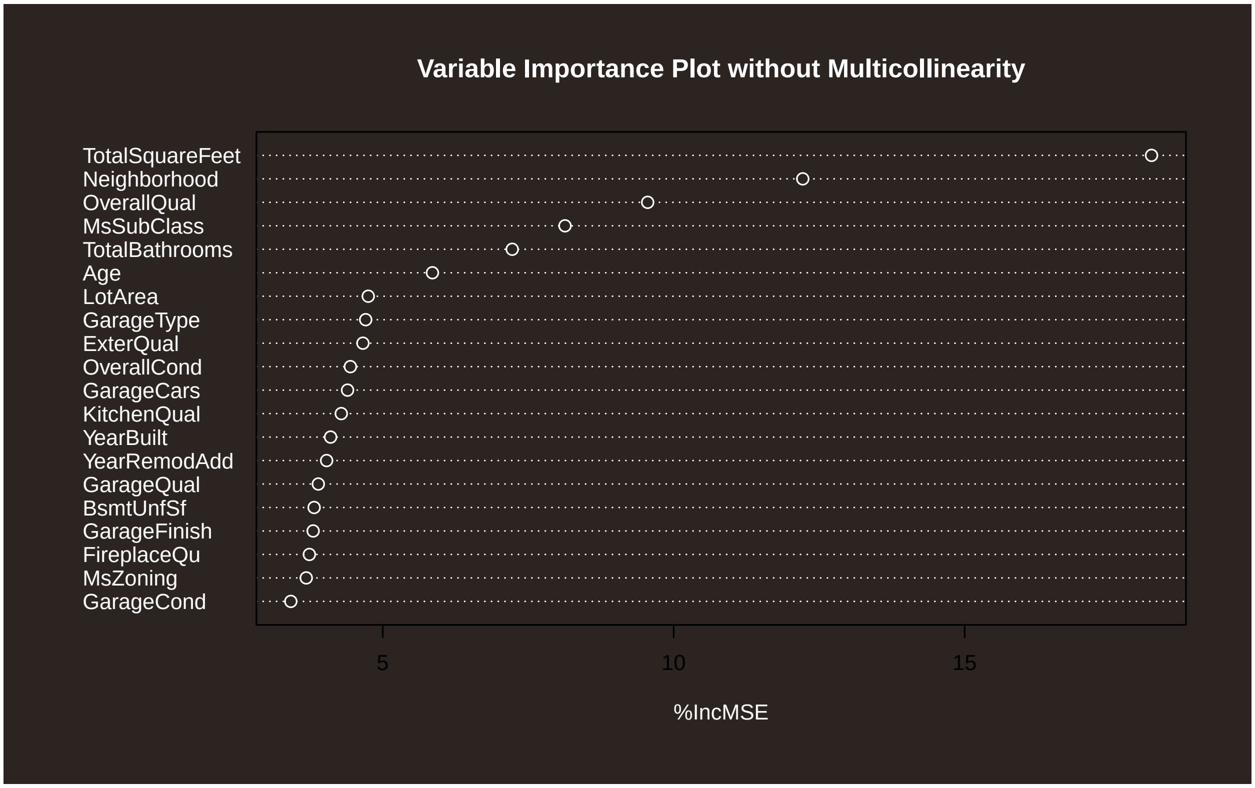
Variables with strong correlations with each other were cross-checked and the variable with a lower correlation to sale price was dropped. Please refer to markdown to view code.

Correlations after multicollinearity variables removed



Multicollinearity significantly reduced.

Random Forest without multicollineariy



Feature selection for final model

The top twenty features by MSE, Mean Square Error. Variables with a high percentage increase in MSE are considered important. MSE is a metric used to evaluate the performance of regression models, therefore, if a variable with a high percentage increase in MSE is removed from our final model, the model performance will significantly decrease, which is not what we want, hence the variable significance.

Please refer to markdown to view code.

6. Modelling

The modelling algorithms I have chosen to use are linear regression, k-nearest neighbors and random forests. Explanations of each of them are given below.

Linear Regression:

lm is a linear modeling technique that assumes a linear relationship between the independent variables and the target variable. It fits a line to the data that minimizes the sum of squared differences between the observed and predicted values. The model is interpretable and provides insights into the impact of each predictor on the target variable through coefficients, however, lm may not capture complex, non-linear relationships present in the data.

k-Nearest Neighbors:

knn is a non-parametric, instance-based algorithm used for both regression and classification tasks. It makes predictions based on the majority class or the average of the values of the k-nearest neighbors to a given data point. The model does not assume any underlying data distribution and can capture non-linear relationships effectively.

Random Forests:

rf is an ensemble learning method that constructs multiple decision trees during training and combines their predictions to make more accurate and robust predictions. Each decision tree in the Random Forests ensemble is trained on a random subset of the training data and a random subset of the features. rf is suitable for both regression and classification tasks and can handle high-dimensional data effectively. It is robust to noise and outliers in the data and can capture complex relationships and interactions between variables.

In summary, Linear Regression is suitable for modeling linear relationships between variables, k-Nearest Neighbors is effective for capturing non-linear relationships and making localized predictions, and Random Forests excel in handling high-dimensional data and capturing complex relationships between variables.

Encode categorical variables for Linear and KNN models

My models are built around three criteria:

- 1. Models with original SalePrice (no transformation or outlier removal)
- 2. Models with log transformed SalePrice (no outlier removal)
- 3. Models with outlier removal (no transformation)

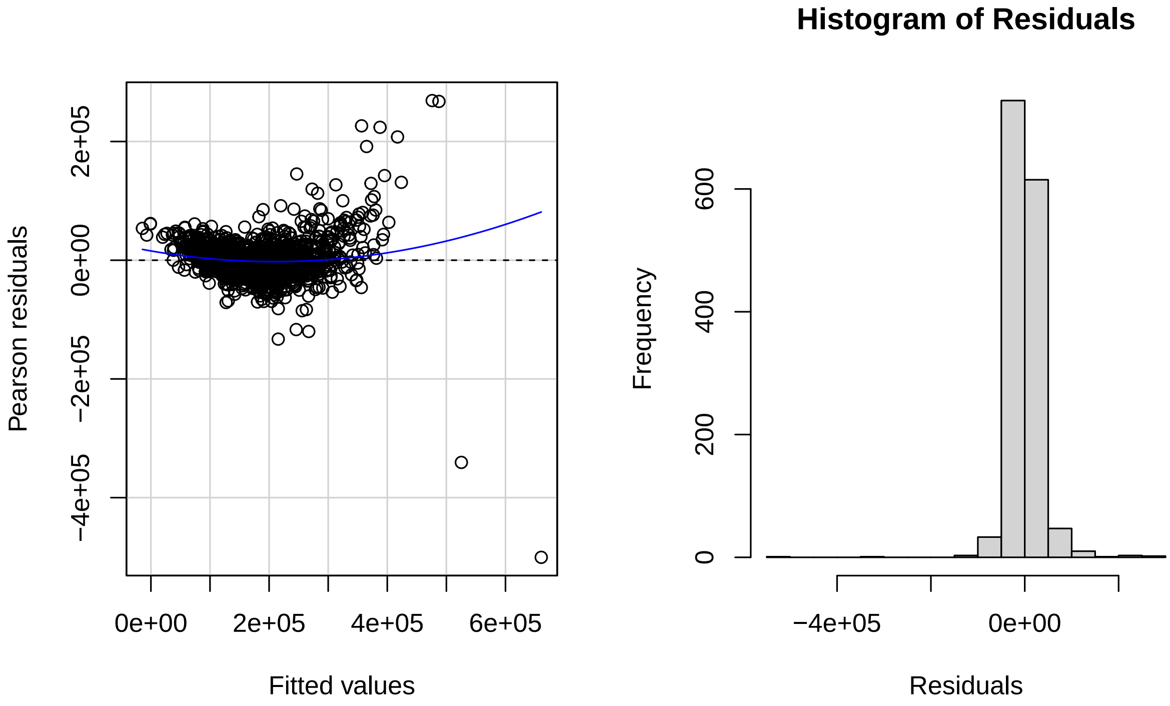
Please refer to markdown to view code.

Models with original SalesPrice

Linear Model 1

Adjusted R-squared was 0.80 with a RSE of 35100.

Spread of residuals



Predictions Predictions made on linear model. Please refer to markdown to view code.

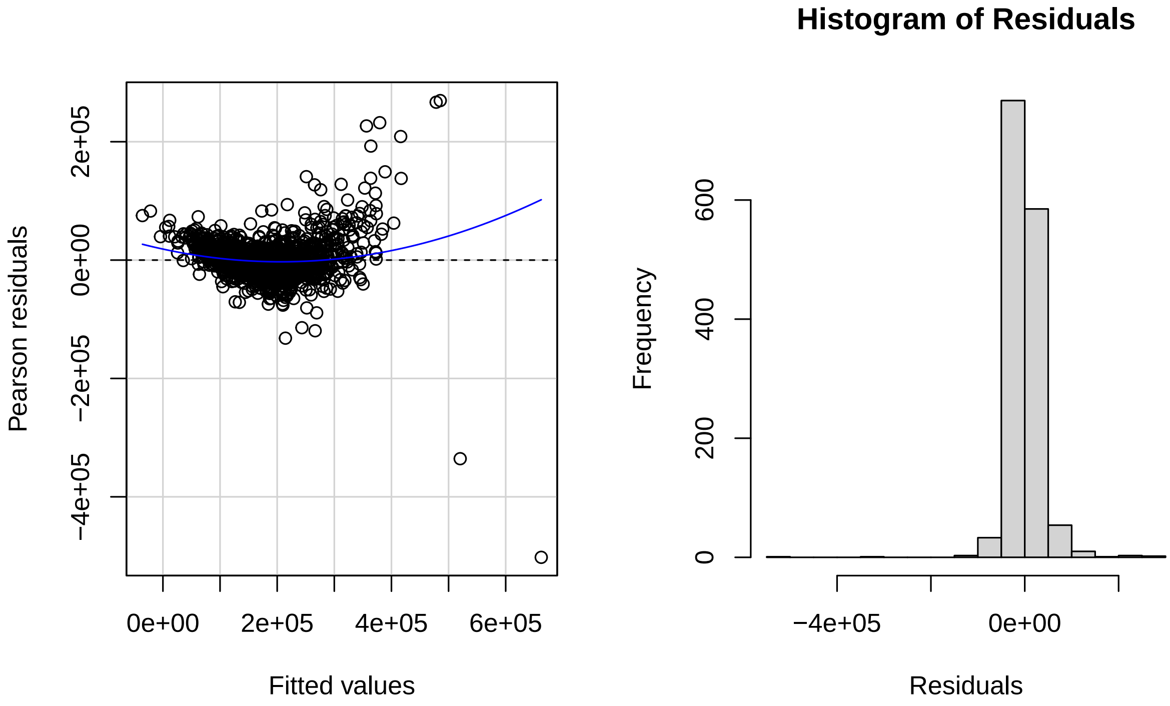
Modify Linear model Anova to determine significant variables.Please refer to markdown to view code.

Linear Model 2

Please refer to markdown to view code.

Adjusted r-squared of 0.80 and RSE of 35620.

Spread of residuals

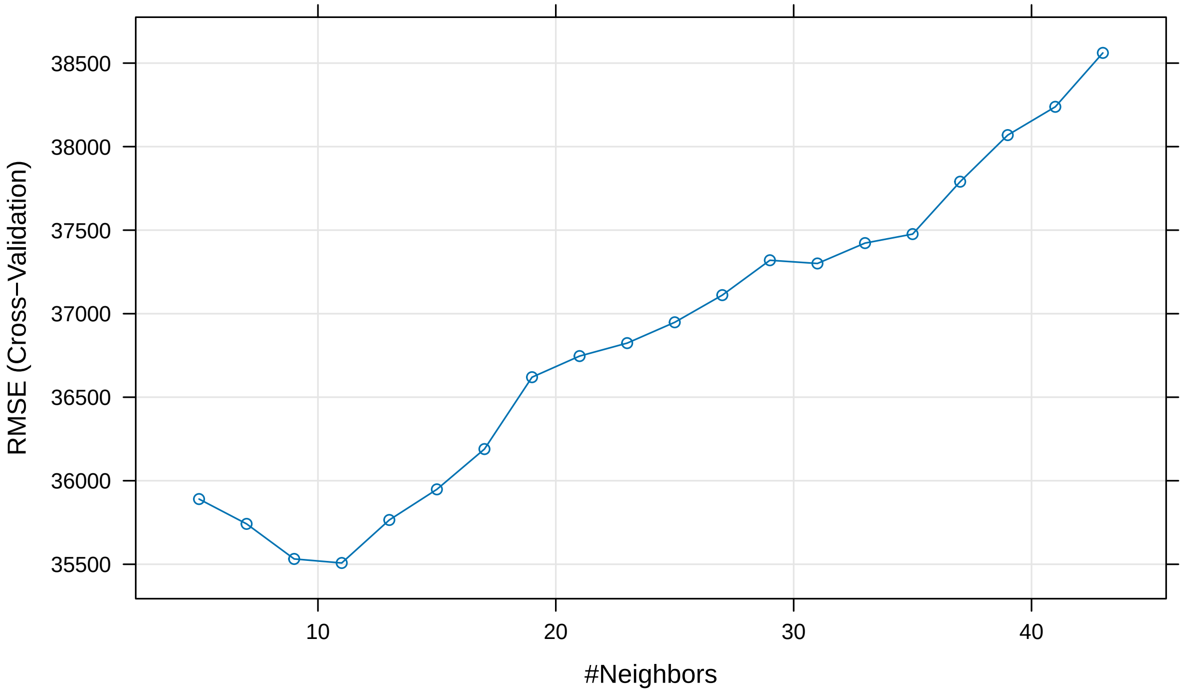


Predictions Please refer to markdown to view code.

KNN Model

Please refer to markdown to view code.

Optimal K



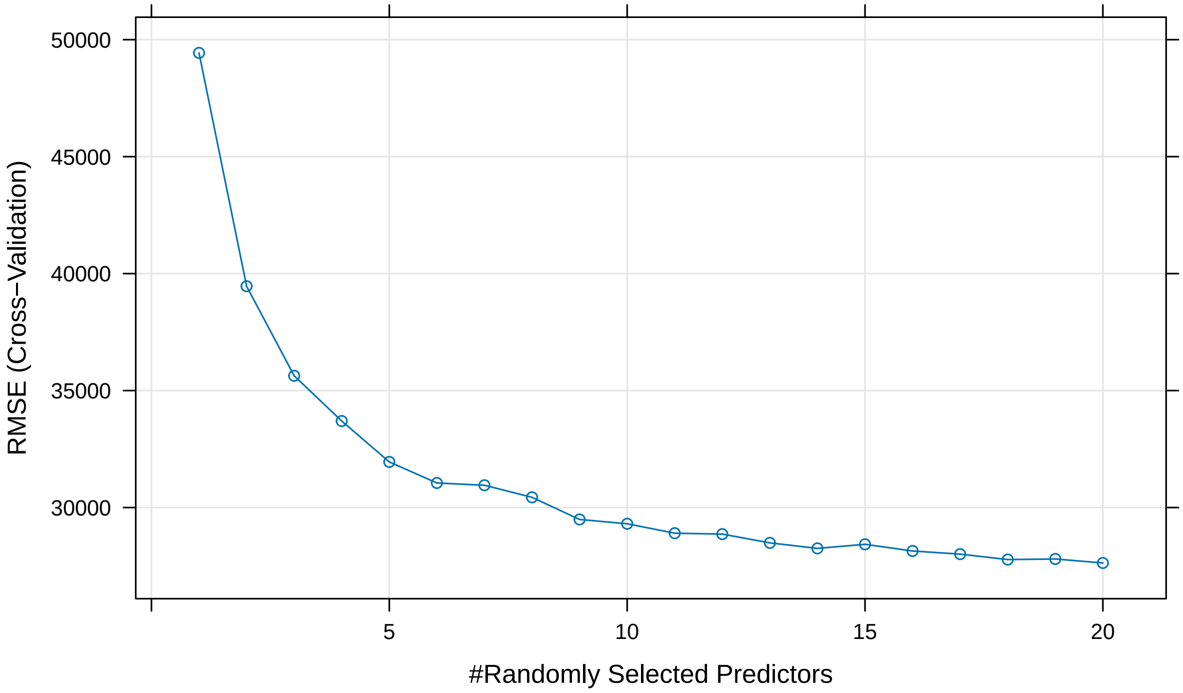
Optimal k selected was k = 11.

Predictions Please refer to markdown to view code.

Random Forest Model

Please refer to markdown to view code.

Optimal RMSE

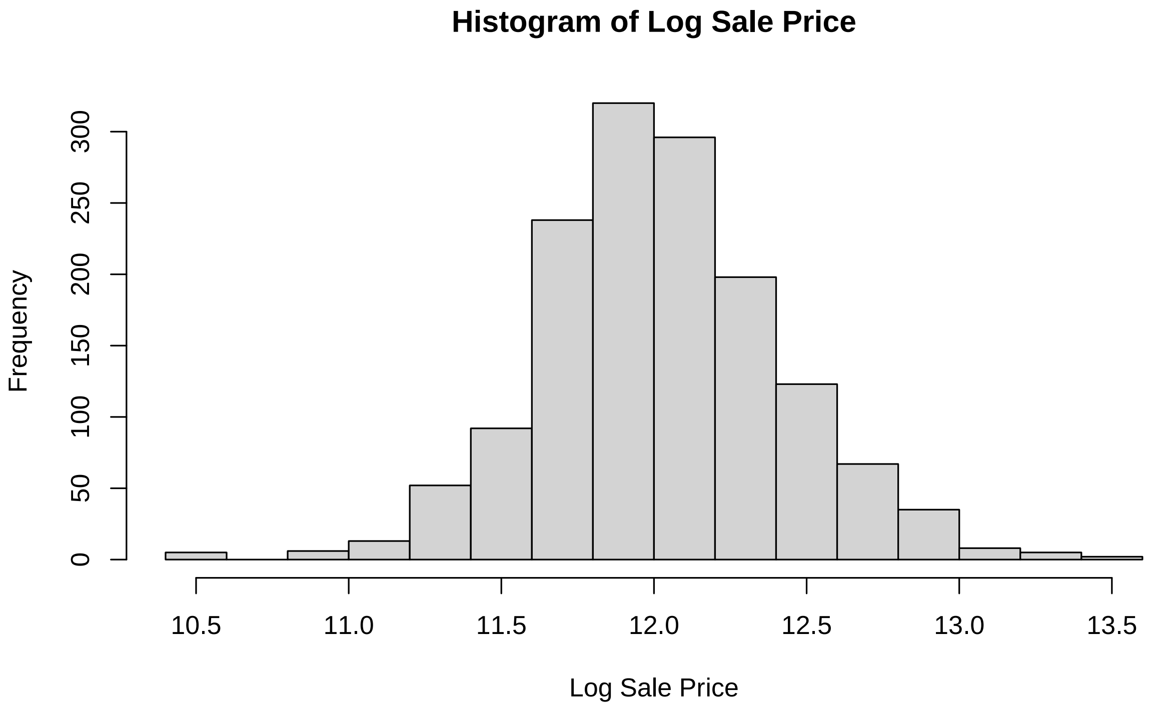


Final value used for model was mtry = 20.

Predictions

Models with Log Sales Price

Histogram of Log Sale Price

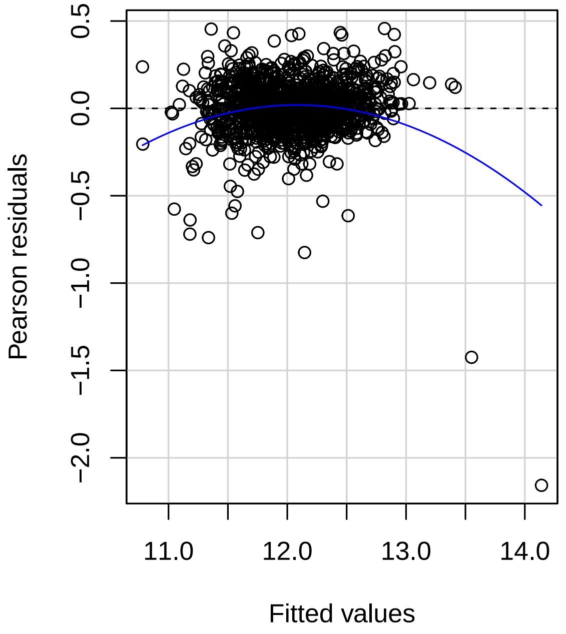


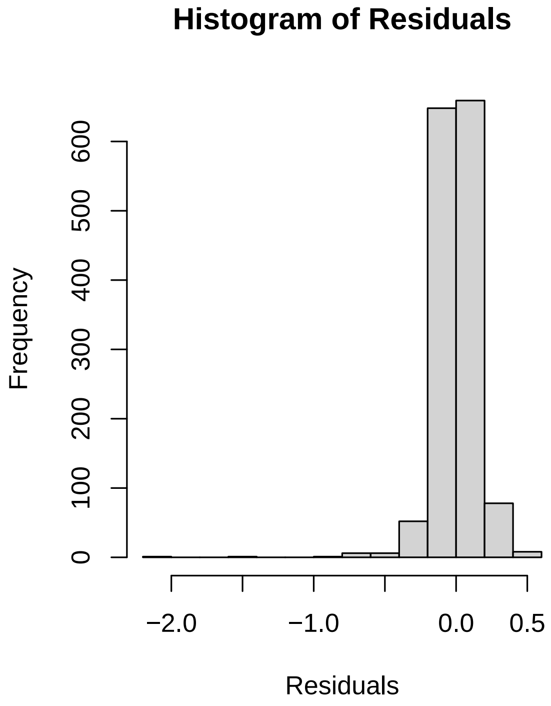
Log transformation is normally distributed.

Linear Model 1

Please refer to markdown to view code. Adjusted r-squared was 0.86 with an RSE of 0.1499

Spread of residuals





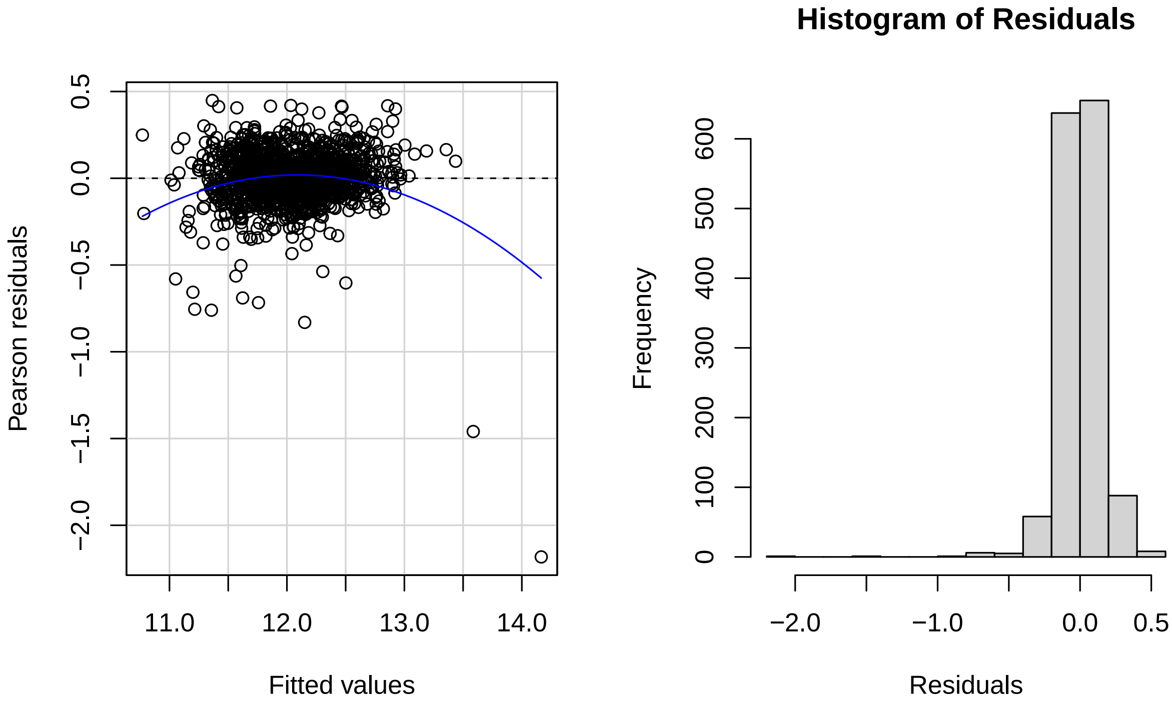
Predictions Please refer to markdown to view code.

Modify Linear model Please refer to markdown to view anova.

Linear Model 2

Please refer to markdown to view code. Asjusted r-squared was 0.86 with an RSE of 0.1513.

Spread of residuals

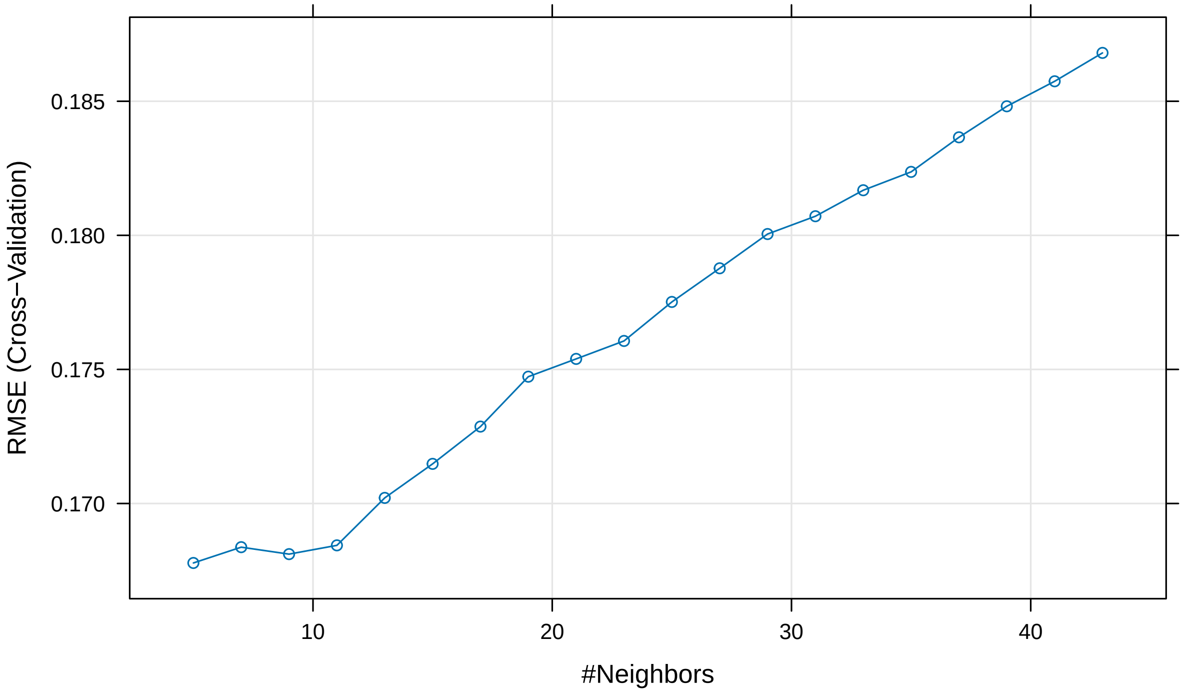


Predictions Please refer to markdown to view code.

KNN Model

Please refer to markdown to view code.

Optimal K



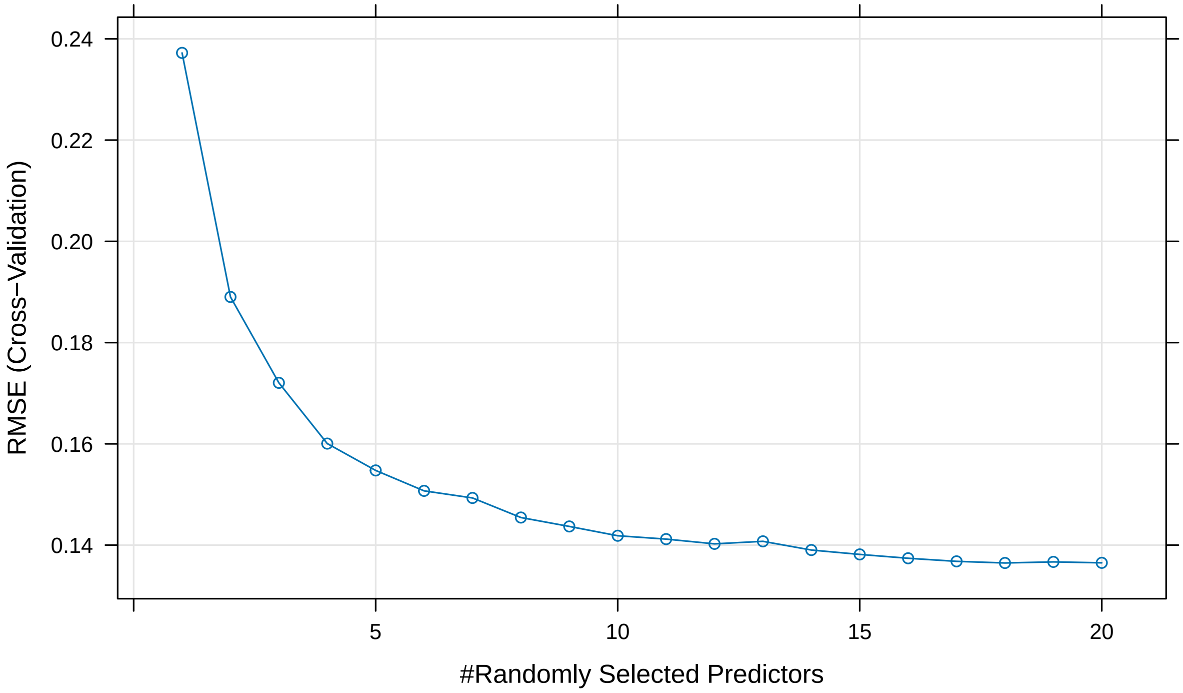
Optimal k selected was k = 5.

Predictions Please refer to markdown to view code.

Random Forest Model

Please refer to markdown to view code.

Optimal RMSE



Final value used for the model was mtry = 18.

Predictions

Models with Sales Price outliers removed

Outlier Detection for LM and KNN

Outliers identified and removed.

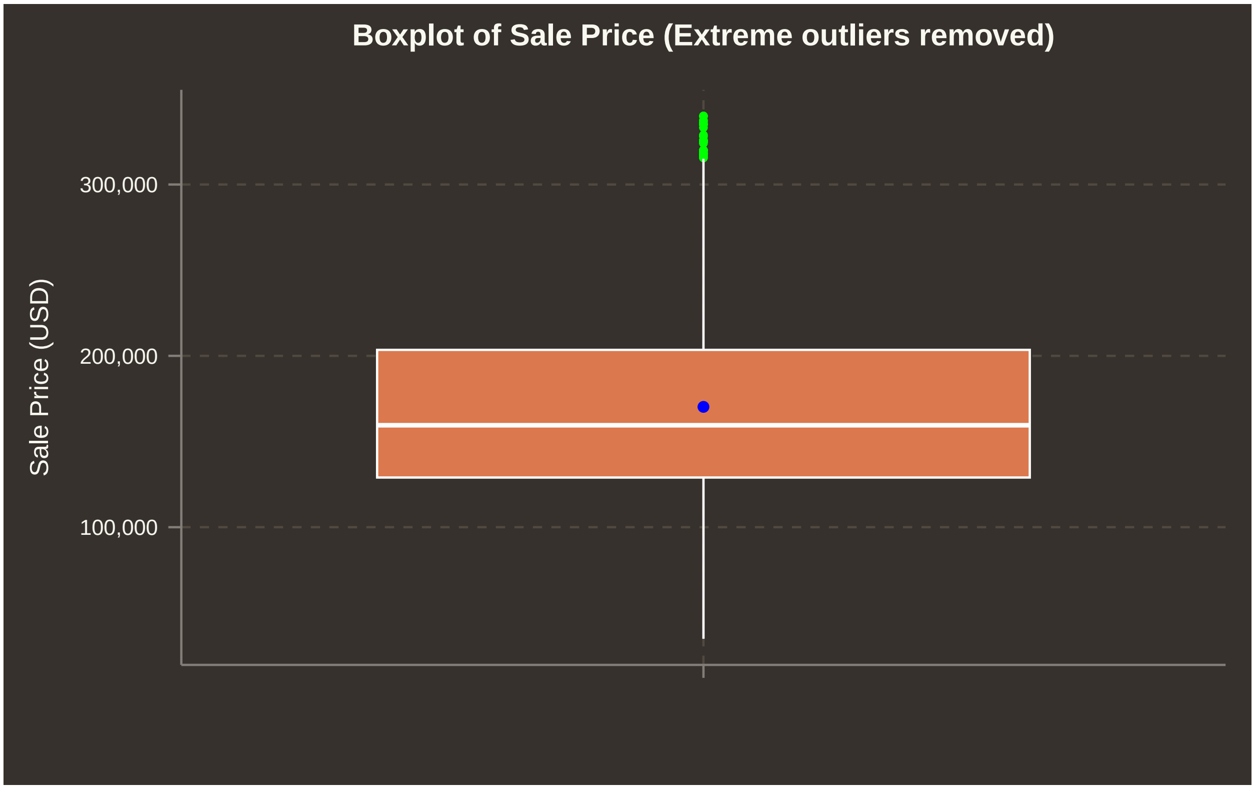
Please refer to markdown for code.

Outlier Detection for Random Forest

Outliers identified and removed.

Please refer to markdown for code.

Distribution of SalePrice without outliers

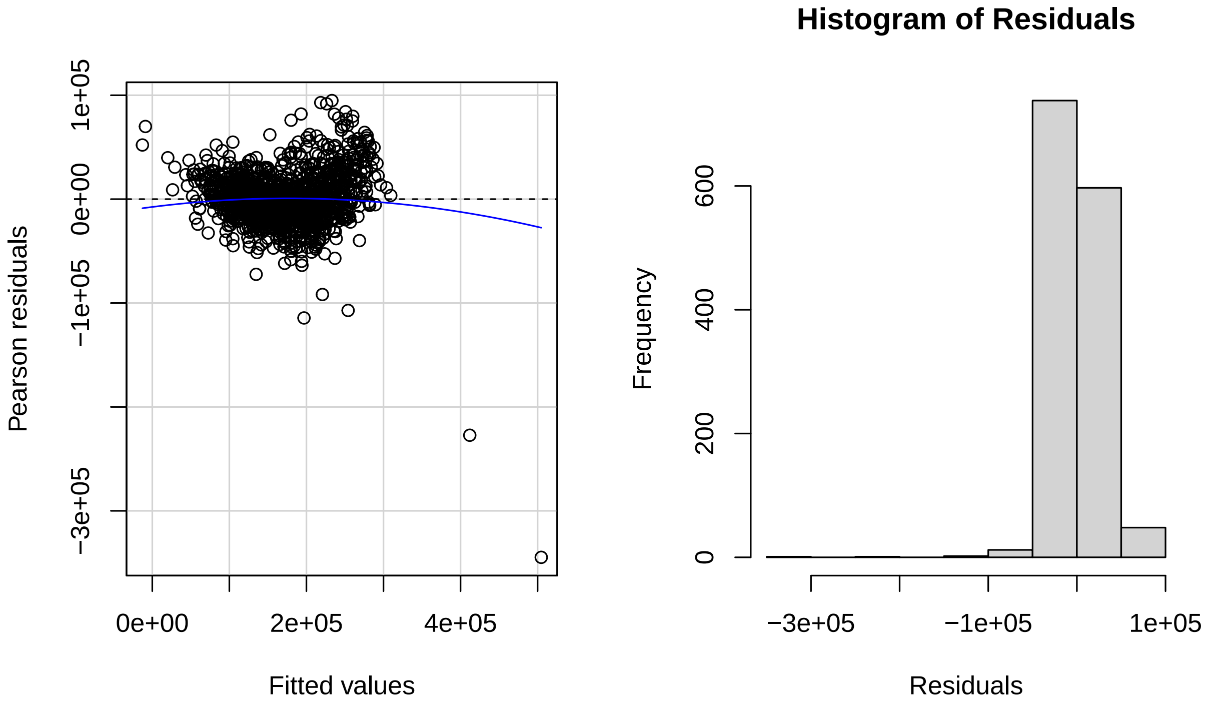


Extreme outliers have been removed.

Linear Model 1

Please refer to markdown for code. Adjusted r-squared was 0.81 with an RSE of 25590.

Spread of residuals



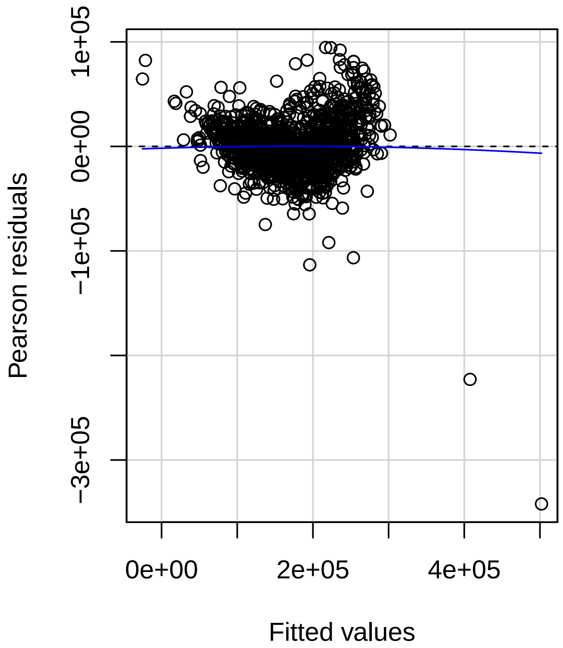
Predictions Please refer to markdown for code.

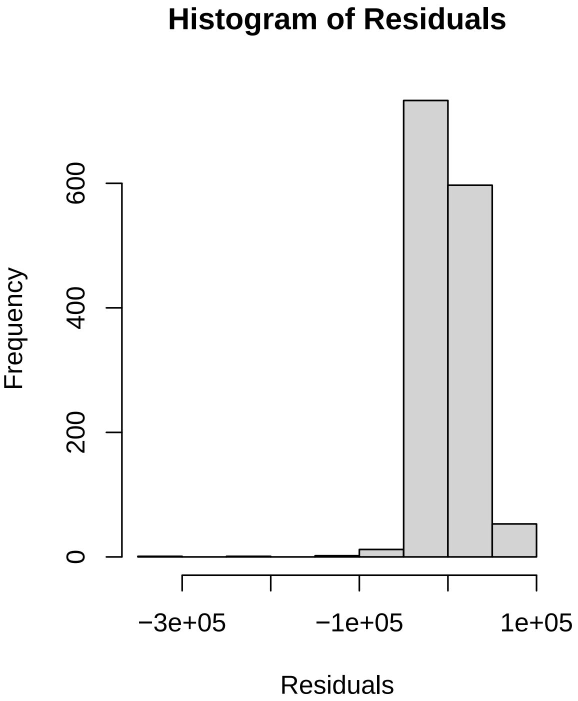
Modify Linear model Please refer to markdown for anova.

Linear Model 2

Please refer to markdown for code. Adjusted r-squared was 0.81 with an RSE of 25830.

Spread of residuals



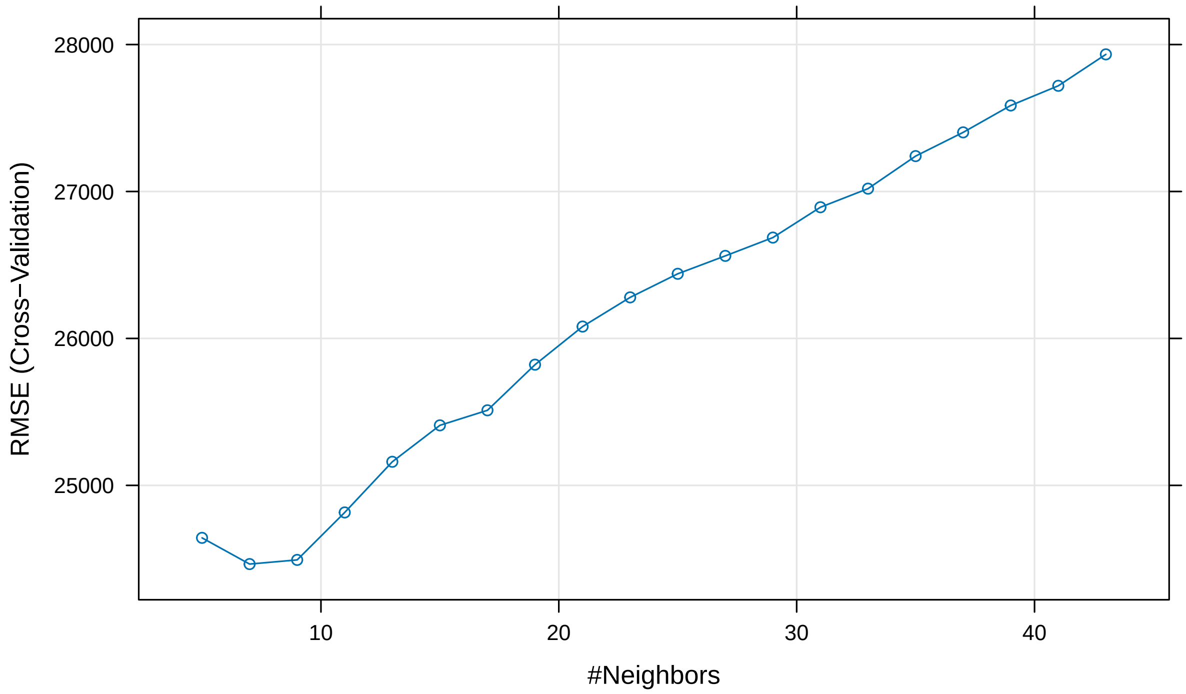


Predictions Please refer to markdown for code.

KNN Model

Please refer to markdown for code.

Optimal K



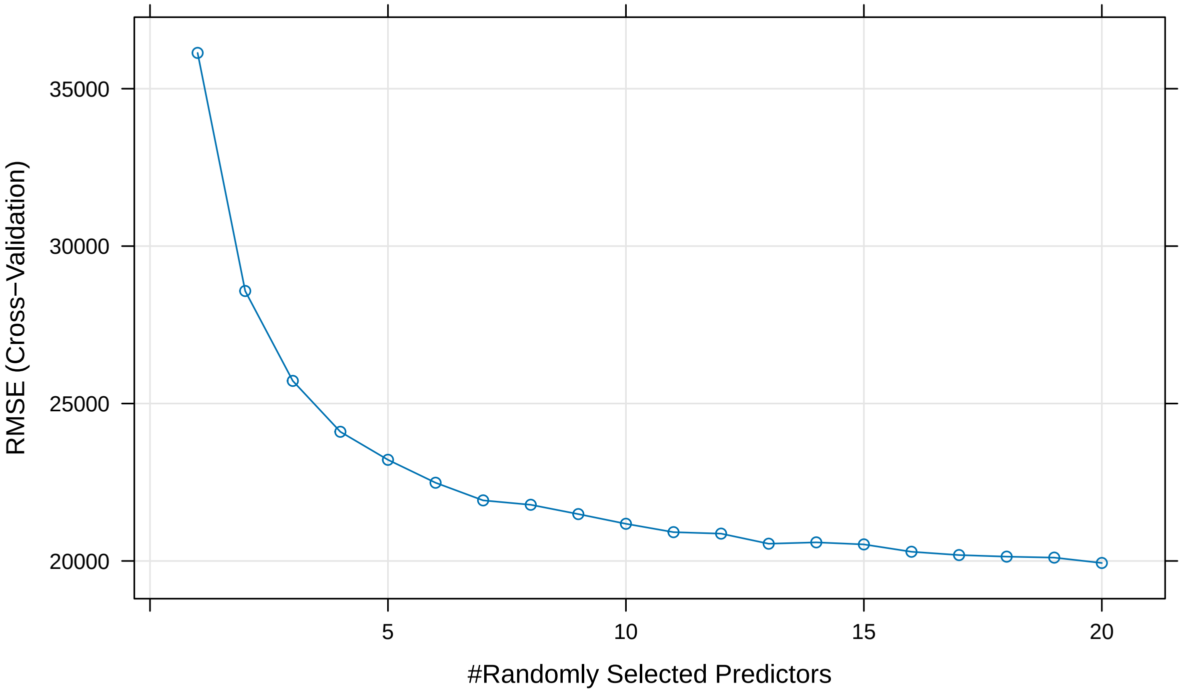
Optimal k selected was k = 7.

Predictions Please refer to markdown for code.

Random Forest Model

Please refer to markdown for code.

Optimal RMSE



Final value selected for model was mtry = 20.

Predictions Please refer to markdown for code.

7. Evaluation

In this section I compare model performance using metrics RMSE, MAE and R-squared.

An explanation of them is given below:

Root Mean Squared Error (RMSE):

RMSE measures the average magnitude of the errors between the predicted values and the actual values. It is calculated by taking the square root of the average of the squared differences between the predicted and actual values. RMSE gives higher weight to large errors, making it sensitive to outliers. Lower RMSE values indicate better model performance, with a value of 0 indicating perfect predictions.

Mean Absolute Error (MAE):

MAE measures the average absolute difference between the predicted values and the actual values. It is calculated by taking the average of the absolute differences between the predicted and actual values. MAE is less sensitive to outliers compared to RMSE since it does not square the errors. Like RMSE, lower MAE values indicate better model performance.

R-squared (R2):

R2 represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. R2 of 1 indicates that the model explains all the variability in the data, while an R2 of 0 indicates that the model does not explain any variability. R2 is a measure of how well the model fits the data relative to a simple average of the dependent variable.

Models with original Sale Price (Evaluation)

	RMSE	MAE	R2
Linear	33508.37	23780.86	0.8380004
Linear Reduced	33062.97	22835.41	0.8352144
KNN	34734.34	24317.96	0.8150152
Random Forest	23219.88	15132.73	0.9204617

Models with Log Sale Price (Evaluation)

	RMSE	MAE	R2
Linear	34684.84	18327.21	0.8173901
Linear Reduced	35116.17	18241.37	0.8116421
KNN	35356.12	25356.06	0.8105782
Random Forest	24347.24	15450.57	0.9166419

Models with Sale Price outliers removed (Evaluation)

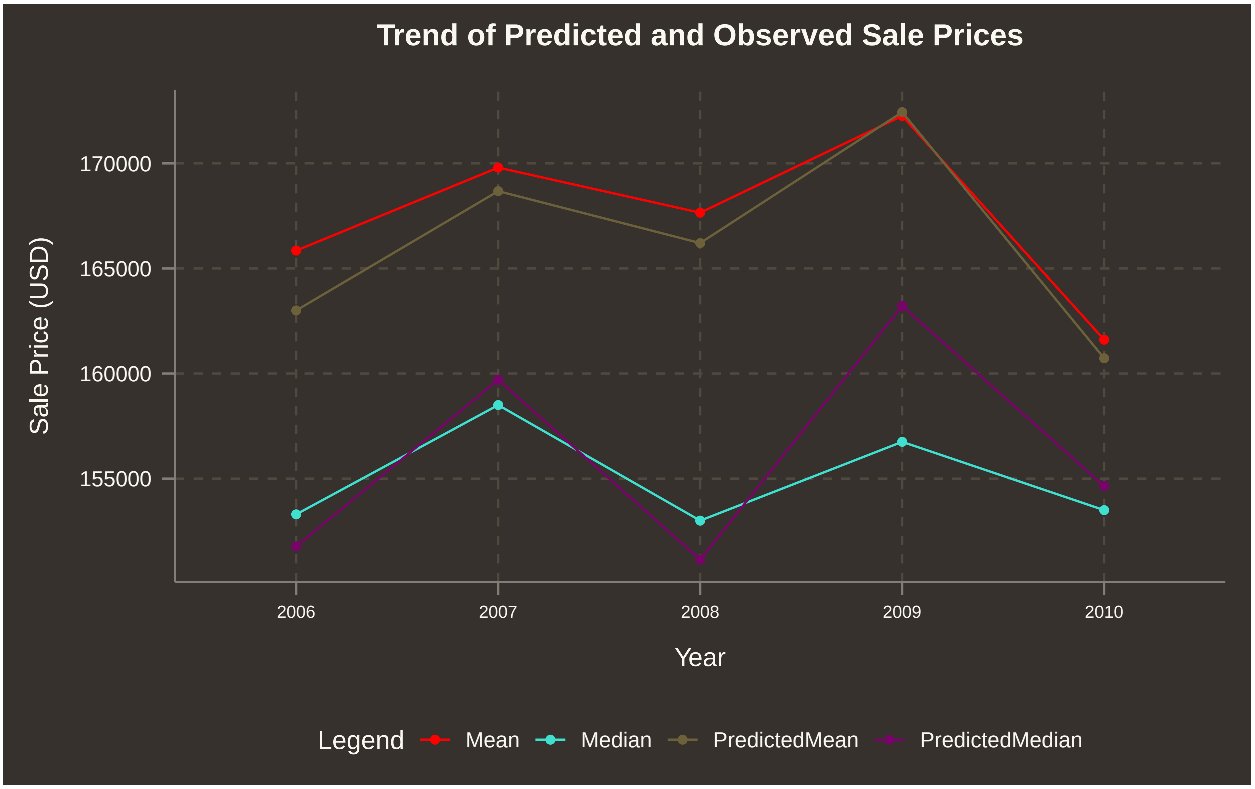
	RMSE	MAE	R2
Linear	25585.85	18734.45	0.8142688
Linear Reduced	26727.89	19733.58	0.8006084
KNN	29997.82	22522.47	0.7456760
Random Forest	19053.02	13561.45	0.9008910

Average RMSE of Model Categories

	Mean_RMSE
Models with original SalePrice	31131.39
Models with Log SalePrice	32376.09
Models with SalePrice no outliers	25341.14

Models that were built on SalePrice without outliers on average performed best. Moving forward, model performance can be improved upon with outlier removal as a baseline.

Trend of Predicted and Observed Sale Prices



The plot above is a replica of the Median/Mean sale price over time but with the predicted values of the best model. Best model picked was the random forest with no outliers in sales price.

Predicted values show very similar pattern to observed values, ie, predicted mean is higher than predicted median.

8. Recommendations and Conclusions

Key findings from my analysis:

Rigourus data pre-processing proved worthwhile as models with low RMSE and high accuracy were produced. Maintaining the data pipeline was particularly challenging because consistency had to ensured between both train and test sets. Not remembering to mirror pre-processing across both datasets can improve very costly and time confusing to fix when working up the pipeline.

EDAs highlighted that Saleprice had strong correlations with neighbourhood and overall quality. More variables could have been plotted to check their correlation however I was most interested in these two variables because those are the factors most home buyers consider when looking for a new home; the neighbourhood/suburb of the property and its build quality, hence it seemed more worthwhile to study them.

Outlier removal definitely had a strong impact on model performance. The best RMSE obtained was a random forest model trained on datasets without outliers. The variables used were the significant variables determined from earlier random forest models. This proved particularly useful in achieving dimensionality reduction and improving model performance.

Improvements that can be made to models:

- Feature engineer more variables:

LandValue and SeasonSold. Land value can be used to study correlation on the value of land the property is built on with sale price. I imagine a number of metrics will have to be taken into account to determine land value include neighbourhood, recreational areas around the lans, commercial areas around the land, demographics around the land etc. Scaling all these metrics into a number that can be assigned as land value would be challenging as well.

SeasonSold can be used to study the finer details between saleprice vs season and number of sales vs season. The time series plot I made already captures most trends but it would be more insightful if the trends were broken down into seasons.

- Build more models:

More models could definitely be built. One such example would be training a models with a combination of log sale price and outlier remova. I expect this would produce very low rmse values since outlier removal on its own already significantly improves model performance. A stricter ANOVA selection could be done when reducing linear models. The five most significant variables from ANOVA could be selected and used on the reduced model. I expect this to produce better rmse values for the linear models. Since linear regression struggles with complex dimensions, strictly reducing dimensions while not straying far away from significance should produce considerably better results. Furthermore, more advanced models such as neural networks could be trained and evaluated.

- Useful findigs:
- Package “ggthemr” for the aesthetics and produce better looking plots.
- plot_missing function was particularly useful in visualising the proportion of missing values.
- Random forests for variable importance made feature selection a lot more efficient.
- Density plots were used as an alternative to histograms where bin and bin-width selection were difficult to determine.

Overall, from my plots and models, you can make recommendations on affordable/expensive neighbourhoods, have an estimate of saleprice on home age, overall quality, totalbathrooms etc, and know which seasons are more affordable to buy a home.

9. References

Boykin, R. (2023) How seasons impact real estate investments, Investopedia. Available at: https://www.investopedia.com/articles/investing/010717/seasons-impact-real-estate-more-you-think.asp (Accessed: 08 May 2024).

Great recession: What it was and what caused it (2023) Investopedia. Available at: https://www.investopedia.com/terms/g/great-recession.asp#:~:text=The%20economic%20slump%20began%20when,and%20derivatives%
(Accessed: 08 May 2024).

Research guides: This Month in business history: The panic of 1873 (2021) The Panic of 1873 - This Month in Business History - Research Guides at Library of Congress. Available at: https://guides.loc.gov/this-month-in-business-history/september/panic-of-1873#:~:text=The%20Panic%20of%201873%20triggered,stock%20market%20crash%
(Accessed: 08 May 2024).