11/21/2024

Predicting House Prices: A Regression Analysis

University Of Waterloo

MSE 609: Quantitative Data Analysis

Kurt Anand
Lily Hijazi
Majid Taherkhani
Maryam Shirinchi
Nima Nafezi

Introduction

The real estate market is complex and significantly affects individual decisions and the overall economy. Predicting house pricing accurately can be challenging, but highly valuable as it provides important insights to buyers, sellers, investors, and decision makers, helping to better understand market dynamics and affordability, and making well-informed decisions.

The primary goals of this project are to identify the key property features influencing sale prices and to develop a regression model that predicts house prices with high accuracy. Utilizing Ames Housing Dataset, 2006-2010 [1], and the programming language R, we employed a regression approach to detect the relationship between sale price and several property features. This analysis enabled us to identify significant variables, negligible factors, and potential interdependencies. By selecting the most impactful variables, we developed an optimal model that minimizes prediction error and has a high R-squared value, as measured by root-mean-square logarithmic error.

We started by checking the effectiveness of different features such as location, number of bedrooms, lot area, and the capacity of the garage on the sale price of houses in Ames, Iowa, USA. After which we developed a regression model to predict house prices in Ames using the key property features. We trained our model and evaluated its performance using test data and then provided a function to input the features and output the predicted house price.

1. Description of Data

1.1 Key Variables:

The primary data set contains 83 columns [2].

- **Numerical:**

For our analysis, we focused on the following 33 key numerical variables that we believe influence housing prices:

Sale Price (Target variable): Refers to the final sale price of the property in dollars.	Bedroom AbvGr: Number of bedrooms above ground.
Lot Frontage: Linear feet of street connected to the property.	Kitchen AbvGr: Number of kitchens above ground.
Lot Area: Total lot size in square feet.	TotRms AbvGrd: Total number of rooms above ground.
Year Built: Original construction year of the house.	Fireplaces: Number of fireplaces.
Year Remod/Add: Year of remodeling or additions.	Garage Yr Blt: Year the garage was built.
BsmtFin SF 1: Square feet of finished area in basement (Type 1).	Garage Cars: Size of garage in car capacity.
BsmtFin SF 2: Square feet of finished area in basement (Type 2).	Garage Area: Size of the garage in square feet.
Bsmt Unf SF: Square feet of unfinished basement area.	Wood Deck SF: Square feet of wood decking.
Total Bsmt SF: Total square feet of basement area (finished and unfinished).	Open Porch SF: Square feet of open porch.

1st Flr SF: Square feet of the first floor.	Enclosed Porch: Square feet of enclosed porch.
2nd Flr SF: Square feet of the second floor.	3Ssn Porch: Square feet of three-season porch.
Low Qual Fin SF: Square feet of low-quality finished space.	Pool Area: Size of the pool area in square feet.
Gr Liv Area: Above ground living area in square feet.	Misc Val: Value of miscellaneous features.
Bsmt Full Bath: Number of full bathrooms in the basement.	Mo Sold: Month the property was sold.
Bsmt Half Bath: Number of half bathrooms in the basement.	Yr Sold: Year the property was sold.
Full Bath: Number of full bathrooms in the house.	MS.SubClass: Identifies the type of dwelling involved in the sale
Half Bath: Number of half bathrooms in the house.	

- **Categorical:**

We have removed the least important categorical variables due to the below reasoning:

1. Utilities: Most homes in the dataset have the same level of utility access, so there is minimal variation in this feature, making it less useful. 2. Street: This variable represents the type of street access (e.g. paved or gravel). Since most homes are on paved streets, it often does not contribute much to model performance. 3. Pool.QC (Pool Quality): Very few homes in the dataset have pools, resulting in a high proportion of missing values, thereby making it less relevant for most models. 4. Misc.Feature: Describes additional features (e.g. sheds, tennis courts), which are rare and hence usually contribute very little to overall predictions. 5. Condition.2: Represents a second proximity condition like railroads or highways. Most properties only have a single condition in Condition.1, so Condition.2 often adds little additional value. 6. Land.Slope: Describes the slope of the property. Most homes in the Ames dataset are on flat land, so there is very minimal variation in this feature. 7. Roof.Matl (Roof Material): Nearly all homes have asphalt singles, so there is little variance here and hence making it less predictive. 8. Order and PID: These are unique identities for each home and hence they do not contain any information relevant to the sale price. Below are the categorical variables we used:

MS.Zoning: Identifies the general zoning classification of the sale	Bsmt.Exposure: Refers to walkout or garden level walls
Land.Contour: Flatness of the property	BsmtFin.Type.1: Type 1 finished square feet
Lot.Config: Lot configuration (ex. Corner lot, Cul-de-sac))	BsmtFin.Type.2: Rating of basement finished area (if multiple types) (ex. Unfinished)
Neighborhood: Physical locations within Ames city limits	Heating: Type of heating
Condition.1: Proximity to various conditions	Heating.QC: Heating quality and condition
Bldg.Type: Type of dwelling (ex. Detached, Townhouse)	Central.Air: Central air conditioning (Yes, No)
House.Style: Style of dwelling (ex. One story, Two Story)	Electrical: Electrical system
Roof.Style: Type of roof	Kitchen.Qual: Kitchen quality

Exterior.1st: Exterior covering on house	Functional: Home functionality (Assume typical unless deductions are warranted)
Exterior.2nd: Exterior covering on house (if more than one material)	Fireplace.Qu: Fireplace quality
Mas.Vnr.Type: Masonry veneer type	Garage.Type: Garage location
Exter.Qual: Evaluates the quality of the material on the exterior	Garage.Finish: Interior finish of the garage
Exter.Cond: Evaluates the present condition of the material on the exterior	Garage.Qual: Garage quality
Foundation: Type of foundation (ex. Wood, Stone)	Garage.Cond: Garage condition
Bsmt.Qual: Evaluates the height of the basement	Paved.Drive: Paved driveway
Bsmt.Cond: Evaluates the general condition of the basement	Fence: Fence quality

1.2 Data Source:

- **Data Source**: We used the Ames Housing dataset from Kaggle platform, which was compiled by Dean De Cock, Professor of Statistics at Truman State University, for use in data science education. The dataset offers comprehensive information about residential properties sold in Ames, Iowa from 2006 to 2010 [1]. Each record in the dataset represents a single house and includes numerous features related to the property's characteristics, conditions, and amenities. The dataset is publicly available on Kaggle and does not require special permissions for use.
- **Data Reliability**: The dataset reliability is ensured as it originates from Ames Assessor's Office [3] which is a government agency in Ames, Iowa, responsible for valuing real properties within the city. They maintain accurate and reliable records used for property tax calculations, ensuring a high level of data quality.
- **Data Vetting**:
  - **Origin**: The data comes from the Assessor's Office, a reliable public agency that values property. This ensures reliability and consistency in the data collection process.
  - **Data Dictionary**: A detailed data dictionary accompanies the dataset, providing clear explanations for each variable and its possible values, which helps in feature selection and preprocessing.
  - **Data Size and Representativeness**: The dataset is large and captures various features of residential properties, enabling in-depth analysis and modeling.
  - **Potential Bias**: There are no potential biases in the data set we used. However, the data may reflect historical trends, potentially limiting its applicability to current market conditions. Additionally, the dataset is specific to Ames, Iowa, and its findings may not generalize to other regions.

- **Data Quality**: We had some missing values, and we discussed the best way to deal with these values. After checking the number of missing values, we decided to remove those columns and rows as shown in Data Cleaning & Preprocessing section below.
- **Data Cleaning & Preprocessing**:
    - Handle Missing Values:
        - For some of the histograms, we used na.omit ( ) to remove missing values of the respective column and then used hist ( ) to graph the histogram. For example, to draw the histogram of the sale prices of the houses in Ames, use hist (na.omit(SalePrice)). For other histograms, we did not use the na.omit ( ) function, so hist ( ) automatically removes missing values, but using na.omit serves as a good practice to handle missing values after conducting EDA each time. Barplots have handled missing values the same way as histograms.
        - For scatterplots between the sale price and each of the predictor variables, what we did is define a new data frame that is a subset of the original Ames dataset, except that there are only two columns, one of which is the SalePrice column and is identical to the sale price in the Ames data set. The other column is identical to the predictor column in the Ames dataset. We now take a subset of this new dataset, where any row with at least one missing value will be removed. Now, plot the two variables in the data set using the plot( ) function where the column derived from the SalePrice column is a dependent variable.
        - Columns with more than 20% missing values were removed. (Alley, Fireplace Qu, and Fence).
        - Columns with less than 20% missing values: used interpolation for numerical variables, and mode for categorical variables.
        - By calculation, there should not be more than 586 missing values (as there are 2930 observations).
    - Standardize and Categorize:
        - Categorical columns were retained in their original form and not converted into numeric as their categorical nature was relevant to the analysis. However, some numerical columns were converted into categorical types (e.g. Pool Area) to better represent their impact as discrete categories rather than continuous values, enhancing interpretability and feature relevance.
- **Ethical considerations**: Although the dataset doesn't contain personally identifiable information (PII) and sensitive information, we did handle it responsibly and ethically, in accordance with Kaggle's data usage policies [4].

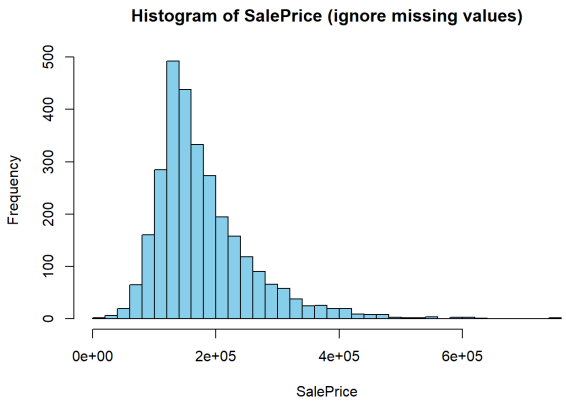
2. Descriptive Data Analysis

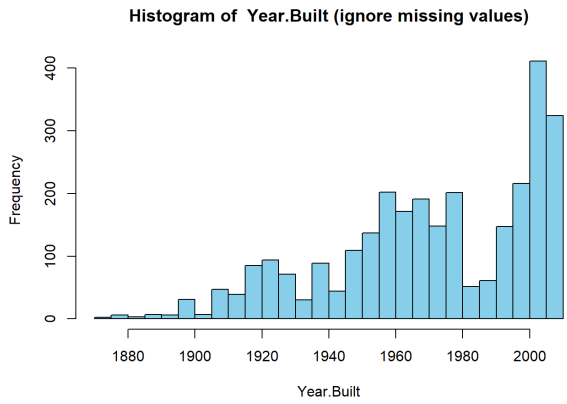
In order to have a more detailed understanding of the variables in the dataset, we decided to perform Exploratory Data Analysis on the housing price dataset. We created histograms of all the numerical variables in the dataset and also created scatterplots for each pair between variables: Bar graphs were included to compare the frequency of each category on the categorical variables. Box plots were created to compare Sale Price across categories.

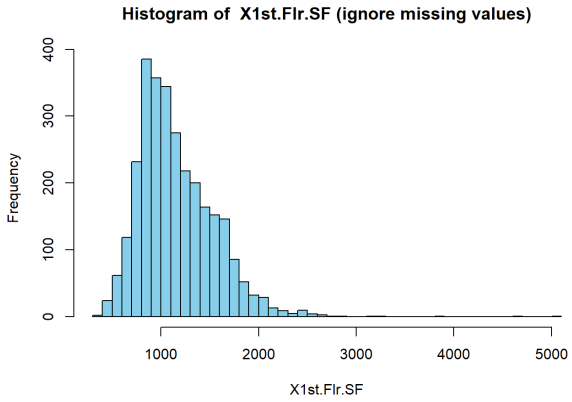
We have analyzed the distributions for each of the numerical columns, including the SalePrice (which is the target variable for this project), by using basic summary statistics, histograms, scatterplots, and boxplots. The summary statistics table included in Appendix A shows that the target variable SalePrice has a range of [12789 755000], mean of 180796, median of 160000, and no missing values. Year Built has a range of [1872 2010], mean of 1971, median of 1973, and no missing values. X1st.Flr.SF has a range of [334 5095], mean of 1159.6, median of 1084, and no missing values. Gr.Liv.Area has a range of [334 5642], mean of 1500, median of 1442, and no missing values. If the number of missing values is not mentioned, then there are no missing values.

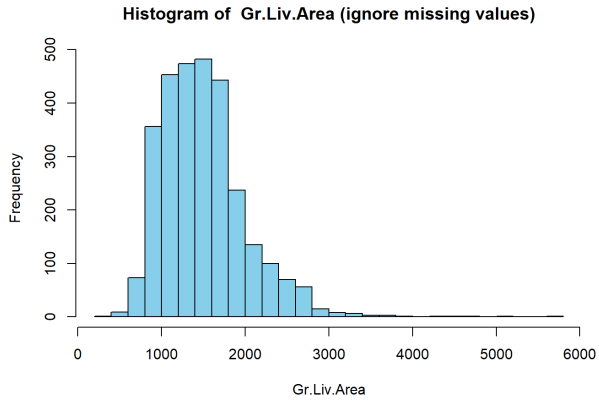
2.1 Visualizations:

Histograms: We constructed histograms for key variables that are numerical to understand their distribution, detect skewness, and identify outliers. Of the key variables that are numerical, we omit missing values and then compute a plot identifying the correlation between the Sale Price and each of the key variables that are numerical.

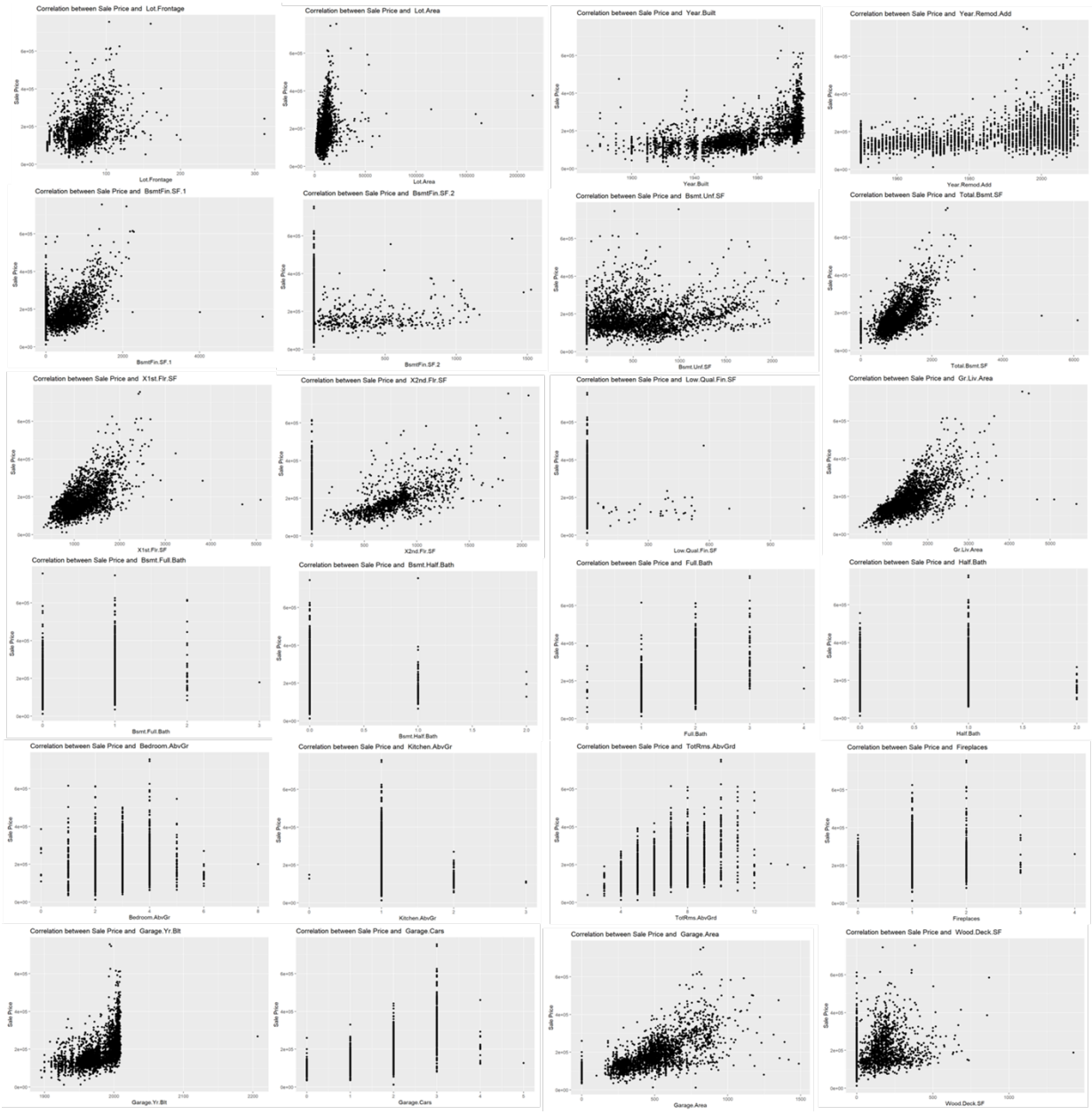


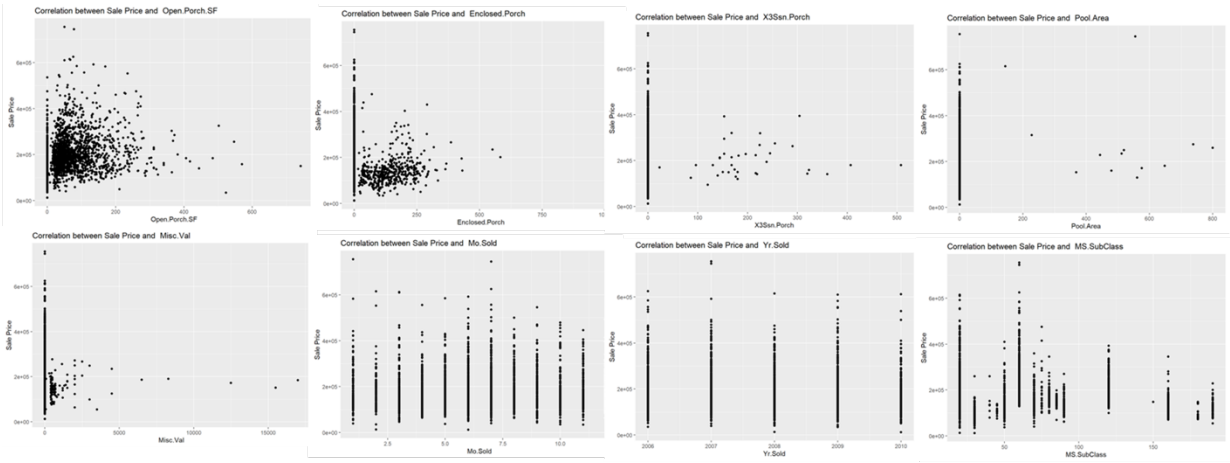






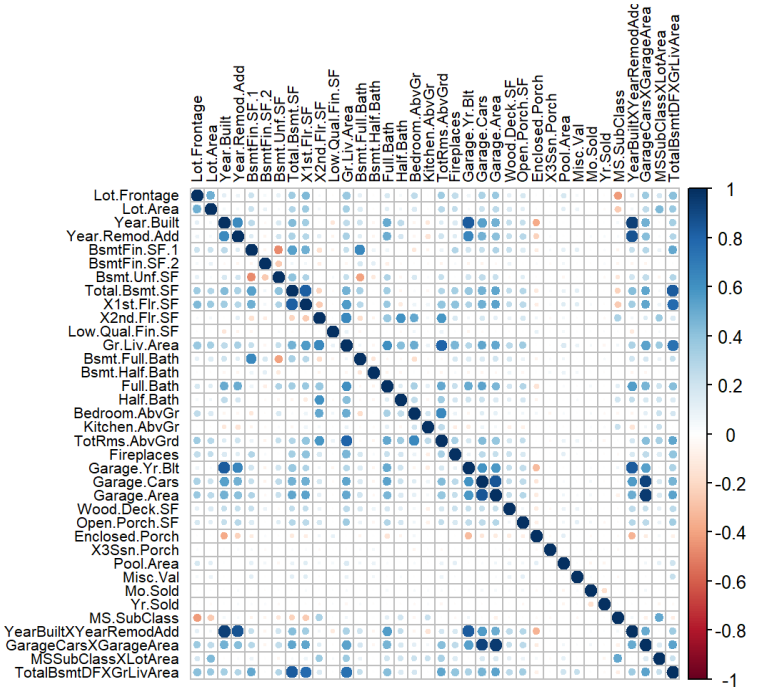
Scatter Plots

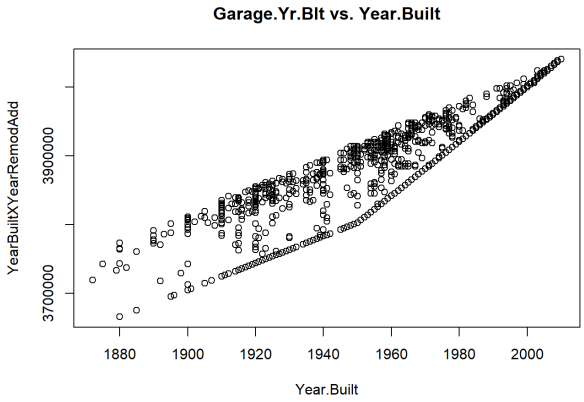


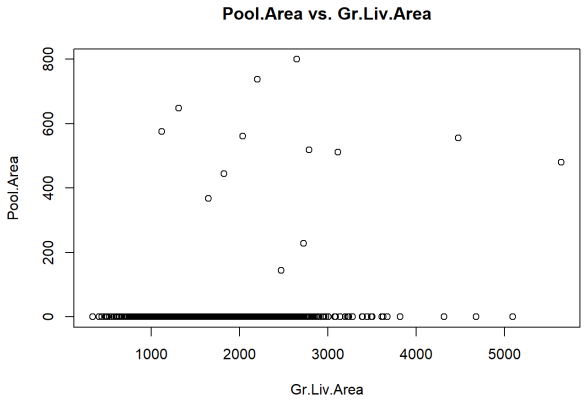


Correlation Analysis

The below heat map shows the correlation for the numerical variables alongside the interaction terms where both of the variables in the interaction terms are numerical. The correlation matrix helps identify which two variables are very correlated and which two variables have weak correlation. For example, it is evident that Year.Built and YearBuiltXRemodelYear have a strong correlation with no outliers, while Pool.Area and Gr.Liv.Area have a weak correlation as shown below.





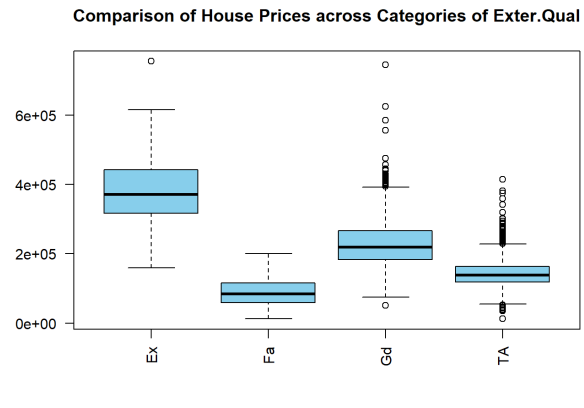


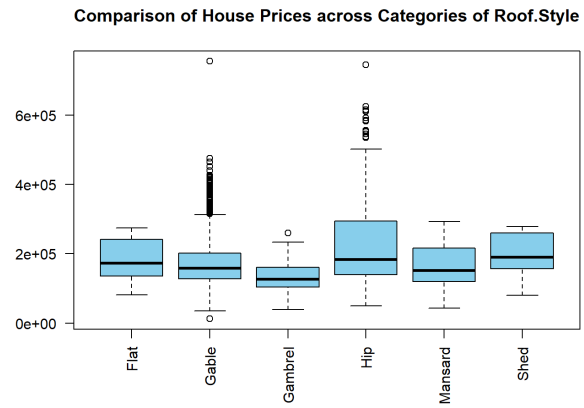
Group Comparison:

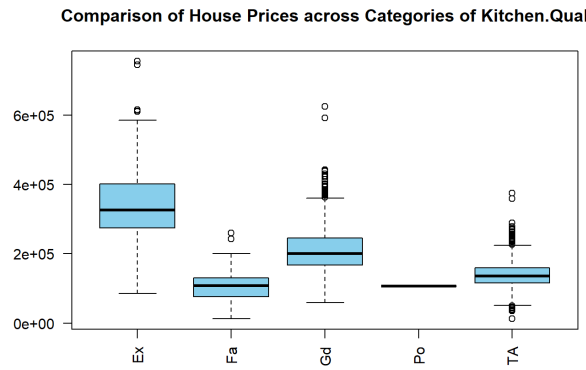
Box plots & Bar plots were created to compare Sale Price across categories.

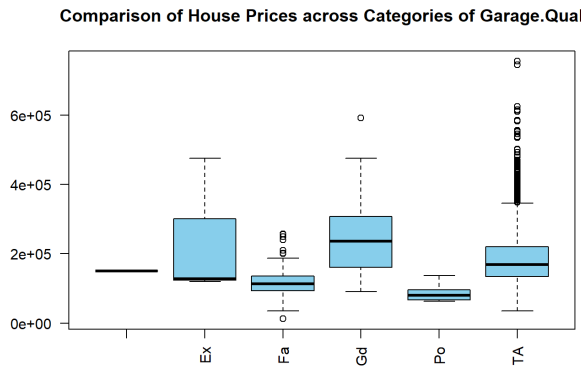
Box Plot Observation: We can visualize the distribution of sale price of houses of a certain category. For example, the sale prices of houses with Exter.Qual of Ex is almost symmetrically distributed although very slightly positively skewed, which could be influenced by an outlier larger than 600000. The sale prices of houses with Exter.Qual of Fa are positively skewed because the lower whisker appears to be of smaller length than the upper whisker, but there are no outliers observed here. The sale prices of houses with Exter.Qual of Gd and the sale prices of houses Exter.Qual of TA are both positively skewed, because of the number of outliers that are very large, even both the lower and upper whisker are of the same lengths. Based on the box plot for sale prices of houses across categories of Exter.Qual, it is very evident that houses with Exter.Qual of Ex have a much higher mean sale price than the houses of any other Exter.Qual.

Note: If one of the boxplots has a missing label, then this boxplot indicates the distribution of sale prices across all houses that have a missing value for this column.

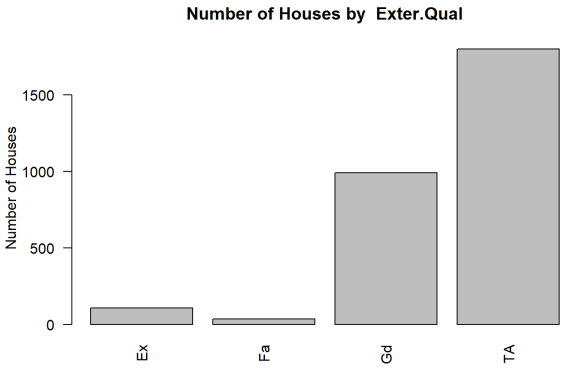


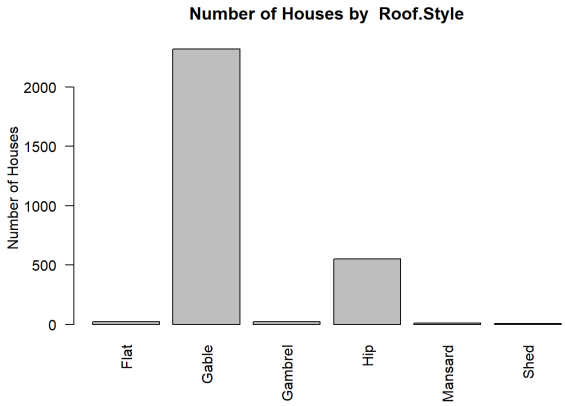


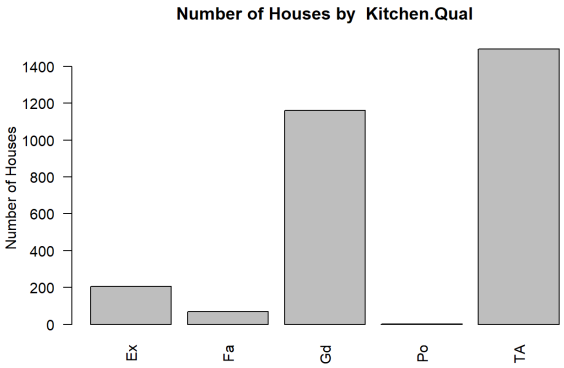


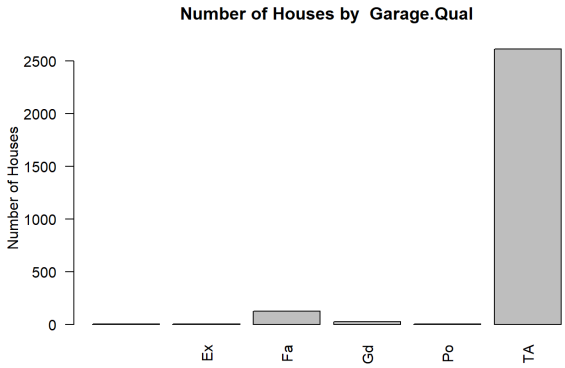


Bar Plots Observation: The number of houses of Exter.Qual of TA is significantly higher at larger than 1500, compared to the number of houses of any other categories. The number of houses of Roof.Style of Gable is significantly higher at larger than 200, compared to the number of houses of any other categories. The number of houses of Kitchen.Qual of TA is significantly higher at larger than 1400, compared to the number of houses of any other categories. The second highest bin is of Gd at approximately 1200 houses. The number of houses of Garage.Qual of TA is significantly larger at higher than 2500, compared to the number of houses of any other categories.



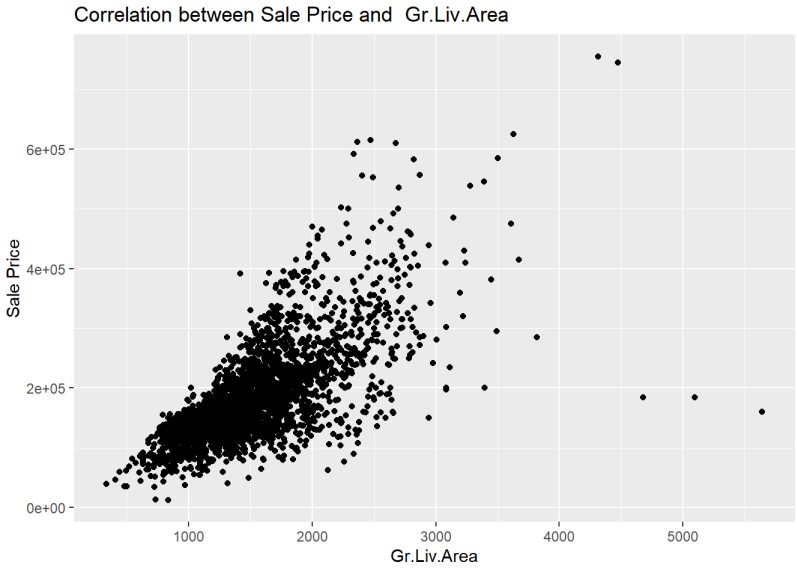




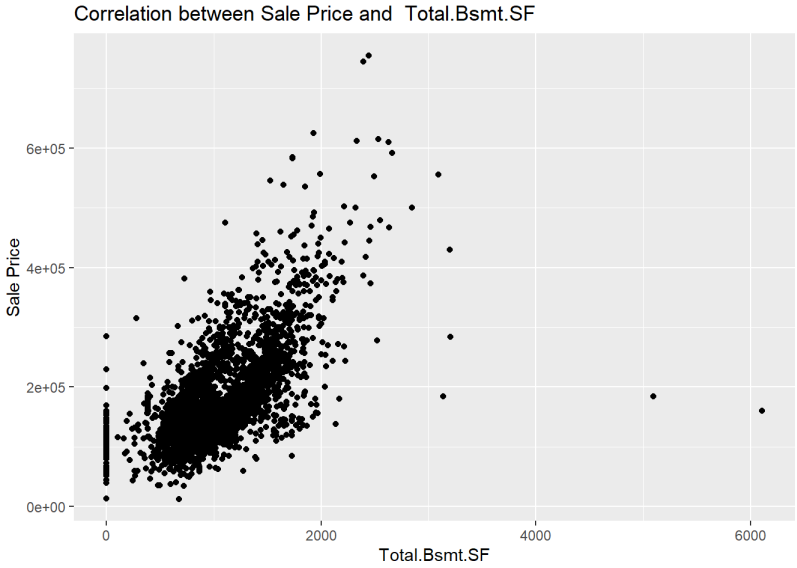


2.2 Patterns

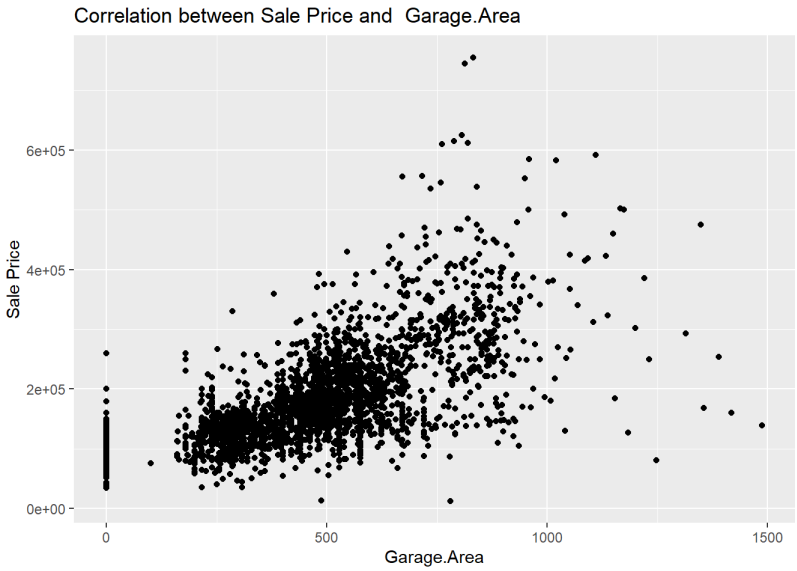
- **Pattern Observed**: A strong, positive correlation between Sale Price and Gr.Liv.Area with some outliers with Gr.Liv.Area above 4000 and sale price between 200000. Very large homes with relatively low sale prices.



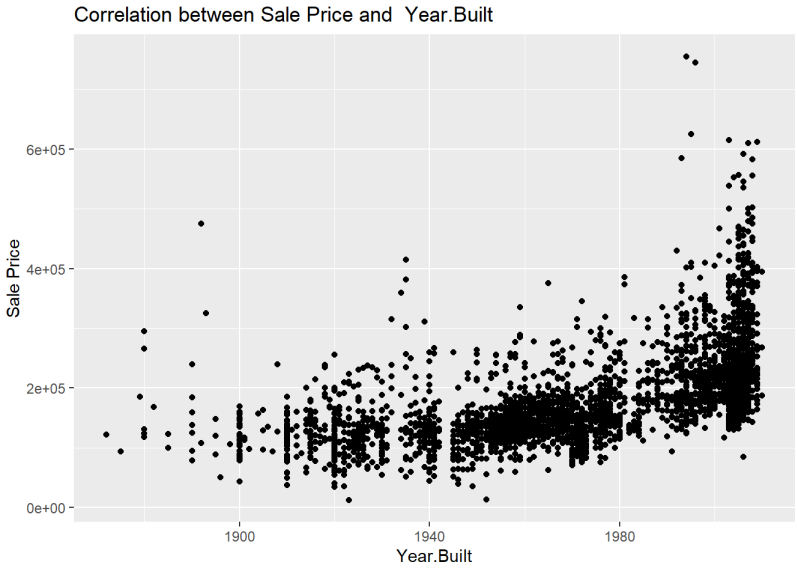
- **Pattern Observed**: A moderate to strong, positive linear correlation between Sale Price and Total.Bsmt.SF with a couple of outliers with Total.Bsmt.SF larger than 4000. Homes with larger basements are often priced higher.



- **Pattern Observed**: There appears to be a moderate to strong, positive linear relationship between sale price and Garage.Area. Homes with large garages that can fit a greater number of cars are often priced higher.



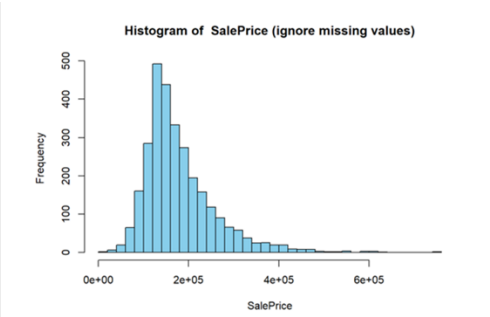
- **Pattern Observed**: There appears to be a moderate, non-linear correlation between Sale Price and Year. Built. Newer homes tend to sell for higher prices.

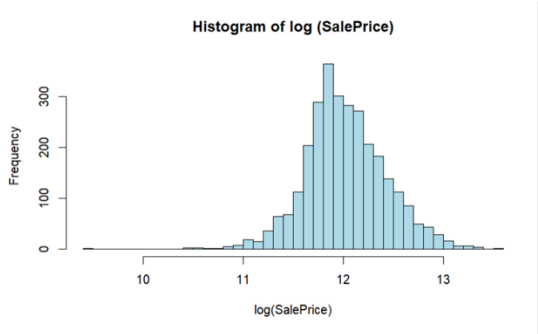


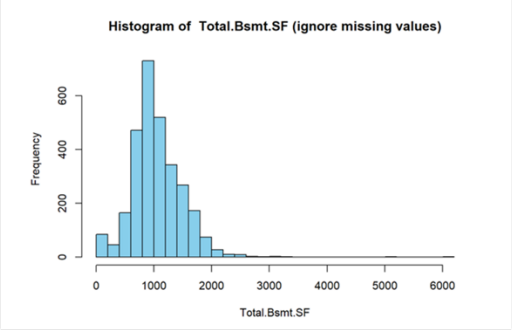
Detailed correlation analysis for all variables used is enclosed in Appendix A (HTML File).

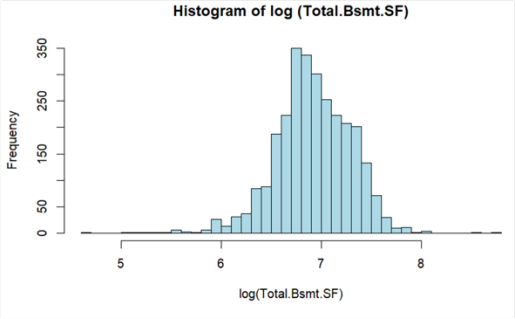
2.3 Feature Engineering:

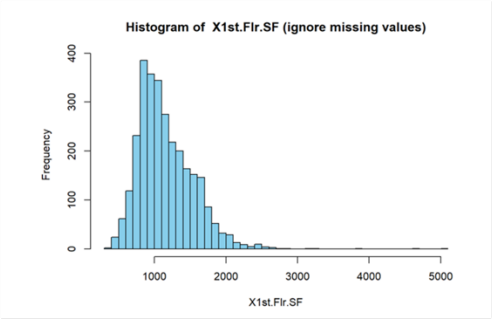
Log, square root and inverse transformation are applied to skewed numerical features to reduce skewness and approximate a normal distribution, making the data more suitable for linear modeling. It compresses large values and spreads out smaller ones, stabilizing variance and improving model performance. For example, we took log for SalePrice as it has a right-skewed distribution. Most houses fall within a lower price range, with fewer high-value properties.

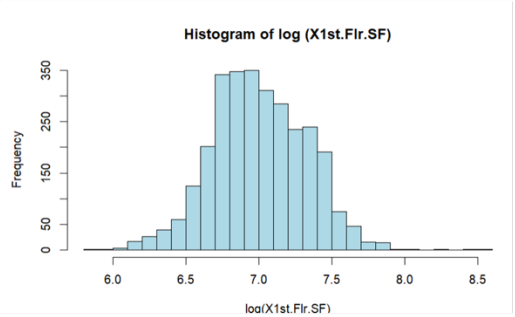

Histogram of SalePrice (ignore missing values)
Frequency
0
100
200
300
400
500
0e+00
2e+05
4e+05
6e+05
SalePrice

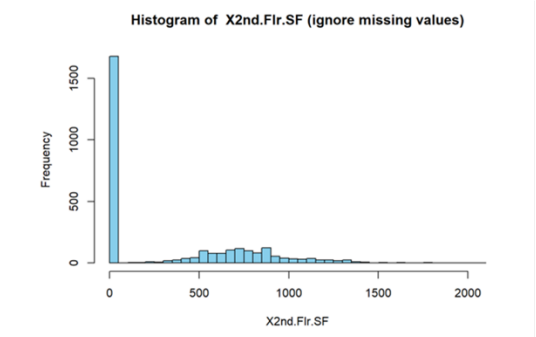

Histogram of log (SalePrice)
Frequency
0
100
200
300
10
11
12
13
log(SalePrice)

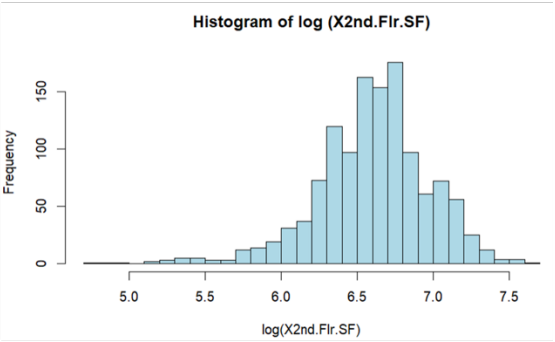

Histogram of Total.Bsmt.SF (ignore missing values)
Frequency
0
200
400
600
0
1000
2000
3000
4000
5000
6000
Total.Bsmt.SF

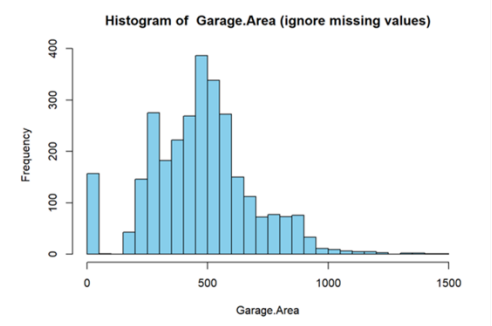

Histogram of log (Total.Bsmt.SF)
Frequency
0
50
150
250
350
5
6
7
8
log(Total.Bsmt.SF)

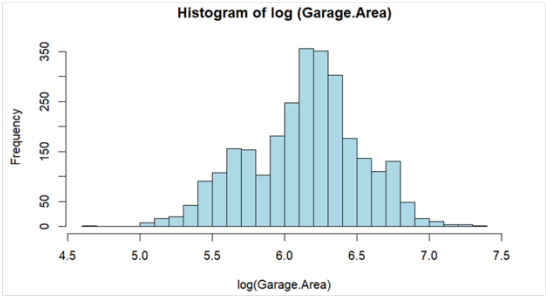

Histogram of X1st.Flr.SF (ignore missing values)
Frequency
0
100
200
300
400
1000
2000
3000
4000
5000
X1st.Flr.SF


Histogram of log (X1st.Flr.SF)
Frequency
0
50
150
250
350
6.0
6.5
7.0
7.5
8.0
8.5


Histogram of X2nd.Flr.SF (ignore missing values)
Frequency
0
500
1000
1500
0
500
1000
1500
2000
X2nd.Flr.SF


Histogram of log (X2nd.Flr.SF)
Frequency
0
50
100
150
5.0
5.5
6.0
6.5
7.0
7.5
log(X2nd.Flr.SF)


Histogram of Garage.Area (ignore missing values)
Frequency
0
100
200
300
400
0
500
1000
1500
Garage.Area


Histogram of log (Garage.Area)
Frequency
0
50
150
250
350
4.5
5.0
5.5
6.0
6.5
7.0
7.5
log(Garage.Area)

2.4 Interaction Terms:

In order to represent correlations between variables that could have a multiplicative effect on housing prices, we created some interaction terms. These new features help the model understand how certain factors work together, which individual variables might not show on their own.

- Encoding Variables:
  Exter.Qual (Exterior Quality) and Kitchen.Qual (Kitchen Quality) were converted to ordinal factors with ordered levels: Poor (Po), Fair (Fa), Typical (TA), Good (Gd), Excellent (Ex).
- Interaction Features Created

1- OverallQual_GrLivArea: Interaction of Overall.Qual (Overall Quality) and Gr.Liv.Area (Above Ground Living Area).
2- ExterQual_OverallQual: Product of Exter.Qual and Overall.Qual.
3- GarageCars_GarageArea: Product of Garage.Cars (Number of Garage Cars) and Garage.Area.
4- TotalBsmtSF_GrLivArea: Product of Total.Bsmt.SF (Total Basement Area) and Gr.Liv.Area.
5- KitchenQual_OverallQual: Product of Kitchen.Qual and Overall.Qual.

3. Question of Interest

3.1 Research Questions

- Q1: Which property features have the most effect on sale price? Compare the effectiveness of the features on sale price.
- Q2: Can a regression model be developed and trained to predict house sale prices? How accurate is it? Which features have the most effect on predicting the price?

3.2 Data Relevance

The dataset is large and captures various features of residential properties. It is comprehensive allowing for the development and training of regression models. In summary, Ames Housing dataset provides a solid framework allowing for effective analyses of feature impact on sale price, comparisons, and building a housing price prediction model.

4. Initial Model and Estimation

4.1 Model Selection: Our initial model was selected based on the nature of the data. With some choices including a linear regression model and a LASSO regression model because of their effectiveness in recognizing relationships between our response variable and the predictors. We chose LASSO- Least Absolute Shrinkage and Selection Operator- as it is essentially useful when there are so many predictors (in our case 83 variables) and some of them might be irrelevant or redundant as it performs variable selection by shrinking less important coefficients to zero. This will lead to having a simpler model.

- **Feature Standardization:** To ensure that all predictors are on the same scale, which is crucial for LASSO regression, numerical predictor variables (excluding SalePrice) were standardized using the scale function.
- **Target Variable and Predictor Matrix**: The target variable SalePrice is the response variable. A sparse model matrix was created for the predictors using sparse.model.matrix. This transformation handles categorical variables effectively by creating dummy variables.
- **Fitting the LASSO model:** As mentioned above, LASSO technique is used here that adds an L1-penalty term to the linear regression objective, shrinking some coefficients to zero. Lambda (λ) is the regularization parameter in LASSO regression that controls the strength of the penalty applied to the coefficients of the predictors.

In this project, cv.glmnet is used to perform cross-validation and identify the best value of the penalty parameter λ. And the final model is fitted the with the optimal λ value (lambda.min). The seed (set.seed(60)) is used to ensure that the cross-validation process in cv.glmnet generates consistent results each time the code is executed. The related codes are in Appendix B.

4.3 Estimation Strategy:

For LASSO regression, cross-validation (CV) is often used to choose the optimal regularization parameter (lambda) that balances model complexity and predictive accuracy. Using R, we try to find the most appropriate lambda by testing the effectiveness of our model across different folds of data (considering the overfitting to be minimized).`

5. Interpretation of Results

5.1 Results Summary:

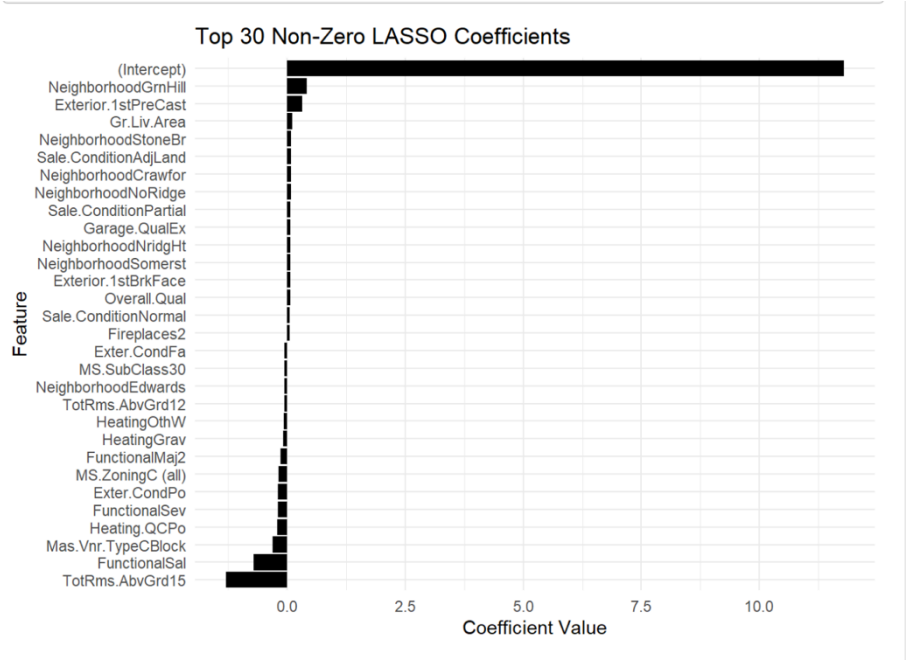
The regression model produced coefficients for various variables, indicating their estimated impact on the dependent variable (e.g., house price or another outcome). Below is a summary of key observations from the model:

- **Key Drivers:** Variables such as Overall.Qual (0.059), Gr.Liv.Area (0.101), and specific neighborhood features like NeighborhoodGrnHill (0.407) have large positive coefficients, suggesting a strong positive influence on the outcome.
- **Negative Impacts:** Some variables, such as MS.ZoningC (all) (-0.182) and TotRms.AbvGrd15 (-1.303), have significantly negative coefficients, indicating a large decrease in the outcome when these conditions apply.
- **Interaction Terms:** Interaction terms like ExterQual_OverallQual (0.017) and GarageCars_GarageArea (0.022) suggest that the combined effects of these variables contribute positively to the outcome.

5.1.1 Visualizing Key Findings:

Scatter Plot: Predicted vs. Actual Sale Prices: The scatter plot compares the predicted sale prices from the model with the actual sale prices in the test dataset. The dashed diagonal line represents the ideal scenario where the predicted prices match the actual prices perfectly. This scatter plot demonstrates that the model achieves a good level of accuracy, particularly for the majority of houses in the dataset.

Bar Chart: Top 30 Non-Zero LASSO Coefficients: This bar chart below visualizes the top 30 non-zero coefficients identified by the LASSO regression model, highlighting the most influential features for predicting house prices. The features are ranked by the magnitude of their coefficients, with positive and negative values included.



Based on the bar chart, we can identify the most influential features affecting house prices, with neighborhood, living area, and overall quality being strong positive contributors, while functional issues, zoning restrictions, and poor material choices negatively impact prices.

- Importance of Location: Neighborhood variables dominate the positive coefficients, reaffirming that location is a critical determinant of house prices.
- Quality and Size: Features like Gr.Liv.Area, Overall.Qual, and Garage.QualEx highlight the importance of structural quality and size in influencing house prices positively.
- Negative Influences: Functional issues (FunctionalSal), undesirable zoning (MS.ZoningC (all)), and poor material choices (Mas.Vnr.TypeCBlock) are major detractors.

5.2 Coefficient Interpretation:

The coefficients from the LASSO model were analyzed, and predictors with non-zero coefficients were retained, indicating their contribution to predicting Sale Price. For non-zero coefficients the output showed variables that had meaningful effects on house prices after regularization. Here is the coefficients table:

## (Intercept)	1.178899e+01
## MS.SubClass30	-5.866333e-02
## MS.SubClass85	4.300302e-04
## MS.SubClass90	-4.766695e-03
## MS.SubClass150	-3.589651e-02
## MS.SubClass160	-2.500405e-02
## MS.ZoningC (all)	-1.825075e-01
## MS.ZoningI (all)	-4.062225e-02
## MS.ZoningRM	-3.593389e-02
## Lot.Frontage	-4.312894e-04
## Lot.Area	3.434025e-02
## Lot.ShapeIR2	2.113413e-04
## Lot.ShapeIR3	-4.279309e-02
## Land.ContourHLS	1.653315e-02
## Land.ContourLvl	7.227761e-03
## Lot.ConfigCulDSac	1.462564e-02
## Lot.ConfigFR2	-1.996120e-02
## Lot.ConfigFR3	-1.754571e-03
## NeighborhoodBlueste	1.908023e-02
## NeighborhoodClearCr	2.428300e-02
## NeighborhoodCrawfor	7.886679e-02
## NeighborhoodEdwards	-6.014320e-02
## NeighborhoodGreens	9.427975e-03

## NeighborhoodGrnHill	4.074673e-01
## NeighborhoodIDOTRR	-1.254556e-02
## NeighborhoodMeadowV	-4.850103e-02
## NeighborhoodNAmes	-1.073236e-02
## NeighborhoodNoRidge	7.850303e-02
## NeighborhoodNridgHt	6.648413e-02
## NeighborhoodNWAmes	-6.863186e-04
## NeighborhoodOldTown	-4.510187e-02
## NeighborhoodSawyerW	-2.969498e-03
## NeighborhoodSomerst	6.432410e-02
## NeighborhoodStoneBr	8.551460e-02
## Condition.1Feedr	-3.439818e-04
## Condition.1Norm	4.385999e-02
## Condition.1PosA	2.209933e-02
## Condition.1PosN	1.716763e-02
## Bldg.TypeDuplex	-1.467922e-05
## House.Style2.5Fin	-1.324881e-02
## Overall.Qual	5.948254e-02
## Overall.Cond	3.624283e-02
## Year.Built1900-1909	-1.350460e-02
## Year.Built1910-1919	-3.626114e-03
## Year.Built1920-1929	-1.826361e-03
## Year.Built1980-1989	2.177033e-03
## Year.Built1990-1999	2.715473e-02
## Year.Built2000-2009	4.411168e-02
## Year.Remod.Add	1.621081e-02
## Roof.StyleMansard	-2.595831e-02
## Exterior.1stBrkFace	5.991743e-02
## Exterior.1stPreCast	3.229203e-01
## Exterior.1stStucco	-9.854947e-03
## Exterior.1stVinylSd	1.767186e-03
## Exterior.1stWd Sdng	-2.447656e-03
## Exterior.2ndMetalSd	4.106744e-03
## Exterior.2ndPreCast	1.514754e-05
## Mas.Vnr.TypeBrkCmn	-1.449729e-02
## Mas.Vnr.TypeCBlock	-3.106270e-01
## Mas.Vnr.TypeStone	1.036027e-04
## Mas.Vnr.Area	2.252472e-03
## Exter.CondFa	-5.746020e-02
## Exter.CondPo	-1.917868e-01

## Exter.CondTA	5.174724e-03
## FoundationPConc	1.401091e-02
## Bsmt.QualEx	4.712433e-02
## Bsmt.QualFa	-4.106697e-02
## Bsmt.QualTA	-1.111661e-02
## Bsmt.CondFa	-1.062534e-02
## Bsmt.CondGd	5.776084e-03
## Bsmt.ExposureAv	9.326060e-03
## Bsmt.ExposureGd	4.991275e-02
## Bsmt.ExposureNo	-7.696424e-03
## BsmtFin.Type.1ALQ	6.191595e-04
## BsmtFin.Type.1LwQ	-8.294287e-03
## BsmtFin.Type.1Rec	-8.029616e-03
## BsmtFin.SF.1	2.549183e-02
## BsmtFin.Type.2BLQ	-4.872666e-03
## Bsmt.Unf.SF	-1.958135e-03
## Total.Bsmt.SF	1.029045e-02
## HeatingGasW	2.791982e-02
## HeatingGrav	-8.805498e-02
## HeatingOthW	-7.666544e-02
## HeatingWall	-3.524615e-02
## Heating.QCFa	-1.074643e-02
## Heating.QCPo	-2.110457e-01
## Heating.QCTA	-1.836140e-02
## Central.AirY	5.342710e-02
## X1st.Flr.SF	3.505098e-02
## Low.Qual.Fin.SF	-2.995281e-04
## Gr.Liv.Area	1.015498e-01
## Bsmt.Full.Bath	1.291607e-02
## Full.Bath	1.232642e-02
## Half.Bath	6.477307e-03
## Bedroom.AbvGr3	-9.010757e-04
## Bedroom.AbvGr4	1.778483e-03
## Kitchen.AbvGr	-1.083912e-02
## Kitchen.Qual.Q	3.525992e-02
## TotRms.AbvGrd3	-2.163027e-02
## TotRms.AbvGrd5	-3.585734e-03
## TotRms.AbvGrd8	1.810844e-03
## TotRms.AbvGrd11	-2.007268e-02
## TotRms.AbvGrd12	-6.199793e-02

## TotRms.AbvGrd15	-1.302572e+00
## FunctionalMaj2	-1.393805e-01
## FunctionalSal	-7.119765e-01
## FunctionalSev	-1.974174e-01
## FunctionalTyp	4.450767e-02
## Fireplaces1	2.127557e-02
## Fireplaces2	5.587869e-02
## Fireplaces3	-4.346847e-02
## Garage.TypeAttchd	4.998173e-03
## Garage.TypeBasment	-2.850776e-02
## Garage.TypeCarPort	-4.597701e-02
## Garage.Yr.Blt	1.045546e-02
## Garage.FinishFin	6.852404e-03
## Garage.FinishUnf	-1.921317e-03
## Garage.Cars	9.583264e-03
## Garage.Area	1.891536e-03
## Garage.QualEx	6.982361e-02
## Garage.QualGd	2.564456e-02
## Garage.CondFa	-3.112095e-02
## Paved.DriveY	3.421622e-02
## Wood.Deck.SF	4.916825e-03
## Open.Porch.SF	1.171950e-03
## Screen.Porch	9.826243e-03
## Yr.Sold2009	-3.009089e-03
## Sale.TypeCon	4.989288e-02
## Sale.TypeConLI	-1.944501e-02
## Sale.TypeNew	1.135904e-02
## Sale.ConditionAdjLand	8.089959e-02
## Sale.ConditionAlloca	1.435208e-02
## Sale.ConditionNormal	5.700950e-02
## Sale.ConditionPartial	7.196661e-02
## ExterQual_OverallQual	1.674753e-02
## GarageCars_GarageArea	2.243987e-02
## TotalBsmtSF_GrLivArea	8.294259e-03
## KitchenQual_OverallQual	1.890444e-02

- **Positive Coefficients:**
  - Overall.Qual (0.059): For each unit increase in overall quality, the dependent variable increases by 0.059 units, holding other factors constant. This aligns with expectations since higher quality typically increases value.

- Gr.Liv.Area (0.101): A 1-unit increase in ground living area corresponds to a 0.101 increase in the dependent variable, indicating that larger living spaces are valued higher.
- NeighborhoodGrnHill (0.407): Being in the "GrnHill" neighborhood increases the dependent variable by 0.407 compared to the baseline neighborhood. This suggests "GrnHill" is a highly desirable location.

- **Negative Coefficients:**

- MS.ZoningC (all) (-0.182): Properties in zoning classification "C" are associated with a 0.182 decrease in the dependent variable compared to the baseline zone. This could indicate lower property value in this zoning classification.
- TotRms.AbvGrd15 (-1.303): Having 15 rooms above ground significantly reduces the dependent variable by 1.303 units, which may indicate inefficiency or lack of demand for homes with excessive rooms.

- **Unexpected Results:**

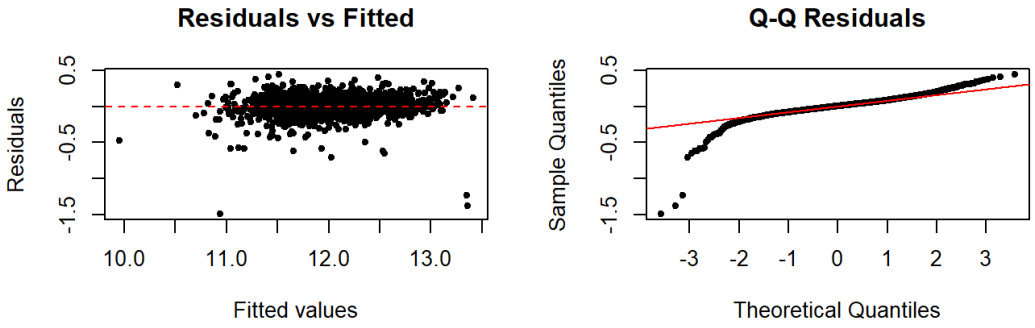
- Fireplaces3 (-0.043): While more fireplaces are typically expected to increase value, the negative coefficient for 3 fireplaces might suggest diminishing returns or that such homes are less common and less desirable.
- Lot.Frontage (-0.0004): A negative coefficient for lot frontage is surprising, as larger lot frontage is often considered desirable. This might be due to confounding effects from other variables or multicollinearity.

- **Interaction Terms:**

- ExterQual_OverallQual (0.017): The positive coefficient suggests that external quality and overall quality have a synergistic effect, amplifying each other's contribution to the dependent variable.
- GarageCars_GarageArea (0.022): Larger garages with more car capacity add value, but the coefficient indicates a modest impact relative to other variables.

5.3 Model Validation:

- **Evaluation of Model Performance:** After fitting the model, we evaluated the model's performance by using MSE metric.
  - **MSE**: The Mean Squared Error (MSE) measures the average squared difference between predicted and actual values. The MSE of 517,291,113.46 corresponds to an RMSE of approximately $22,745, which is a reasonable error given the wide range of house prices and the complexity of the dataset. The result demonstrates good predictive performance compared to a baseline model and reflects a balance between accuracy and interpretability offered by LASSO regression.
  - **Residual Diagnostics:** To better understand the MSE and evaluate the model's assumptions, we analyzed the residuals, which represent the differences between the actual and predicted values.



Residual diagnostics help identify patterns, assess model performance, and highlight potential areas for improvement.

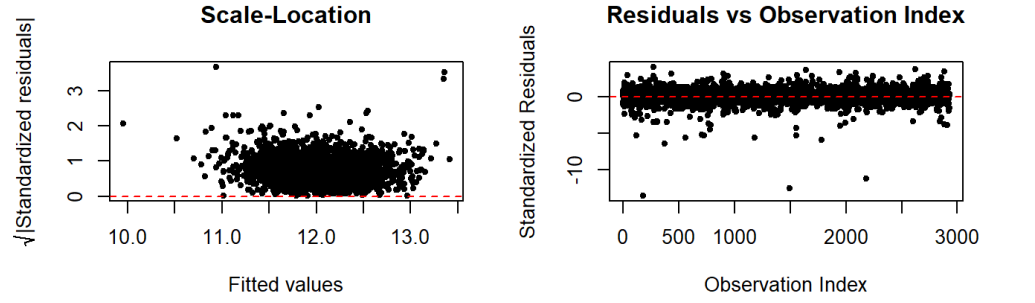
1- Residuals vs. Fitted Values: This plot checks for patterns in the residuals.

Ideally, residuals should be randomly scattered around the horizontal line at zero, indicating no systematic errors. In the left below plot, the residuals appear centered around zero, which indicates that the model captures the linear relationship between the predictors and response reasonably well.

2. Q-Q Plot of Residuals: The Q-Q plot checks whether residuals are normally distributed. Residuals should align closely with the diagonal line. In the Q-Q plot (the right below) most points follow the line, but deviations in the tails suggest some non-normality in the residual distribution.

3. Scale-Location Plot (Spread-Location Plot): This plot checks the homoscedasticity of the residuals. The points should be spread equally along the range of fitted values without a systematic pattern. The left below plot shows some variability, which might indicate slight heteroscedasticity.

4. Residuals vs. Observation Index: This plot identifies trends or outliers in residuals through observations. Standardized residuals should be randomly scattered within a range of [−3,3]. The right below plot shows a fairly random scatter of residuals, suggesting that there is no clear pattern, so the assumption of independence seems to be met.



- **Prediction Prices:**

  1. Predicting Log-Transformed Prices: The LASSO model generated predictions for the log-transformed SalePrice variable (predicted_log_prices). These predictions represent the natural logarithm of house prices, as the target variable was log-transformed during preprocessing to stabilize variance and reduce skewness.

  2. Back-Transformation to Original Scale: To interpret the predicted values in their original scale (house prices), the log-transformed predictions were exponentiated:

  $$\text{Predicted\_Prices} = \exp(\text{Predicted\_Log\_Prices})$$

  This back-transformation converts the logarithmic predictions into actual house prices, making them directly comparable to the original SalePrice values.

  The table below showcases a selection of predicted prices compared to the actual prices.

  To ensure fairness and provide a clear comparison, the table below presents the first 5 rows of prediction results alongside their corresponding actual values.

Observation	Actual Price ($)	Predicted Price ($)
1	215000	214121.62
2	105000	118512.33
3	172000	157313.12
4	244000	264222.42
5	189900	179918.87

6. Limitations and Further Research:

6.1 Model Limitations:

- **LASSO assumes that the observations are independent**. Any dependence structures, such as spatial correlation between houses, could affect the results. Incorporating domain knowledge about location-based dependencies or neighborhood effects is important.
- **LASSO can be sensitive to outliers**, which can possibly influence the selection of variables and coefficients. Robust preprocessing, including outlier removal or transformation, is essential to mitigate this limitation
- **LASSO assumes a linear relationship between the predictors and the response**. Any non-linear relationships present in the data may not be captured well unless non-linear features (e.g., polynomial terms or interactions) are explicitly included in the model.
- **Interpretation can be challenging** because LASSO reduces the number of features by setting some coefficients to zero. While this simplifies the model, it can make it hard to explain why certain variables were selected or ignored, especially in complex relationships.

6.2 Future Research:

- **Data Quality and Preprocessing:**
  - ✓ Investigate the presence of outliers and influential points. Remove or transform them if appropriate.
  - ✓ Check for multicollinearity among predictors and use techniques like PCA or regularization if necessary.
- **Model Refinement:**
  - ✓ Explore transformations of the dependent variable (e.g., log-transformation) to reduce heteroscedasticity and improve normality.
  - ✓ Incorporate polynomial terms or interaction effects to capture potential non-linear relationships.
- **Alternative Models:**
  - ✓ Classification Models: Use logistic regression or multinomial classification to predict price ranges instead of exact prices.
  - ✓ Nonlinear Models: Explore tree-based models (e.g., Random Forests, Gradient Boosting) or neural networks to capture complex patterns.
- **Additional Data:** If feasible, collect additional predictors that could explain the variance in the dependent variable more effectively.

7. Division of Labour

Task	Team member(s)
Proposal & Final Report Compilation	Primary: Lily & Maryam
Data Collection	Primary: All
Data Cleaning & Preprocessing	Primary: Majid & Maryam
Descriptive Data Analysis: Visualization & patterns	Primary: Kurt, Nima & Lily
Correlation Analysis & Comparison	Primary: Kurt & Nima
Development of Model & Interpretation of Results	Primary: Majid & Nima

References

[1] De Cock, Dean. "🏠ames House Price Prediction Regression 🌎🚗🚂." *Kaggle*, 20 July 2022, www.kaggle.com/datasets/nabilabdul/ames-house-price-prediction-regression.

[2] De Cock,Dean. "🏠ames House Price Prediction Regression 🌎🚗🚂." Kaggle, 20 July 2022, www.kaggle.com/datasets/nabilabdul/ames-house-price-prediction-regression?select=data_description.txt.

[3] City Assessor | City of Ames, IA, www.cityofames.org/government/departments-divisions-a-h/city-assessor. Accessed 16 Nov. 2024.

[4] "Acceptable Use Policy." Kaggle, www.kaggle.com/aup. Accessed 16 Nov. 2024.