UNIVERSITÀ DEGLI STUDI DI PADOVA

Department of Mathematics

Master's Degree in Data Science

House Price Prediction: Advanced Regression Techniques

Onur Alp Güvercin, 2072249

Su Doğan, 2071957

Işıkay Karakuş, 2071938

Academic Year 2022/2023

House Price Prediction Report

Statistical Learning Final Exam Project

1. Aim of the Project

House pricing is of utmost importance in our world, as it impacts individuals, families, and the economy at large. The purchase of a house represents a significant financial decision, often constituting a substantial portion of one's wealth. The fluctuation of house prices directly influences personal financial stability and the ability to build assets. Moreover, the real estate market, driven by house prices, contributes significantly to economic activity, including property sales, construction, and associated industries.

The aim of our project is to develop a model to predict house pricing. This project holds great importance considering the large number of people buying houses every year. Housing affordability is a critical issue, and accurate prediction models can help address it. By analyzing market trends and evaluating property values, our model aims to provide individuals, sellers, and industry professionals with valuable insights for making informed decisions. The project contributes to understanding housing market dynamics, assisting with financial planning, and promoting fair and transparent practices in the real estate sector.

Through our research project, our primary objective is to illuminate the key factors that influence house prices, discern the most impactful among them, and formulate an effective estimation framework for predicting the price of a house based on its specific characteristics.

2. Dataset Description

For this project, we have selected the "House Prices - Advanced Regression Techniques" dataset, which consists of 3 csv files (train, test, and sample submission) and can be found on the Kaggle platform at the following link: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

Our dataset provides a comprehensive collection of housing-related information, including various attributes and characteristics of houses. This dataset encompasses a wide range of factors that can potentially influence house prices, such as location, size, number of bedrooms, year that it was built, garage quality,, and other relevant variables. The dataset offers a rich source of information that enables us to analyze and explore the relationships between these factors and house prices, ultimately facilitating a deeper understanding of the housing market dynamics.

2.1.Data Components

In the train dataset, there are 1401 entries, while the test and sample datasets consist of 1459 entries each. Among these, the train dataset contains 434 null values, the test dataset contains 1481 null values, and the sample dataset contains no null values. When we examined the quantities of null values, we recognized that there might be a specific reason behind their absence. Consequently, we made the decision not to remove all the rows containing null values in order to preserve potentially significant information within the dataset.

The columns of our dataset are given below with their explanation:

- Bedroom: Bedrooms above grade (does NOT include basement bedrooms). This column contains numerical data indicating the number of bedrooms above ground level (excluding basement).
- Kitchen: Kitchens above grade. This column contains numerical data indicating the number of kitchens above ground level.
- KitchenQual: Kitchen quality. This column contains categorical data representing the quality rating of the kitchen, such as "Ex" for excellent or "Gd" for good.
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms). This column contains numerical data indicating the total number of rooms (excluding bathrooms) above ground level.
- Functional: Home functionality (Assume typical unless deductions are warranted). This column contains categorical data indicating the level of functionality of the house, such as "Typ" for typical or "Mod" for moderate.
- Fireplaces: Number of fireplaces. This column contains numerical data indicating the number of fireplaces in the house.
- FireplaceQu: Fireplace quality. This column contains categorical data representing the quality rating of the fireplaces, such as "Ex" for excellent or "Gd" for good.
- GarageType: Garage location. This column contains categorical data indicating the location of the garage, such as "Attached" or "Detached".
- GarageYrBlt: Year garage was built. This column contains numerical data indicating the year the garage was built.
- GarageFinish: Interior finish of the garage. This column contains categorical data representing the interior finish of the garage, such as "Unf" for unfinished or "Fin" for finished.
- GarageCars: Size of garage in car capacity. This column contains numerical data indicating the capacity of the garage in terms of the number of cars it can hold.
- GarageArea: Size of garage in square feet. This column contains numerical data indicating the area of the garage in square feet.
- GarageQual: Garage quality. This column contains categorical data representing the quality rating of the garage, such as "Ex" for excellent or "Gd" for good.
- GarageCond: Garage condition. This column contains categorical data representing the condition rating of the garage, such as "Ex" for excellent or "Gd" for good.
- PavedDrive: Paved driveway. This column contains categorical data indicating the type of driveway, such as "Y" for paved or "N" for not paved.
- WoodDeckSF: Wood deck area in square feet. This column contains numerical data indicating the area of the wood deck in square feet.
- OpenPorchSF: Open porch area in square feet. This column contains numerical data indicating the area of the open porch in square feet.
- EnclosedPorch: Enclosed porch area in square feet. This column contains numerical data indicating the area of the enclosed porch in square feet.
- 3SsnPorch: Three-season porch area in square feet. This column contains numerical data indicating the area of the three-season porch in square feet.
- ScreenPorch: Screen porch area in square feet. This column contains numerical data indicating the area of the screen porch in square feet.
- PoolArea: Pool area in square feet. This column contains numerical data indicating the area of the pool in square feet.
- PoolQC: Pool quality. This column contains categorical data representing the quality rating of the pool, such as "Ex" for excellent or "Gd" for good.
- Fence: Fence quality. This column contains categorical data representing the quality rating of the fence, such as "GdPrv" for good privacy or "MnWw" for minimum wood/wire.

- MiscFeature: Miscellaneous feature not covered in other categories. This column contains categorical data indicating miscellaneous features of the property, such as "Shed" or "TenC" (tennis court).
- MiscVal: $Value of miscellaneous feature. This column contains numerical data indicating the value of the miscellaneous feature in dollars.
- MoSold: Month Sold (MM). This column contains numerical data indicating the month the house was sold.
- YrSold: Year Sold (YYYY). This column contains numerical data indicating the year the house was sold.
- SaleType: Type of sale. This column contains categorical data indicating the type of sale, such as "WD" for warranty deed or "New" for new home.
- SaleCondition: Condition of sale. This column contains categorical data indicating the condition of the sale, such as "Normal" or "Abnorml" (abnormal).

With the summary statistics, we obtained valuable insights into the dataset.

LotFrontage: The 'LotFrontage' column represents the linear feet of the street connected to the property. The minimum value is 21.00, indicating a property with a relatively small street frontage. The maximum value is 200.00, suggesting a property with a large street frontage. The mean value is 68.65, and the median (50th percentile) is 68.00, indicating that the distribution is approximately symmetric. There are 7 missing values (NA's) in this column.

LotArea: The 'LotArea' column denotes the lot size in square feet. The minimum value is 7.171, indicating a small lot size. The maximum value is 10.944, representing a large lot size. The mean value is 9.070, and the median is 9.144, suggesting a roughly symmetric distribution. There are no missing values in this column.

1stFlrSF: The '1stFlrSF' column represents the square footage of the first floor. The minimum value is 5.814, indicating a property with a small first-floor area. The maximum value is 8.536, representing a property with a large first-floor area. The mean value is 6.998, and the median is 6.984, suggesting a roughly symmetric distribution. There are 7 missing values (NA's) in this column.

2ndFlrSF: The '2ndFlrSF' column indicates the square footage of the second floor. The minimum value is 0.000, indicating properties without a second floor. The maximum value is 7.530, representing properties with a relatively large second-floor area. The mean value is 2.814, and the median is 0.000, suggesting a distribution skewed towards properties without a second floor. There are 7 missing values (NA's) in this column.

LowQualFinSF: The 'LowQualFinSF' column represents the low-quality finished square footage across all floors. The minimum value is 0.000, indicating properties without low-quality finished areas. The maximum value is 2.076, representing properties with a relatively large low-quality finished area. The mean value is 0.025, and the median is 0.000, suggesting a distribution skewed towards properties without low-quality finished areas. There are 7 missing values (NA's) in this column.

GrLivArea: The 'GrLivArea' column denotes the above-grade (ground) living area square footage. The minimum value is 5.814, indicating properties with a small above-grade living area. The maximum value is 8.536, representing properties with a large above-grade living area. The mean value is 7.255, and the median is 7.270, suggesting a roughly symmetric distribution. There are 7 missing values (NA's) in this column.

BsmtFinSF1: The 'BsmtFinSF1' column represents the type 1 finished square feet in the basement. The minimum value is 0.000, indicating properties without a type 1 finished basement area. The maximum value is

8.297, representing properties with a relatively large type 1 finished basement area. The mean value is 4.208, and the median is 5.893, suggesting a distribution skewed towards properties with lower type 1 finished basement areas. There are 8 missing values (NA's) in this column.

BsmtFinSF2: The 'BsmtFinSF2' column indicates the type 2 finished square feet in the basement. The minimum value is 0.000, indicating properties without a type 2 finished basement area. The maximum value is 2.120, representing properties with a relatively large type 2 finished basement area. The mean value is 0.2248, and the median is 0.000, suggesting a distribution skewed towards properties without type 2 finished basement areas. There are 8 missing values (NA's) in this column.

BsmtUnfSF: The 'BsmtUnfSF' column represents the unfinished square feet of the basement area. The minimum value is 0.000, indicating properties without an unfinished basement area. The maximum value is 7.757, representing properties with a relatively large unfinished basement area. The mean value is 5.628, and the median is 6.154, suggesting a distribution skewed towards properties with lower unfinished basement areas. There are 8 missing values (NA's) in this column.

TotalBsmtSF: The 'TotalBsmtSF' column indicates the total square feet of the basement area. The minimum value is 0.000, indicating properties without a basement. The maximum value is 5.095, representing properties with a relatively large total basement area. The mean value is 4.208, and the median is 5.893, suggesting a distribution skewed towards properties with lower total basement areas. There are 8 missing values (NA's) in this

X1stFlrSF: The 'X1stFlrSF' column represents the square footage of the first floor. The minimum value is 5.814, indicating a property with a small first-floor area. The maximum value is 8.536, representing a property with a large first-floor area. The mean value is 6.998, and the median is 6.984, suggesting a roughly symmetric distribution. There are 7 missing values (NA's) in this column.

X2ndFlrSF: The 'X2ndFlrSF' column indicates the square footage of the second floor. The minimum value is 0.000, indicating properties without a second floor. The maximum value is 7.530, representing properties with a relatively large second-floor area. The mean value is 2.814, and the median is 0.000, suggesting a distribution skewed towards properties without a second floor. There are 7 missing values (NA's) in this column.

LowQualFinSF: The 'LowQualFinSF' column represents the low-quality finished square footage across all floors. The minimum value is 0.000, indicating properties without low-quality finished areas. The maximum value is 2.076, representing properties with a relatively large low-quality finished area. The mean value is 0.025, and the median is 0.000, suggesting a distribution skewed towards properties without low-quality finished areas. There are 7 missing values (NA's) in this column.

GrLivArea: The 'GrLivArea' column denotes the above-grade (ground) living area square footage. The minimum value is 5.814, indicating properties with a small above-grade living area. The maximum value is 8.536, representing properties with a large above-grade living area. The mean value is 7.255, and the median is 7.270, suggesting a roughly symmetric distribution. There are 7 missing values (NA's) in this column.

BsmtFullBath: The 'BsmtFullBath' column represents the number of full bathrooms in the basement. The minimum value is 0.000, indicating properties without a full bathroom in the basement. The maximum value is 3.000, representing properties with a relatively high number of full bathrooms in the basement. The mean value is 0.4241, and the median is 0.000, suggesting a distribution skewed towards properties without a full bathroom in the basement. There are 7 missing values (NA's) in this column.

BsmtHalfBath: The 'BsmtHalfBath' column indicates the number of half bathrooms in the basement. The minimum value is 0.000, indicating properties without a half bathroom in the basement. The maximum value is 0.7413, representing properties with a relatively high number of half bathrooms in the basement. The mean value is 0.0315, and the median is 0.000, suggesting a distribution skewed towards properties without a half bathroom in the basement. There are 7 missing values (NA's) in this column.

FullBath: The 'FullBath' column represents the number of full bathrooms above grade. The minimum value is 0.000, indicating properties without a full bathroom above grade. The maximum value is 4.000, representing properties with a relatively high number of full bathrooms above grade. The mean value is 1.563, and the median is 2.000, suggesting a distribution skewed towards properties with fewer full bathrooms above grade. There are 7 missing values (NA's) in this column.

HalfBath: The 'HalfBath' column indicates the number of half bathrooms above grade. The minimum value is 0.000, indicating properties without a half bathroom above grade. The maximum value is 2.000, representing properties with a relatively high number of half bathrooms above grade. The mean value is 0.3778, and the median is 0.000, suggesting a distribution skewed towards properties with fewer half bathrooms above grade. There are 7 missing values (NA's) in this column.

BedroomAbvGr: The 'BedroomAbvGr' column represents the number of bedrooms above grade. The minimum value is 0.000, indicating properties without bedrooms above grade. The maximum value is 8.000, representing properties with a relatively high number of bedrooms above grade. The mean value is 2.856, and the median is 3.000, suggesting a distribution slightly skewed towards properties with fewer bedrooms above grade. There are 7 missing values (NA's) in this column.

KitchenAbvGr: The 'KitchenAbvGr' column indicates the number of kitchens above grade. The minimum value is 0.000, indicating properties without kitchens above grade. The maximum value is 0.8697, representing properties with a relatively high number of kitchens above grade. The mean value is 0.5358, and the median is 0.5266, suggesting a distribution skewed towards properties with fewer kitchens above grade. There are 7 missing values (NA's) in this column.

TotRmsAbvGrd: The 'TotRmsAbvGrd' column represents the total number of rooms above grade (excluding bathrooms). The minimum value is 1.099, indicating properties with a small number of rooms above grade. The maximum value is 2.773, representing properties with a relatively high number of rooms above grade. The mean value is 1.984, and the median is 1.946, suggesting a roughly symmetric distribution. There are 7 missing values (NA's) in this column.

GarageYrBlt: The 'GarageYrBlt' column denotes the year the garage was built. The minimum value is 0, indicating properties without a garage. The maximum value is 2207, which seems to be an outlier and might be a data entry error. The mean value is 1864, and the median is 1977, suggesting a right-skewed distribution. There are 7 missing values (NA's) in this column.

GarageCars: The 'GarageCars' column represents the number of cars that can be accommodated in the garage. The minimum value is 0.000, indicating properties without a garage. The maximum value is 5.000, representing properties with a garage that can accommodate a relatively high number of cars. The mean value is 1.758, and the median is 2.000, suggesting a distribution slightly skewed towards properties with fewer garage parking spaces. There are 7 missing values (NA's) in this column.

GarageArea: The 'GarageArea' column indicates the garage area in square feet. The minimum value is 0.000, indicating properties without a garage. The maximum value is 1488.000, representing properties with a

relatively large garage area. The mean value is 469.6, and the median is 478.0, suggesting a roughly symmetric distribution. There are 7 missing values (NA's) in this column.

3. Data Cleaning and Filtering

In this section, we present a description of the pre-processing activities carried out on our dataset, namely the "House Price Prediction: Advanced Regression Techniques". The dataset encompasses a multitude of variables, and our objective was to eliminate outliers while comprehending the nature and characteristics of each variable, differentiating between numeric, categorical, and integer types.

Within the House Prices dataset, we encountered three distinct datasets: train, test, and sample. Initially, we conducted an examination of the structure of these datasets using the 'str()' function. Through this analysis, we determined that the "train" dataset comprises 81 variables, consisting of 38 integers and 43 characters. The "test" dataset encompasses 80 variables, with 37 integers and 43 characters. Lastly, the "sample" dataset comprises 2 variables, featuring 1 integer and 1 numeric type.

We proceeded with our analysis by iteratively examining each column and generating plots between the variables and the SalePrice column to identify any underlying relationships. Through this observation, we manually identified potential outliers and subsequently eliminated them to mitigate model complexity and overfitting.

Next, to streamline and expedite the data cleaning process, we decided to merge the train and test datasets. To accomplish this, we introduced a new column named "isTrain" in both the train and test datasets. For instances originating from the train dataset, the "isTrain" column was set to "1", while for instances originating from the test dataset, it was set to "0". Additionally, we appended a "SalePrice" column to the test dataset and assigned "NA" as its value. Finally, we combined the train and test datasets using the "rbind" function.

Subsequently, we conducted an examination of missing values within our combined dataset. Surprisingly, we encountered an unexpectedly high number of missing values for certain variables. Considering that these missing values could potentially contain significant information about the dataset, we opted to treat the ones with a high amount of NA as "None" rather than deleting the corresponding rows in the combined dataset.

To handle the remaining missing values in the combined dataset, we utilized the "na.omit()" function to remove rows with any missing values. This step ensured that we had a complete dataset for further analysis.

Next, we identified the numeric columns within the combined dataset. To gain a better understanding of the data distribution, we calculated the skewness of each numeric variable. Variables with skewness above the threshold of "0.75" were transformed to reduce their skewness. Specifically, we applied a logarithmic transformation using the formula log(x + 1) to address excessively skewed features.

As our primary focus was to obtain insights that would be valuable for model training, we proceeded with a train-test split. The combined dataset was divided into a training dataset and a validation dataset. The train dataset was then further split into the training and validation subsets.

4. Exploratory Data Analysis (EDA)

We proceeded by selecting the categorical variables, including factor and character variables, from our train dataset. We converted all these variables into factor variables. This allowed us to calculate the mean SalePrice for each category of the categorical variable and perform an ANOVA test to determine the correlation coefficient.

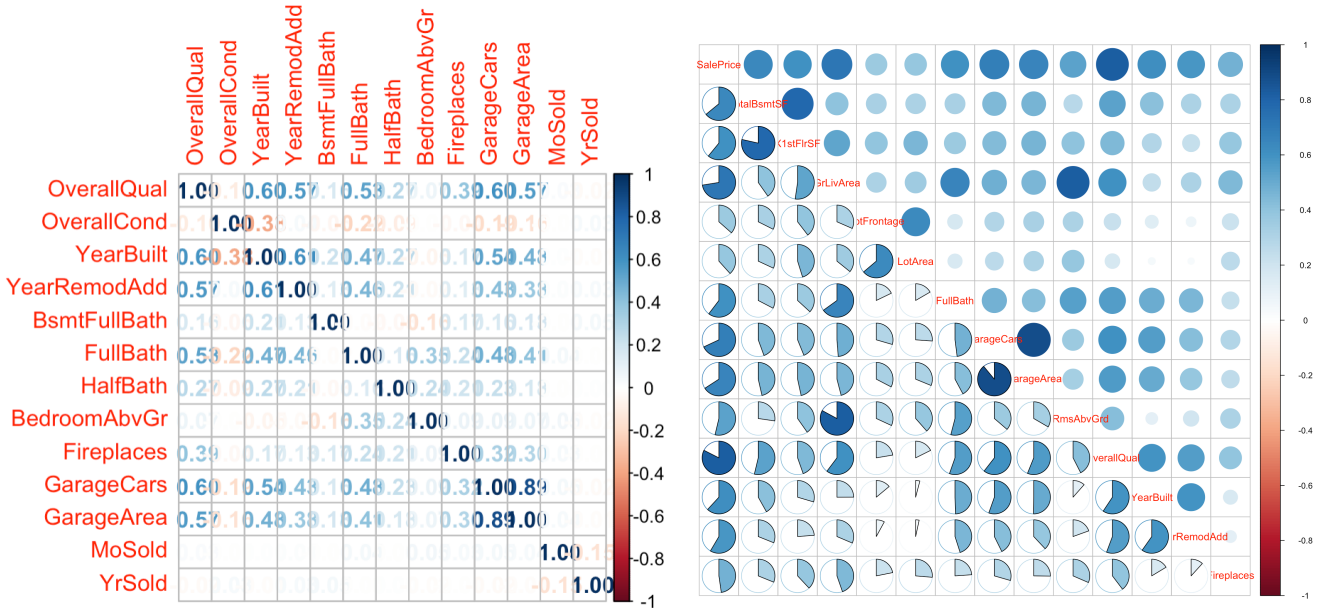
Let's get a little more into detail of what ANOVA is and what we observed by applying it. ANOVA (Analysis of Variance) is a statistical method used to analyze differences between group means and assess whether these differences are statistically significant. In our analysis, we applied ANOVA to calculate the correlation coefficients between the categorical variables and the SalePrice, which serves as the target variable in our regression problem. These correlation coefficients help identify the categorical variables that exhibit a stronger relationship with the SalePrice. A higher correlation coefficient indicates a more significant impact of the categorical variable on the SalePrice and suggests it may be an important predictor.

From our ANOVA analysis, we obtained the top 15 variables with the highest correlation coefficients. Among these, we present the top 5 variables along with an explanation of their influence on the SalePrice:

1. Neighborhood (Correlation coefficient: 0.5822): The neighborhood in which a house is located emerges as the strongest influencer of the SalePrice. Houses in specific neighborhoods tend to command significantly higher or lower prices.
2. ExterQual (Correlation coefficient: 0.4757): The quality of the exterior material significantly affects the SalePrice. Houses with better exterior quality generally have higher prices.
3. BsmtQual (Correlation coefficient: 0.4609): The quality of the basement is another influential factor. Houses with higher-quality basements tend to have higher SalePrices.
4. KitchenQual (Correlation coefficient: 0.4591): The quality of the kitchen plays a crucial role in determining the SalePrice. Houses with better kitchen quality often have higher prices.
5. MasVnrArea (Correlation coefficient: 0.4034): The masonry veneer area in square feet demonstrates a significant correlation with the SalePrice. Houses with larger masonry veneer areas typically command higher prices.

These findings highlight the importance of these variables in predicting the SalePrice and provide insights into the factors that have a substantial influence on housing prices.

After that we created a correlation plot and manually determined the variables with the highest correlation with the SalePrice compared to others, which were: TotalBsmtSF (0.64), X1stFlrSF (0.61), GrLivArea (0.72), FullBath (0.60), GarageCars (0.68), GarageArea (0.66), TotRmsAbvGrd (0.54), OverallQual (0.82), YearBuilt (0.62), YearRemodAdd (0.58), LotFrontage (0.36), LotArea (0.38), and FirePlaces (0.54). The correlation matrix, which aided in the identification of influential variables, is presented in the Figure below.



Afterwards, we utilized these variables to generate an additional correlation matrix, aiming to evaluate their impact on the SalePrice. To enhance the clarity and precision of our analysis, we augmented the correlation matrix by including pie charts that depict the relationships between pairs of variables. During our analysis using the pie chart correlation matrix, we identified several variables that exhibit a strong correlation with each other. The top 5 variables with the highest correlations, listed in descending order, are as follows:

- GarageCars and GarageArea
- TotalBsmtSF and X1stFlrSF
- GrLivArea and TotRmsAbvGrd
- GrLivArea and FullBath
- LotFrontage and LotArea

A short recap of the meaning of these variables in our dataset:

- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- TotalBsmtSF: Total square feet of basement area
- X1stFlrSF: The first floor area (in square feet) of a house
- GrLivArea: Above grade (ground) living area square feet
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- FullBath: Full bathrooms above grade
- LotFrontage: The width of the lot at the front portion facing the street
- LotArea: Lot size in square feet

Subsequently, to better understand the variables that strongly impact the SalePrice, we categorized them as follows: Categorical variables with high influence: SaleCondition, Neighborhood, Fence, GarageQual, ExterQual, BsmtQual, KitchenQual, MasVnrArea, GarageFinish, GarageType, Foundation, FireplaceQu, HeatingQC, BsmtFinType1, MasVnrType, Exterior1st, Exterior2nd, MSZoning.

Numeric variables with high influence: GrLivArea, GarageCars, GarageArea. Integer variables with high influence: OverallQual, YrSold, OverallCond. This categorization allows us to focus our analysis on these key variables and their significant role in predicting the SalePrice.

In addition to the preceding analysis, we conducted further investigations to delve deeper into the relationships among the variables of interest. This subsequent analysis aimed to provide a more comprehensive understanding of the intricate connections and associations present within the dataset.

To explore the relationship between these variables and the sale price, we utilized for loops to create visual representations, such as bar plots and scatter plots. In addition, we used the ANOVA test and correlation matrix to examine the relationship between selected variables and the sale price. By analyzing the ANOVA test results, correlation matrix, and visualizations, we gained insights into the factors influencing property prices. This comprehensive analysis helps us better understand the dynamics of the housing market and provides valuable insights for buyers, sellers, and real estate professionals.

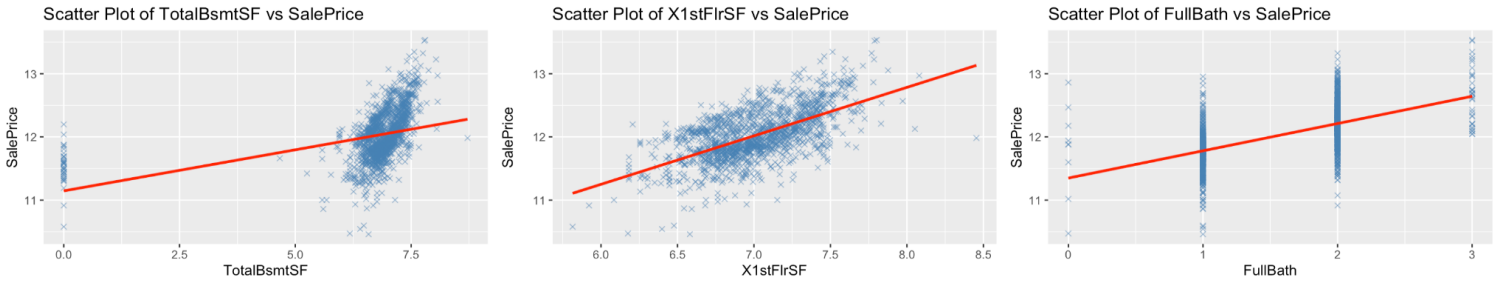
- **TotalBsmtSF:** The variable "TotalBsmtSF" represents the total square footage of the basement area. It has a significant F-value of 730.56 and a p-value of 0.000, indicating a strong correlation with the sale price. This suggests that the size of the basement can have a substantial impact on the property's sale price. Furthermore, it shows significant relationships with other variables such as X1stFlrSF, FullBath, YearBuilt, BsmtExposure, BsmtQual, ExterQual, Condition1, Street, Condition2, LotConfig, LandSlope, RoofStyle, RoofMatl, BsmtFinSF2, BsmtUnfSF, and X2ndFlrSF.
- **X1stFlrSF:** The variable "X1stFlrSF" represents the square footage of the first floor of the house. It has an F-value of 857.99 and a p-value of 0.000, indicating a strong correlation with the sale price. The size of the first floor is an important factor for potential buyers, and it influences the overall livable area of the house. It also shows significant relationships with other variables such as TotalBsmtSF, FullBath, YearBuilt, BsmtExposure, BsmtQual, ExterQual, Condition1, Street, Condition2, LotConfig, LandSlope, RoofStyle, RoofMatl, BsmtFinSF2, BsmtUnfSF, and X2ndFlrSF.
- **X2ndFlrSF:** The variable "X2ndFlrSF" represents the square footage of the second floor of the house. It has an F-value of 777.27 and a p-value of 0.000, indicating a strong correlation with the sale price. As the size of the second floor increases, the sale price tends to increase as well. However, including this variable in the model may lead to redundancy and multicollinearity due to its correlations with other variables like MSSubClass, GrLivArea, HalfBath, BedroomAbvGr, and TotRmsAbvGrd.
- **FullBath:** The variable "FullBath" represents the number of full bathrooms above grade. It has an F-value of 818.30 and a p-value of 0.000, indicating a strong correlation with the sale price. Having multiple full bathrooms can enhance the desirability and functionality of a house, potentially increasing its sale price. Additionally, it shows significant relationships with other variables such as YearBuilt, BsmtExposure, BsmtQual, ExterQual, Condition1, Street, Condition2, LotConfig, LandSlope, RoofStyle, RoofMatl, BsmtFinSF2, BsmtUnfSF, and X2ndFlrSF.
- **YearBuilt:** The variable "YearBuilt" represents the original construction date of the house. It has an F-value of 921.76 and a p-value of 0.000, indicating a strong correlation with the sale price. The age of the house can influence its sale price, with older houses potentially having historical significance or unique architectural features. It is highly correlated with other variables like BsmtExposure, BsmtQual, ExterQual, Condition1, Street, Condition2, LotConfig, LandSlope, RoofStyle, RoofMatl, BsmtFinSF2, BsmtUnfSF, and X2ndFlrSF. However, multicollinearity issues should be addressed when including YearBuilt in the model.
- **BsmtExposure:** The variable "BsmtExposure" represents the walkout or garden level walls of the basement. It has an F-value of 725.10 and a p-value of 0.000, indicating a strong correlation with the sale price. The exposure of the basement to the outside can influence the desirability and value of a property. It shows significant relationships with other variables such as BsmtQual, ExterQual,

Condition1, Street, Condition2, LotConfig, LandSlope, RoofStyle, RoofMatl, BsmtFinSF2, BsmtUnfSF, and X2ndFlrSF.

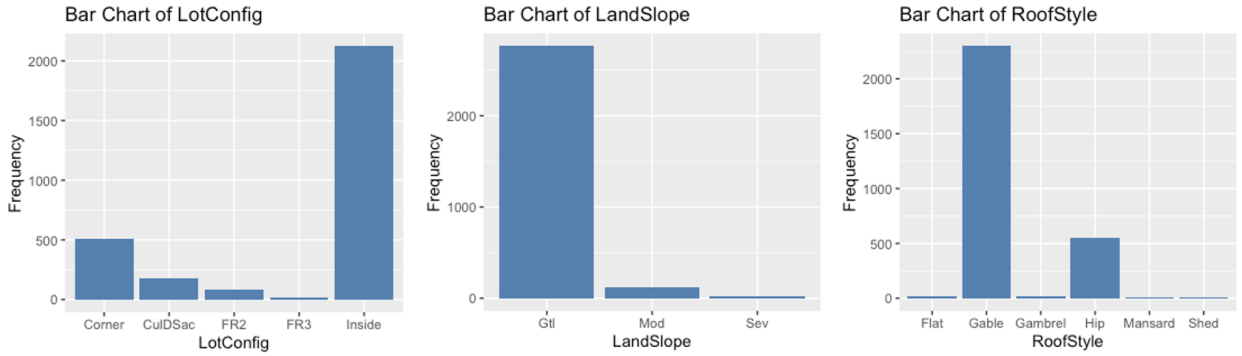
- **BsmtQual:** The variable "BsmtQual" represents the height of the basement. It has an F-value of 1034.70 and a p-value of 0.000, indicating a strong correlation with the sale price. The quality of the basement space is an important factor for potential buyers, and it can significantly impact the sale price. It also shows significant relationships with other variables such as ExterQual, Condition1, Street, Condition2, LotConfig, LandSlope, RoofStyle, RoofMatl, BsmtFinSF2, BsmtUnfSF, and X2ndFlrSF.
- **ExterQual:** The variable "ExterQual" represents the quality of the exterior material. It has an F-value of 1045.71 and a p-value of 0.000, indicating a strong correlation with the sale price. The overall quality of the house's exterior, including the materials used, can influence the sale price. It shows significant relationships with other variables such as Condition1, Street, Condition2, LotConfig, LandSlope, RoofStyle, RoofMatl, BsmtFinSF2, BsmtUnfSF, and X2ndFlrSF.
- **Condition1:** The variable "Condition1" represents the proximity of the property to various conditions. It has an F-value of 638.51 and a p-value of 0.000, indicating a strong correlation with the sale price. The proximity to different conditions, such as parks, roads, or railways, can impact the desirability and value of a property. It also shows significant relationships with other variables such as Street, Condition2, LotConfig, LandSlope, RoofStyle, RoofMatl, BsmtFinSF2, BsmtUnfSF, and X2ndFlrSF.
- **Street**: The variable "Street" represents the type of road access to the property. It has an F-value of 694.33 and a p-value of 0.000, indicating a strong correlation with the sale price. The type of road access, whether it's paved or gravel, can influence the property's desirability and value. However, the frequency distribution of this variable is imbalanced, which may impact the model's performance and should be handled during the model selection phase.
- **Condition2:** The variable "Condition2" represents the proximity to various conditions when more than one condition is present. It has an F-value of 697.75 and a p-value of 0.000, indicating a strong correlation with the sale price. Similar to Condition1, the proximity to different conditions can impact the desirability and value of a property. It shows significant relationships with other variables such as LotConfig, LandSlope, RoofStyle, RoofMatl, BsmtFinSF2, BsmtUnfSF, and X2ndFlrSF.
- **LotConfig**: The variable "LotConfig" represents the lot configuration. It has an F-value of 697.01 and a p-value of 0.000, indicating a strong correlation with the sale price. The configuration of the lot, such as being inside a subdivision or adjacent to various features, can impact the desirability and value of a property. However, the frequency distribution of this variable is imbalanced, which may need to be handled during the model selection process.
- **LandSlope:** The variable "LandSlope" represents the slope of the property. It has an F-value of 692.06 and a p-value of 0.000, indicating a strong correlation with the sale price. The slope of the land can impact the property's aesthetics, drainage, and construction requirements. However, the frequency distribution of this variable is imbalanced, which may lead to imbalanced classes and potentially impact the model's performance. It should be addressed during the model selection process.
- **RoofStyle:** The variable "RoofStyle" represents the style of the roof. It has an F-value of 691.16 and a p-value of 0.000, indicating a strong correlation with the sale price. The style of the roof can contribute to the overall architectural appeal of a house and influence its value. It is highly correlated with the RoofMatl variable.
- **RoofMatl:** The variable "RoofMatl" represents the material used for the roofs of the houses. It has an F-value of 690.44 and a p-value of 0.000, indicating a strong correlation with the sale price. The choice of roofing material can impact the property's aesthetics, durability, and maintenance requirements. However, the frequency distribution of roof materials is imbalanced, which may impact the model's performance and should be addressed during the model selection process.
- **BsmtFinSF2:** The variable "BsmtFinSF2" represents the type 2 finished square feet in the basement. It has an F-value of 702.19 and a p-value of 0.000, indicating a strong correlation with the sale price. As the finished square footage of the basement increases, the sale price tends to be affected. It is

negatively correlated with BsmtUnfSF, indicating that as the finished square footage increases, the unfinished square footage decreases.

- **BsmtUnfSF:** The variable "BsmtUnfSF" represents the unfinished square feet of the basement area. It has an F-value of 693.93 and a p-value of 0.000, indicating a strong correlation with the sale price. The size of the unfinished basement area can impact the overall livable area and potential functionality of the house. It is correlated with TotalBsmtSF and OverallQual, indicating that the sale price of houses slightly increases when the square footage of the basement area increases.

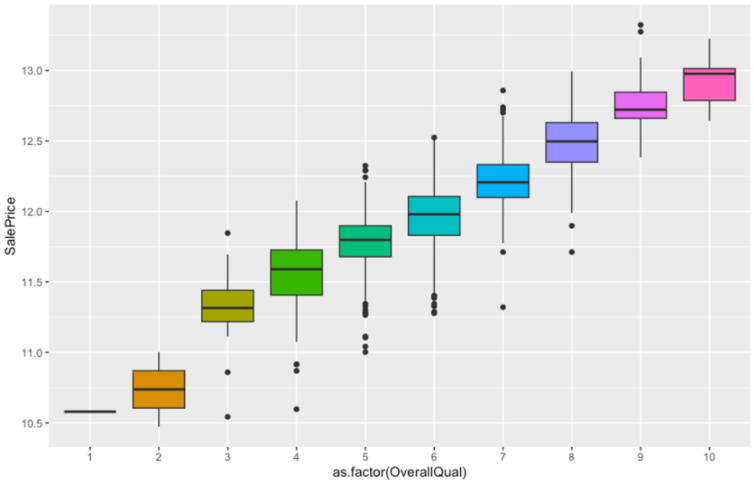


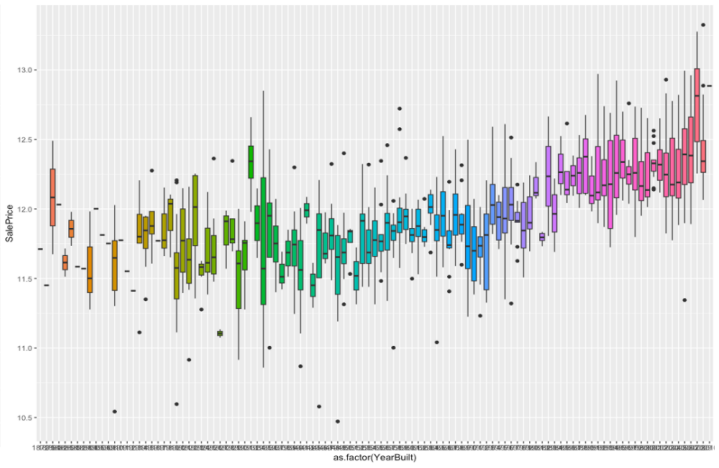
A small example of the scatter plots we created for TotalBsmtSF, X1stFlrSF, and FullBath against SalePrice indicates a clear linear relationship between these variables and the sale price of the houses. The plots show a positive trend, suggesting that as the values of TotalBsmtSF, X1stFlrSF, and FullBath increase, the sale price of the houses also tends to increase. This indicates that these variables have a significant impact on the pricing of the properties.



A glimpse of the categorical graphs that we created to understand the relationship between LotConfig, LandSlope, and RoofStyle against SalePrice reveals that the data is not evenly distributed among the categories. This imbalance in the classes could potentially cause imbalanced classes in our model, which may affect its performance and accuracy.

After identifying the variables deemed important through ANOVA and the correlation matrix, we proceeded to create their bar plots. These plots aimed to visualize the patterns and relationships between SalePrice and each individual variable. By examining these bar charts, we gained valuable insights into the trends and associations, enabling a better understanding of how these variables influence the SalePrice.

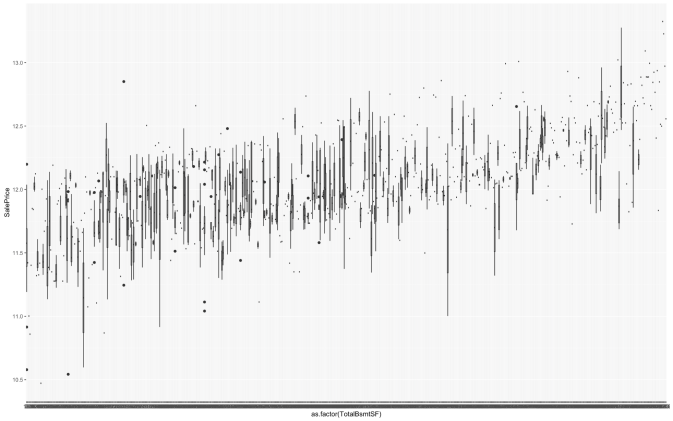


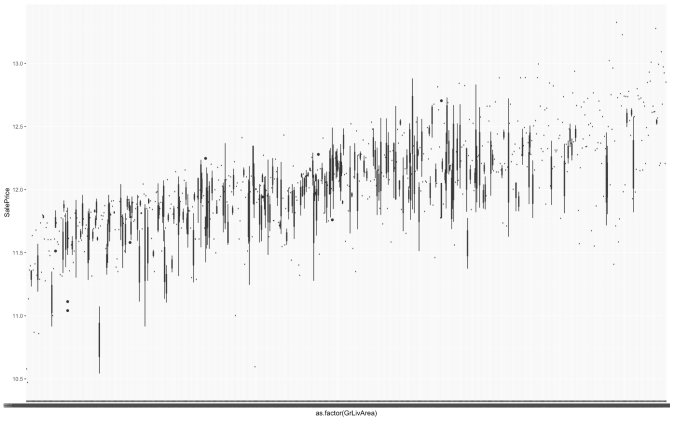


We initially examined whether there was a noticeable gradual increase in certain variables as their values increased. In this regard, we identified four cases where such trends were evident.

The first graph illustrates the relationship between OverallQual and SalePrice. It is apparent that as the quality of the houses improved, the SalePrice also increased. This observation is supported by the median values displayed in the graph.

The second graph, depicting the relationship between YearBuilt and SalePrice, confirms that as the year of construction approached the present day, the SalePrice tended to rise accordingly.

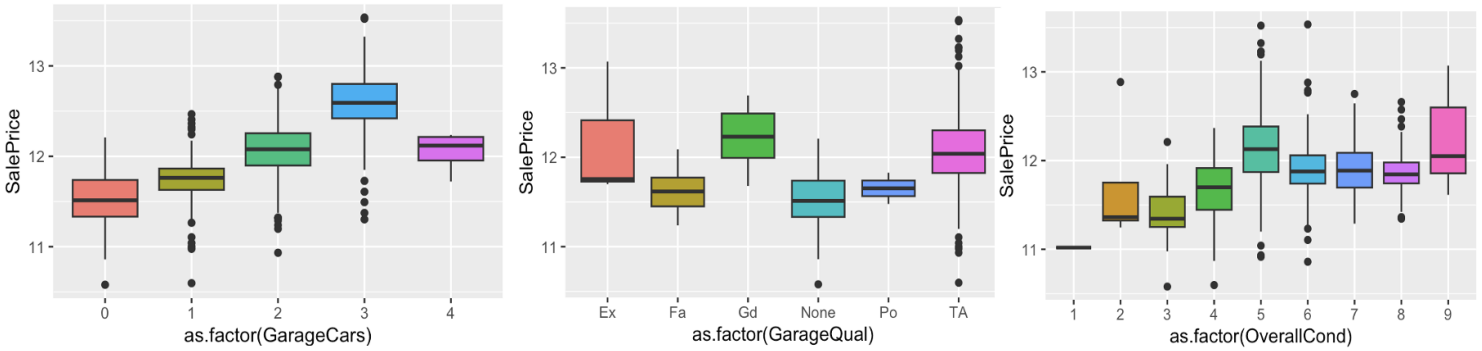




Similarly, the third graph shows that as TotalBsmtSF (total square feet of basement area) increased, the SalePrice exhibited an upward trend.

Lastly, the fourth graph demonstrates the positive correlation between GrLivArea (above-grade living area square feet) and SalePrice. As the size of the living area increased, the corresponding SalePrice also showed an upward trajectory.

Following that, we conducted an analysis to identify any unexpected cases within the dataset. As a result, we discovered three variables that exhibited unexpected patterns, as outlined below:



- **GarageCars**: The number of cars a garage can accommodate was found to be a significant factor affecting the sale price. Curiously, our analysis indicated that houses with a garage capacity for 3 cars had the highest SalePrice, rather than those with a capacity for 4 cars.
- **GarageQual:** The quality of the garage, rated from "Ex" (Excellent) to "Po" (Poor), was also found to impact the sale price. Contrary to expectations, the highest SalePrice was associated with houses having "Good" garage quality, followed by those with an "Average" quality instead of “Excellent”.
- **OverallCond:** The overall condition of the house, rated on a scale of 1 to 10, was found to influence the sale price. Surprisingly, our analysis revealed that houses in "Average" condition (level 5) had the highest SalePrice. This unexpected result contradicts the assumption that houses in "Very Excellent" condition would command higher prices.

5. Model

5.1. Linear Model

In order to fit the linear model, we divided the data into training and validation sets to evaluate the performance of a linear regression model on unseen data. In order to ensure consistency between the training and validation sets, certain features with differing levels in the same variable were excluded from the training set. This step was necessary to prevent errors during the prediction process. By removing these variables, we aimed to maintain the integrity of the model evaluation on the validation set. This approach helped to mitigate potential discrepancies caused by inconsistent variable levels and improve the reliability of the model's performance assessment. We then fitted a linear regression model using the remaining variables in the training data to predict the target variable. We printed the model's summary, which included coefficient estimates and statistical measures. Next, we made predictions on the validation dataset using the fitted model and calculated the mean squared error (MSE) to assess the prediction error.

In the output of our linear model, we examined the coefficient estimates, standard errors, t-values, and p-values to evaluate the significance of these variables. We observed that certain predictors such as MSZoning, LotArea, Street, OverallQual, OverallCond, YearBuilt, YearRemodAdd, GrLivArea, and PoolArea demonstrate statistically significant coefficients with p-values below 0.05. However, other variables like MSSubClass, LotFrontage, Alley, and LotShape have coefficients near zero and high p-values, suggesting a limited impact on the sale price. We also noticed the presence of "NA" values, indicating potential collinearity issues caused by highly correlated predictors.

The high multiple R-squared value of 0.9414 indicates that our model explains approximately 94.14% of the variance in sale prices, suggesting a strong fit. However, it is important to consider the adjusted R-squared value of 0.9276, which accounts for the number of predictor variables used. The residual standard error of 0.106 represents the average difference between observed sale prices and our model's predicted values, providing an assessment of prediction accuracy. The low p-value (less than 2.2e-16) associated with the F-statistic of 67.98 indicates the overall statistical significance of our model. However, we received a warning message about a rank-deficient fit.

When a variable has a leverage of one, it means that a particular observation has a strong influence on the estimated regression coefficients. These influential observations can significantly impact the results of the regression analysis, particularly the estimated coefficients and the fitted values. In our analysis, we encountered difficulties in generating certain plots and received a warning due to the presence of leverage value of one. Consequently, we made the decision to proceed with fitting the model without removing these variables, acknowledging the potential influence they might exert on the analysis and accepting the associated warning.

In a linear model, several crucial assumptions need to be satisfied for reliable results:

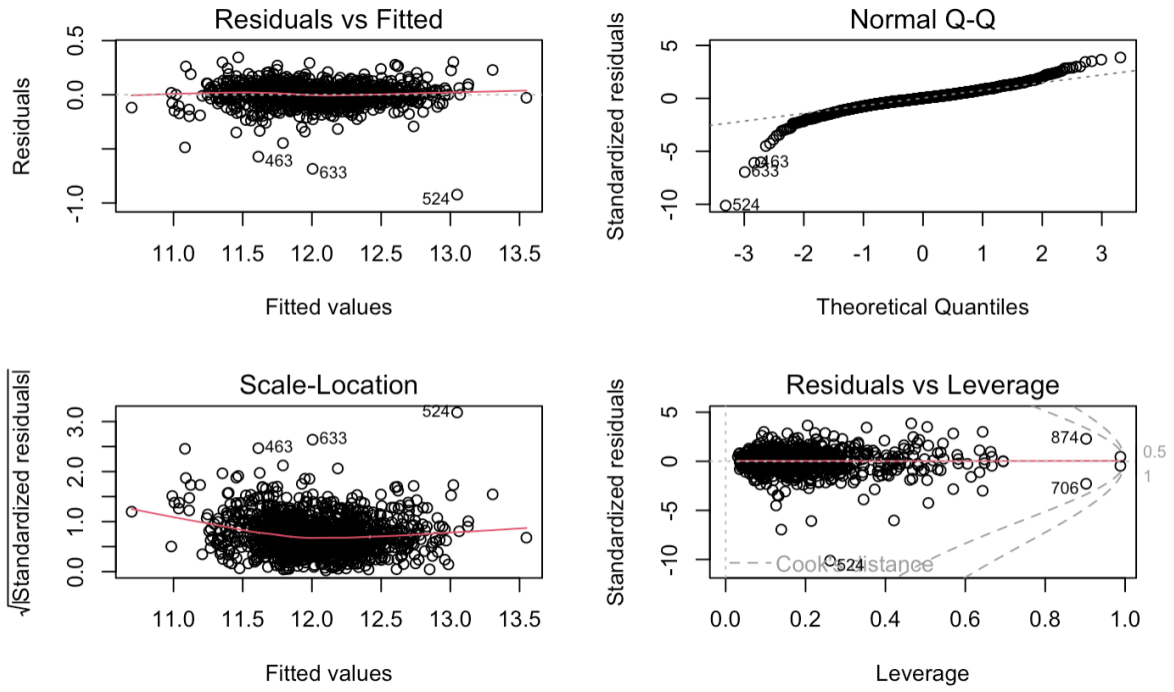
- Linearity: We assume a linear relationship between the predictors and the target variable, signifying a constant and proportional effect.
- Homoscedasticity: We expect the variance of the residuals to remain constant across all levels of the predictors, ensuring consistent spread regardless of predicted values.
- Normality and Independence of Errors: We assume that the residuals follow a normal distribution and are independent, as violations can lead to biased and inefficient estimates.

To assess the validity of our model and identify potential issues, we examined the following plots:

- Residuals vs. Fitted plot: We expect to see the residuals randomly dispersed around the x-axis, indicating adherence to linearity. Patterns or trends in the residuals could suggest violations of the linearity assumption or other model issues.
- Quantile-Quantile (qq) plot: This plot helps us evaluate the normality of the residuals. Significant deviations from a straight line indicate departures from normality, which can affect the validity of our statistical inference and prediction intervals.
- Leverage: We identify high leverage points that have a strong influence on the estimated coefficients. These points, acting as outliers, can impact the overall performance of our model. It is important to identify them to ensure the reliability of our results.

The Residuals vs. Fitted QQ plot is used to assess the adequacy of a regression model. It displays the relationship between the predicted values (fitted values) and the corresponding residuals. An evenly scattered pattern of residuals with no discernible trend indicates that the model adequately captures the relationship between predictors and the response variable. Deviations from this pattern can suggest violations of assumptions such as non-linearity or heteroscedasticity.

In light of the residuals vs. fitted plot, we observe that the points are evenly dispersed both above and below the predicted value line. Consequently, we can conclude that, in this particular case, our assumptions hold true.



The Residuals vs. Fitted Values plot reveals a funnel shape in the distribution of residuals. The majority of residuals appear to be randomly distributed along the line, with only a few residuals noticeably deviating from the line. These deviations suggest potential outliers or influential observations that have a more significant impact on the model's fit. However, the overall pattern of residuals appearing randomly distributed within the funnel shape indicates that the model captures the relationship between the predictors and the response reasonably well.

The Normal Q-Q plot suggests that the residuals approximate a normal distribution, although there are minor deviations from the diagonal reference line. These deviations are expected and considered within an acceptable range. The model's estimates and assumptions based on normality are likely to be reliable.

The Residuals vs. Leverage plot combines information about the residuals (the differences between observed and predicted values) and the leverage of each data point. It helps identify influential points that deviate significantly from the overall trend and have high leverage, potentially impacting the estimated coefficients and model fit.

When interpreting these plots, we consider the proximity of data points to Cook's distance, which is a measure of their influence on the regression model. Cook's distance quantifies the effect that removing a specific data point would have on the regression coefficients and overall model fit. By examining the distance of data points to Cook's distance, we can assess their individual impact on the analysis.

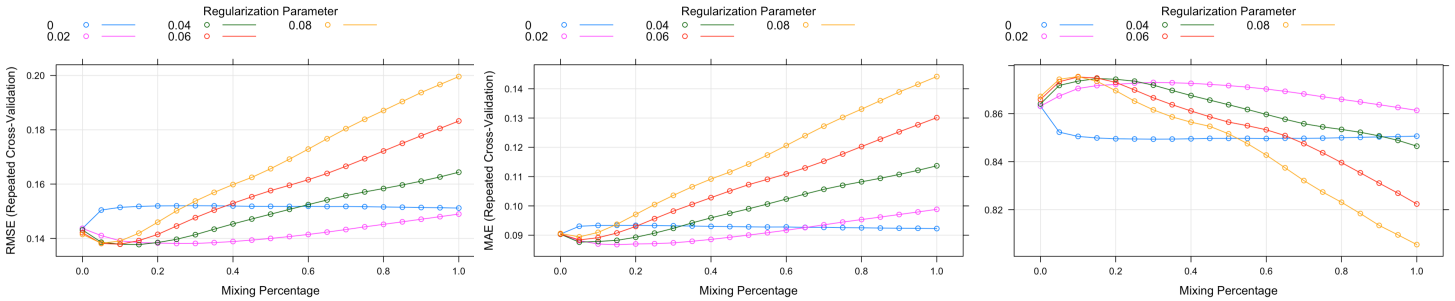
In our Residuals vs. Leverage plot, none of the data points exceed the threshold of Cook's distance, indicating the absence of influential points within our regression model.

5.2. Regularized Linear Model

After fitting the linear model, we moved on to performing regularized linear regression to improve the predictive performance of our model. To achieve this, we created a train control object for cross-validation, specifying the number of folds and repetitions. Next, we defined the range of alpha and lambda values for tuning, which control the degree of regularization and the magnitude of the penalty applied to the regression coefficients. We defined alpha_values ranging from 0 to 1 in increments of 0.05 and lambda_values ranging from 0 to 0.08 in increments of 0.02. This selection allowed us to investigate the effects of both ridge and lasso regularization and capture the impact of regularization in finer detail. The smaller increments were chosen to obtain a more precise understanding of the regularization effects. Using the train function, we fitted the regularized linear model on the train_new2 dataset, incorporating the defined train control object and method. We used the expand.grid function to generate a grid of alpha and lambda values to search over during the model tuning process.

We plotted the coefficients of the model against the logarithm of lambda, which is the regularization parameter. Then, we examined the model's performance by plotting the cross-validated mean squared error (CV RMSE), mean absolute error (CV MAE), and R-squared values against the logarithm of lambda.

Next, we used the regularized linear model to make predictions on a validation dataset (validate2) and calculated the mean squared error (MSE) as a measure of prediction accuracy. Finally, we compared the MSE of the regularized model with that of a non-regularized model, and it was found that the non-regularized model achieved a better MSE score.



RMSE-Mixing Percentage Plot: This plot shows the cross-validated mean squared error (CV RMSE) against the mixing percentage or regularization parameter (lambda). In your case, you have used the "glmnet" method, which performs elastic net regularization with a combination of L1 (Lasso) and L2 (Ridge) regularization. As the mixing percentage or lambda increases (from 0 to 0.08), the model's performance in terms of RMSE worsens. This means that higher levels of regularization are not beneficial for improving the model's predictive accuracy in your case. The lowest RMSE values are observed for the mixing percentages or lambda values of 0 and 0.02, indicating that these levels of regularization result in better model performance in terms of RMSE.

MAE-Mixing Percentage Plot: This plot shows the cross-validated mean absolute error (CV MAE) against the mixing percentage or regularization parameter (lambda). Similar to the RMSE plot, as the mixing percentage or lambda increases, the model's performance in terms of MAE worsens. Higher levels of regularization result in

higher MAE values, indicating poorer model performance. The highest MAE value is observed for the mixing percentage or lambda value of 0.08, suggesting that this level of regularization leads to the worst performance in terms of MAE. The lowest MAE values are observed for the mixing percentages or lambda values of 0 and 0.02, indicating that these levels of regularization result in better model performance in terms of MAE.

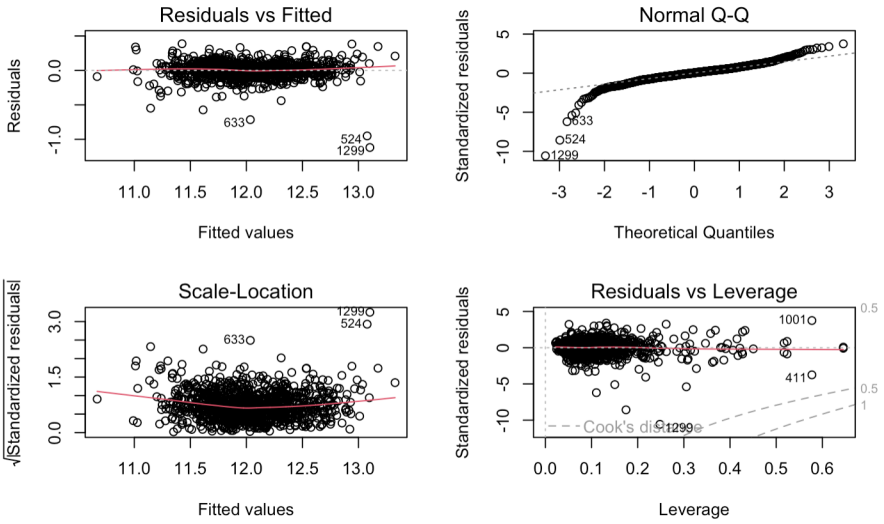
R-Squared-Mixing Percentage Plot: This plot shows the cross-validated R-squared coefficient against the mixing percentage or regularization parameter (lambda). As the mixing percentage or lambda increases, the model's performance in terms of R-squared tends to decrease. Higher levels of regularization result in lower R-squared values, indicating less of the variance in the target variable (SalePrice) is explained by the model. The highest R-squared value is observed for the mixing percentage or lambda value of 0.02, suggesting that this level of regularization leads to the best performance in terms of explaining the variance in the target variable. The lowest R-squared value is observed for the mixing percentage or lambda value of 0.08, indicating that this level of regularization results in the poorest performance in terms of explaining the variance in the target variable.

5.3. Backward Model Selection

For the backward model selection, our aim was to select the model with the highest accuracy based on the subset of variables that we had. Subsequently, we investigated whether including or removing variables would lead to any improvements in our model's performance.

We initially conducted a backward model selection without considering the variables that we removed to calculate the mean squared error in the previous sections. We focused on identifying a subset of predictor variables that exhibited a substantial impact on the target variable, SalePrice.

Subsequently, we aimed to improve the accuracy of our model by selecting the most relevant variables. We achieved this by iteratively removing variables one by one and comparing the performance of the resulting models. This process allowed us to identify the best model, which exhibited higher accuracy compared to the model using all variables (Multiple R-squared: 0.9148, Adjusted R-squared:0.9042, MSE for the linear model: 0.1175047). The model results of the model that we picked is given below.



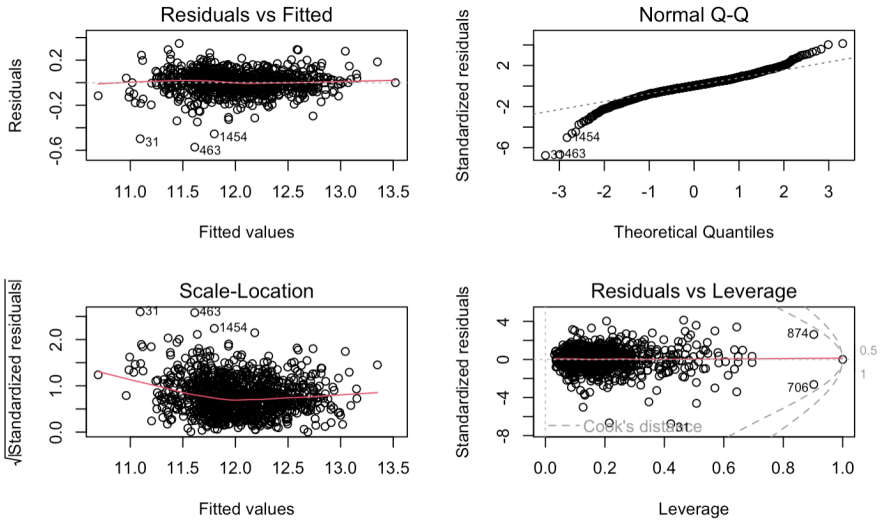
The analysis of the Residuals vs. Fitted Values plot reveals a discernible funnel pattern in the distribution of residuals. Most of the residuals exhibit a scattered distribution that aligns reasonably well with the line, with only a few distinct deviations observed. These deviations point towards potential outliers or influential observations that significantly impact the model's fit. However, it is worth noting that the prevailing trend of residuals appearing randomly scattered within the funnel shape indicates a satisfactory alignment between the model and the relationship between predictors and the response variable.

The analysis of the Normal Q-Q plot indicates that the residuals closely align with a normal distribution, displaying minor deviations from the diagonal reference line. These discrepancies, while present, fall within an acceptable range and do not significantly impact the dependability of the model's estimations or the validity of assumptions regarding normality. Thus, the model's estimates and the underlying assumptions can be regarded as reliable for the given analysis.

In our Residuals vs. Leverage plot, none of the data points exceed the threshold of Cook's distance, indicating the absence of influential points within our regression model.

Next, we turned our attention to handling outliers in the dataset. Outliers are data points that deviate significantly from the overall pattern. To identify outliers, we set a threshold of 0.70 for the residuals, which are the differences between the actual and predicted values. Any observation with a residual exceeding this threshold was considered an outlier.

By removing the outliers from the dataset, we created a new version of the linear model. The summary statistics of this revised model revealed the impact of outlier removal on its performance (Multiple R-squared: 0.9512, Adjusted R-squared: 0.9396). We also made predictions on a validation dataset and calculated the mean squared error (MSE) as a measure of prediction accuracy. Notably, the removal of outliers led to an improvement in the model's performance, as evident from the summary statistics and the lower MSE value (MSE for the linear model: 0.1151298) as can also be seen below.



6. Prediction

In the prediction part of our analysis, we focused on predicting the prices of houses in the test dataset. We used our trained model to make these predictions. By applying the trained model to the test data, we obtained the predicted prices. Since our model was trained using logarithmic values, we converted the predicted prices back to their original scale. This allowed us to interpret the predicted prices in a more meaningful way.

We then updated the test dataset by adding the predicted prices which allowed us to see the predicted prices alongside other relevant information for each house in the test dataset. By reviewing the modified test dataset, we could assess the predicted house prices based on our trained model. This step provided valuable insights into the expected prices for the houses in the test dataset.

7. Conclusion

In this data analysis project, our goal was to predict house prices based on various features. We began by exploring the dataset, conducting ANOVA analysis, and examining the correlation matrix to gain insights into the relationships between variables and the sale price. We checked the influence of different variables on the sale price through these initial analyses, which provided valuable information about the significant predictors and their correlations with the target variable.

Next, we employed the backward model selection technique to determine the most influential variables for predicting house prices. Through iterative model refinement, we identified the best model that demonstrated higher accuracy compared to the model using variables. The selected model showed improved performance, as evidenced by the lower mean squared error (MSE) and the model summary statistics.

Furthermore, we addressed outliers by setting a threshold and removing observations with residuals exceeding this threshold. This outlier elimination step led to further improvement in the linear model's performance, as indicated by the reduced MSE and visual examination of the model results.

Finally, we used the trained model to predict house prices on the test dataset. By converting the predicted logarithmic prices back to their original scale, we obtained meaningful predictions. The modified test dataset provided insights into the predicted prices for each house. The findings we obtained from this data analysis offer valuable insights for understanding and estimating house prices based on the provided dataset.