House Price Prediction

Welcome to Boston!

CONTENTS

Introduction

Preprocessing

EDA (Exploratory Data Analysis)

Testing/Training

Model Building

Variable Selection

Residual Analysis

Conclusion

Introduction

- Our data source: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data?select=train.csv).
    - This dataset provides information on the housing prices as well as the details about the location, lot area, condition, year built, sale type, etc.
- Presenting you the software (RStudio) to analyze and predict the cost of housing in a particular area.
- The aim of this project is to forecast the house prices so as to minimize the problems faced by the customer.

Introduction

- The proposed solution features the Multiple Linear Regression Model.
  - Multiple Linear Regression algorithm can be used to help investors to invest in an appropriate estate according to their desired requirements.
- After preprocessing, the model will be split and tested according to the 80/20 rule. 80% of data for training, test on remaining 20%.
- Feed the system with property information based on data, and the system will predict the estimated price of this house
- The following slides detail our work. Enjoy!

Problem Description

- Real estate property values are strongly correlated with our economy.
- Nowadays, we see applications of Machine Learning and Artificial Intelligence in most of the domains but for a long time, the real estate industry was quite slow in adapting Data Science and Machine Learning for problem-solving and improving their processes.
- So, the objective of this study is to apply machine learning to forecast the selling values of houses based on a variety of economic attributes

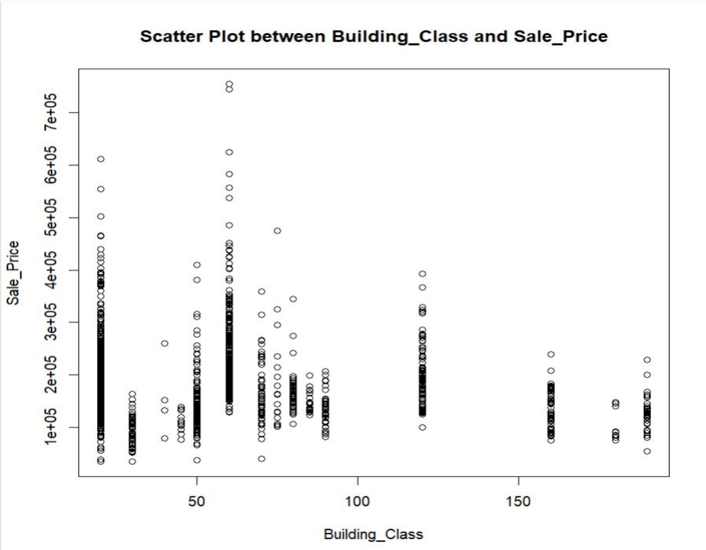
Data Description

- “House Prices - Advanced Regression Techniques”
- From Boston, MA!
- This dataset provides information on the housing prices as well as the details about the location, lot area, condition, year built, sale type, etc.
- Initial Data Dimensions - 1459 Rows and 81 Columns
- Number of Categorical Variables: 43
- Number of Numerical Variables: 38

Data Preprocessing

- Start with the Unique value check and remove those variables.
- For certain character variables, replace N/A with a string.
  - E.g. if basement_condition is N/A for a house, replace with string “No_Basement.”
- Remove remaining character values with N/A above 75%
- Check rows for presence of N/A values
  - Replace N/A with the mode metric value for categorical variables
  - Replace N/A with the mean metric value for numeric variables

Preprocessing



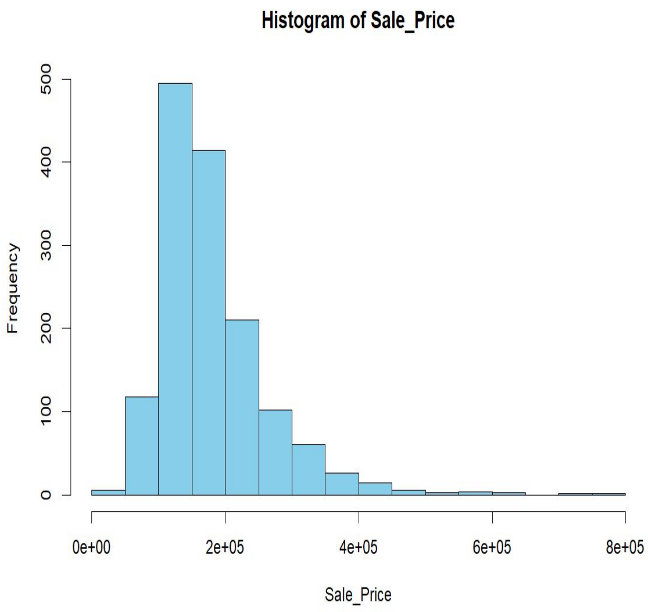
- Convert ordinal values to categorical non-ordinal using as.factor.
    - Ordinal variables have an ordering (1, 2, 3). As.factor removes rank from each category

More on preprocessing

- Remove biased columns with dominant class
    - Biased columns are columns where 95% of results or more are the same value.
    - Lack of data for presence of other distinct values makes those harder to predict
- For all variables with skewness >|2|, apply a log transformation
    - This process lowered the skewness to acceptable range for most but not all the variables
    - Most important for target variable (sale price) to have skewness < |1|
- Perform correlation analysis to check for highly correlated variables (we chose 0.8 as the cut-off. No variables have values greater than 0.8

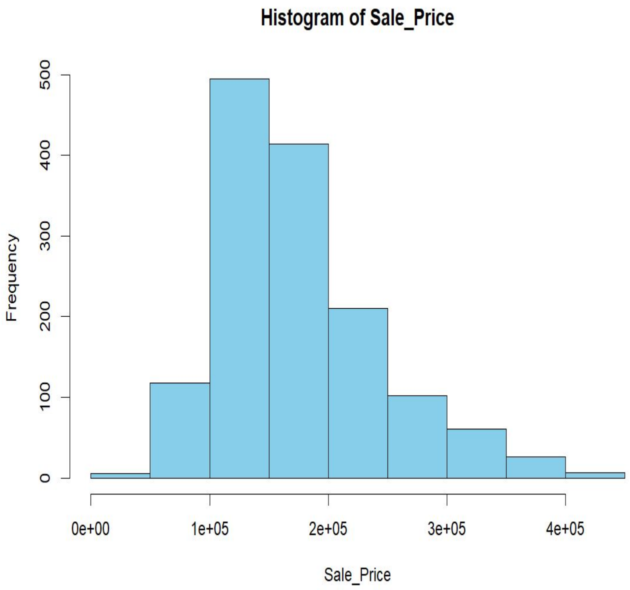
TARGET VARIABLE ANALYSIS

Histogram of the Target Variable before removing outliers



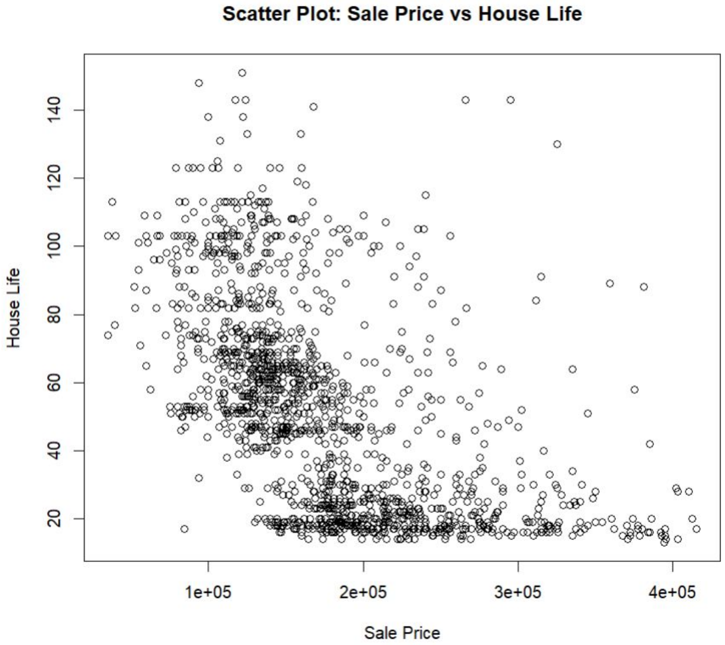
Skewness = 1.877893

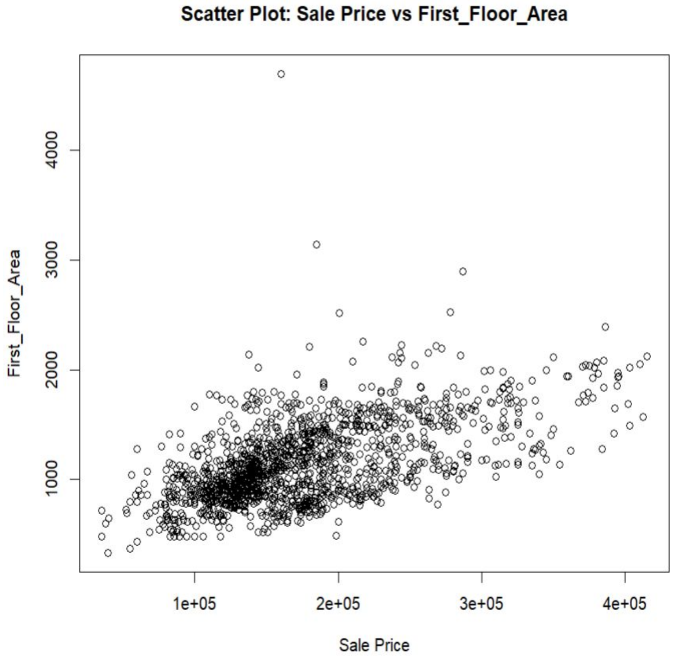
Histogram of the Target Variable after removing outliers



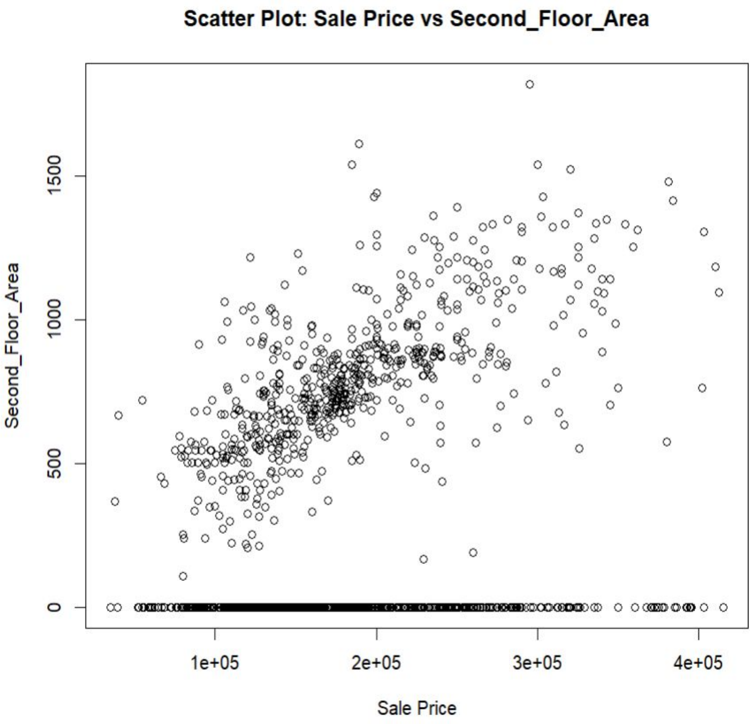
Skewness = 0.9986061

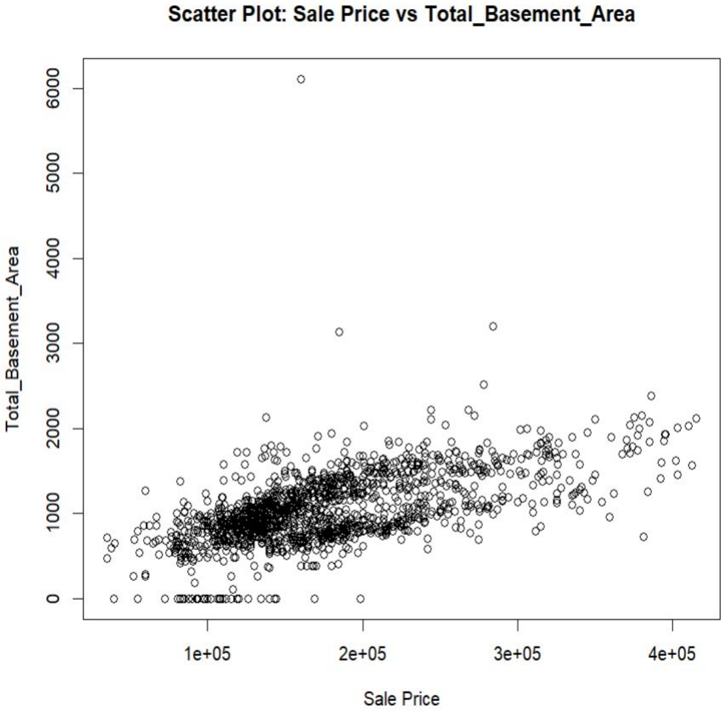
Exploratory Data Analysis



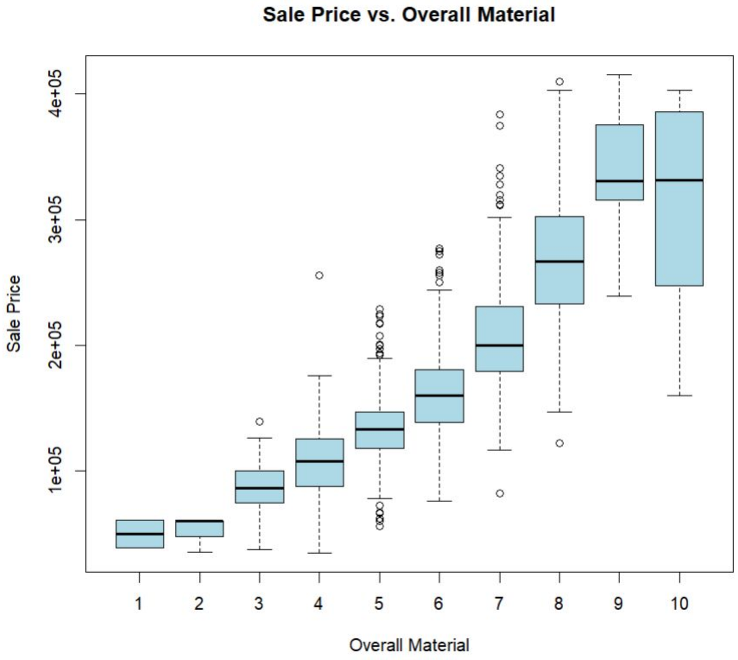


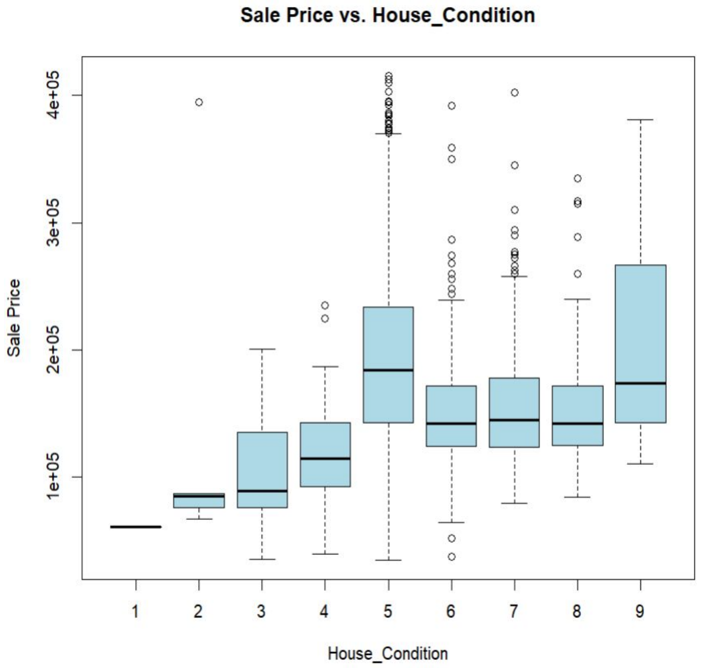
More exploration on numeric variables



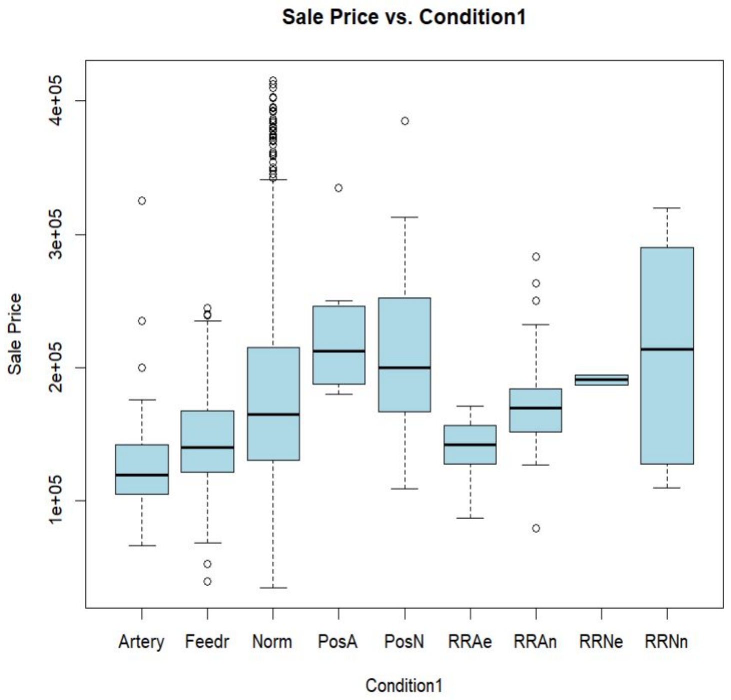


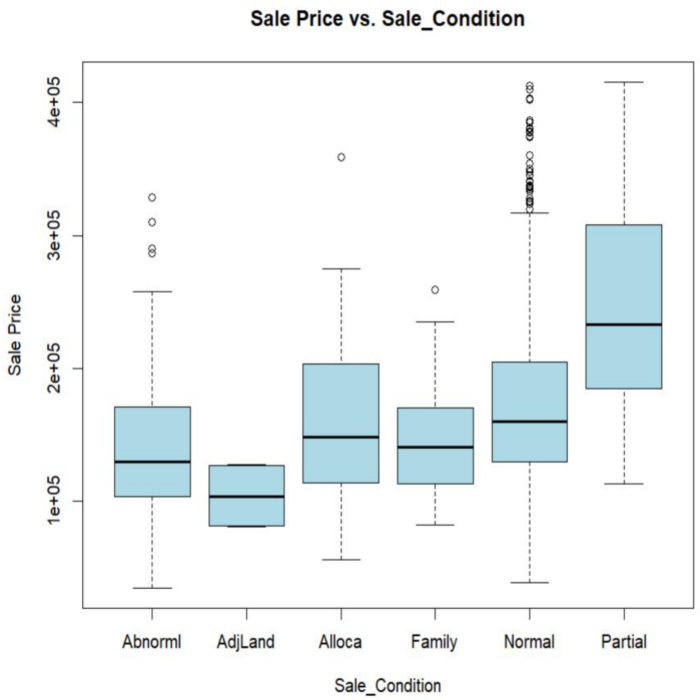
Now time to explore categorical data





More exploration on factors



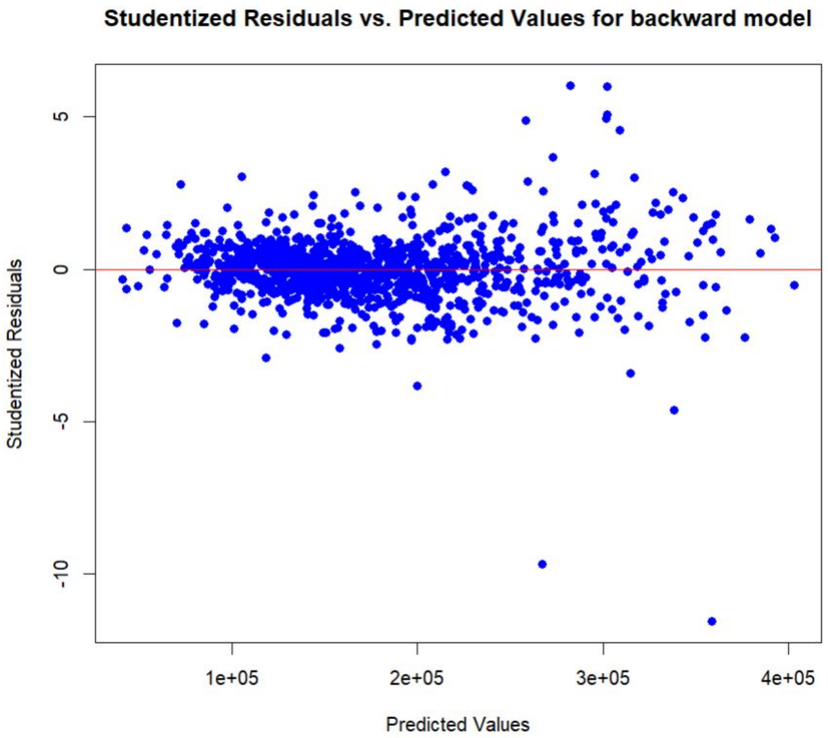


Applying Linear Regression

- Split the data into 80% training, 20% testing.
- Apply linear regression model with all variables.
- "Apply two variable selection procedures to find an optimal subset of independent variables to predict" House Price.
  - We chose backward selection and forward selection
  - See if MSE reduced, Adjusted R-squared increased.
- Comparing the MSE, MAE and Adjusted R-Squared values.

	Model	MSE	MAE	R_Squared	Adjusted_R_Squared
1	Initial_model	706094381	17780.28	0.9297326	0.9081733
2	backward_model	687252927	17477.80	0.9251316	0.9152572
3	forward_selection	706094381	17780.28	0.9297326	0.9081733

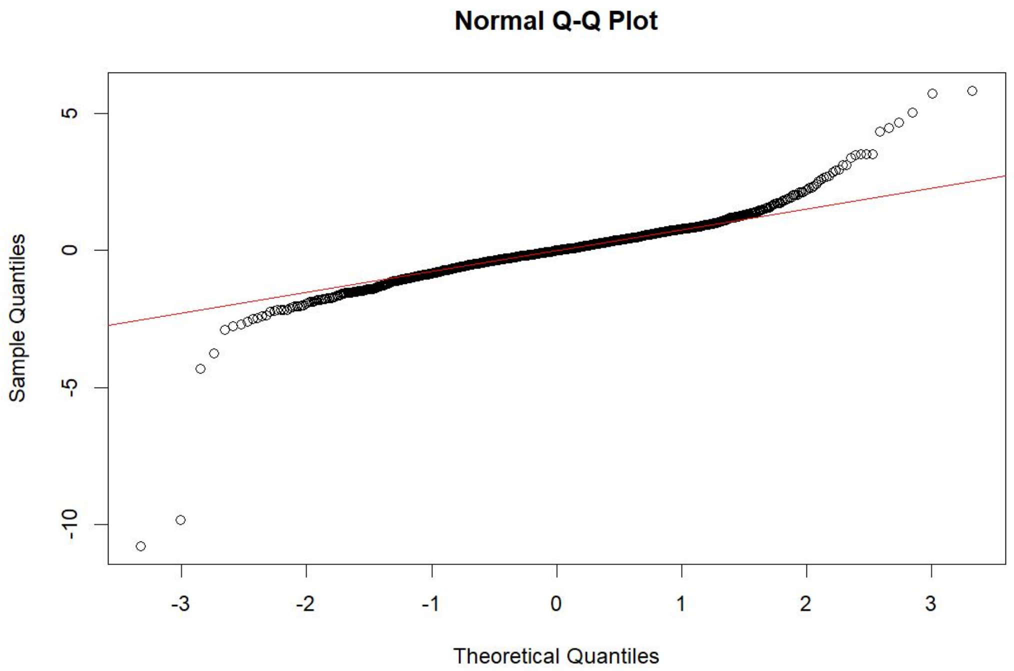
Residual Analysis of selected final model



Residuals are randomly scattered around zero. No pattern detected.

Spread of residuals is roughly constant hence there is presence of Homoscedasticity.

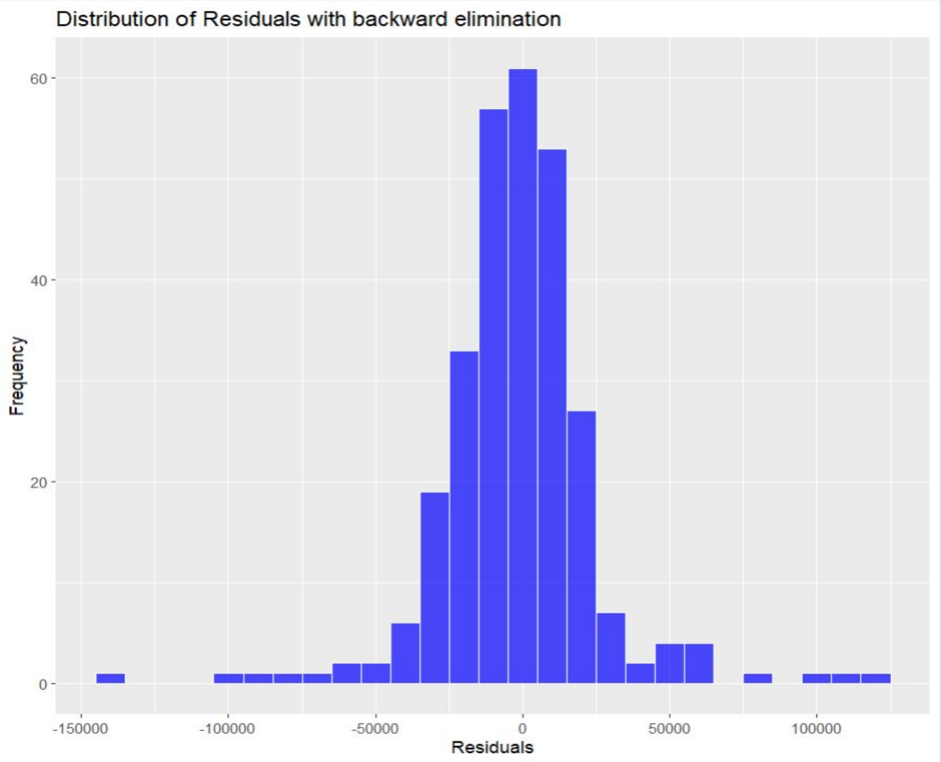
Normal Q-Q Plot



Most of the residuals following the straight line.

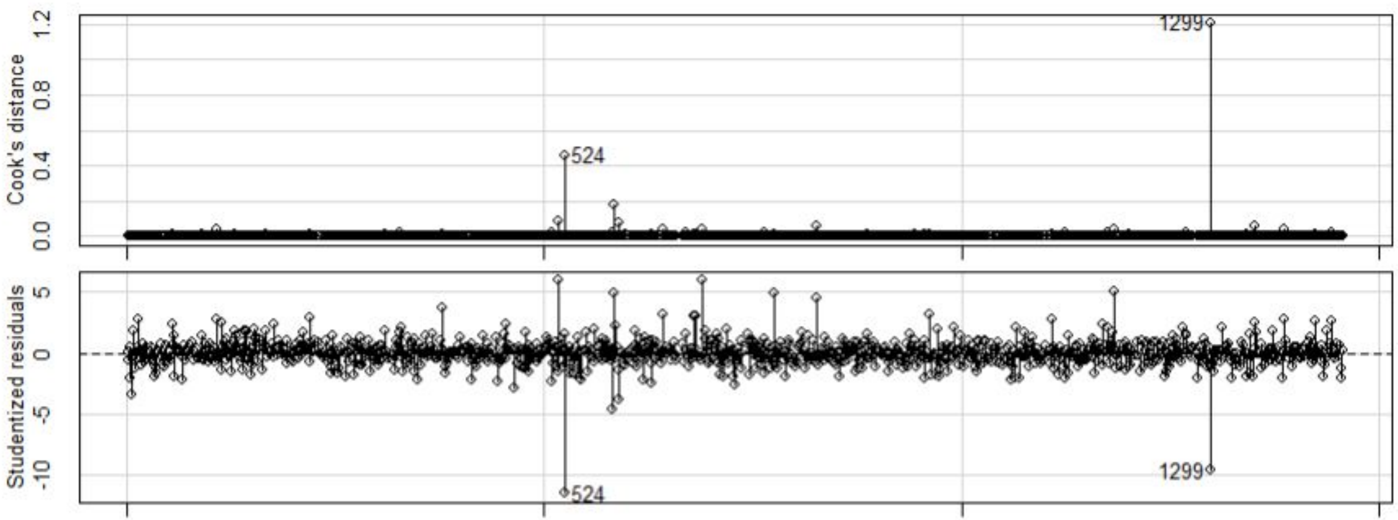
Few points bending upward on the right may cause a positive skewness in the residuals.

Distribution plot for the residuals



We can see residuals are normally distributed with a skewness value of 0.05426359.

Check for influential points



Checking if assumptions are right

	Variable	P_Value
1	House_Life	1.942902e-05
2	Total_Basement_Area	4.750816e-02
3	Second_Floor_Area	2.504519e-03
4	House_Condition8	2.317924e-02
5	House_Condition5	7.949387e-07
6	Overall_Material9	7.863021e-21
7	Sale_ConditionNormal	3.286395e-04

Here we can say our assumptions in the EDA part are true that the variables we chose for EDA will be significant.

FUTURE SCOPE

- Many features can be added to make the system more widely acceptable.
- One of the key goals for the future is to add a database of real estate from more cities, which will allow the user to explore more areas and make a specific decision.
- Some metrics that were biased (with one value more than 95% of the data) were removed. Collect more data with varied values for that metric to keep in the models
- Other factors, such as the state of the economy, inflation, that will affect property prices can be added.

CONCLUSION

- Our data provided a good model for predicting the house price. We deleted many metrics and still retained a good model, proving that some metrics are more impactful than others when determining the final price. Our model explains more than 90% of the variation in house price.

- Data science techniques are a valuable tool to efficiently predict prices. This helps buyers and customers make wise financial decisions on whether to sell or buy a property at a certain price. If applied to the real word, our model can also mitigate the risk of property investment.

THANKS!

Any questions?