STA 141C FINAL PROJECT

HOUSE PRICE PREDICTION

Winter 2025

Professor: Jairo Fuquene Patino

Group Member Contribution:

Nick Fitzpatrick: Code (Data Exploration, Model Fitting, Model Evaluation), Part V, VI

Yanying He: Code (Model Fitting), Part VIII

Minyi Liu: Code (Data Exploration), Part II, III, IV, V

Daniel Ly: Code (Visualization)

Hancy Zou: Part I, VII

I. Introduction

Accurately predicting house prices is essential due to its direct impact on individual financial decisions and broader economic conditions. The purpose of this project is to develop robust predictive models for accurately estimating house prices, an important concern for home buyers, sellers, and real estate professionals. By analyzing key features—such as land characteristics, structural details, location attributes, and interior designs—we aim to determine which factors most significantly impact home values. Specifically, our objectives are to identify critical predictors, compare predictive accuracies across different modeling techniques, and evaluate the extent to which these models reveal the relationships between house features and pricing. Employing methods like Linear Regression, Polynomial Regression, Lasso Regression, Decision Trees, and Random Forest, we explore both interpretability and predictive power to derive actionable insights from house pricing data.

II. Research Problems

This project aims to develop a predictive model for house prices based on various features of homes. Accurately predicting house prices is crucial for buyers, sellers, and real estate professionals. This project aims to explore the key factors influencing housing prices and build a model that can provide reliable price estimates. The questions we want to explore are the following:

1. We primarily examine how land characteristics of the house, utility infrastructure, structural features, and interior design influence house prices. What are the key factors that have the greatest impact on house values?
2. How accurately can house prices be predicted based on these features?
3. How do different models compare in terms of predictive performance for this problem?

By addressing these questions, this project will offer valuable insights into the factors influencing house prices and the performance of various predictive models. The findings will enhance price estimation accuracy, aiding informed decision-making in the real estate market.

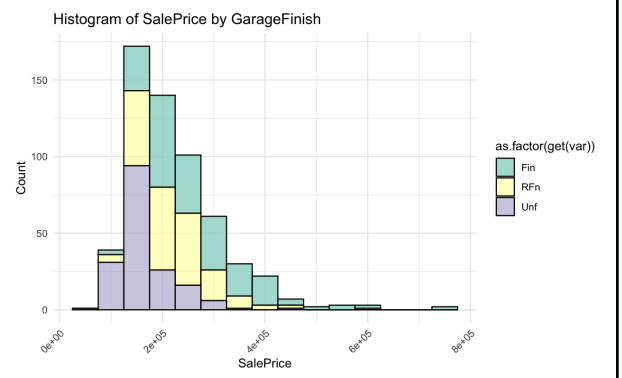
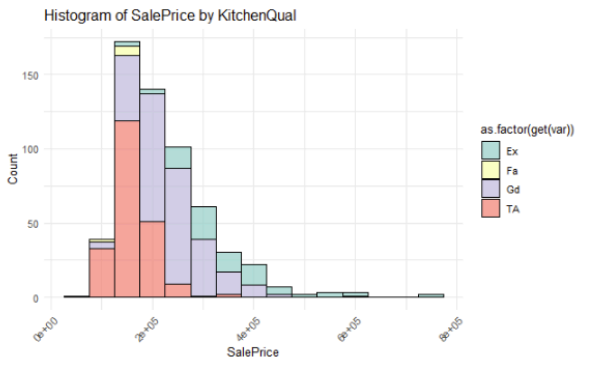
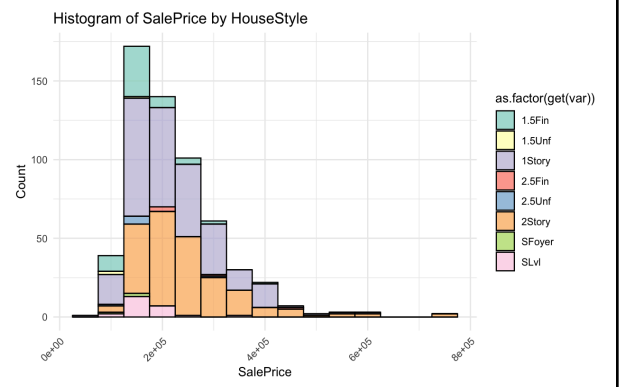
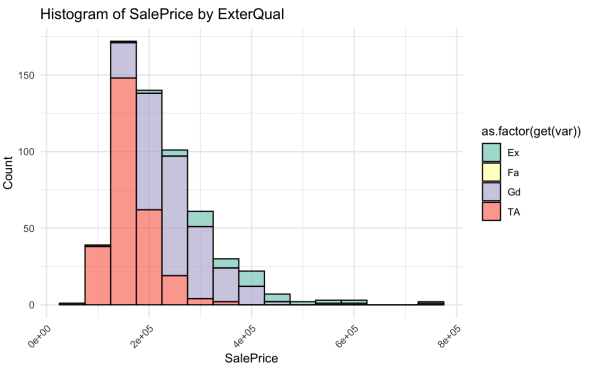
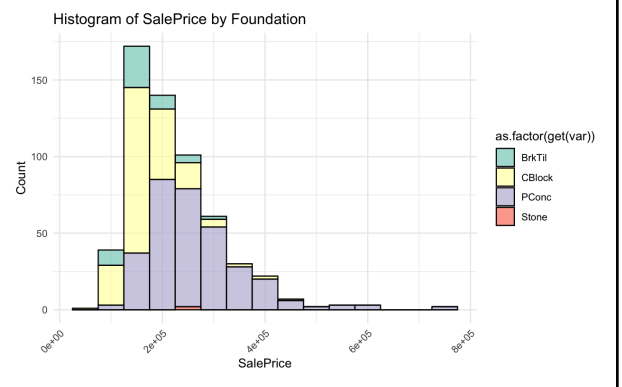
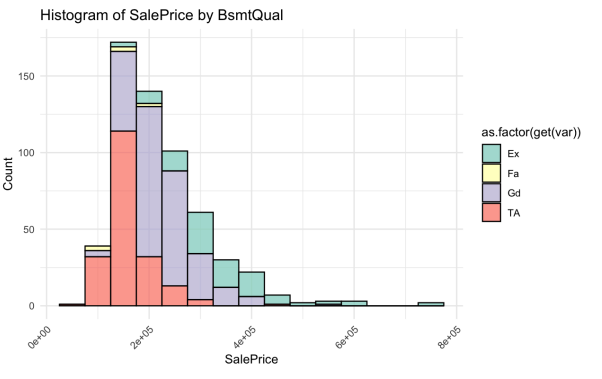
III. Data Information

We plan on using the house prices data set and are trying to predict the sale price of a sold property. The dataset provided a training dataset and a test dataset. The target variable to predict is SalePrice. The main predictor variables are divided into four categories as the following:

1. **Land Characteristics of the House:** LotFrontage, LotArea, LotShape, LandContour, Utilities, LotConfig, LandSlope, PoolArea, PoolQC, Fence, MiscFeature
2. **House Location Characteristics:** MSZoning, Street, Alley, Neighborhood, Condition1, Condition2, MoSold, YrSold, SaleType, SaleCondition
3. **Structural Characteristics:** MSSubClass, BldgType, HouseStyle, OverallQual, OverallCond, YearBuilt, YearRemodAdd, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, MasVnrArea, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinSF1, BsmtFinType2, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, Heating, HeatingQC, CentralAir, Electrical, GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond, PavedDrive

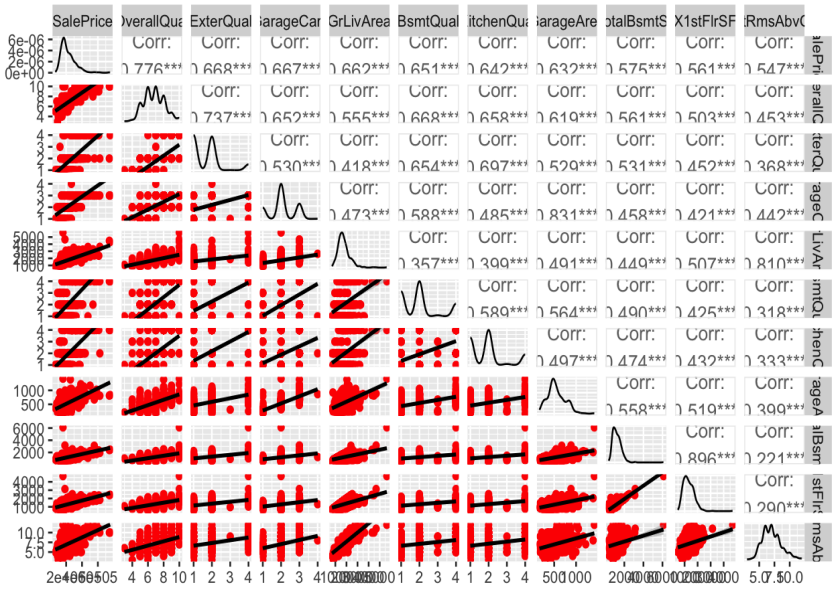
4. **Interior Designs:** 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, Bedroom, Kitchen, KitchenQual, TotRmsAbvGrd, Functional, Fireplaces, FireplaceQu, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch

IV. Data Visualization

Strong Categorical Predictors (nominal)	Strong Categorical Predictors (ordinal)
	
	
	

On the left are the top three nominal variables, and on the right are the top three ordinal variables that appear to be the most influential compared to all other categorical variables. While the other variables

have multiple levels or types, in most cases, only one level or type is present. Therefore, we have decided to use only these six categorical variables, combined with the significant numerical variables, to fit our prediction models.



After performing one-hot encoding on the categorical variables and combining them with the numeric variables, we conducted a correlation analysis to identify the most significant relationships. From this analysis, we selected the top ten variables most strongly correlated with the SalePrice response variable. These variables were chosen based on their ability to explain variations in SalePrice. The figure on the left shows the relationships between these variables and SalePrice, highlighting the strength and direction of their correlations.

From the figure, we can observe several key patterns. Higher overall quality ratings are strongly associated with higher sale prices, reflecting a clear positive linear trend. Homes with better exterior quality ratings tend to have higher sale prices, emphasizing the importance of exterior quality in home valuation. The number of garage spaces is also positively correlated with sale price, indicating that homes with more garage capacity generally sell for higher prices. Larger above-ground living areas are strongly linked to higher sale prices, making this one of the most influential features. Similarly, higher basement quality ratings are associated with higher sale prices, showing that the condition of the basement impacts perceived home value.

Homes with better-rated kitchens tend to sell for higher prices, underscoring the role of kitchen quality in determining home value. Larger garage areas also contribute to higher sale prices, aligning with the importance of space and utility. Greater total basement square footage is associated with higher sale prices, suggesting that the total size of a home matters. Larger first-floor square footage generally leads to higher sale prices, reflecting the value placed on usable space on the main level. Lastly, while more above-ground rooms tend to result in higher sale prices, this relationship is weaker than the connection with the total living area.

V. Models

1. Linear Regression

After fitting the linear regression model, we evaluated the assumptions of normality and constant variance of residuals using the Shaprio-Wilk normality test and the studentized Breusch-Pagan test. Both tests failed, therefore the assumptions of normality of residuals and homoskedasticity were violated. Due to these failures, we cannot rely on the model for statistical inference. We can use the coefficients as estimates but we cannot draw reliable conclusions about the significance of the individual predictors. We cannot determine which variables have the most impact on the sale price.We applied the Box-Cox transformation to the sale price in an attempt to improve the normality and variance assumptions of the residuals. However, the transformation did not significantly improve the heteroskedasticity or normality of the residuals.

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              33.048348   0.049689 665.099  < 2e-16 ***
OverallQual               0.515940   0.065572   7.868 1.92e-14 ***
ExterQual                 0.102000   0.062519   1.632  0.10336
GarageCars                0.211216   0.070776   2.984  0.00297 **
GrLivArea                 0.855875   0.083272  10.278  < 2e-16 ***
BsmtQual                  0.232406   0.053115   4.376 1.45e-05 ***
KitchenQual               0.213625   0.052772   4.048 5.91e-05 ***
GarageArea                0.076165   0.069189   1.101  0.27146
TotalBsmtSF               0.461529   0.101641   4.541 6.89e-06 ***
X1stFlrSF                -0.003879   0.090148  -0.043  0.96570
TotRmsAbvGrd             -0.085386   0.067786  -1.260  0.20834
GarageCars:GrLivArea      0.036618   0.045015   0.813  0.41630
GrLivArea:TotalBsmtSF    -0.141807   0.027037  -5.245 2.24e-07 ***
OverallQual:KitchenQual  -0.025747   0.041051  -0.627  0.53079
GrLivArea:TotRmsAbvGrd   -0.111905   0.048609  -2.302  0.02170 *
BsmtQual:TotalBsmtSF     -0.031705   0.049293  -0.643  0.52036
X1stFlrSF:TotRmsAbvGrd   -0.076451   0.064563  -1.184  0.23687
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8205 on 549 degrees of freedom
Multiple R-squared:  0.8227,    Adjusted R-squared:  0.8175
F-statistic: 159.2 on 16 and 549 DF,  p-value: < 2.2e-16
```

2. Polynomial Regression

After fitting the polynomial regression model, we found that poly(GrLivArea, 2)1, poly(GarageCars, 2)1, poly(TotalBsmtSF, 2)1, OverallQual, ExterQual, KitchenQual, and poly(GrLivArea, 2)2:GarageCars are the significant variables in this model.

Multiple R-squared = 0.7805, and Adjusted R-squared = 0.7742, indicating that the model explains about 78.05% of the variance in the response variable. The F-statistic = 122 with a p-value < 2.2e-16, suggesting that the overall model is statistically significant.

However, the Shapiro-Wilk Test for Normality yields W = 0.97432 and a p-value = 2.101e-08. Since the p-value is very small (< 0.05), we reject the null hypothesis of normality. This indicates that the residuals are not normally distributed, which can affect the reliability of statistical inference.

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      32.93015    0.06022 546.845  < 2e-16 ***
poly(GarageCars, 2)1              5.72315    2.10602   2.718  0.00679 **
poly(GarageCars, 2)2             -0.70092    1.38319  -0.507  0.61254
poly(OverallQual, 2)1            16.24568    1.67280   9.712  < 2e-16 ***
poly(OverallQual, 2)2            -0.21306    1.32393  -0.161  0.87221
TotalBsmtSF                       0.26912    0.10862   2.478  0.01353 *
ExterQual                         0.15182    0.06931   2.191  0.02891 *
KitchenQual                       0.25687    0.05887   4.364 1.53e-05 ***
GarageArea                        0.11091    0.07755   1.430  0.15323
X1stFlrSF                         0.30460    0.09384   3.246  0.00124 **
TotRmsAbvGrd                      0.39945    0.04883   8.180 1.99e-15 ***
poly(GarageCars, 2)1:TotalBsmtSF  0.95007    1.82618   0.520  0.60310
poly(GarageCars, 2)2:TotalBsmtSF  1.23787    1.61466   0.767  0.44362
KitchenQual:OverallQual          -0.08531    0.05227  -1.632  0.10325
TotRmsAbvGrd:GarageCars           0.09206    0.05229   1.760  0.07890 .
TotalBsmtSF:BsmtQual             -0.10737    0.05204  -2.063  0.03956 *
X1stFlrSF:TotRmsAbvGrd           -0.30382    0.04392  -6.917 1.29e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9127 on 549 degrees of freedom
Multiple R-squared:  0.7805,    Adjusted R-squared:  0.7742
F-statistic:   122 on 16 and 549 DF,  p-value: < 2.2e-16

[1] "Shapiro-Wilk Test for Normality (Polynomial Regression):"

        Shapiro-Wilk normality test

data:  residuals(model_poly)
W = 0.97432, p-value = 2.101e-08

[1] "Breusch-Pagan Test for Heteroskedasticity (Polynomial Regression):"

        studentized Breusch-Pagan test

data:  model_poly
BP = 28.394, df = 16, p-value = 0.02836
```

Additionally, the Breusch-Pagan Test for Heteroskedasticity results in BP = 28.394, df = 16, and a p-value = 0.02836. Since the p-value is less than 0.05, we reject the null hypothesis of homoskedasticity. This means the model suffers from heteroskedasticity, indicating that the variance is not constant across observations.

Although the model has a high R-squared value and includes multiple significant predictors, it fails both the normality and heteroskedasticity tests. This means the assumptions required for reliable statistical inference are violated. Since both tests failed, we cannot rely on this model for statistical inference.

3. Lasso Regression

After fitting the lasso regression model, we evaluated the optimal lambda using cross-validation. At the optimal value of lambda, the model shrinks some of the coefficients to zero, which eliminates them from the model. Since the linear model failed the assumptions of normality of residuals and constant variance, the lasso regression model fails as well. Therefore, we cannot infer the causal effects of the predictors on sale price.

```
Optimal lambda: 0.1332473

Coefficients at lambda.min (optimal lambda):
11 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)  32.85372335
OverallQual   0.69199226
ExterQual     0.03842481
GarageCars    0.43787255
GrLivArea     0.37293867
BsmtQual      0.16355991
KitchenQual   0.16873524
GarageArea    .
TotalBsmtSF   .
X1stFlrSF     0.08794116
TotRmsAbvGrd  .
```

4. Regression Tree

```
n= 566

node), split, n, deviance, yval
      * denotes terminal node

 1) root 566 2084.17500 32.74657
   2) OverallQual< -0.2453613 250  330.31380 31.27676
     4) GrLivArea< -0.9203613 67   51.54455 30.23897 *
     5) GrLivArea>=-0.9203613 183  180.19110 31.65672
      10) GarageCars< -0.8654034 50   50.46348 30.92486 *
      11) GarageCars>=-0.8654034 133   92.87884 31.93185 *
   3) OverallQual>=-0.2453613 316  786.49380 33.90939
     6) OverallQual< 0.5257741 159  268.17830 32.98215
      12) GrLivArea< 0.3940442 114  140.04660 32.50571
        24) TotalBsmtSF< 0.5572502 95   86.91023 32.24039 *
        25) TotalBsmtSF>=0.5572502 19   13.01362 33.83227 *
      13) GrLivArea>=0.3940442 45   36.69724 34.18913 *
     7) OverallQual>=0.5257741 157  243.16160 34.84845
      14) GrLivArea< -0.1923688 36   31.29632 33.89726 *
      15) GrLivArea>=-0.1923688 121  169.60270 35.13145
        30) GarageCars< 0.656558 41   47.79374 34.48216 *
        31) GarageCars>=0.656558 80   95.66533 35.46422 *
```

Based on the tree on the left shows that more significant variables appear at the top of the tree, indicating their higher predictive power. The tree prioritizes OverallQual as the most important feature for prediction. Other important factors influencing predictions include: GrLivArea, GarageCars, and TotalBsmtSF. The regression tree model resulted in a tree with 31 nodes and multiple splits, primarily based off of OverallQual, GrLivArea, and GarageCars. The tree's structure provides an easy way to interpet how different feature impact sale price predictions

5. Random Forest

The random forest model was trained using 500 trees with 3 variables considered at each split. The model achieved a mean square residual of 0.76 and explained 79% of the variance of the sale price. We created this model in the hopes of reducing overfitting of the decision tree to provide better generalization.

```
Call:
 randomForest(formula = SalePrice ~ ., data = train_data_clean[,
c(top_vars, "SalePrice")], ntree = 500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.7578483
                    % Var explained: 79.42
```

VI. Model Evaluation & Selection

As shown in the table on the right, all models demonstrate relatively high predictive performance. However, traditional regression-based models such as linear regression, polynomial regression, and lasso regression fail to meet key assumptions, particularly the normality of residuals. Because assumption violations can lead to biased or inefficient predictions, we focus on tree-based models, which do not rely on such assumptions and can effectively handle non-linearity and complex interactions among features.

Linear Regression	MSE for Linear Model: 0.5262387 RMSE for Linear Model: 0.7254231 Success Rate (within ±5%): 97.36842 %
Polynomial Regression	MSE for Polynomial Model: 0.6926505 RMSE for Polynomial Model: 0.8322563 Success Rate for Polynomial Model (within ±5%): 92.10526 %
Lasso	MSE for Lasso Model: 0.6806493 RMSE for Lasso Model: 0.8250147 Success Rate for Lasso Model (within ±5%): 95.61404 %
Regression Tree	MSE for Tree Model: 0.5875826 RMSE for Tree Model: 0.7665394 Success Rate for Tree Model (within ±5%): 95.61404 %
Random Forest	MSE for Random Forest Model: 0.1338438 RMSE for Random Forest Model: 0.3658466 Success Rate for Random Forest Model (within ±5%): 100 %

Among the models, the Random Forest algorithm stands out, achieving the highest predictive accuracy, the lowest MSE, and the lowest RMSE. This superior performance can be attributed to its ensemble learning approach, which reduces overfitting and improves generalization by averaging multiple decision trees. Therefore, our results highlight the effectiveness of ensemble methods like Random Forest for robust and reliable house price predictions in real estate analytics.

VII. Result

Through extensive analysis and comparative evaluation of multiple models, we found that the Random Forest model performed superiorly, achieving the highest predictive accuracy, explaining approximately 79% of the variance in sale price, and demonstrated the lowest mean squared error (MSE = 0.76). Models such as Linear, Polynomial, and Lasso Regression faced critical assumptions violations (e.g., non-normal residuals and heteroskedasticity), limiting their inferential reliability despite decent predictive performances. Decision Tree modeling provided good interpretability but was more susceptible to overfitting. Random Forest stood out due to its ability to handle complexity and reduce overfitting effectively. Therefore, leveraging ensemble methods like Random Forest is highly recommended for reliable house price predictions in real estate analytics.

VIII. Conclusion

Our research found that the random forest model performs better in predicting housing prices, providing more accurate predictions and handling complex data. Traditional regression models (Linear regression, Lasso regression) can help us explain the relationships between some variables, but they cannot meet the statistical assumptions. Therefore, the applicability of traditional regression models is limited. Decision tree models, although intuitive and easy to interpret, are prone to overfitting problems. Our research found that the price of a house is significantly related to its characteristics. For example, the overall quality of the house, living area, garage size, and basement quality are key factors affecting house prices. Our research can provide important reference for both sellers and buyers in the real estate market. In the future, we can explore more machine learning techniques to further improve the accuracy and reliability of housing price predictions.

Appendix

Dataset: https://www.kaggle.com/c/house-prices-advanced-regression-techniques