Linear Regression: Housing Price Prediction

BUS 310 Business Analytics III: Predictive and Prescriptive Business Analytics

Spring 2025 | March 11, 2025

I. Introduction

This paper aims to explore linear regression application in predicting housing prices. The sample dataset is from sold properties in a residential area in Iowa from 2006-2010. This research aims to (1) determine most significant predictors in house prices and (2) determine which regression predicting model performs best. This research is important for real estate participants, including homebuyers, investors, sellers, and agents because accurate housing price predictions have significant implications. By identifying the most significant predictors of house prices, this study provides insights into the key factors that drive property values, helping buyers and sellers make informed decisions.

Previous researchers have explored and confirmed the importance and effectiveness of machine learning in house price prediction by developing and testing regression models. These models were drawn upon foundational real estate characteristics affecting house prices such as residence usability, number of rooms, building properties, floor factors, geographic location, and transportation for value estimation (Zhang, 2021). For example, McCluskey and Borst did research employing Geographically Weighted Regression analysis to identify submarkets, underscoring the importance of segmenting the real estate market by geographic location (McCluskey & Borst, 2011).

Zhang pointed out that traditional multiple regression or feature models are gradually being replaced because of emerging technologies and inherent issues that prevent it from always being the optimal modelling approach (Zhang, 2023). However, in some other research, simple multiple linear regression has been found to be suitable for predicting house prices. Kaushal and Shankar' research conducted a multivariate linear regression model to predict property prices and resulted in a 84.5% accuracy (Kaushal & Shankar, 2021).

For this research, two simple regression models using Mean Squared Error (MSE) and R-squared are trained, tested, and compared to see which one is more effective in predicting house prices. The first model selects 10 most correlated predictors with SalePrice, while model 2 selects all predictors before determining the most significant ones. This paper is structured as follows: data processing steps, model testing, results from analyses, discussion on implications of the findings, and conclusions with study limitations and future research suggestions.

Through this research, the two proposed models are found to be not ideal for predicting house prices. However, model 1 can be used as a base for developing a more refined regression model. The findings from Model 1 are still helpful in accurately forecasting property prices, providing valuable insights into the factors influencing property price, and suggesting factors to improve on a property to ensure competitive housing prices.

II. Data Section

This dataset consists of final house prices for each property sold from a residential area in Ames, Iowa within 2006 - 2010. The dataset consists of the houses and their features, location, information on rooms and square footage, time sold, sale conditions, and prices, presented as both numerical and categorical variables. There are 81 variables with 2919 observations in total in this initial dataset. The dependent variable of this dataset is SalePrice, and throughout this research, SalePrice will be used as a target variable in all regression models and visualizations.

Before analyzing, we processed the data by organizing the data: handling missing values, and processing categorical variables. After converting some variables into dummies, we have a new total of 389 variables. We then carried out an Exploratory Data Analysis.

In the first regression model, we picked out 10 numerical variables most correlated with the SalePrice. After carefully examining, we decided to drop 4 variables, resulting in 6 variables to consider. In the second regression model, we ran the model on all variables and dropped the insignificant ones, resulting in 38 variables.

1. Data Organization

a. Handling Missing Values

Firstly, we identified the missing data by calculating the number of null entries in each column. To achieve this, we used .isnull().sum() function. There were a total of 34 columns with missing values. We then imputed the numerical missing data by assigning the mean and the categorical/object data by assigning the most frequent value. To verify that there are no more missing values, we confirmed with the second .isnull().sum().

b. Categorical Variable Processing

We proceeded the process with preparing categorical data for a later machine learning model. As machine learning models work better with numerical data, we are converting categorical columns into new columns with numerical values that represent each unique category. We first extracted and converted the categorical columns (columns with ‘category’ or ‘object’ data type) into a list. We used the “one-hot encoding” process to create dummy variables from the list with pd.get_dummies(df[cat_vars_col], drop_first=True), then concat these new columns back to the original dataframe. For specification, we used “drop_first=True” to avoid redundancy by removing the first category of each column.

c. Numerical Variable Processing

For later usage of the important numerical data, we created a backup of the dataset. In this section, we converted year and month sold, year built, and MSSubClass from numeric variables into dummy variables. It allows the model to show the differences between individual years, months, and SubClass instead of creating a strictly linear numeric relationship. Those features do not follow a linear numeric relationship with the target SalePrice. Instead, each unique year or subclass can act more like a discrete category, rather than a measure on a continuous numeric scale. We used pd.get_dummies() to perform "one-hot encoding," but instead of dropping the first category for each column, we are keeping all categories by setting "drop_first=False". From that, we retained dummy variables for available categories, which provides better insights in the regression model. By preserving each unique year or subclass as its own column, the model can more transparently compare price differences across all categories. Although dropping the first category helps reduce collinearity, we want to keep all potential categories for interpretative clarity and to avoid losing relevant information. Converting numerical variables to dummy variables is also useful in real estate where houses built or sold in different years or months may reflect economic and market trends.

2. Exploratory Data Analysis

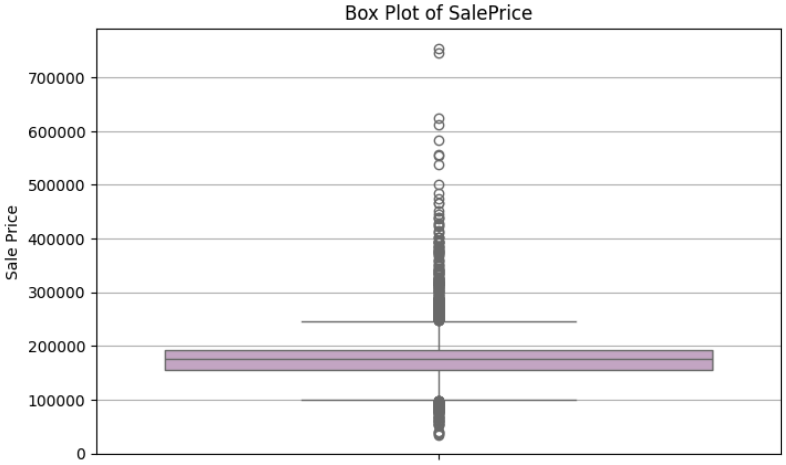


Figure 1: Box Plot of SalePrice

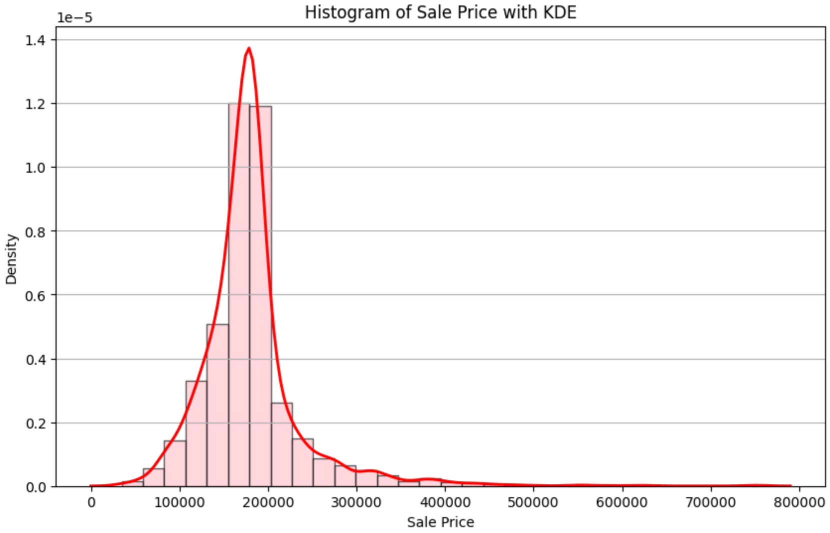


Figure 2: Histogram of SalePrice and KDE

From these 2 charts, the typical sale price appears to be around $175,000 - $200,000. There are few cases of high prices ranging from $300,000 to $750,000. These two charts show a distinct market separation of a dominant "typical" segment and a luxury segment.

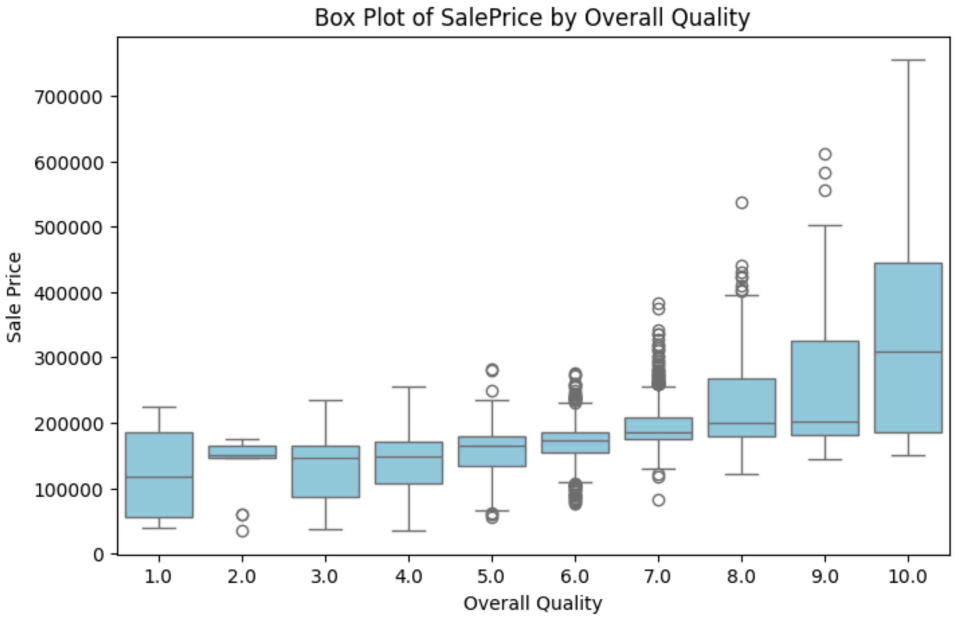


Figure 3: Box Plot of SalePrice by Overall Quality

This chart supports the claim of the separation of 2 market segments: typical and luxury. We can further separate these 2 segments into 3 types of homes by their overall quality: lower quality homes (1-4) with median prices around $100,000-$150,000, mid-quality homes (5-7) with median prices around $150,000-$200,000, and high-quality homes (8-10) with median prices around $200,000-$300,000 and above.

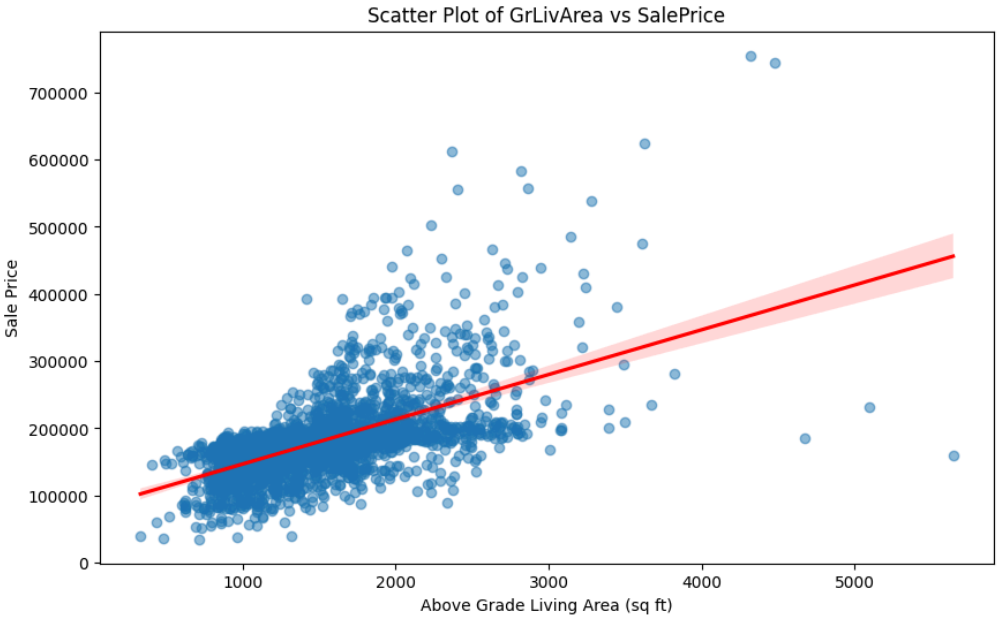


Figure 4: Scatter Plot of GrLiveArea versus SalePrice

We can see a positive linear relationship between GrLivArea and SalePrice, which is indicated by the upward-sloping regression line. House prices significantly increase as living area expands with most houses priced in the range of $100,000 to $300,000 and the area of approximately 1,000 to 2,500 square feet. The outliers above 3,000 square feet achieve higher prices exceeding $400,000. This observation is relevant to the findings that highlight living area as a significant predictor of house prices (Basysyar and Dwilestari, 2022). Pricing patterns highlight market preferences, especially in the periods of economic recovery when larger houses often attract higher valuations. The observed data reflect market behavior during the recovery period following the 2008 financial crisis, where property attributes like size played a crucial role in determining house values.

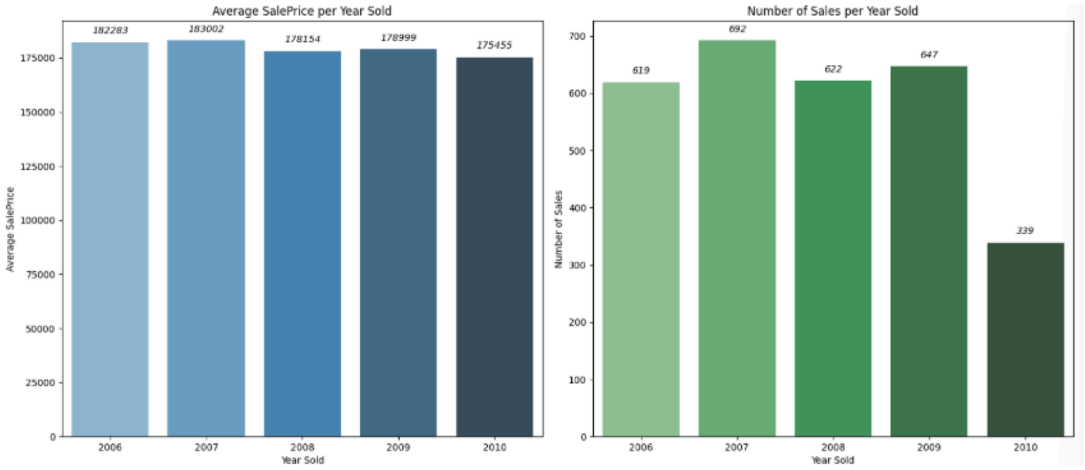


Figure 5: Bar charts of Average SalePrice and number of Sales per Year Sold

The impact of categorical variables on housing prices was shown through bar charts above. The figure showed average SalePrice and number of sales by year, where minor fluctuations can reflect the conditions of the market over time. The average SalePrice bar chart indicates minimal yearly variation in average prices, holding consistently around $175,000. This is the steadiness in market valuation even though the economy changed dramatically during these years such as the financial crisis in 2008. The number in the sales chart reveals a decline in sales volume, especially in 2010. This reduction reflected the lingering impacts of the financial crisis. These charts indicate price resilience despite the decrease in volume, which gives a point to a possible market contraction.

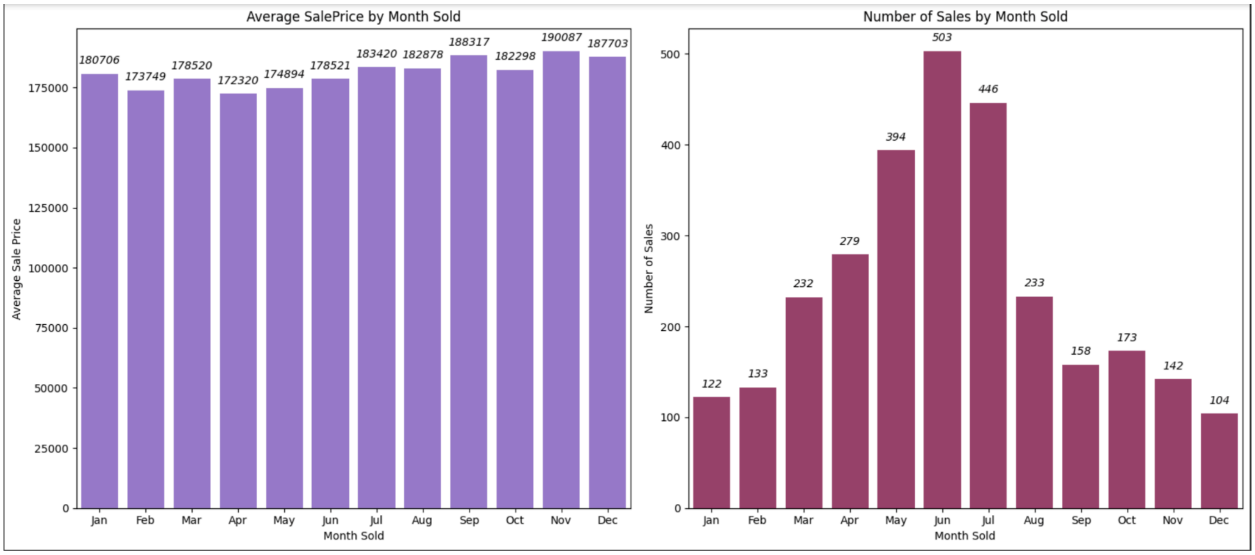


Figure 6: Bar charts of Average SalePrice and number of Sales per Month Sold

The average house prices stayed stable throughout the year, fluctuating around $180,000. The Number of Sales bar chart shows variability in sales volume with a consistent increase in the first half of the year. The transactions increased from approximately 100 sales in the beginning of the year to a peak of 5 times in the summer. However, the volume declined in the second half and ended up with the lowest number of sales in December. This distinct seasonal pattern, peaking during summer, shows that market activity boosted in the middle of the year even though the house prices negligibly changed.

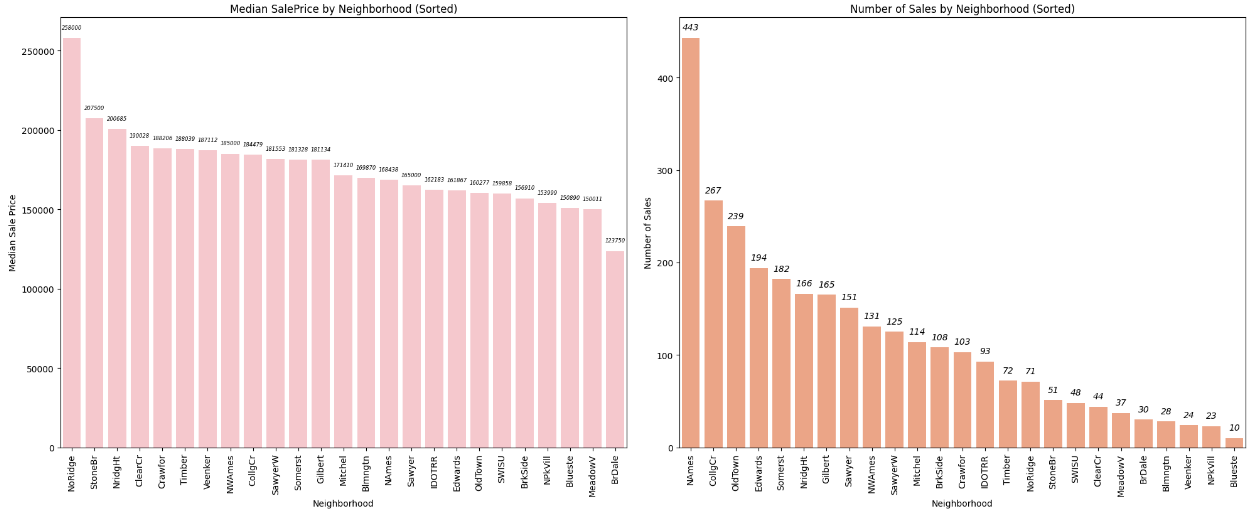


Figure 7: Bar charts of Median SalePrice and number of Sales by Neighborhood

The median SalePrice and the number of sales across different neighborhoods are sorted in descending order. StoneBr is the neighborhood with the highest median price exceeding $230,000, which outperformed the others. Conversely, BrDale represents the lowest value with just half of the top with $123,750. It is interesting that neighborhoods with higher transaction volumes do not command the median prices. NAmes leads with 443 sales, which is significantly higher than 276 of the second highest, CollgCr's. Buyer preferences might prioritize affordability rather than exclusivity, highlighting the factors determining house prices beyond only location (Thamarai and Malarvizhi, 2020).

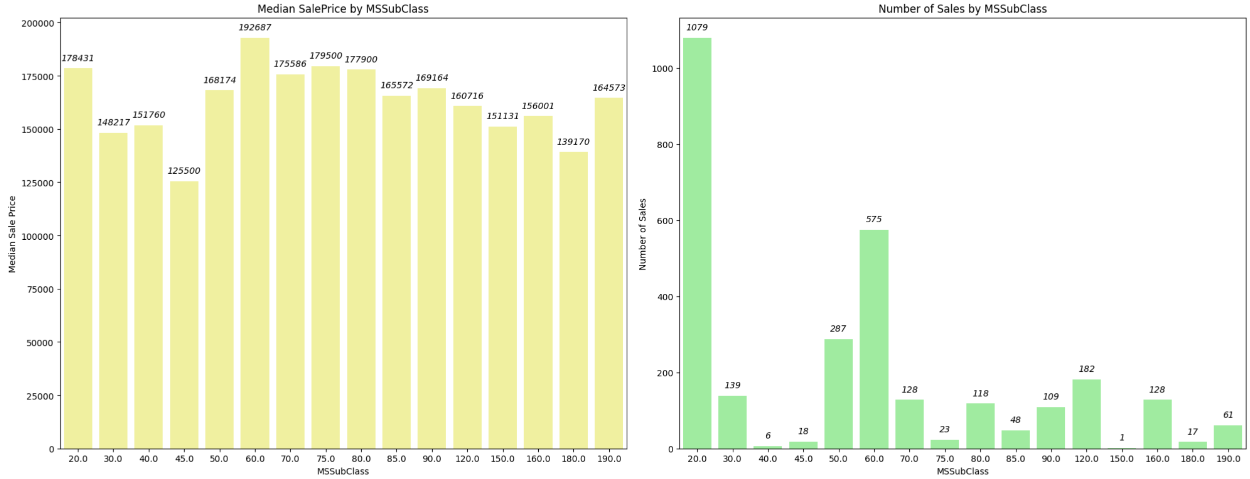


Figure 8: Bar charts of Median SalePrice and number of Sales by MSSubClass

There is a significant concentration of sale on houses with a 20.0 and 60.0 MSSubClass. The market is dominated by class 20.0 dwellings with 1,079 sales and a median price of $178,431. Class 60.0 is the second dominated class, which has both the highest median price of $192,687 and second-highest sales volume of 575 units. This extreme concentration on only 2 categories shows that there is a clear preference, and factors beyond price related to property characteristics, age of the house, etc are strongly influencing buyer decisions.

In conclusion, based on the exploratory data analysis, it appears that certain factors, such as GrLivArea, Overall Quality, Year Sold, Month Sold, Neighborhood, and MSSubClass, may have a significant influence on house prices. These variables show potential as key predictors of housing market trends, and they will be tested in the subsequent analysis.

III. Methods Section

The purpose of this study is to discover the relationship between a continuous variable, which is the house price, and other determinants; therefore, the ordinary least squares (OLS) was chosen as the appropriate approach for this regression analysis. Additionally, this study tests and compares two models to determine the significant independent variables affecting house prices. Model 1 regresses SalePrice on top 10 most correlated predictors, and removes all insignificant predictors imposing collinearity problems, and Model 2 is the regression on all predictors then removes the statistically insignificant ones.

1. Model 1

This model first extracts the top 10 predictors that demonstrate the highest correlation coefficients with SalePrice. The correlation among these variables is inspected by a heatmap, as well as through more measures of multicollinearity such as the Variance Inflation Factor (VIF). In addition, VIF values exceeding 10 may indicate substantial overlap between 2 or more variables, risking unstable regression coefficients and inflated standard errors ((Thamarai and Malarvizhi, 2020). For improving the model stability, highly collinear that larger than 0.8 or statistically non-significant predictors are excluded.

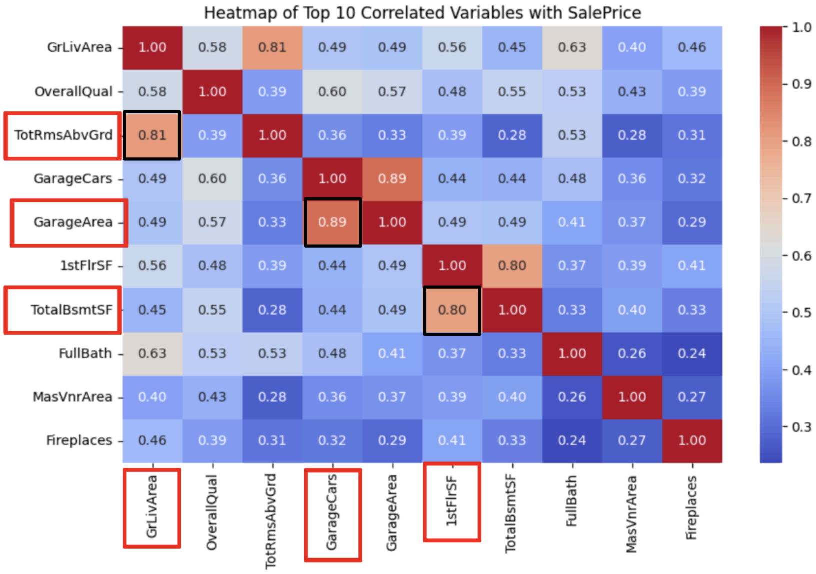
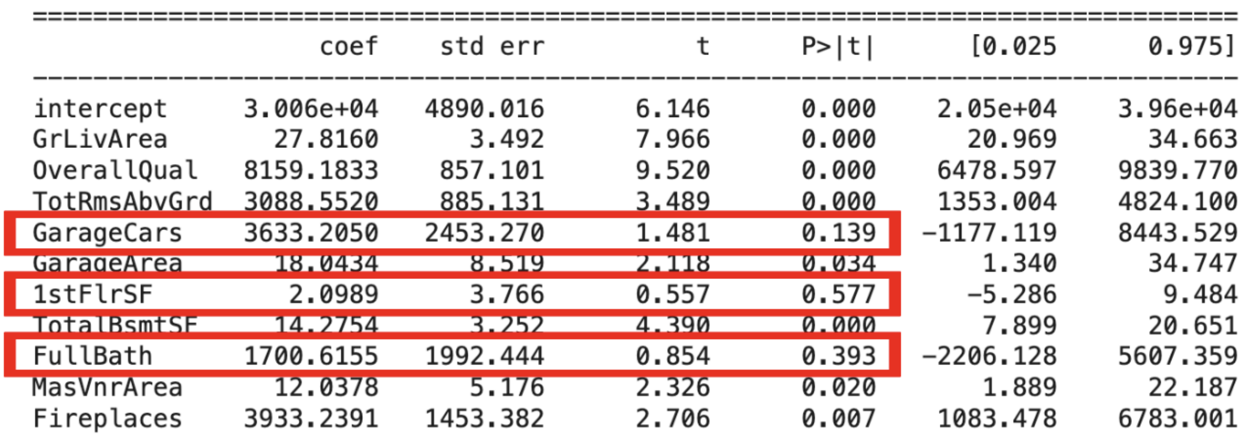


Figure 9: Heatmap of top 10 correlated variables with SalePrice



```
==================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
intercept    3.006e+04   4890.016      6.146      0.000    2.05e+04    3.96e+04
GrLivArea      27.8160      3.492      7.966      0.000      20.969      34.663
OverallQual  8159.1833    857.101      9.520      0.000    6478.597    9839.770
TotRmsAbvGrd 3088.5520    885.131      3.489      0.000    1353.004    4824.100
GarageCars   3633.2050   2453.270      1.481      0.139   -1177.119    8443.529
GarageArea     18.0434      8.519      2.118      0.034       1.340      34.747
1stFlrSF        2.0989      3.766      0.557      0.577      -5.286       9.484
TotalBsmtSF    14.2754      3.252      4.390      0.000       7.899      20.651
FullBath     1700.6155   1992.444      0.854      0.393   -2206.128    5607.359
MasVnrArea     12.0378      5.176      2.326      0.020       1.889      22.187
Fireplaces   3933.2391   1453.382      2.706      0.007    1083.478    6783.001
```

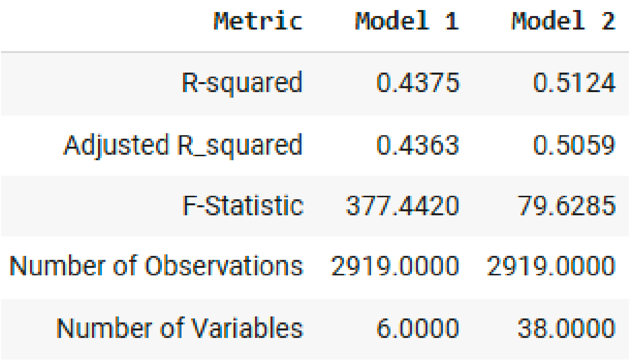
The heatmap reveals a strong correlation between GrLiveArea and TotRmsAbvGrd, with a correlation coefficient of 0.81. This can be attributed to the fact that GrLiveArea represents the square footage of the above-grade living area, while TotRmsAbvGrd, which measures the total number of rooms above grade (excluding bathrooms), can be considered a subset of GrLiveArea. Additionally, high correlations are observed between GarageCars and GarageArea (0.89), as well as between TotalBsmtSF and 1stFlrSF (0.80), indicating overlapping variables. Moreover, model 1 aims to produce an equation highlighting the most influential factors without overburdening the regression with noises. As a result, four independent variables, which are GarageCars (size of garage in cars capacity), TotRmsAbvGrd (total room above ground not including bathrooms), 1stFloor (size of the first floor in square feet), and FullBath (full bathrooms above grade), were dropped among the ten initial predictors with the highest correlation with the sales price.

2. Model 2

This model selected all predictors from the dataset while excluding the target variable, SalePrice. The OLS regression model showed all the correlations between all predictors and the SalePrice. These predictors were then assessed by the p-value, and values with p-value > 0.05 were dropped as they are statistically insignificant. We took the refined list of predictors and built the second OLS model to enhance interpretability and stability. Model 2 ensures that only the most meaningful predictors are taken into account, and it is more accurate when trying to pinpoint the statistically significant variables.

Both regression models went through 80:20 split training and testing phases. The models' performance was evaluated through scatter plots comparing actual and predicted SalePrice and quantitatively by MSE and R^2 values. Model 2 showed a more accurate prediction than Model 1, highlighting the benefits of considering a larger set of predictors.

IV. Results

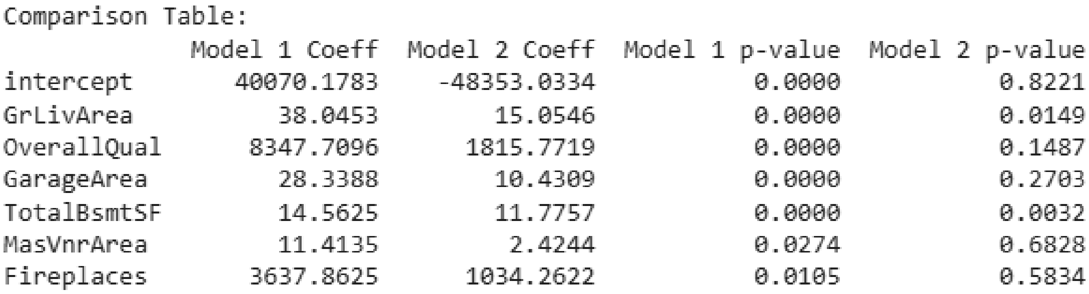


Metric	Model 1	Model 2
R-squared	0.4375	0.5124
Adjusted R_squared	0.4363	0.5059
F-Statistic	377.4420	79.6285
Number of Observations	2919.0000	2919.0000
Number of Variables	6.0000	38.0000

Figure 10: Comparison between Model 1 vs Model 2 Summary Statistics

Overall, model 1 has the R-squared of 0.4375, the high F-stat of 377.44, and a total of 6 statistically significant predictors, including GrLiveArea, OverallQual, GarageArea, TotalBsmtSF, MasVnrArea, and Fireplaces. This suggests that this set of 6 independent variables provides a relatively good fit for the data and explains 43.75% of the variance in house prices. Additionally, the adjusted R-squared is extremely close to the R-squared, signaling that the model does not inflate the R-squared simply by adding more variables. However, the simplicity of model 1, while efficient, may also risk omitted variable bias. It is possible that important predictors are missing from the model, which could lead to an incomplete explanation of the variance in house prices.

Meanwhile, model 2 has the R-squared of 0.5124, the F-stat of 79.62, and includes 38 independent variables. This suggests that model 2 explains a larger portion of the variance in house prices, but the significantly lower F-statistic compared to model 1 indicates a weaker overall model fit. The inclusion of a large number of predictors likely results in a more complex model, but it also raises concerns about overfitting. With 38 predictors, many of which may be irrelevant or highly correlated, the model's ability to generalize to new data is questionable. Additionally, the relatively low adjusted R-squared for model 2 compared to its R-squared indicates that the inclusion of many additional variables is inflating the R-squared without meaningfully improving the model's predictive power.



```
Comparison Table:
             Model 1 Coeff  Model 2 Coeff  Model 1 p-value  Model 2 p-value
intercept       40070.1783    -48353.0334           0.0000           0.8221
GrLivArea          38.0453        15.0546           0.0000           0.0149
OverallQual      8347.7096      1815.7719           0.0000           0.1487
GarageArea         28.3388        10.4309           0.0000           0.2703
TotalBsmtSF        14.5625        11.7757           0.0000           0.0032
MasVnrArea         11.4135         2.4244           0.0274           0.6828
Fireplaces       3637.8625      1034.2622           0.0105           0.5834
```

Figure 12: Comparison of Variables in Model 1 Performance and that in Model 2 Performance

The table compares the set of variables in Model 1 with their performance in Model 2, highlighting notable changes in coefficients and p-values. GrLivArea, which aligns with the hypothesis from the EDA, remained statistically significant in both models, though its coefficient dropped from 38.0453 to 15.0546, and its p-value increased from 0 to 0.0149 in Model 2. While its significance was maintained, the impact of GrLivArea weakened when additional predictors were included. This means that in Model 1, if GrLivArea increases by one additional square foot, the House Price will increase by $38.05, holding other included determinants constant. While in Model 2, the same one square foot increase in GrLivArea will result in a smaller price increase of $15.05, holding other examined independent variables constant. Similarly, TotalBsmtSF, which was not explicitly examined in the EDA, showed stable performance, remaining significant with relatively smaller changes in its coefficient. In Model 1, if TotalBsmtSF increases by one additional square foot, the House Price will increase by $14.56, controlling other included variables. While in Model 2, a one square foot increase in TotalBsmtSF will increase the House Price by $11.78, controlling other examined variables. In contrast, OverallQual, another variable hypothesized to be important based on the EDA, remained significant in Model 1 but lost its statistical significance in Model 2. In Model 1, if OverallQual increases by one additional unit (on its rating scale), the House Price will increase by $8,347.71, holding other examined variables constant. While in Model 2, though this effect is no longer statistically significant (p=0.1487), a one unit increase in OverallQual is associated with a much smaller increase of $1,815.77 in House Price, controlling other variables constant.

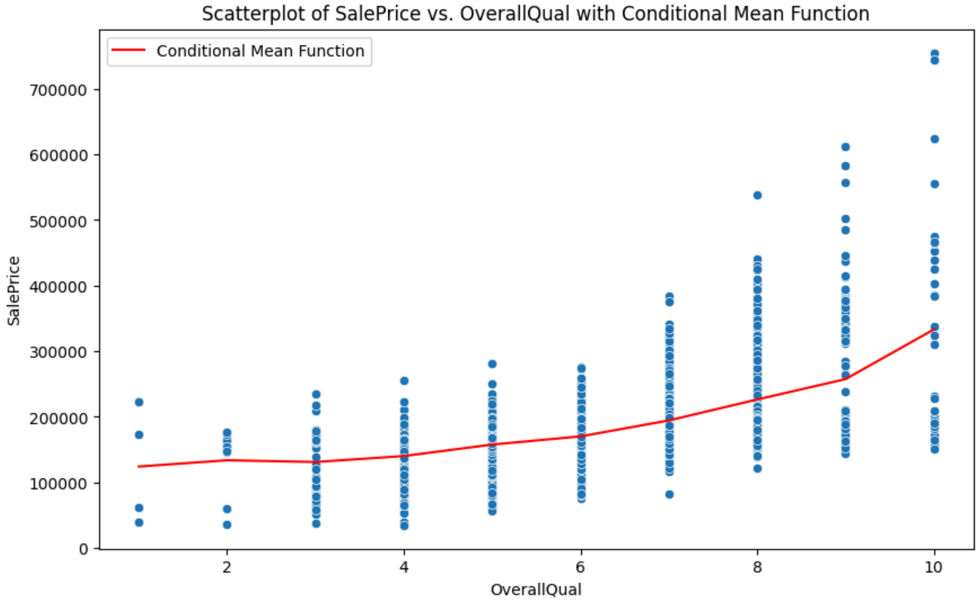
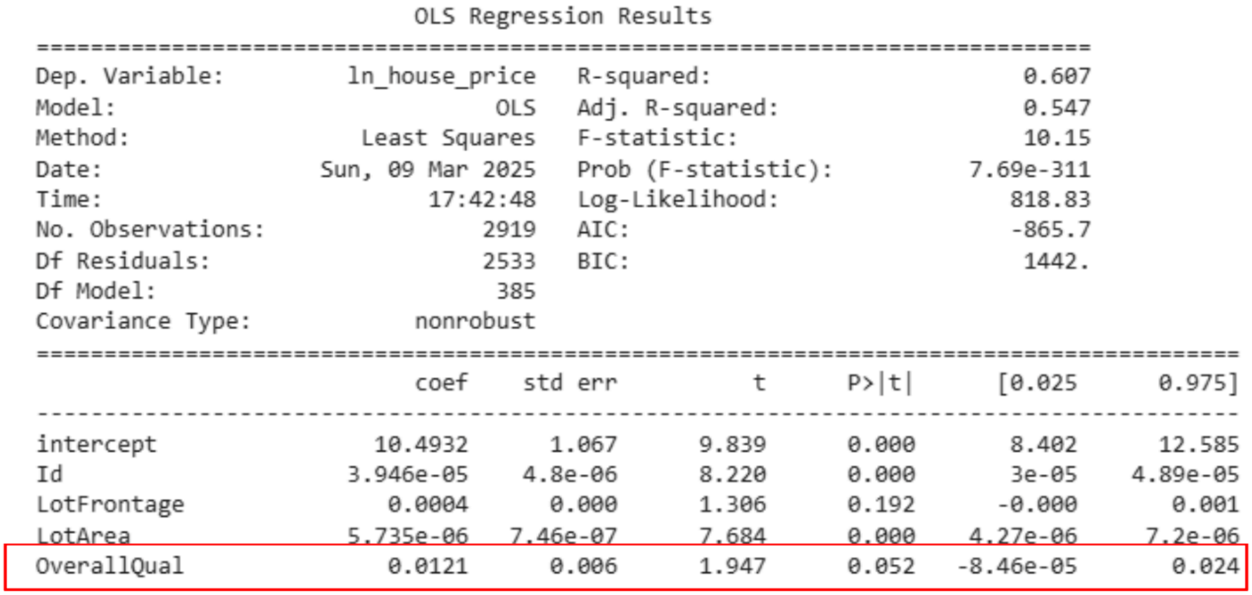


Figure 14: Scatter Plot of SalePrice vs OverallQual with CMF

To further examine the relationship between OverallQual and house price, we created a scatter plot with the Conditional Mean Function. The plot reveals a non-linear, upward-curving relationship between OverallQual and Sale Price, suggesting that increases in OverallQual lead to a multiplicative rather than additive rise in house prices. To better capture this non-linearity, we transformed the dependent variable by using the natural log of house price. However, despite this adjustment, the regression analysis, which regresses the natural log of house price on all predictors, still found the coefficient for OverallQual to be statistically insignificant.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          ln_house_price   R-squared:                       0.607
Model:                             OLS   Adj. R-squared:                  0.547
Method:                  Least Squares   F-statistic:                     10.15
Date:                 Sun, 09 Mar 2025   Prob (F-statistic):          7.69e-311
Time:                         17:42:48   Log-Likelihood:                 818.83
No. Observations:                 2919   AIC:                            -865.7
Df Residuals:                     2533   BIC:                             1442.
Df Model:                          385
Covariance Type:             nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
intercept                10.4932      1.067      9.839      0.000       8.402      12.585
Id                     3.946e-05    4.8e-06      8.220      0.000       3e-05    4.89e-05
LotFrontage               0.0004      0.000      1.306      0.192      -0.000       0.001
LotArea                5.735e-06   7.46e-07      7.684      0.000    4.27e-06     7.2e-06
OverallQual               0.0121      0.006      1.947      0.052   -8.46e-05       0.024
```

These results indicate that, except for GrLivArea and TotalBsmtSF, the remaining four variables in Model 1 became statistically insignificant in Model 2. This shift contrasts with the hypothesis made in the EDA, where these variables were expected to be important predictors. The inconsistency in significance underscores the potential issue with the approach of selecting variables based on their high correlation with SalePrice, as this method may introduce omitted variable bias. In other words, by relying on correlation alone, important variables may be left out, leading to overstated relationships in Model 1. When additional predictors were included in Model 2, they accounted for some of the omitted variance, causing previously significant variables to lose their impact or become overshadowed by interrelated features (Thamarai & Malarvizhi, 2020).

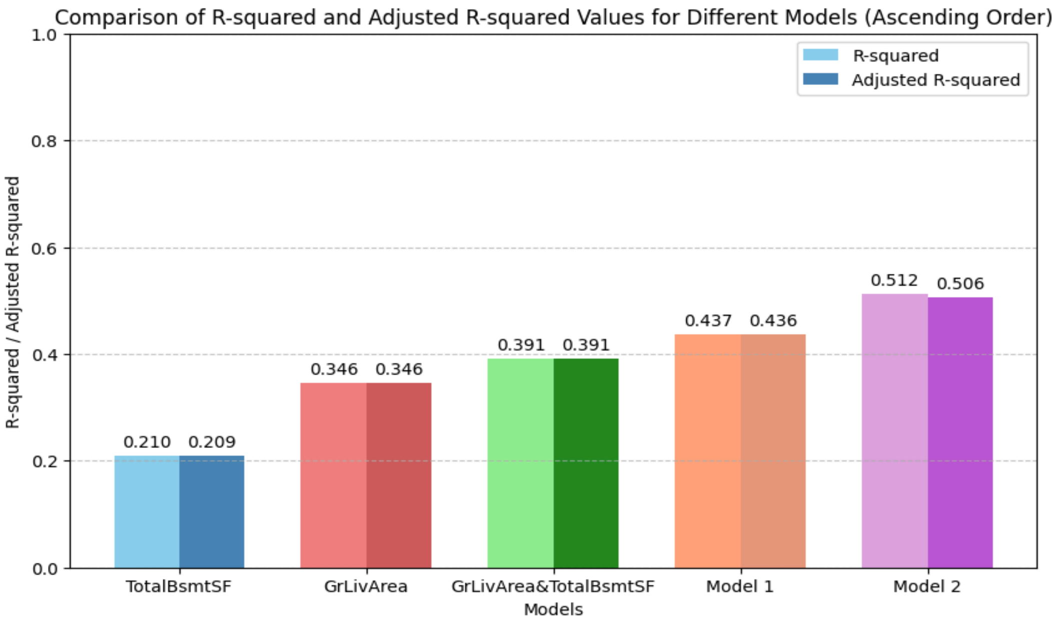
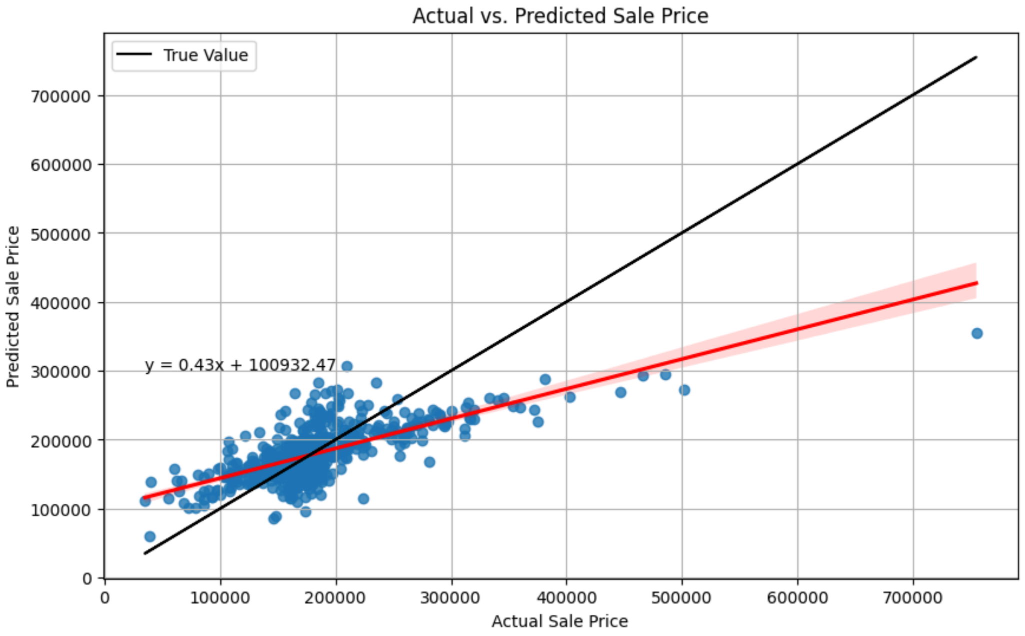
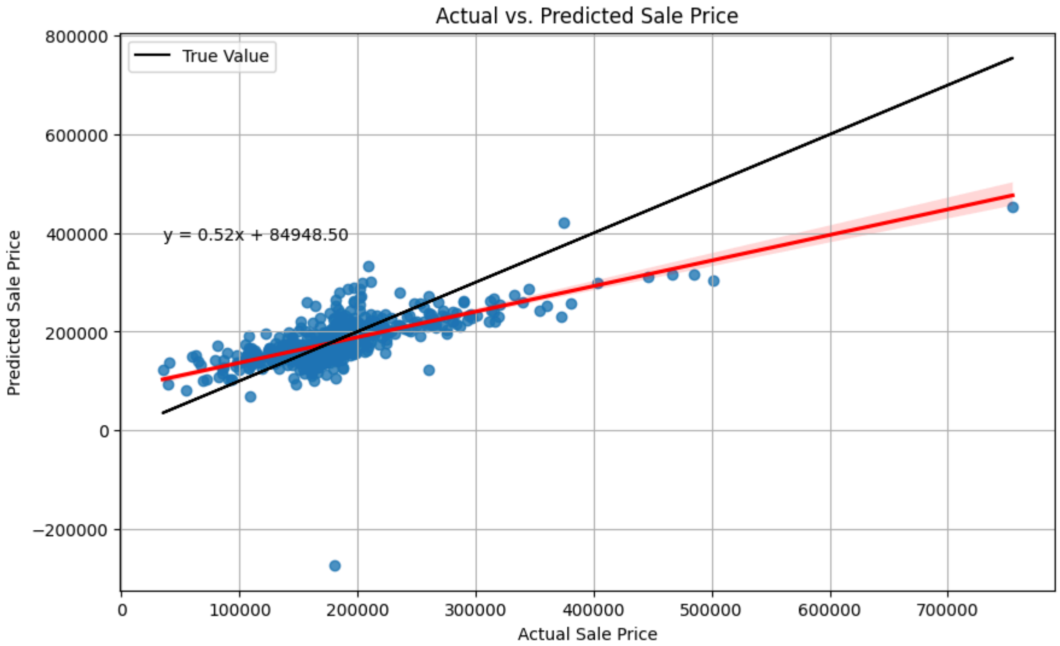


Figure 15: Performance of R-squared and Adjusted R-squared for each Model

We also conducted three additional regression tests. The first regression models the sale price as a function of TotalBsmtSF, the second models the sale price as a function of GrLivArea, and the third models the sale price as a function of both TotalBsmtSF and GrLivArea. These tests were performed using both Model 1 and Model 2, and we observed how the R-squared and adjusted R-squared values changed across the different models to assess the performance over time.



Model 1



Model 2

Figure 13: Comparison of Actual and Predicted Sale Price in Model 1 and Model 2

The scatter plots compare predicted versus actual house prices for both models, displaying fitted regression lines and equations, as well as the ideal y=x line where the predicted price equals the actual price. While Model 1 achieves a lower Mean Squared Error overall, performance differences emerge across price tiers. Both models predict adequately for homes under $300,000. However, for properties above $300,000, Model 2's slope aligns more closely to the ideal 1:1 line, indicating stronger performance in higher-priced segments. Despite this, Model 2 shows a major shortcoming through an extreme negative outlier, yielding a predicted value of roughly -$200,000 for a home actually selling at $200,000. This severe error highlights potential issues, necessitating additional scrutiny.

In conclusion, neither Model 1 nor Model 2 is ideal for predicting house prices. Model 1 may suffer from omitted variable bias due to its reliance on a smaller set of variables, while Model 2's inclusion of too many predictors introduces the risk of overfitting and noise. To identify the optimal model, it would be more effective to start with Model 1 as a base and selectively add relevant predictors based on their contribution to the model's predictive power, rather than simply including all available predictors.

V. Conclusion

Although both models are not ideal and have flaws, we believe that Model 1 performs better than Model 2 as it can serve as the foundation for a more balanced approach. One of our findings from Model 1 suggests that the two significant factors in predicting house price are GrLivArea (Above grade (ground) living area square feet) and Total Basement Square Footage. To put this information into practice, this research is proposing some suggestions for real estate participants. For buyers, it is important to ensure the house has larger above-ground living areas and finished basements before purchasing to invest in a high value property and higher resale price. When considering between two properties, homebuyers should choose the one with a larger above-ground living area, because it is the strongest predictor of the price. With this in mind, sellers when listing the property should highlight the usable living space area and the basement. Additionally, before selling the house, sellers should also consider refinishing the basement and expanding the above-ground living space by building a backyard shed, converting existing attics into living areas, etc. This maximizes the sale price and ensures to catch buyers' attention. Our finding is also applicable for real-estate agents, as they can use these two predictors to price a property. By conducting comparative market analysis on ground living area and basement area, agents can use the insights to list the accurate and competitive pricing for each property.

It is important to acknowledge that there are limitations in this study. One of them is the limited generalizability. While the dataset offers a mass variety of observations, it is restricted to residential property sales in Ames, Iowa. Therefore, there are specific factors such as location, transportation, and market preferences that make it limited to generalize with other regions or states. Another limitation associated with the study is the utilization of a relatively outdated dataset covering the housing price between 2006 and 2010. Given the significant changes in housing prices and consumer behavior over the past 15 years, the findings may not fully reflect current market trends. Additionally, both models are proved to be not ideal, with model 1 having omitted variable bias and model 2 having risks of noises and overfitting variables. Further research should be conducted with a more recent dataset covering house prices of more regions within the United States, and a more refined model should be built based on the groundwork of Model 1.

VI. References

Basysyar, F. M., & Dwilestari, G. (2022). House price prediction using exploratory data analysis and machine learning with feature selection. *Acadlore Transactions on AI and Machine Learning, 1*(1), 11-21. https://doi.org/10.56578/ataiml010103

Kaushal, A., & Shankar, A. (2021). House price prediction using multiple linear regression. *Proceedings of the International Conference on Innovative Computing & Communication (ICICC), 2021.* https://ssrn.com/abstract=3833734 or http://dx.doi.org/10.2139/ssrn.3833734

Zhang, L. (2023). Housing price prediction using machine learning algorithm. *Journal of World Economy, 2(3),* 18–26. https://www.pioneerpublisher.com/jwe/article/view/392

McCluskey, W. J., & Borst, R. A. (2011). Detecting and validating residential housing submarkets: A geostatistical approach for use in mass appraisal. *International Journal of Housing Markets and Analysis, 4(3),* 290-318.

Thamarai, M., & Malarvizhi, S. P. (2020). House price prediction modeling using machine learning. *I.J. Information Engineering and Electronic Business, 12*(2), 15-20.

Zhang, Q. (2021). Housing price prediction based on multiple linear regression. *Scientific Programming, 2021,* 7678931. https://doi.org/10.1155/2021/7678931