Assessing Home Price Determinants: A Regression Analysis of Ames Housing Data

Ryan Croce Keely Kinnane Sai Teja Yapuram Kavindu Wellalage

2024-04-26

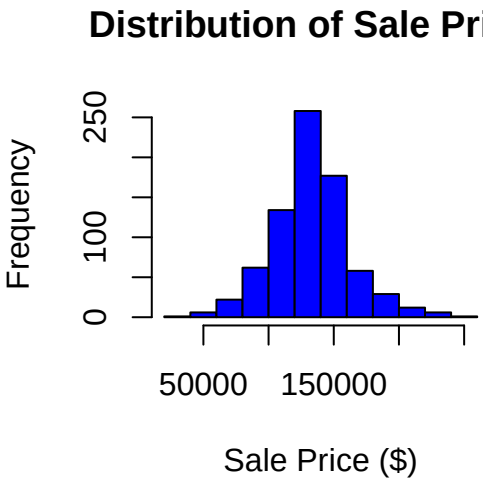
Section 1: Introduction

This project aims to explore the factors influencing the sales price of homes in Ames, Iowa, using a subset of the Ames Housing dataset. The dataset, curated by Dean De Cock from Truman State University, contains information on 766 homes with 28 variables. The subset was created by following these steps:

- Removed large houses with an above-ground living area of more than 1500 square feet.
- Filtered the data set to include houses that had a `Normal` sales condition.
- Included the top five neighborhoods with the most houses for sale.
- Removed rows with missing values.
- Renamed a few variables to make their meaning clearer.

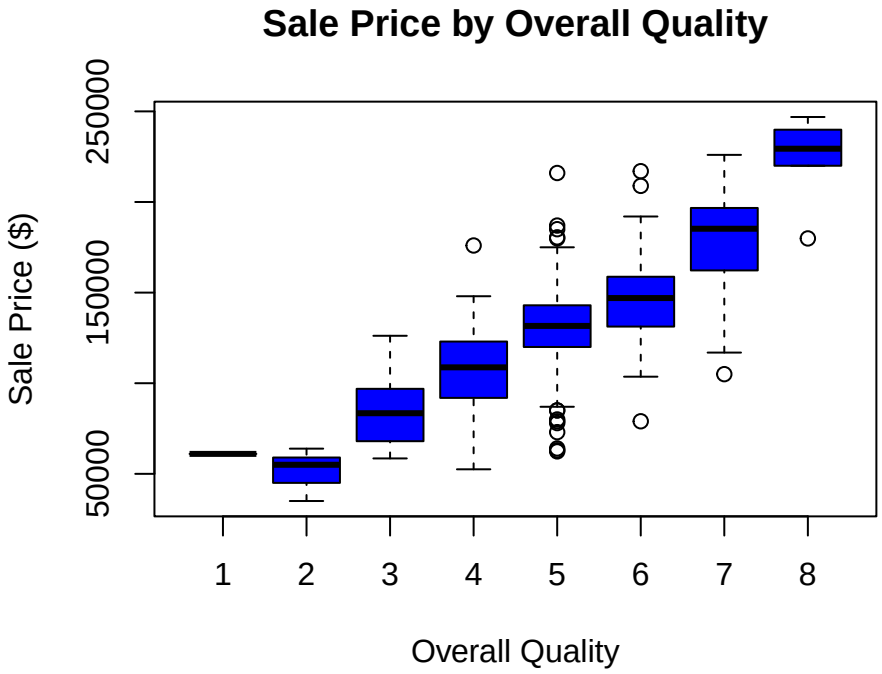
Our goal is to identify the most relevant predictors of home prices and develop an accurate predictive model. This information can be valuable for various stakeholders in the housing market, such as home sellers, policymakers, and real estate professionals.Identifying the most relevant predictors of home prices can help policymakers and urban planners in developing strategies to promote affordable housing, improve neighborhood amenities, and support sustainable urban development. Developing an accurate predictive model for home prices can assist in automated valuation processes, real estate portfolio management, and risk assessment in the lending industry. Overall, this study aims to provide actionable insights into the Ames, Iowa housing market, which can benefit a wide range of stakeholders and contribute to data-driven decision-making in the real estate sector.

To begin our analysis, we conducted exploratory data analysis on key variables and their relationships with the sale price. The distribution of sale prices is right-skewed, and we observed positive relationships between overall quality, year built, and above-grade living area with the sale price.

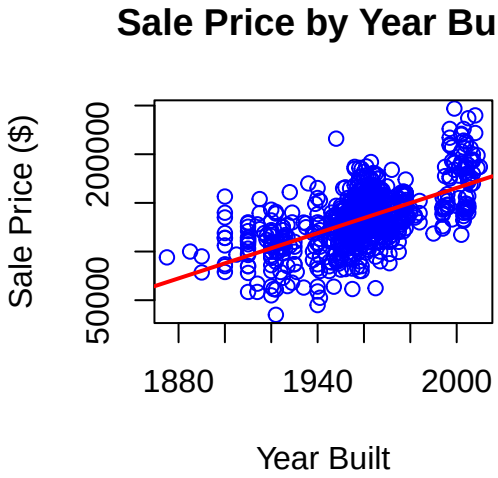


The histogram shows the overall distribution of sale prices in the subset of data. We can observe that the

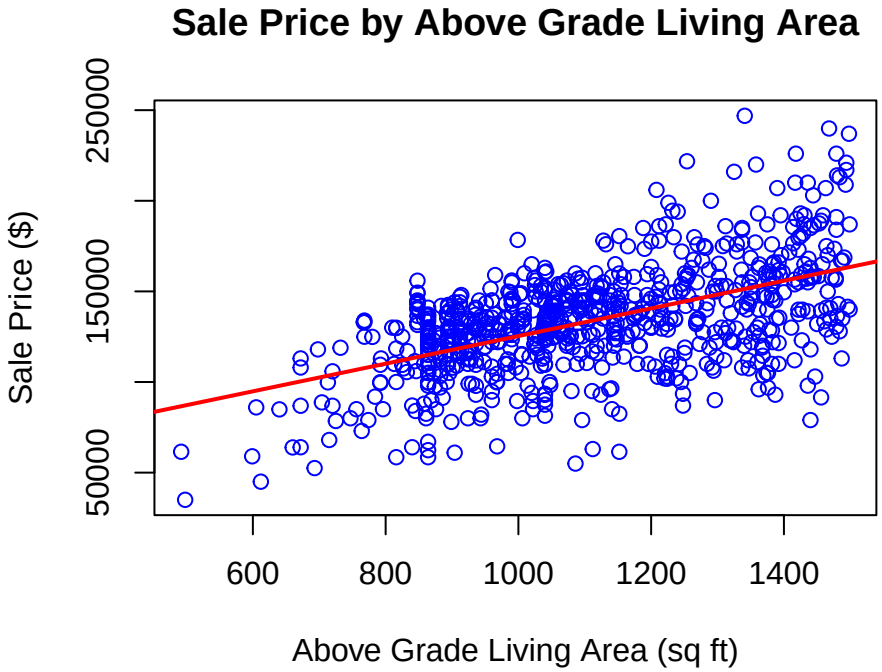
sale prices are right-skewed, indicating that there are fewer homes with very high prices compared to the majority of the homes.



The boxplot displays how sale prices vary based on the overall quality rating of the houses in the subset. We can see a clear positive relationship between overall quality and sale price, with higher quality homes generally having higher sale prices.



The scatterplot with a regression line shows the trend in sale prices over time based on the year the houses were built, using the subset of data. We can observe a general increase in sale prices for newer houses, indicating that the age of the house may have an impact on its value.



The scatterplot with a regression line demonstrates the relationship between the above-grade living area and sale prices in the subset of data. We can see a positive correlation, suggesting that larger living areas are associated with higher sale prices.

Based on these initial findings, we hypothesize that the variables `LotArea`, `Neighborhood`, `OverallQual`, and `OverallCond` are the most relevant predictors for the response variable `SalePrice`. By leveraging the relationships between the predictor variables and the response variable, we aim to create a model that can accurately estimate the sale price of a home based on its features. An accurate predictive model can assist in pricing strategies, investment decisions, and market analysis.

The primary goal of creating a model for the Ames Housing dataset is to accurately predict home prices based on the most relevant variables. By identifying the key factors that influence sale prices, the model aims to provide a reliable estimate of a home's value given its specific characteristics. Achieving this goal involves several sub-objectives:

- `Identifying the most important predictors`: The model should be able to determine which variables have the strongest impact on home prices, allowing for a focused analysis of the key drivers in the housing market.
- `Quantifying the relationship between predictors and sale prices`: The model should provide meaningful coefficients or feature importances that quantify the effect of each predictor variable on the sale price, enabling a clear understanding of how different factors contribute to home values.
- `Minimizing prediction errors`: The model should be optimized to reduce the difference between predicted and actual sale prices, ensuring that the estimates are as accurate as possible. This involves selecting an appropriate model architecture, tuning hyperparameters, and validating the model's performance on unseen data.
- `Providing interpretable insights`: The model should be interpretable, allowing for clear communication of the relationships between predictors and sale prices to stakeholders. This may involve using techniques such as feature importance plots, partial dependence plots, or other visualizations to present the model's findings in an accessible manner.

By achieving these goals, the model can serve as a valuable tool for homeowners, buyers, real estate professionals, and policymakers in the Ames, Iowa housing market. It can assist in pricing decisions, investment strategies, and policy formulation, ultimately supporting a more informed and efficient housing market.

In the following sections, we will develop and evaluate models to test this hypothesis and identify the most important factors influencing home prices in the subset of the Ames Housing dataset.

Section 2: Regression Analysis

The data set was split into a training and test set, so that we could perform valid inferences about the relationship of certain predictors with the sales price of homes. We also ensured that the neighborhoods in the data set were appropriately identified as categorical predictors. We then began the process of variable selection by using the best subset method with AIC and adjusted R^2 criteria for variable selection. The choice of the following selection procedure was due to the fact that it's an exhaustive method that can check all possible models. The two models chosen from the search algorithm were then assessed using the rmseloocv, which gives us an estimate of the test rmse, to determine the best fitting model using this method.

Table 1: RMSEloocv for Quality Criterion

Criterion	RMSEloocv
AIC	12417.59
Adjusted R^2	12465.33
Hypothesized Model	19399.57

We hypothesized that the variables `LotArea`, `Neighborhood`, `OverallQual`, and `OverallCond` are the most relevant predictors for the response variable 'SalePrice and would be chosen during variable selection. The model chosen through AIC criteria was far larger and consequently had a lower test RMSE.

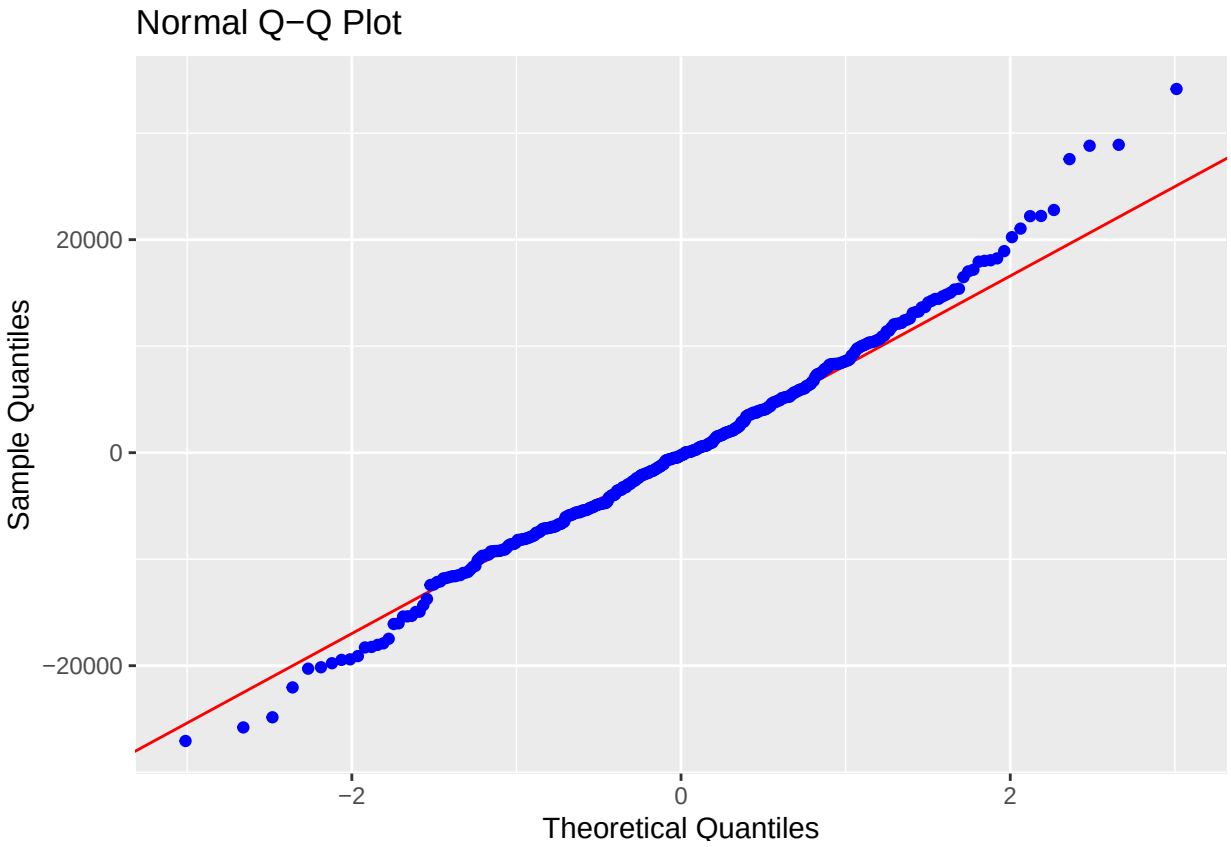
Once chosen, the model was fit to the test data for further analysis. We first checked whether there were issues with collinearity since this would decrease the power of our hypothesis test. The condition number was 910, which seems to be due to the relationship between the categorical predictors. The VIFs, as shown in the table below, were all below 5, except for NeighborhoodOldTown, which again likely has to do with the relationship between the categorical predictors. We should be cautious about the large condition number, but since the VIFs are relatively low, we're not losing too much power relative to the uncorrelated case.

Table 2: VIF Values for Regression Variables

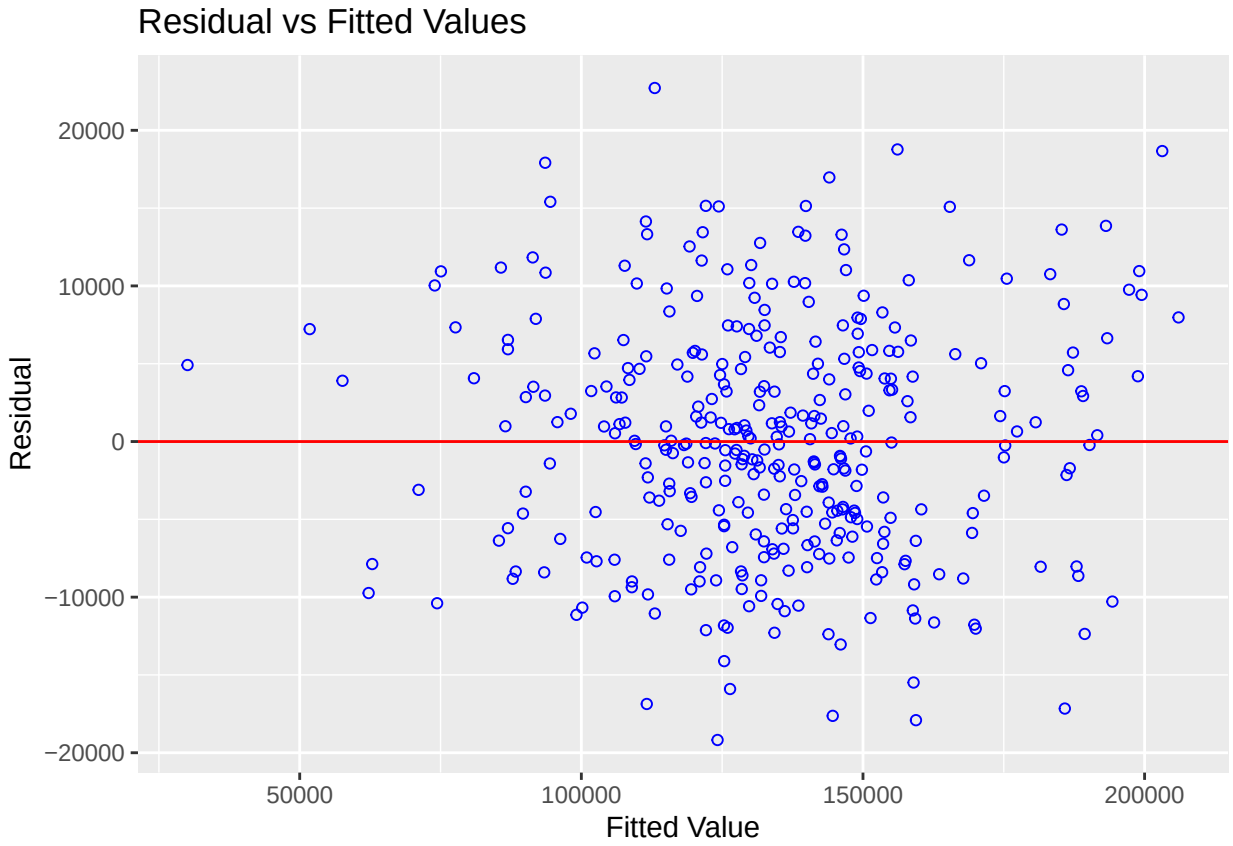
Variable	VIF
LotArea	1.300574
NeighborhoodEdwards	2.649184
NeighborhoodNAmes	3.966050
NeighborhoodOldTown	5.012429
NeighborhoodSawyer	2.275655
OverallQual	1.993924
OverallCond	1.701585
YearBuilt	3.572031
YearRemodAdd	1.845922
BsmtFinSF1	1.913897
TotalBsmtSF	1.882527
GrLivArea	3.111682
BsmtFullBath	1.679272
BsmtHalfBath	1.112376
FullBath	1.772713
BedroomAbvGr	1.996536
KitchenAbvGr	1.570560
TotRmsAbvGrd	2.874547
Fireplaces	1.344416

Variable	VIF
GarageCars	1.335036
WoodDeckSF	1.170969
OpenPorchSF	1.164659
ScreenPorch	1.146769

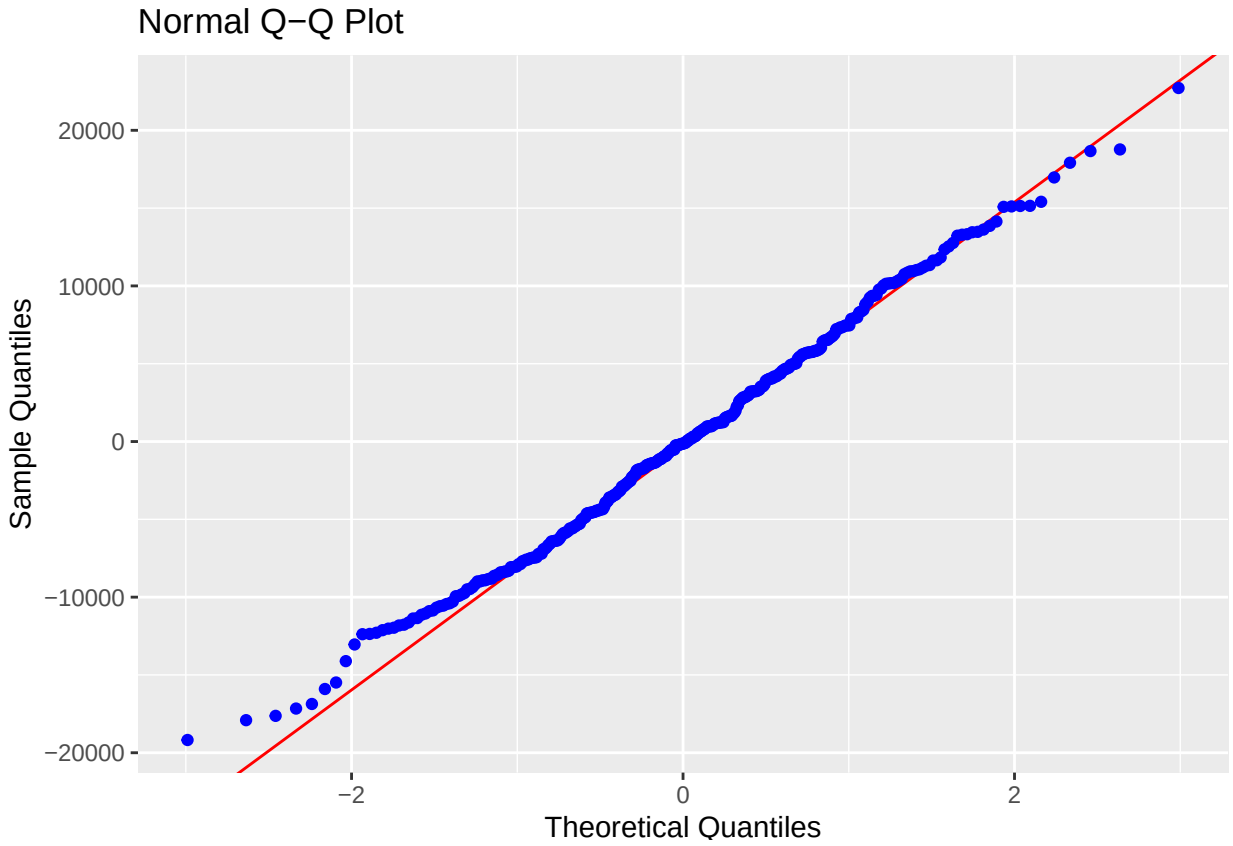
The LINE assumptions were checked using the following plots and tests to determine if there were any violations. Our first analysis before checking for any highly influential observations showed that only the normality assumption was violated. The Shapiro-wilks test had a p-value of .03, leading us to reject the null hypothesis at the 5% significance level and conclude that the errors did not follow a normal distribution. The Q-Q plot also showed moderate discrepancies from the line at the tails, as seen in the plot below.



The highly influential observations were then removed from the model and the model assumptions were checked again. We concluded that the linearity assumption wasn't violated by checking the fitted versus residual plot, which showed that the residuals were centered around zero. The spread of the residuals remains constant, which suggests that the constant variance assumption holds.



The Breush-Pagan test was used to validate what was suggested by the fitted versus residual plot. The reported p-value from the BP-test was .465, so we fail to reject the null hypothesis at any reasonable significance level and conclude that the errors are homoscedastic. The Q-Q plot was then checked after removing the highly influential points, giving us the following plot. Observe that the tails closely follow the 45 degree line, suggesting normality is not violated. The p-value from the shapiro-wilk test was .35, so we again fail to reject the null hypothesis at any reasonable significance level and conclude that the errors follow a normal distribution.



The final model after checking for collinearity, removing highly influential observations, and verifying the LINE assumptions:

```
library(jtools)
summ(model_fix_aic)
```

```
## MODEL INFO:
## Observations: 358
## Dependent Variable: SalePrice
## Type: OLS linear regression
##
## MODEL FIT:
## F(23,334) = 194.94, p = 0.00
## R² = 0.93
## Adj. R² = 0.93
##
## Standard errors: OLS
## ------------------------------------------------------------------
##                                   Est.       S.E.   t val.      p
## ------------------------- ------------- ---------- -------- ------
## (Intercept)                -1140327.95   76642.20   -14.88   0.00
## LotArea                          1.24       0.19     6.59   0.00
## NeighborhoodEdwards          -6632.58    1994.20    -3.33   0.00
## NeighborhoodNAmes             4015.77    1642.98     2.44   0.02
## NeighborhoodOldTown             76.93    2401.15     0.03   0.97
## NeighborhoodSawyer            3675.31    1873.68     1.96   0.05
## OverallQual                   6755.23     614.61    10.99   0.00
## OverallCond                   4651.67     466.54     9.97   0.00
```

```
## YearBuilt                        452.35      34.68    13.04   0.00
## YearRemodAdd                     121.07      27.26     4.44   0.00
## BsmtFinSf1                         5.04       1.82     2.77   0.01
## TotalBsmtSf                       12.90       1.80     7.16   0.00
## GrLivArea                         36.51       3.45    10.59   0.00
## BsmtFullBath                    4320.65    1101.04     3.92   0.00
## BsmtHalfBath                    2167.45    1686.99     1.28   0.20
## FullBath                        5622.58    1514.72     3.71   0.00
## BedroomAbvGr                   -1248.01    1004.36    -1.24   0.21
## KitchenAbvGr                   -7163.52    3435.59    -2.09   0.04
## TotRmsAbvGrd                    1867.20     735.36     2.54   0.01
## Fireplaces                      4221.25     914.26     4.62   0.00
## GarageCars                      6027.75     746.92     8.07   0.00
## WoodDeckSf                        13.93       3.81     3.66   0.00
## OpenPorchSf                       -0.88       7.92    -0.11   0.91
## ScreenPorch                        1.67       9.23     0.18   0.86
## ---------------------------------------------------------------------
```

Section 3: Discussion

The model, chosen through a thorough variable selection process that satisfies all the LINE conditions, should give us confidence that our test will accept and reject correctly. The lack of significant collinearity means our model's power isn't drastically affected, which also means our inferences from our tests should be valid. Policymakers and business professionals in the housing industry can use our model to clarify the relationship of certain factors with sales price like lot area and material quality.

An F-test for the significance of the model had an F-statistic of 194.9 with a p-value less than 2.2×10^{-16}. Therefore, we reject the null hypothesis that all the coefficients are zero. We conclude at the 5% significance level that at least one of the predictors in our model has a significant linear relationship with the sales price of homes, given that the other predictors are in the model. The R^2 for our model was .93, which indicates that 93% of the observed variation in sales price is explained by the predictor variables.

A nested comparison model lets us verify whether the inclusion of the categorical variable Neighborhood categories are significant, where College Creek is the default reference level. The value of the F-test statistic is 15.842. The p-value of the test is 6.948×10^{-12}, so we will reject the null hypothesis at the $\alpha = 0.05$ significance level. As such, we reject the null hypothesis and conclude that at least one of the neighborhood categories is significant.

A major benefit of our model is understanding what factors have significant linear relationships that can be important to professionals who want to focus on key factors. For example, looking at the overall quality predictor, we can run a t-test and notice that the overall quality of materials has a significant linear relationship with the sales price, with the other predictors in the model. The t-value was 10.9 and the p-value is less than 2e-16. The table in section 2 shows the significance of individual predictors for reference.

When selecting our model during the variable selection process, we checked the test RMSE by use of the RMSEloocv to determine which model best fit the data. Therefore, we can say that our model was the best for the chosen quality criterion and gave us the smallest test RMSE. Our model can also be used for predictions about sales price using chosen values for sets of predictors. For example, we can take an observation from the data set, which had a sales price of $119,000, and modify the overall quality, year built, and lot area, and build a 99% confidence interval for the mean price of homes with these specific factors and a 99% prediction interval for an individual home.

Suppose the overall quality of the materials increased from 6 to 8, then the 99% confidence interval for the mean sales price would be the following:

```
##         fit      lwr      upr
## 51 141989.1 136224.8 147753.3
```

The prediction interval for an individual home with the same properties:

```
predict(model_fix_aic, newdata = observation, interval = 'prediction', level = .99)
```

```
##         fit      lwr      upr
## 51 141989.1 121073.3 162904.8
```

Section 4: Limitations

The major limitation of our model is the reduction of observations that were flagged as highly influential. The reduction in observations can impact the the power of our tests. This is the cost of attempting to build a model that is not only for predictions, but to provide valuable inferences to policy makers and professionals. The model condition number of 910 is another cause for concern but, as stated in section 2, may not have too much of an impact on the power of the model. If prediction is preferred, then future models can be built on the total data set instead of splitting it for valid inference.

Section 5: Conclusions

In this project, we aimed to explore the factors influencing the sales price of homes. By leveraging the relationships between the predictor variables and sales price, we built a model that can estimate the average sale price for a home, for a given set of predictor values. Our analysis involved a thorough variable selection process, ensuring a model that fits the model under the specified quality criteria. The LINE conditions were satisfied ensuring that we have valid inferences about the relationship between certain predictors and the sales price of homes. Key results from our process of model selection and from the model itself are the following: the final model explained 93% of the observed variation in sale price, an F-test showed the significance of our model with at least one of the predictors having a significant linear relationship, individual t-test revealed that several predictors have significant linear relationships with sale price.

#Code Appendix

```
#Libraries used
library(leaps)
library(olsrr)
library(faraway)

#Grab the data set
data = read.csv('ames_housing.csv')

#Converting Neighborhood into a categorical predictor

data$Neighborhood = as.factor(data$Neighborhood)

#Data Splitting for variable selection

# Random seed for reproducability

set.seed(42)

#Randomly sample 50% of the obsercation for training
train = sample(1:nrow(data), round(.5*nrow(data)))

test = -train

data_train = data[train,]
data_test = data[test,]
```

```r
#Best subset selection for ajdr2
n = nrow(data_train)


mod_subsets = summary(regsubsets(SalePrice ~., data=data_train, nvmax=28))
coef_names = colnames(mod_subsets$which)

best_r2_ind = which.max(mod_subsets$adjr2)

coef_names[mod_subsets$which[best_r2_ind,]]

#Best subset selection AIC
p = ncol(mod_subsets$which)
mod_aic = n * log(mod_subsets$rss / n) + 2 * (2:p)
best_aic_ind = which.min(mod_aic)
coef_names[mod_subsets$which[best_aic_ind,]]

# Checking Test RMSE

subset_model_aic = lm(SalePrice ~ LotArea + Neighborhood + OverallQual + OverallCond + YearBuilt + Yearl

subset_model_r2 = lm(SalePrice ~ LotArea + Neighborhood + OverallQual + OverallCond + YearBuilt + YearR

calc_loocv_rmse = function(model) {
  sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2))
}

calc_loocv_rmse(subset_model_aic)

calc_loocv_rmse(subset_model_r2)

calc_loocv_rmse(hypothesis_model)

#Model Built with Test data

test_model_aic = lm(SalePrice ~ LotArea + Neighborhood + OverallQual + OverallCond + YearBuilt + YearRe


#Condition Number and Condition Index
round(ols_eigen_cindex(test_model_aic)[,1:2], 4)

#Test VIFs
vif(test_model_aic)

#Fitted Vs Residual Plot
ols_plot_resid_fit(test_model_aic)

#Breush Pagan Test
bptest(test_model_aic)

#Q-Q plot
ols_plot_resid_qq(test_model_aic)
```

```r
#Shapiro Wilk Test

shapiro.test(resid(test_model_aic))

# High leverage points

# Check for high leverage points
which(hatvalues(test_model_aic) > 2 * mean(hatvalues(test_model_aic)))

# Check for highly influential points
which(cooks.distance(test_model_aic) > 4 / length(cooks.distance(test_model_aic)))

# ids for non-influential observations
noninfluential_ids = which(
    cooks.distance(test_model_aic) <= 4 / length(cooks.distance(test_model_aic)))

#Model without high inf

model_fix_aic = lm(SalePrice ~ LotArea + Neighborhood + OverallQual + OverallCond + YearBuilt + YearRem

# Nested model for the categorical variable F-test

restricted_neighborhood = lm(SalePrice ~ LotArea + OverallQual + OverallCond + YearBuilt + YearRemodAdd

anova(restricted_neighborhood, model_fix_aic)

#Check Signficance of regression and individual t-tests

summary(test_model_aic)

#Creating an observation for CI and prediction]
observation = data_test[30, ]

observation$OverallQual = 8

# CI .99
predict(model_fix_aic, newdata = observation, interval = 'confidence', level = .99)

# Prediction .99


predict(model_fix_aic, newdata = observation, interval = 'prediction', level = .99)
```