Predicting House Prices in Ames, Iowa

Aryan, Devesh, Rupesh, Micah

[Date]

Introduction

Understanding factors that influence house prices is crucial for buyers, sellers, and real estate agents. This study aims to develop a predictive model for house prices using data from Ames, Iowa. By analyzing attributes such as overall quality, living area size, and neighborhood, we can identify key drivers of house prices. This knowledge assists in determining fair prices, setting competitive prices, and providing better advice, thereby offering insights into market dynamics and property value.

Data Preparation

```r
# Load and combine datasets
train_data <- read.csv("train.csv")
test_data <- read.csv("test.csv")
combined_data <- bind_rows(train_data %>% mutate(Dataset="Train"), test_data
%>% mutate(Dataset="Test"))
names(combined_data)[names(combined_data) == "X1stFlrSF"] <- "FirstFlrSF"
names(combined_data)[names(combined_data) == "X2ndFlrSF"] <- "SecondFlrSF"
```

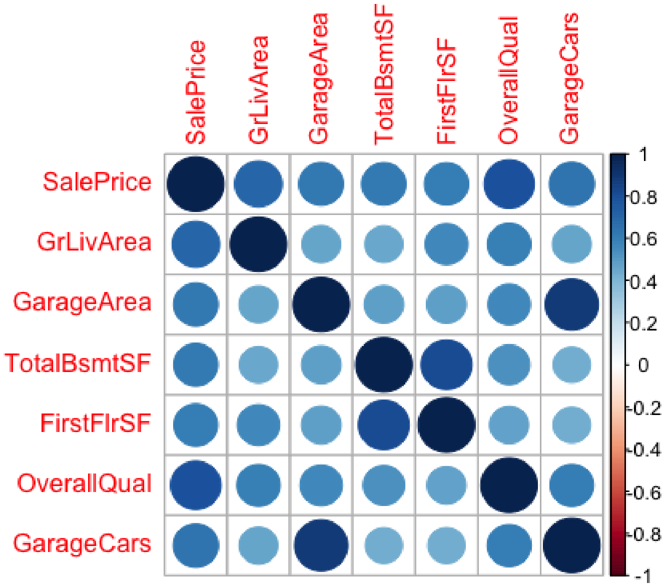
See Appendix 1 for summary.

Exploratory Data Analysis (EDA)

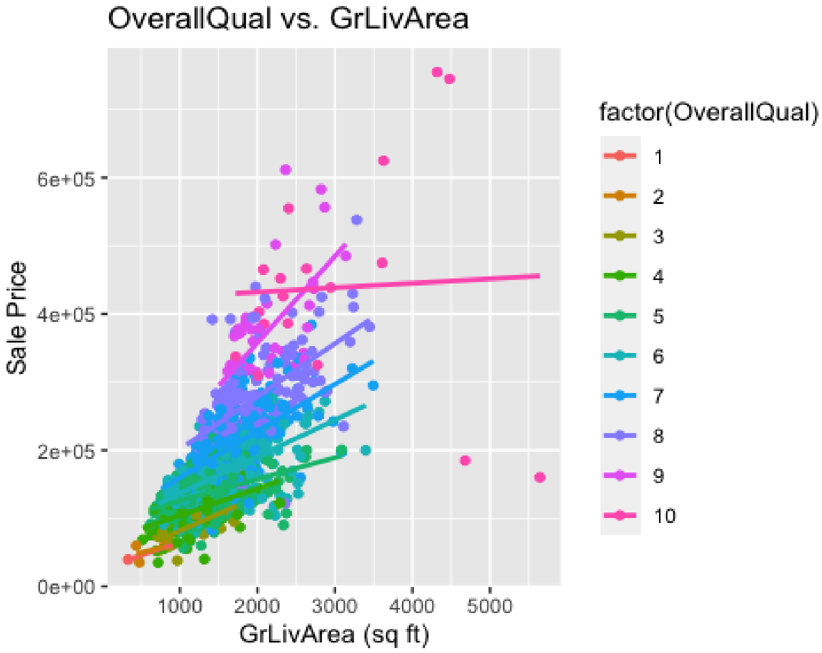
Analyze correlations and interactions.

```r
# Select columns and subset data
selected_columns <- c("SalePrice", "GrLivArea", "GarageArea", "TotalBsmtSF",
"FirstFlrSF", "OverallQual", "KitchenQual", "GarageCars")
eda_data <- combined_data %>% filter(Dataset == "Train") %>% select(all_of(se
lected_columns))

# Correlation matrix
cor_matrix <- cor(eda_data %>% select_if(is.numeric), use = "complete.obs")
corrplot(cor_matrix, method = "circle")
```



```
# Interaction plot
ggplot(eda_data, aes(x = GrLivArea, y = SalePrice, color = factor(OverallQual
))) + geom_point() + geom_smooth(method = "lm", se = FALSE) + labs(title = "O
verallQual vs. GrLivArea", x = "GrLivArea (sq ft)", y = "Sale Price")
```



Model Fitting

Fit models and perform an ESS test.

```
# Manual model fitting
X <- model.matrix(SalePrice ~ GrLivArea + OverallQual, data = eda_data)
```

```
y <- eda_data$SalePrice
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
print(beta_hat)

##                      [,1]
## (Intercept) -104092.66964
## GrLivArea        55.86223
## OverallQual   32849.04744

# ESS Test
X1 <- model.matrix(SalePrice ~ GrLivArea + OverallQual + GarageCars, data = e
da_data)
y_hat1 <- X1 %*% solve(t(X1) %*% X1) %*% t(X1) %*% y
RSS1 <- sum((y - y_hat1)^2)

X2 <- model.matrix(SalePrice ~ GrLivArea + OverallQual, data = eda_data)
y_hat2 <- X2 %*% solve(t(X2) %*% X2) %*% t(X2) %*% y
RSS2 <- sum((y - y_hat2)^2)

ESS <- RSS2 - RSS1
df <- nrow(eda_data) - ncol(X1)
F_stat <- ESS / (RSS1 / df)
p_value <- 1 - pf(F_stat, 1, df)

print(paste("ESS:", ESS))

## [1] "ESS: 229241894491.302"

print(paste("F-statistic:", F_stat))

## [1] "F-statistic: 138.921428939754"

print(paste("p-value:", p_value))

## [1] "p-value: 0"
```

Model Diagnostics

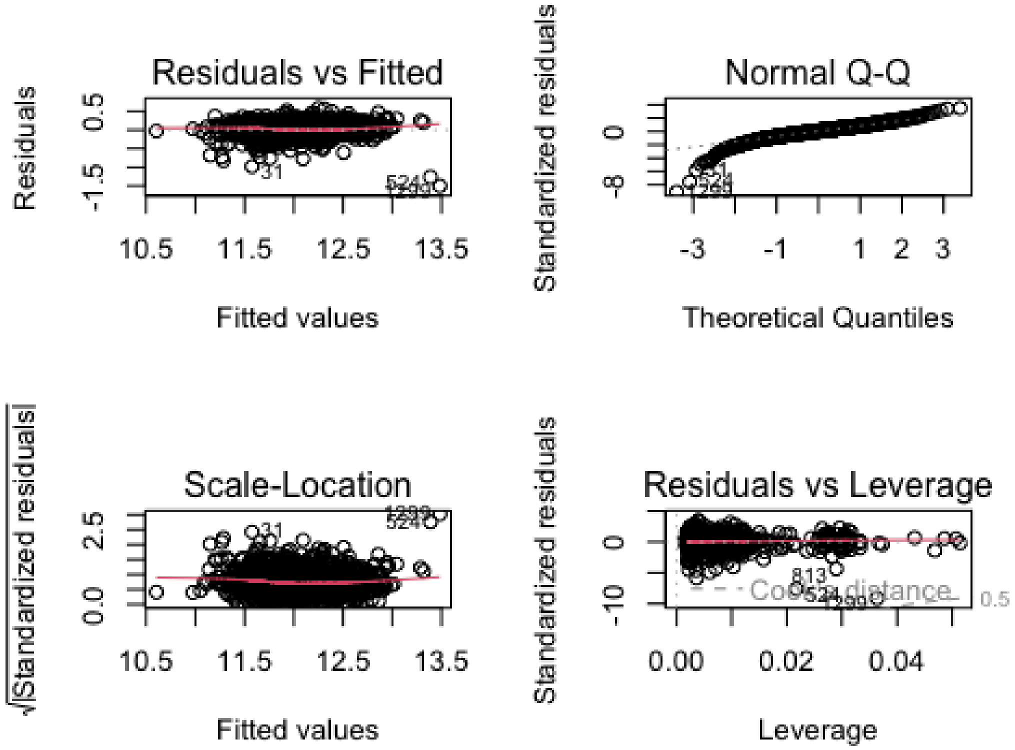
Check diagnostics and choose the final model.

```
# Transformations and model selection
combined_data$LogSalePrice <- log(combined_data$SalePrice)
combined_data$SqrtGrLivArea <- sqrt(combined_data$GrLivArea)
combined_data$LogTotalBsmtSF <- log1p(combined_data$TotalBsmtSF)

train_data <- combined_data %>% filter(Dataset == "Train")
final_model <- lm(LogSalePrice ~ SqrtGrLivArea + OverallQual + LogTotalBsmtSF
+ GarageCars + KitchenQual + YearBuilt, data = train_data)

# Diagnostic plots
```

```
par(mfrow = c(2, 2))
plot(final_model)
```



```
summary(final_model)$r.squared

## [1] 0.8243248

f <- summary(final_model)$fstatistic
p <- pf(f[1],f[2],f[3],lower.tail=F)
f

##   value   numdf   dendf
##  851.07    8.00 1451.00

p

## value
##     0
```

Executive Summary

Objective: This study aims to develop a predictive model for house prices using data from Ames, Iowa. By analyzing various attributes mentioned below, we can identify key drivers of house prices. This knowledge assists in determining fair prices, setting competitive prices, and providing better advice to stakeholders.

Key Predictors: The final model uses the following key predictors:

- **Living Area Size (GrLivArea)**: The total living area of the house.
- **Overall Quality (OverallQual)**: An overall rating of the house's quality.
- **Basement Area (TotalBsmtSF)**: The total area of the basement.
- **Garage Capacity (GarageCars)**: The number of cars that can fit in the garage.
- **Kitchen Quality (KitchenQual)**: The quality rating of the kitchen.
- **Year Built (YearBuilt)**: The year the house was built.

Modeling Process:

1. **Data Collection and Preparation**: We combined training and test datasets to create a comprehensive dataset for analysis. Key features were selected based on their relevance to predicting house prices.
2. **Exploratory Data Analysis (EDA)**: We examined correlations between predictors and identified significant interactions between them. For instance, the interaction between living area size and overall quality showed a strong correlation with sale prices.
3. **Model Fitting**: We manually fit a linear regression model using key predictors. This involved calculating coefficients that describe the relationship between each predictor and the sale price.
4. **Transformations**: To improve model accuracy, we applied transformations such as taking the logarithm of sale prices and basement areas, and the square root of living area sizes.
5. **Model Diagnostics**: We checked the model's assumptions and validated its performance using diagnostic plots. The final model explained approximately 82.4% of the variability in house prices.

Key Findings:

- **Living Area Size**: Larger living areas are associated with higher house prices. Specifically, a one-unit increase in the square root of the living area increases the log-transformed sale price by 0.022.
- **Overall Quality**: Higher overall quality ratings significantly increase house prices. A one-unit increase in overall quality leads to an 8.1% increase in sale price.
- **Basement Area**: Larger basement areas contribute positively to house prices. A one-unit increase in the log of the total basement area increases the log-transformed sale price by 0.037.
- **Garage Capacity**: Houses with larger garages tend to sell for higher prices. Each additional car capacity in the garage increases the log-transformed sale price by 0.083.
- **Kitchen Quality**: Kitchen quality has a substantial impact on house prices. Houses with fair kitchen quality have a 24.2% lower sale price compared to those with excellent kitchen quality.
- **Year Built**: Newer houses tend to sell for higher prices. Each additional year since the house was built increases the log-transformed sale price by 0.0024.

Limitations:

- **Data Quality**: Missing or inaccurate data could affect the model's accuracy.
- **Generalizability**: Results may not generalize to other regions beyond Ames, Iowa.
- **Multicollinearity**: High correlations between predictors can lead to instability in coefficient estimates.
- **Non-linear Relationships**: Some relationships between predictors and sale prices may not be fully captured by linear transformations.

Conclusion: The model provides valuable insights into the key factors driving house prices in Ames, Iowa. Stakeholders such as buyers, sellers, and real estate agents can use this information to make informed decisions about pricing and marketing properties. While the model is robust, care should be taken to consider its limitations when applying the findings to other contexts.

Apendix

Apendix 1

```
str(combined_data)

## 'data.frame':    2919 obs. of  85 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : chr  "RL" "RL" "RL" "RL" ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120
7420 ...
##  $ Street       : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr  NA NA NA NA ...
##  $ LotShape     : chr  "Reg" "Reg" "IR1" "IR1" ...
##  $ LandContour  : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : chr  "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ LotConfig    : chr  "Inside" "FR2" "Inside" "Corner" ...
##  $ LandSlope    : chr  "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood : chr  "CollgCr" "Veenker" "CollgCr" "Crawfor" ...
##  $ Condition1   : chr  "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition2   : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType     : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr  "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939
...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950
...
##  $ RoofStyle    : chr  "Gable" "Gable" "Gable" "Gable" ...
##  $ RoofMatl     : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...
##  $ Exterior2nd   : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...
##  $ MasVnrType    : chr  "BrkFace" "None" "BrkFace" "None" ...
##  $ MasVnrArea    : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual     : chr  "Gd" "TA" "Gd" "TA" ...
##  $ ExterCond     : chr  "TA" "TA" "TA" "TA" ...
##  $ Foundation    : chr  "PConc" "CBlock" "PConc" "BrkTil" ...
##  $ BsmtQual      : chr  "Gd" "Gd" "Gd" "TA" ...
##  $ BsmtCond      : chr  "TA" "TA" "TA" "Gd" ...
##  $ BsmtExposure  : chr  "No" "Gd" "Mn" "No" ...
##  $ BsmtFinType1  : chr  "GLQ" "ALQ" "GLQ" "ALQ" ...
##  $ BsmtFinSF1    : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2  : chr  "Unf" "Unf" "Unf" "Unf" ...
##  $ BsmtFinSF2    : int  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF     : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF   : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
##  $ Heating       : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC     : chr  "Ex" "Ex" "Ex" "Gd" ...
##  $ CentralAir    : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical    : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ FirstFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ SecondFlrSF   : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea     : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077
...
##  $ BsmtFullBath  : int  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath  : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath      : int  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath      : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr  : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr  : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual   : chr  "Gd" "TA" "Gd" "Gd" ...
##  $ TotRmsAbvGrd  : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional    : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces    : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu   : chr  NA "TA" "TA" "Gd" ...
##  $ GarageType    : chr  "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt   : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939
...
##  $ GarageFinish  : chr  "RFn" "RFn" "RFn" "Unf" ...
##  $ GarageCars    : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea    : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual    : chr  "TA" "TA" "TA" "TA" ...
##  $ GarageCond    : chr  "TA" "TA" "TA" "TA" ...
##  $ PavedDrive    : chr  "Y" "Y" "Y" "Y" ...
##  $ WoodDeckSF    : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF   : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch : int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch    : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : chr  NA NA NA NA ...
##  $ Fence        : chr  NA NA NA NA ...
##  $ MiscFeature  : chr  NA NA NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008
...
##  $ SaleType     : chr  "WD" "WD" "WD" "WD" ...
##  $ SaleCondition : chr  "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 2
00000 129900 118000 ...
##  $ Dataset      : chr  "Train" "Train" "Train" "Train" ...
##  $ LogSalePrice  : num  12.2 12.1 12.3 11.8 12.4 ...
##  $ SqrtGrLivArea : num  41.4 35.5 42.3 41.4 46.9 ...
##  $ LogTotalBsmtSF: num  6.75 7.14 6.83 6.63 7.04 ...
```

Appendix 2

```
summary(final_model)

##
## Call:
## lm(formula = LogSalePrice ~ SqrtGrLivArea + OverallQual + LogTotalBsmtSF +
##     GarageCars + KitchenQual + YearBuilt, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49954 -0.08150  0.01145  0.09662  0.58463
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.7100369  0.3932509  14.520  < 2e-16 ***
## SqrtGrLivArea   0.0223966  0.0009102  24.606  < 2e-16 ***
## OverallQual     0.0808372  0.0055661  14.523  < 2e-16 ***
## LogTotalBsmtSF  0.0368037  0.0040929   8.992  < 2e-16 ***
## GarageCars      0.0826878  0.0080476  10.275  < 2e-16 ***
## KitchenQualFa  -0.2415198  0.0363615  -6.642 4.36e-11 ***
## KitchenQualGd  -0.1139894  0.0193099  -5.903 4.43e-09 ***
## KitchenQualTA  -0.1768440  0.0221043  -8.000 2.52e-15 ***
## YearBuilt       0.0023882  0.0002005  11.910  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1679 on 1451 degrees of freedom
## Multiple R-squared:  0.8243, Adjusted R-squared:  0.8234
## F-statistic: 851.1 on 8 and 1451 DF,  p-value: < 2.2e-16
```