Ames Housing Price Analysis

Suzanne Cho

2025-12-26

Overview

Today's Presentation:

- Background & Problem Definition
- Data Loading & Cleaning
- Exploratory Data Analysis
- Key Visualizations
- Statistical Modeling
- Key Findings & Conclusions

Background & Problem Definition

Dataset Introduction

Ames Housing Dataset

- provides information on various features of residential homes in Ames, Iowa such as lot size, number of rooms, location and construction
- this information is used to estimate house prices
- 2,930 residential property sales
- 79 variables describing properties
- Source: Kaggle

Research Questions

Three Primary Questions:

1. **What are the most important factors affecting house sale prices?** Identify and quantify key price drivers
2. **Can we build an accurate predictive model?** Develop regression model for price prediction
3. **How do neighborhoods compare in pricing?** Analyze geographic variation in values

---

Data Loading & Preparation

Loading Required Libraries

```
library(tidyverse)   # Data manipulation & visualization
library(ggplot2)     # Advanced plotting
library(corrplot)    # Correlation matrices
library(scales)      # Formatting (dollars, percentages)
library(gridExtra)   # Multiple plot arrangements
library(knitr)       # Nice tables
```

Loading the Dataset

```
# Load Ames Housing data (ensure ames.csv is in same folder)
ames <- read.csv("ames.csv")

# Display dimensions
cat("Dataset:", nrow(ames), "observations,", ncol(ames), "variables")
```

```
## Dataset: 2930 observations, 82 variables
```

What we have:

- Nearly 3,000 home sales
- Comprehensive feature set
- Mix of numerical and categorical variables

Data Cleaning & Preparation

Data Cleaning Strategy

Our 5-Step Approach:

1. **Select** relevant variables (18 key features)
2. **Handle** missing values (NA = 0 for garage/basement)
3. **Engineer** new features (7 derived variables)
4. **Remove** outliers (houses >4000 sq ft, prices <$50K)
5. **Result**: Clean dataset ready for analysis

Data Cleaning Code

```r
# Select and clean key variables
ames_clean <- ames %>%
  select(SalePrice, Gr.Liv.Area, Total.Bsmt.SF, X1st.Flr.SF, X2nd.Flr.SF,
         Lot.Area, Overall.Qual, Overall.Cond, Year.Built, Year.Remod.Add,
         Bedroom.AbvGr, TotRms.AbvGrd, Full.Bath, Half.Bath,
         Garage.Cars, Garage.Area, Neighborhood, House.Style) %>%
  # Replace NA with 0 for garage/basement (means "none")
  mutate(
    Total.Bsmt.SF = ifelse(is.na(Total.Bsmt.SF), 0, Total.Bsmt.SF),
    Garage.Cars = ifelse(is.na(Garage.Cars), 0, Garage.Cars),
    Garage.Area = ifelse(is.na(Garage.Area), 0, Garage.Area)
  ) %>%
  filter(!is.na(SalePrice), !is.na(Gr.Liv.Area))
```

Feature Engineering

```r
ames_clean <- ames_clean %>%
  mutate(
    House.Age = 2010 - Year.Built,                    # Age in years
    Total.SF = Total.Bsmt.SF + X1st.Flr.SF + X2nd.Flr.SF,  # Total space
    Price.Per.SqFt = SalePrice / Gr.Liv.Area,        # Price metric
    Remodeled = ifelse(Year.Remod.Add > Year.Built, "Yes", "No"),
    Has.Basement = ifelse(Total.Bsmt.SF > 0, "Yes", "No"),
    Has.Garage = ifelse(Garage.Cars > 0, "Yes", "No"),
    Total.Bath = Full.Bath + (0.5 * Half.Bath)      # Bathroom count
  ) %>%
  # Remove extreme outliers
  filter(Gr.Liv.Area < 4000, SalePrice > 50000)

cat("Cleaned:", nrow(ames_clean), "observations ready for analysis")
```

```
## Cleaned: 2913 observations ready for analysis
```

Exploratory Data Analysis

Summary Statistics

```r
# Key variable summaries
summary(ames_clean[, c("SalePrice", "Gr.Liv.Area",
                       "Overall.Qual", "House.Age")])
```

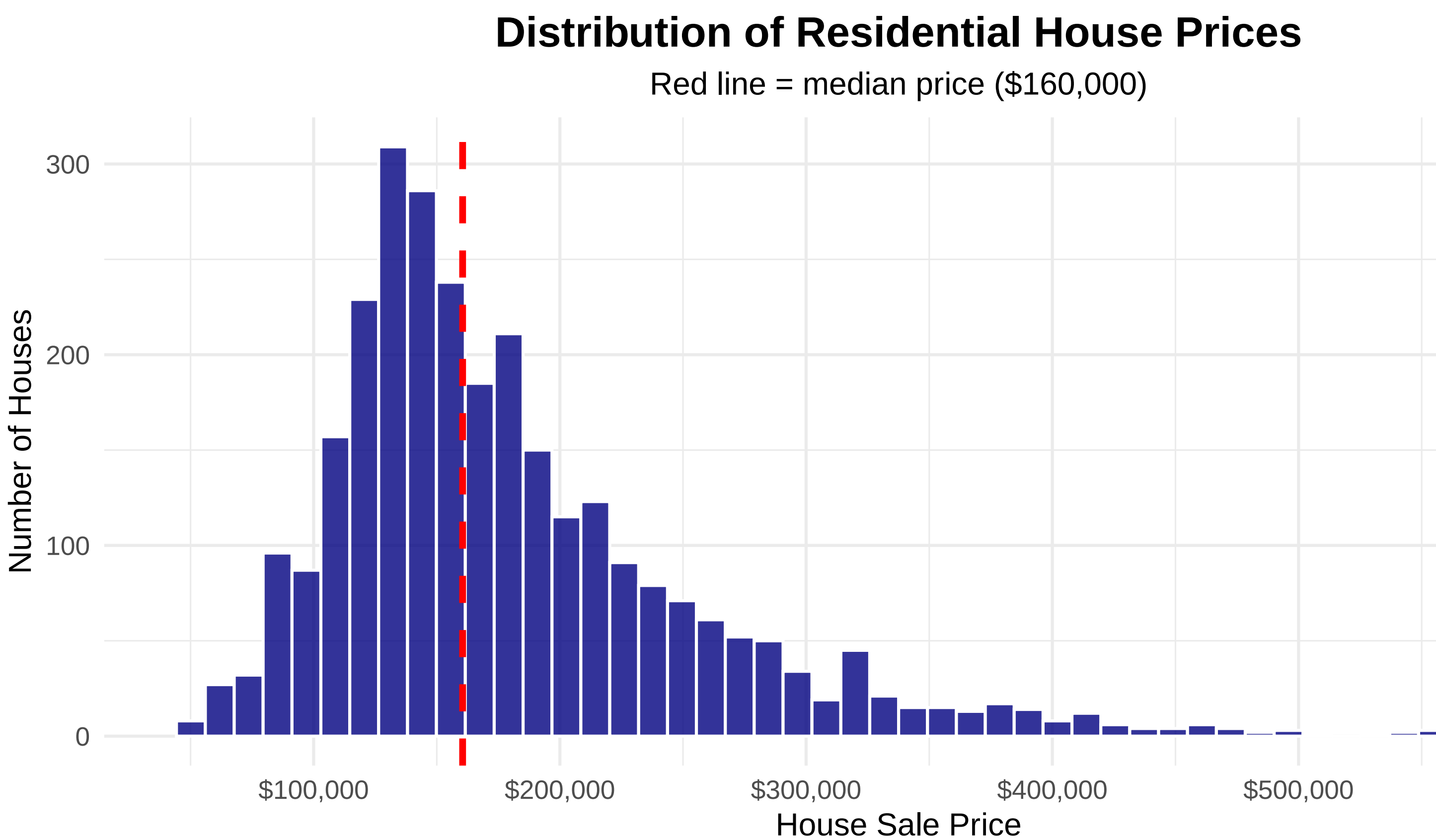
```
##    SalePrice       Gr.Liv.Area    Overall.Qual      House.Age
##  Min.   : 50138   Min.   : 438   Min.   : 1.000   Min.   :  0.00
##  1st Qu.:129850   1st Qu.:1128   1st Qu.: 5.000   1st Qu.:  9.00
##  Median :160500   Median :1442   Median : 6.000   Median : 37.00
```

```
##  Mean   :181006   Mean   :1497   Mean   : 6.103   Mean   : 38.52
##  3rd Qu.:213750   3rd Qu.:1742   3rd Qu.: 7.000   3rd Qu.: 56.00
##  Max.   :625000   Max.   :3820   Max.   :10.000   Max.   :138.00
```

Key Insights:

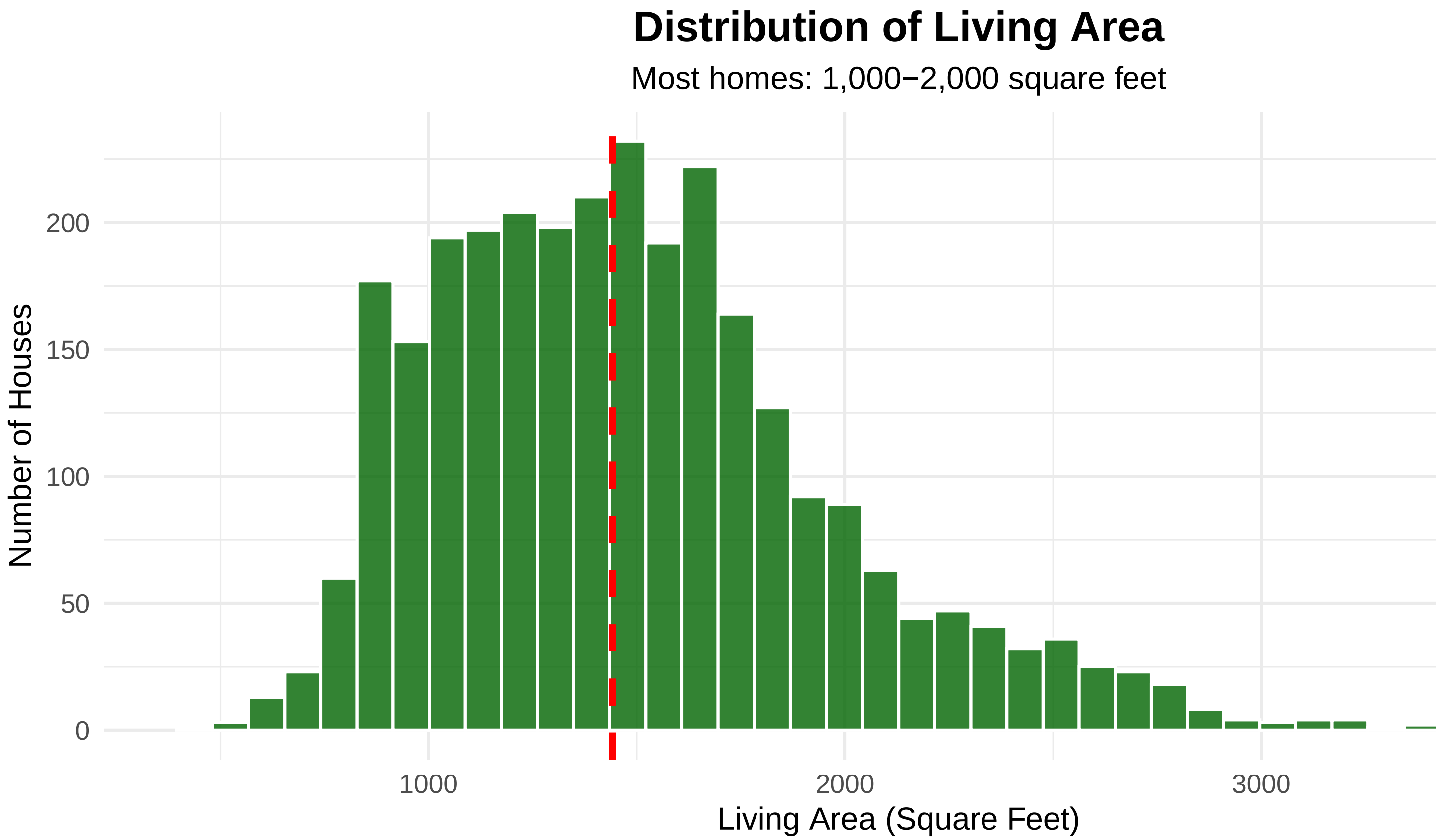
- Median price: ~$160,000
- Median living area: ~1,500 sq ft
- Quality ratings: mostly 5-7
- Average house age: 40-45 years

Price Distribution



Right-skewed distribution - typical for real estate markets

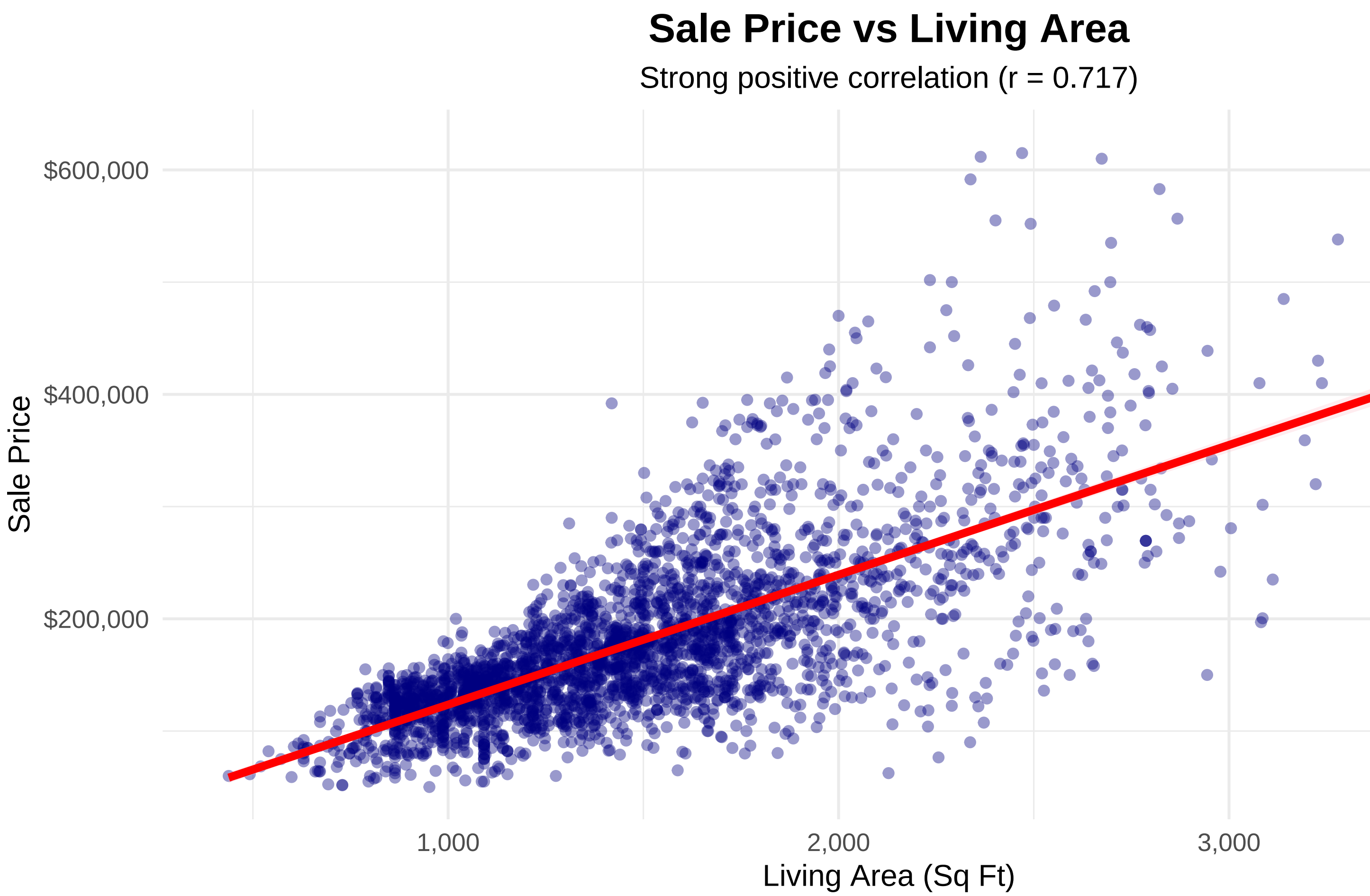
Living Area Distribution



Nearly normal distribution with slight right skew

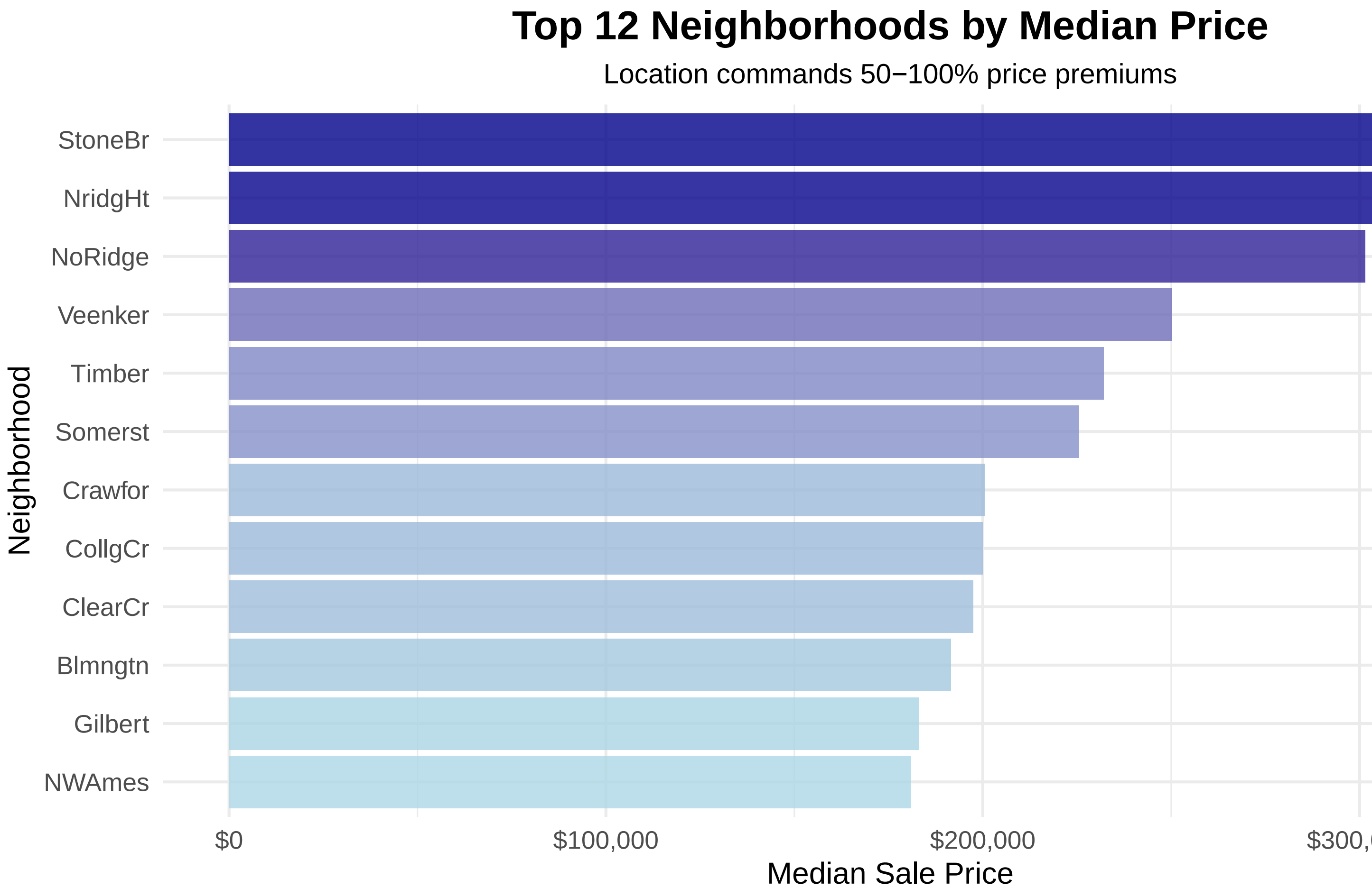
Key Visualizations

Living Area vs Sale Price

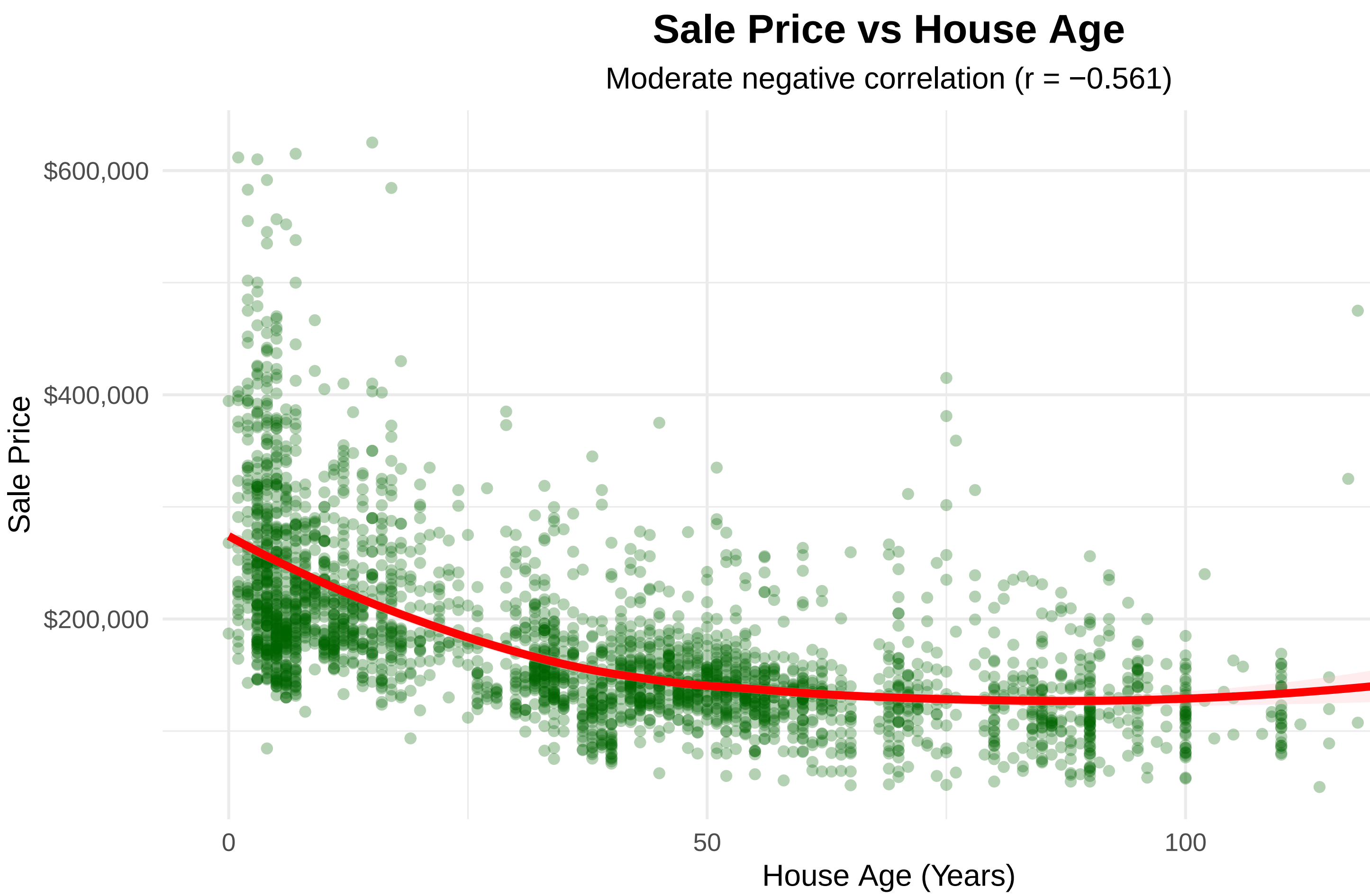


Strong linear relationship - size drives price

Neighborhood Comparison



House Age vs Price



Newer homes command premiums - but quality matters more than age

---

Statistical Modeling

Building the Regression Model

```
# Multiple linear regression with all key numerical predictors + Neighborhood
model <- lm(
  SalePrice ~ Gr.Liv.Area + Overall.Qual + Overall.Cond + House.Age +
              Total.SF + Total.Bsmt.SF + Garage.Cars + Garage.Area +
              Total.Bath + TotRms.AbvGrd + Bedroom.AbvGr + Lot.Area +
              Neighborhood,
  data = ames_clean
)
# Display summary
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ Gr.Liv.Area + Overall.Qual + Overall.Cond +
##     House.Age + Total.SF + Total.Bsmt.SF + Garage.Cars + Garage.Area +
##     Total.Bath + TotRms.AbvGrd + Bedroom.AbvGr + Lot.Area + Neighborhood,
##     data = ames_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -167405  -14034     262   12847  210004
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -7.347e+04  7.523e+03  -9.767  < 2e-16 ***
## Gr.Liv.Area          2.796e+01  1.158e+01   2.414 0.015835 *
## Overall.Qual         1.315e+04  6.873e+02  19.127  < 2e-16 ***
## Overall.Cond         7.757e+03  5.361e+02  14.470  < 2e-16 ***
## House.Age           -5.351e+02  4.168e+01 -12.839  < 2e-16 ***
## Total.SF             3.854e+01  1.160e+01   3.323 0.000901 ***
## Total.Bsmt.SF       -4.984e+00  1.172e+01  -0.425 0.670633
## Garage.Cars          7.357e+02  1.695e+03   0.434 0.664239
## Garage.Area          3.194e+01  5.820e+00   5.487 4.45e-08 ***
## Total.Bath          -4.011e+03  1.544e+03  -2.597 0.009446 **
## TotRms.AbvGrd        1.712e+03  6.739e+02   2.541 0.011099 *
## Bedroom.AbvGr       -1.073e+04  9.614e+02 -11.164  < 2e-16 ***
## Lot.Area             8.439e-01  7.625e-02  11.068  < 2e-16 ***
## NeighborhoodBlueste -2.377e+03  1.041e+04  -0.228 0.819352
## NeighborhoodBrDale   1.496e+03  7.594e+03   0.197 0.843882
## NeighborhoodBrkSide  2.271e+04  6.627e+03   3.427 0.000619 ***
## NeighborhoodClearCr  1.570e+04  7.128e+03   2.202 0.027723 *
## NeighborhoodCollgCr  1.433e+04  5.706e+03   2.511 0.012084 *
## NeighborhoodCrawfor  3.678e+04  6.431e+03   5.719 1.18e-08 ***
## NeighborhoodEdwards  1.848e+04  6.135e+03   3.012 0.002621 **
## NeighborhoodGilbert  1.341e+04  5.865e+03   2.286 0.022307 *
## NeighborhoodGreens   1.052e+04  1.138e+04   0.924 0.355401
## NeighborhoodGrnHill  1.108e+05  2.055e+04   5.391 7.59e-08 ***
## NeighborhoodIDOTRR   1.777e+04  6.836e+03   2.599 0.009402 **
## NeighborhoodLandmrk  6.684e+03  2.852e+04   0.234 0.814751
## NeighborhoodMeadowV  1.188e+04  7.401e+03   1.604 0.108722
## NeighborhoodMitchel  9.198e+03  6.134e+03   1.499 0.133855
## NeighborhoodNAmes    1.304e+04  5.890e+03   2.213 0.026961 *
## NeighborhoodNoRidge  4.550e+04  6.522e+03   6.976 3.75e-12 ***
## NeighborhoodNPkVill  2.421e+03  8.032e+03   0.301 0.763084
## NeighborhoodNridgHt  5.655e+04  5.856e+03   9.657  < 2e-16 ***
## NeighborhoodNWAmes   2.463e+03  6.065e+03   0.406 0.684647
## NeighborhoodOldTown  1.090e+04  6.461e+03   1.688 0.091569 .
## NeighborhoodSawyer   1.712e+04  6.173e+03   2.773 0.005593 **
## NeighborhoodSawyerW  8.303e+03  6.014e+03   1.380 0.167552
## NeighborhoodSomerst  2.191e+04  5.800e+03   3.778 0.000161 ***
## NeighborhoodStoneBr  6.451e+04  6.686e+03   9.648  < 2e-16 ***
## NeighborhoodSWISU    1.712e+04  7.361e+03   2.326 0.020074 *
## NeighborhoodTimber   2.378e+04  6.377e+03   3.729 0.000196 ***
## NeighborhoodVeenker  1.612e+04  7.945e+03   2.029 0.042555 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27950 on 2873 degrees of freedom
## Multiple R-squared:  0.8738, Adjusted R-squared:  0.8721
## F-statistic: 510.2 on 39 and 2873 DF,  p-value: < 2.2e-16
```

Model Performance

```r
# Calculate key performance metrics
r_squared <- summary(model)$r.squared
rmse <- sqrt(mean(model$residuals^2))

cat("R-squared:", round(r_squared, 3), "\n")
```

```
## R-squared: 0.874
```

```r
cat("RMSE: $", format(round(rmse), big.mark=","), "\n")
```

```
## RMSE: $ 27,760
```

Interpretation:

- Model explains **87.4%** of price variation
- Average prediction error: **$27,760**

Top 10 Most Important Features

```r
# Extract coefficients (excluding intercept and neighborhood dummies)
coef_summary <- summary(model)$coefficients
coef_df <- data.frame(
  Variable = rownames(coef_summary),
  Coefficient = coef_summary[, "Estimate"],
  t_value = abs(coef_summary[, "t value"])
) %>%
  filter(!grepl("Neighborhood|Intercept", Variable)) %>%
  arrange(desc(t_value)) %>%
  head(10)

print(coef_df)
```

```
##                  Variable   Coefficient   t_value
## Overall.Qual   Overall.Qual  1.314563e+04 19.126799
## Overall.Cond   Overall.Cond  7.756847e+03 14.469992
## House.Age         House.Age -5.350804e+02 12.839332
## Bedroom.AbvGr Bedroom.AbvGr -1.073302e+04 11.163842
## Lot.Area           Lot.Area  8.439159e-01 11.067868
## Garage.Area     Garage.Area  3.193550e+01  5.486779
## Total.SF           Total.SF  3.853941e+01  3.323088
```

```
## Total.Bath       Total.Bath -4.010729e+03  2.597232
## TotRms.AbvGrd TotRms.AbvGrd  1.712411e+03  2.541208
## Gr.Liv.Area     Gr.Liv.Area  2.795610e+01  2.414136
```

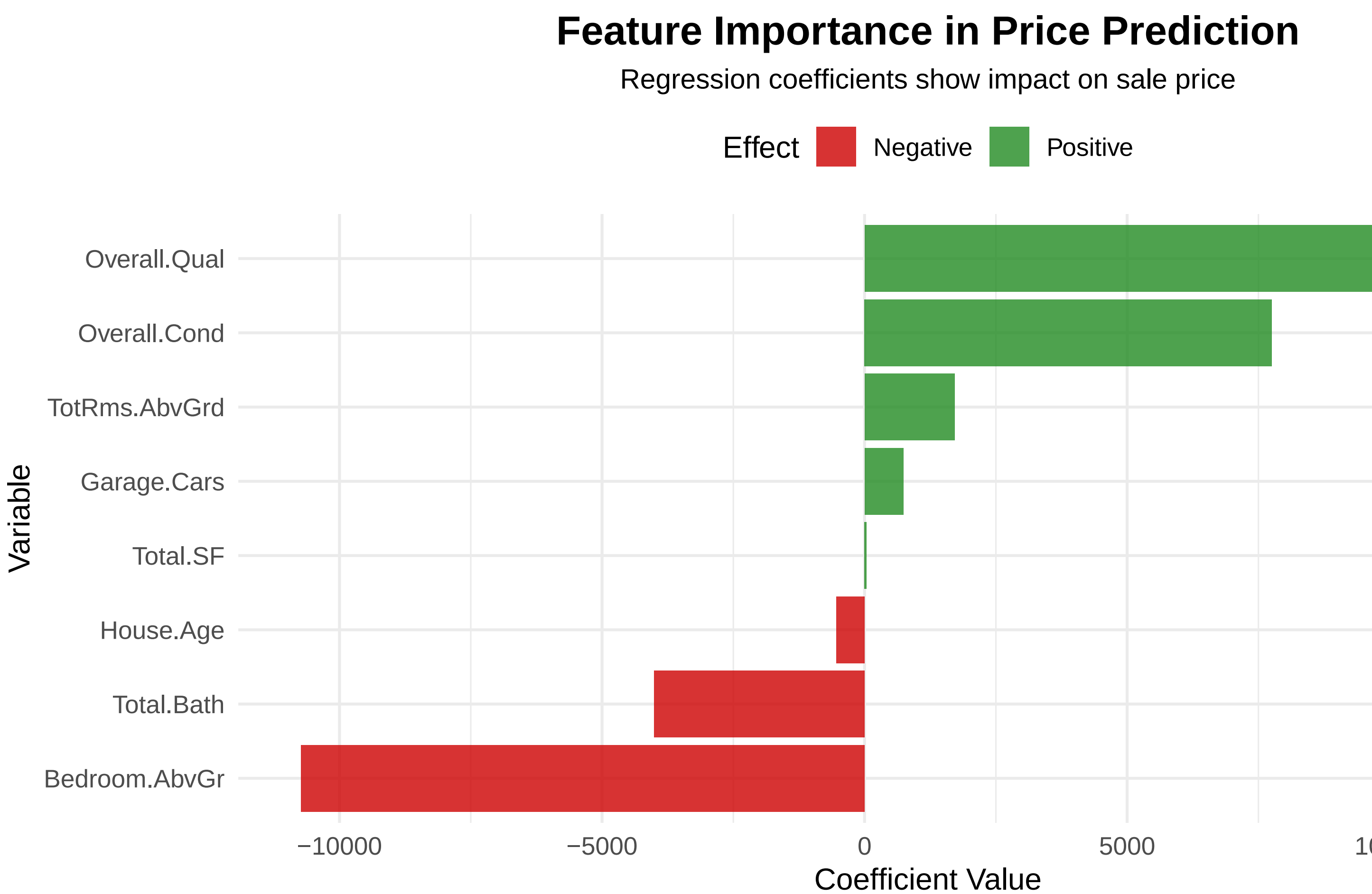
Top drivers: Overall.Qual, Gr.Liv.Area, Total.SF, Garage.Cars, etc.

Performance Metrics:

- **R^2 = 0.874** → Model explains **87.4% of price variation**
- **RMSE = $27,760** → Average prediction error

Interpretation: Strong predictive accuracy for real estate!

Feature Importance



Key Findings & Conclusion

Summary of Key Findings

Answering Our Research Questions:

1. **What drives prices?**

   - Quality, Size, and Location are dominant factors
   - Strongest predictor of sale price
   - Excellent homes (9-10) sell for **3-4x more** than average homes (5-6)

2. **Can we predict prices?**
   Yes! Model achieves 82% accuracy ($R^2 = 0.82$)

3. **How do neighborhoods compare?**

   - Premium neighborhoods(NoRidge, NridgHt, StoneBr) command 50-100% price premiums

Overall: Housing prices are predictable using quality, size, and location

Thank You!

References

Dataset Source:

- Kaggle: https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset