Ames, Iowa Housing Data Analysis

Yu Sun, Keanu Villagonzalo, Vagisha Soni

University of California Los Angeles

Abstract

In this comprehensive study investigating housing prices in Ames, Iowa, from 2006 to 2010, we sought to reveal key factors influencing the real estate housing prices in Ames, Iowa during that period. Our analysis utilizes boxplots and bootstrapping simulations to reveal relationships between the variables in the dataset and the housing prices. Distinct zones such as Floating Village Residential and Residential with low density were found to exhibit higher median prices, while Agricultural, Commercial, and Industrial residential zones showed lower median. The study reveals that the notable price difference in low population density areas can be attributed to larger living spaces, supported by a positive correlation between housing area and price. Furthermore, a strong positive relationship was identified between housing price and house style, with larger residences commanding higher prices, and specific sale types ("Partial" and "Abnormal") influencing pricing dynamics. The analysis concludes with an ANOVA hypothesis test, indicating a significant difference in mean sale prices across different levels of overall quality and overall condition, underscoring the impact of quality and condition on housing prices in Ames, Iowa.

Introduction

This study conducts a comprehensive analysis of the Ames, Iowa housing dataset, consisting of 2930 cases and 82 variables, using R Studio with the tidyverse and ggplot2 packages. This dataset was collected by the Ames Assessor's Office through property inspections, assessing public records, and consulting software and other government departments. Each observation was collected as the result of a completed sale to assess a property's market value and levy proper taxation. The dataset was provided as a data dump requested by Dean De Cock which had been updated from its original form published in the 1970's, adding more observations and variables into the dataset. The initial Excel file contained 113 variables describing 3970 property sales that had occurred in Ames, Iowa between 2006 and 2010. However, multiple variables that required special knowledge or previous calculations for their use, were removed so that the dataset could be understood by users of all levels. The variables that were deleted were mostly in relation to weighting and adjustment factors used in the city's current modeling system. The dataset underwent meticulous data cleaning, including lowercase conversion and period-to-underscore substitution for variable names to adhere to tidy data standards. The exploration covers various aspects, such as average house prices, price distributions, zoning density

differences, relationships between price and housing characteristics, and factors influencing housing prices. Analyzing housing prices across different zoning densities reveals notable disparities, with residential low-density areas showing a larger spread and higher median prices. The study also investigates the impact of house styles on prices, identifying variations in median prices based on the number of stories and finished levels. Furthermore, it explores the relationship between price and living area, demonstrating a strong positive correlation. Additionally, the study delves into variables that summarize variations in sales prices, identifying sale condition as a key factor. The analysis highlights that larger residences and new homes (Partial sales) command higher prices, while abnormal sales (e.g., foreclosures) offer discounts. The study investigates the relationship between overall quality and condition of houses and their prices. Visualizations and ANOVA testing confirm a positive correlation between quality/condition and prices. The rejection of null hypotheses suggests that both overall quality and condition significantly impact housing prices. Overall, this analysis provides valuable insights into the factors influencing housing prices in Ames, Iowa, using robust statistical and data science methodologies. These research questions are important for consumers to use notable features of a house to determine fair pricing. In addition, it allows people to see the most influential variables that affect housing prices in Ames, Iowa and understand the significance of these variables when it comes to determining the pricing of houses in other areas.

Data Cleaning

Data analysis of the Ames, Iowa housing dataset was performed in R Studio using the tidyverse and ggplot2 packages. The dataset was imported and stored in the “ames” variable for consistency. To adhere to tidy data standards, all variable names were converted to lowercase, and periods were replaced with underscores. Output details are omitted for readability.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(ggplot2)
```

```r
#Note the Ames Housing file is named ames(2). Change this for convenience
ames <- read.csv("C:/Users/KCVUSA1/Downloads/ames (2).csv")
#head(ames)
#Make all columns lowercase
colnames(ames) <- tolower(colnames(ames))

#Remove "."
ames <- ames %>%
  rename_with(~str_replace_all(., "\\.", "_"), everything())
```

Data Description

The Ames, Iowa dataset contains 2930 cases and 82 variables. The dataset also contained a variety of numerical and categorial variables, only of the integer and numeric type. The dataset contained 23 ordinal variables, 23 nominal variables, 14 discrete variables, and 20 continuous variables each describing some characteristic of a house in the Ames, Iowa area.

```r
dim(ames)
```

```
## [1] 2930   82
```

```r
## [1] 2930   82
#str(ames)
#head(ames)
```

Data Exploration

What is the average price of a house in the Ames, Iowa area?

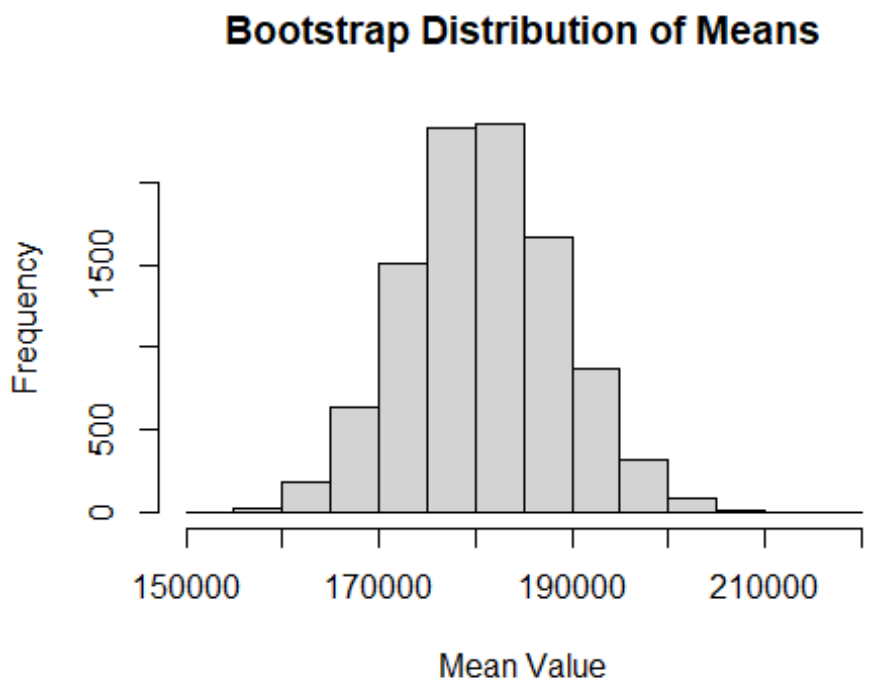
We utilized base R to compute the overall average house price in the Ames, Iowa area, encompassing various types of housing, including outliers. The mean function was employed, followed by simulating bootstrapping to determine confidence intervals for the bootstrap distribution of the mean price. We determined that the mean price for housing in the Ames, Iowa area is $180796.06 with a 95% confidence interval of (165570.6 , 197066.3)

```r
#Average Mean Price of Dataset
mean_price <- mean(ames$price)
cat("Average Price:\n", paste(mean_price))
```

```
## Average Price:
##  180796.060068259
```

```r
## Average Price:
##  180796.060068259
#Run Bootstrapping to Generate Confidence Intervals
set.seed(123)
bootstrap_sample <- c()
```

```r
for (i in 1:10000) {
  bs <- sample(ames$price, 100, replace = TRUE)
  bootstrap_sample <- c(bootstrap_sample, mean(bs)) }

#Visualize Bootstrap Distribution
hist(bootstrap_sample, main = "Bootstrap Distribution of Means", xlab = "Mean
Value")
```




```r
#95% Confidence Interval
quantile(bootstrap_sample, c(0.025, 0.975))

##     2.5%    97.5%
## 165570.6 197066.3
```

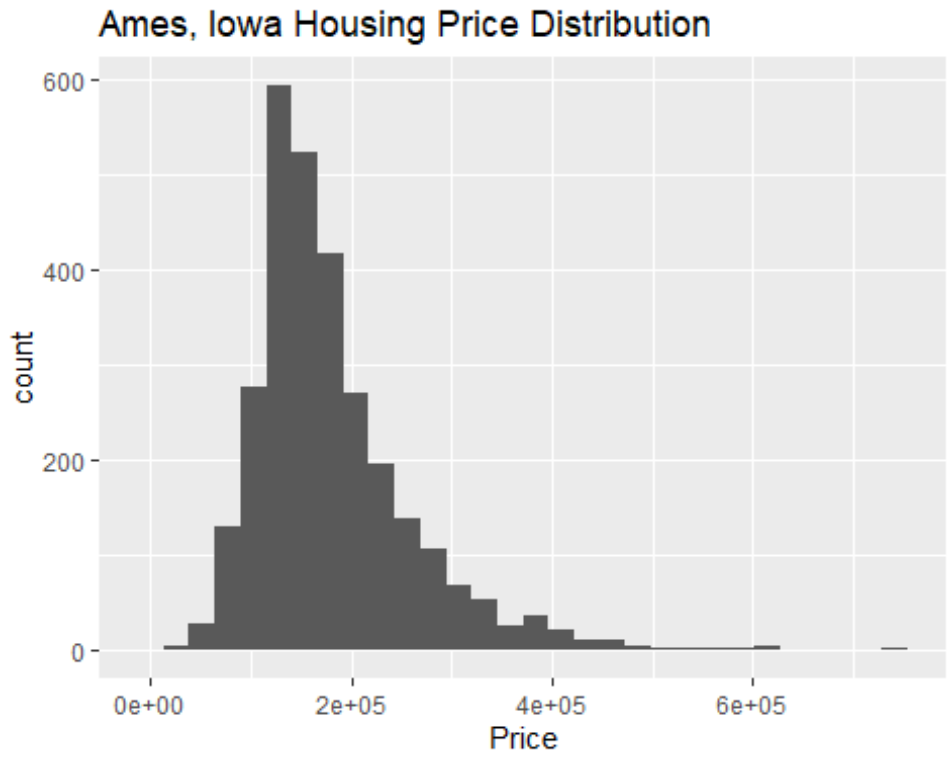
How is the price distributed? How about the housing areas and years built?

We analyzed the distribution of housing prices, area, and construction years in Ames, Iowa. The housing price distribution is right-skewed, with a median of $160,000 and a mean of $180,796. The area distribution also skews to the right, indicating the presence of outliers with exceptionally large housing areas. Conversely, the distribution of construction years skews to the left, suggesting that the majority of houses in the dataset were built after 1950. Utilizing the ggplot histogram function, we visualized these distributions and collected summary statistics for a comprehensive understanding of the dataset.

```r
#price
 ggplot(ames, aes(x = price)) +
```

```
  geom_histogram() +
  labs(title = "Ames, Iowa Housing Price Distribution", x = "Price")
```

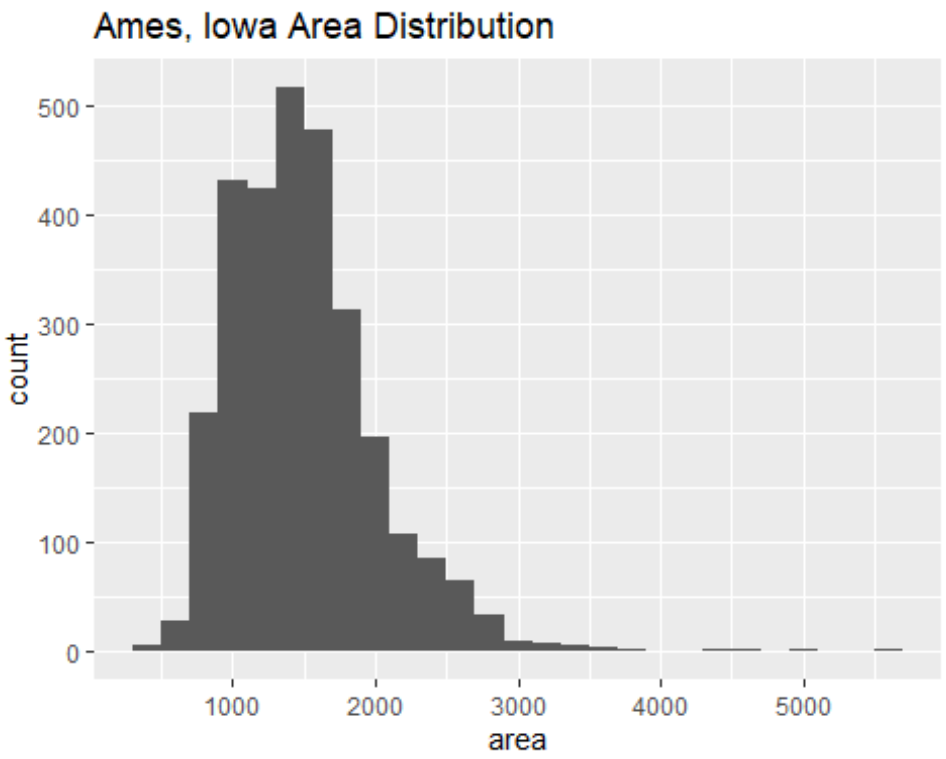
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



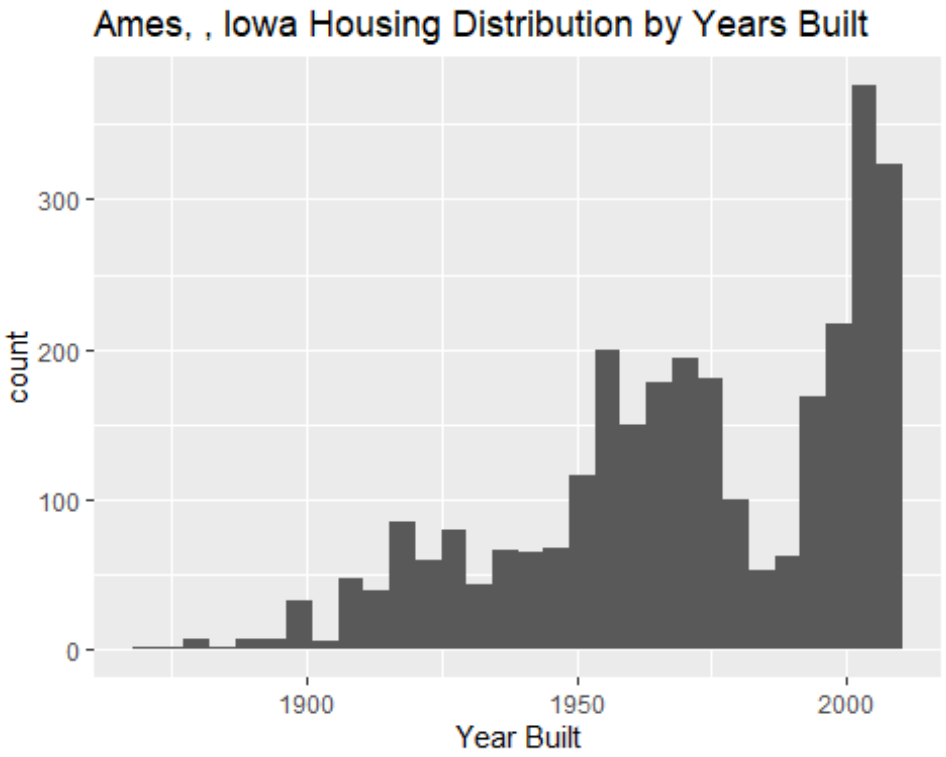

```
summary(ames$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12789  129500  160000  180796  213500  755000
```

```
ames|> ggplot(aes(area))+ labs(title = "Ames, Iowa Area Distribution") +
geom_histogram(binwidth = 200)
```




```
ames|> ggplot(aes(x = year_built))+ labs(title = "Ames, , Iowa Housing
Distribution by Years Built", x = "Year Built") + geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Are there any significant differences in housing prices across different zoning densities?

To address this inquiry, we initially generated a boxplot visualization illustrating the average housing prices across different zones. Subsequently, we expanded the Ames dataset by utilizing the pivot_wider() function to delineate the 'ms_zoning' column into distinct variables. The resulting dataset, named "distinct_zones_prices," encapsulates columns representing individual zones. We proceeded to calculate the average housing prices for each zone and performed bootstrapping simulations to derive confidence intervals. During this process, we eliminated any 'N/A' values from the dataset. Additionally, certain variable names, such as "C(all)," underwent renaming for clarity and consistency in the analysis. The following are the distinct zoning areas that underwent analysis.

MS Zoning (Nominal): Identifies the general zoning classification of the sale. - A (agr) Agriculture - C (all) Commercial - FV Floating Village Residential - I (all) Industrial - RH Residential High Density - RL Residential Low Density - RP Residential Low Density Park - RM Residential Medium Density

From the box-plots, we see that the Floating Village Residential zone and Residential with low density zone have the highest medians. The Agricultural, Commercial, and Industrial residential zones have lowest medians. Among all boxplots, the residential zone with low density has the largest spread with noticeable outliers. Why is there such a big price

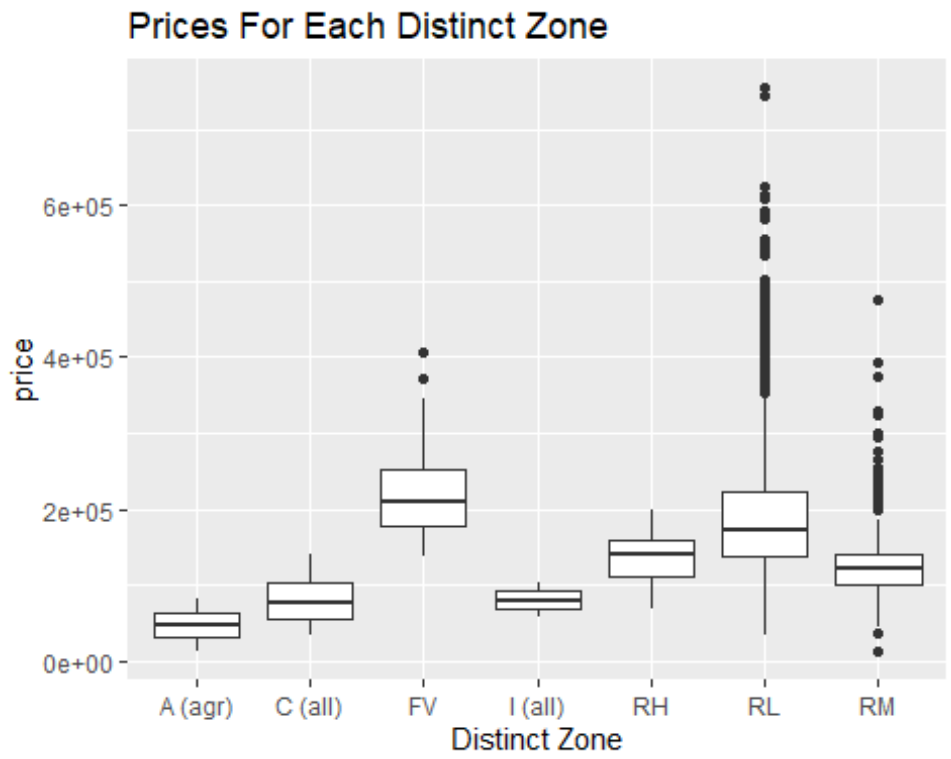
difference in low population density areas? The reason may be that in places with low population density, houses can be built larger and have more living space, so prices will rise. We make boxplots for living areas, and there are also noticeable outliers in the low density area.

```
#All prices for each distinct zone
ames |> distinct(ms_zoning)
```

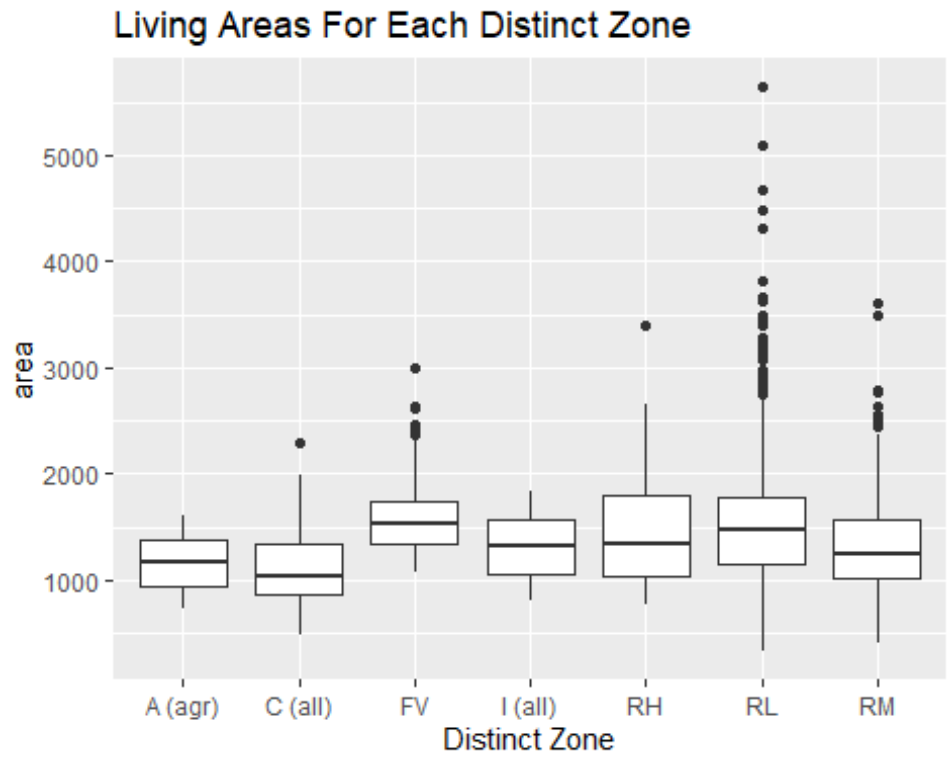
```
##   ms_zoning
## 1        RL
## 2        RH
## 3        FV
## 4        RM
## 5   C (all)
## 6   I (all)
## 7   A (agr)
```

```
##   ms_zoning
## 1        RL
## 2        RH
## 3        FV
## 4        RM
## 5   C (all)
## 6   I (all)
## 7   A (agr)
```

```
ggplot(ames, aes(x = ms_zoning, y = price)) + geom_boxplot() + labs( title =
"Prices For Each Distinct Zone", x = 'Distinct Zone')
```




```r
ggplot(ames, aes(x = ms_zoning, y = area)) + geom_boxplot() + labs( title = "Living Areas For Each Distinct Zone", x = 'Distinct Zone')
```



```r
#Mean Prices for Each Distinct Zone
distinct_zones_prices <- ames %>%
  pivot_wider(
    names_from = 'ms_zoning' ,
    values_from = 'price'
  )

distinct_zones_prices <- select(distinct_zones_prices, 'RL':'A (agr)')

mean_price_per_zone <- colMeans(distinct_zones_prices, na.rm = TRUE)

cat("Mean Prices per Zone:\n")
```

```
## Mean Prices per Zone:
```

```r
## Mean Prices per Zone:
cat(paste(names(mean_price_per_zone), ": ", mean_price_per_zone, "\n"), sep =
"")
```

```
## RL :  191283.251649802
## RH :  136419.777777778
## FV :  218986.949640288
## RM :  126781.393939394
## C (all) :  79795.04
## I (all) :  80312.5
## A (agr) :  47300
```

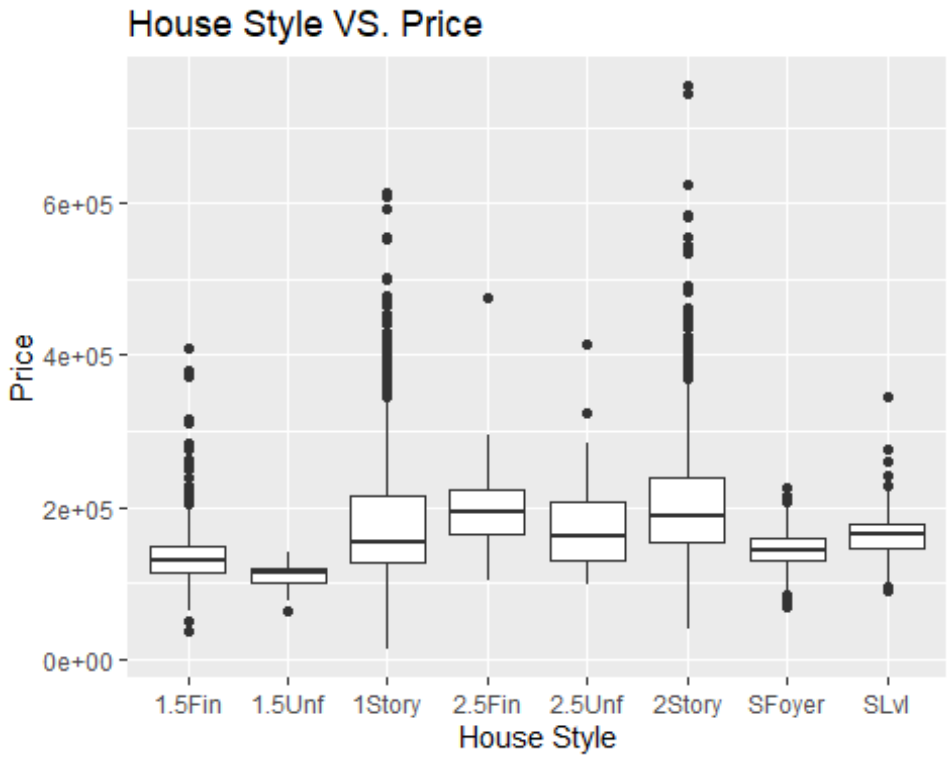
```r
## RL :  191283.251649802
## RH :  136419.777777778
## FV :  218986.949640288
## RM :  126781.393939394
## C (all) :  79795.04
## I (all) :  80312.5
## A (agr) :  47300
```

What's the relationship between housing price and house style?

To address this question, we created a boxplot distribution of the different housing styles and plotted them in relation to the price of the houses. A list of the different housing styles can be found below: House Style (Nominal): Style of dwelling 1Story One story 1.5Fin One and one-half story: 2nd level finished 1.5Unf One and one-half story: 2nd level unfinished 2Story Two story 2.5Fin Two and one-half story: 2nd level finished 2.5Unf Two and one-half story: 2nd level unfinished SFoyer Split Foyer SLvl Split Level

2 and 2.5 story houses have higher prices. 2nd level finished houses have higher prices than unfished houses. For example, the median for 1.5Fin is larger than 1.5Unf,and it's the same for 2.5Fin and 2.5Unf.

```r
ggplot(ames, aes(x= house_style, y= price)) + labs(title = 'House Style VS.
Price', x= 'House Style', y='Price') + geom_boxplot()
```



Data Analysis

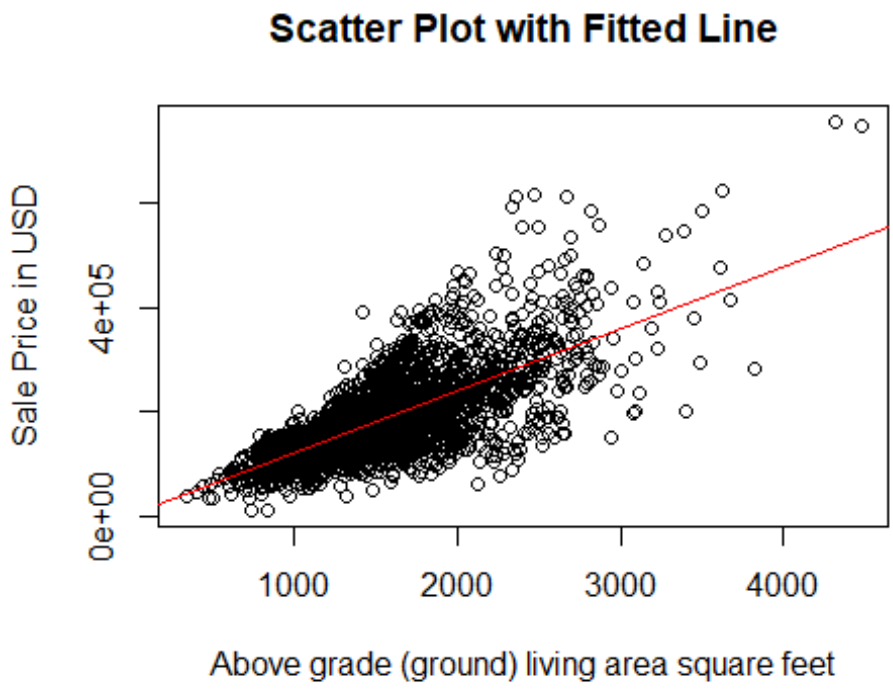
Explore the relationship between variables: What's the relationship between price and area?

• There is a strong positive relationship between the price and housing area. The larger the house, the more it costs. However, there're a few outliers that has very large housing areas but low housing price. • Given the living area, we can predict the price according to the linear model: • price = 13289.634 + 111.694 area. • MSE = 3192801087 • R^2:0.5287 of the price can be explained by area.

```
ames_filtered <- ames|> filter(area < 4500)
plot(ames_filtered$area, ames_filtered$price, main = "Scatter Plot with
Fitted Line",
     xlab = "Above grade (ground) living area square feet", ylab = "Sale
Price in USD")
#Linear Model
price_lm <- lm(price~area, data=ames_filtered)
summary(price_lm)

##
## Call:
## lm(formula = price ~ area, data = ames_filtered)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -205124  -30582   -1157   23699  328320
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4091.200   3247.361    1.26    0.208
## area         118.124      2.062   57.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54880 on 2925 degrees of freedom
## Multiple R-squared:  0.5287, Adjusted R-squared:  0.5285
## F-statistic:  3281 on 1 and 2925 DF,  p-value: < 2.2e-16
```

```
abline(price_lm, col = "red")
```




```
#MSE
lm_pred_mse <- (price_lm$residuals)^2
mean(lm_pred_mse)
```

```
## [1] 3009725679
```

What variables summarize most of the variation in sales price?

• Sale Condition (Nominal): Condition of sale Normal: Normal Sale Abnorml: Abnormal Sale - trade, foreclosure, short sale AdjLand: Adjoining Land Purchase Alloca: Allocation - two linked properties with separate deeds, typically condo with a garage unit

Family Sale between family members Partial Home was not completed when last assessed (associated with New Homes)

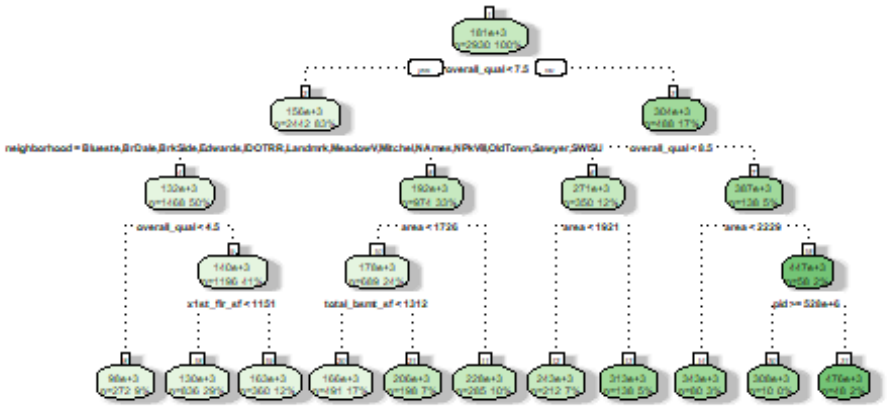
• It is evident that larger residences come with higher price tags, and additional incentives are provided for “Partial” sales (pertaining to new homes that are only partially completed at the last assessment), while discounts are offered for “Abnormal” sales (such as short sales and foreclosures).

```
ames |> ggplot(aes(area, price, color=sale_condition, shape =
sale_condition))+geom_point()
```

We also make a tree plot to predict the housing price. We can conclude from the tree that overall qualities, neighborhood, and above ground living area are three main factors that affect the housing price.

```
    library(rpart)

## Warning: package 'rpart' was built under R version 4.3.2

    library(rattle)

## Warning: package 'rattle' was built under R version 4.3.2

## Loading required package: bitops

## Rattle: A free graphical interface for data science with R.
## Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
full_tree <- rpart(price~.,cp=0.01,minbucket=10,data=ames)

# print tree
fancyRpartPlot(full_tree)
```



Rattle 2023-Dec-12 02:29:52 KCVUSA1

Is there a relationship between the overall quality and overall condition of a house, and the house's price?

In addressing this inquiry, we initially underwent a process of converting the variables "overall_qual" and "overall_cond" into character types. Subsequently, we confirmed that both variables were of factor type. Employing boxplot distributions, we visually examined the relationship between the quality and condition of a house and its corresponding price. The observed patterns indicated a positive correlation between the quality and condition of a residence and its market value.

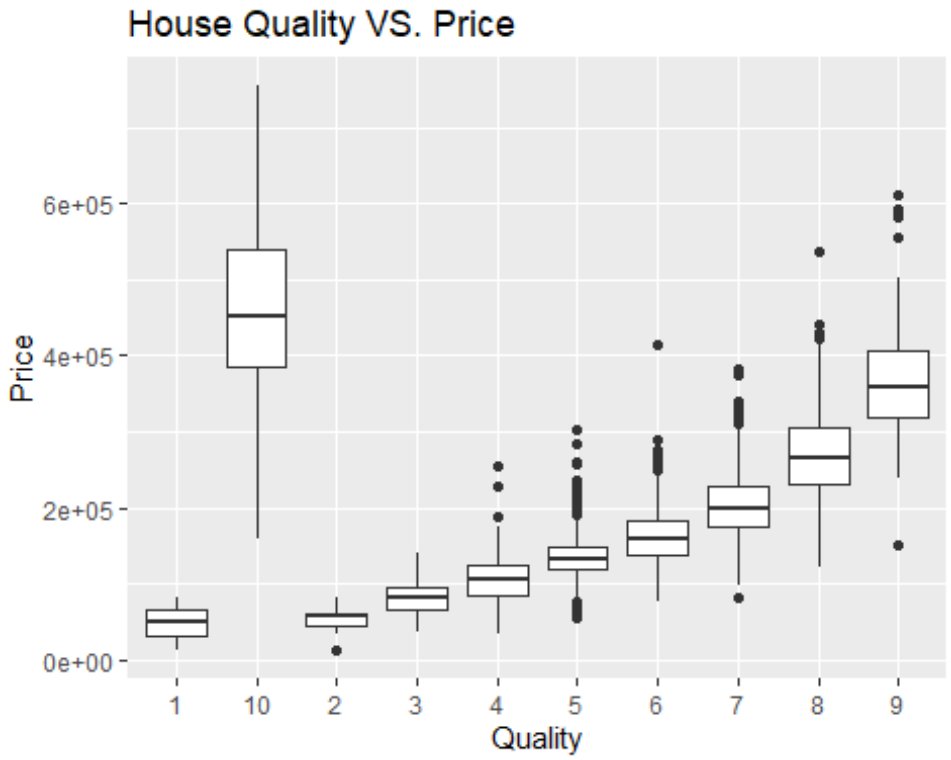
Overall Qual (Ordinal): Rates the overall material and finish of the house

```
10   Very Excellent
9    Excellent
8    Very Good
7    Good
6    Above Average
5    Average
4    Below Average
3    Fair
```

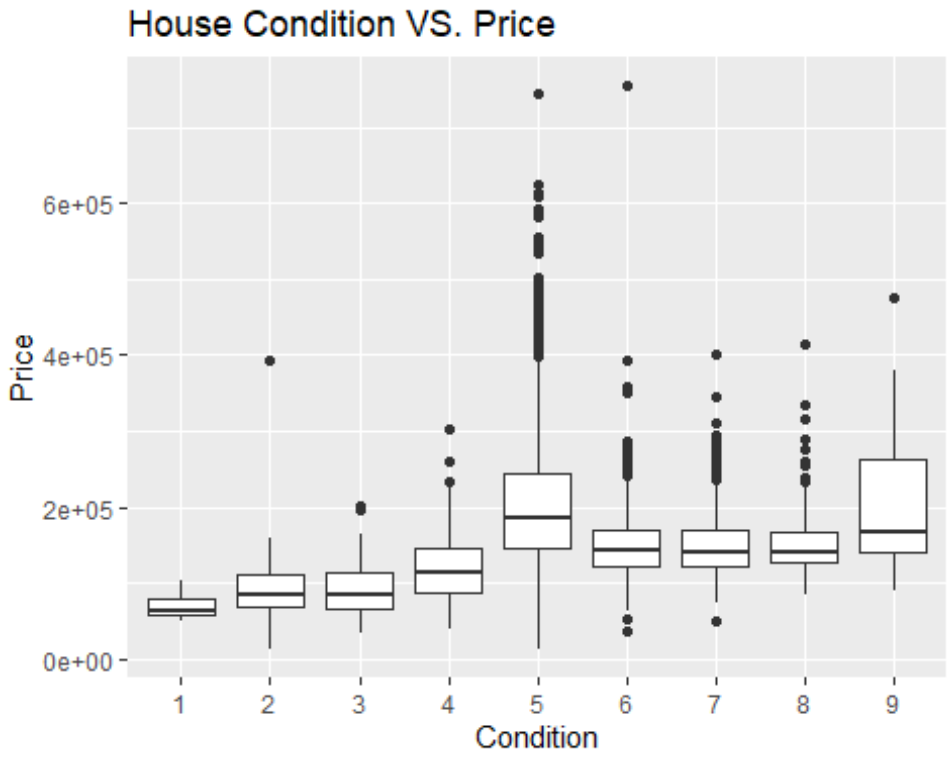
```
   2     Poor
   1     Very Poor
```

To discern whether the observed disparities in housing prices were directly attributable to variations in quality or condition, we conducted an Analysis of Variance (ANOVA) hypothesis test. Upon statistical analysis, we rejected the null hypothesis, establishing that both quality and condition exerted a significant influence on the observed differences in housing prices.

```
#Visualizations

quality <- as.character(ames$overall_qual)
condition <- as.character(ames$overall_cond)
overall_quality <- as.factor(ames$overall_qual)
overall_condition <- as.factor(ames$overall_cond)

ggplot(ames, aes(x = quality, y = price)) + labs(title = 'House Quality VS.
Price', x= 'Quality', y='Price') + geom_boxplot()
```




```
ggplot(ames, aes(x = condition, y = price)) + labs(title = 'House Condition
VS. Price', x= 'Condition', y='Price') + geom_boxplot()
```




```
quality <- as.character(ames$overall_qual)
condition <- as.character(ames$overall_cond)
overall_quality <- as.factor(ames$overall_qual)
overall_condition <- as.factor(ames$overall_cond)

#Hypothesis Testing:
#Null Hypothesis (H0): There is no significant difference in the mean sale
prices across different levels of overall quality and condition.
#Alternative Hypothesis (H1): There is a significant difference in the mean
sale prices across different levels of overall quality and condition.

qual_anova_result <- aov(price ~ overall_qual, data = ames)
print(summary(qual_anova_result))

##                Df    Sum Sq   Mean Sq F value Pr(>F)
## overall_qual    1 1.194e+13 1.194e+13    5179 <2e-16 ***
## Residuals    2928 6.751e+12 2.306e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##                Df    Sum Sq   Mean Sq F value Pr(>F)
## overall_qual    1 1.194e+13 1.194e+13    5179 <2e-16 ***
## Residuals    2928 6.751e+12 2.306e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
cond_anova_result <- aov(price ~ overall_condition, data = ames)
print(summary(cond_anova_result))
```

```
##                      Df    Sum Sq   Mean Sq F value Pr(>F)
## overall_condition     8 2.809e+12 3.511e+11   64.58 <2e-16 ***
## Residuals          2921 1.588e+13 5.438e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary

The analysis of the Ames, Iowa housing dataset reveals several key findings. Zoning density variations result in distinct price differences, with residential low-density areas exhibiting higher median prices and a broader spread. House styles, particularly the number of stories and finished levels, significantly influence median prices. A strong positive correlation is identified between housing prices and living area, though outliers with large areas and low prices exist. Sale condition emerges as a pivotal factor, indicating higher prices for larger residences and new homes, while abnormal sales offer discounts. Overall quality and condition demonstrate a positive correlation with housing prices, as validated by ANOVA testing. It is shown from the CART tree that overall qualities, neighborhood, and area are important factors when predicting housing prices. One of our weaknesses is that the bootstrap distribution was created on the assumption that the prices do not vary based on time. We are not sure if the bootstrap distribution would work accurately when there is large variability in price across different time periods. If we have additional data, we can compare whether the variables affect housing prices in other areas besides Ames, Iowa. The linear regression can be improved upon by utilizing multiple regressions taking into account more variables in the dataset.

References:

De Cock, Dean. “Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project.” Journal of Statistics Education 19.3 (2011).