# **Appendix: R Code and Visualizations for House Price Prediction**

### **Mohit Maithani**

ALY 6040 - Data Mining  
Instructor: Justin Grosz

## **1. Introduction**

This appendix contains the **complete R code and visualizations** used for the analysis, exploration, and modeling of house prices in King County. The models include **Multiple Linear Regression, Decision Tree, and Random Forest**, and the visualizations support key findings in the main report. All figures are labeled for reference.

## **2. Load Required Libraries**

library(ggplot2)

library(dplyr)

library(corrplot)

library(rpart)

library(rpart.plot)

library(caret)

library(randomForest)

## **3. Load and Preprocess Data**

kc\_data <- read.csv("kc\_house\_data.csv", stringsAsFactors = FALSE)

kc\_data <- kc\_data[!duplicated(kc\_data$id), ] # Remove duplicates

kc\_data$zipcode <- as.factor(kc\_data$zipcode) # Convert Zipcode to Factor

kc\_data$log\_price <- log(kc\_data$price) # Log Transformation of Price

## **4. Encode Zipcode into Clusters**

zipcode\_median\_price <- kc\_data %>%

group\_by(zipcode) %>%

summarize(median\_price = median(price)) %>%

arrange(median\_price)

zipcode\_median\_price$zipcode\_cluster <- as.numeric(factor(zipcode\_median\_price$median\_price,

levels = unique(zipcode\_median\_price$median\_price)))

kc\_data <- merge(kc\_data, zipcode\_median\_price, by = "zipcode")

kc\_data$zipcode <- NULL # Remove Original Zipcode Column

## **5. Train-Test Split**

set.seed(123)

trainIndex <- createDataPartition(kc\_data$log\_price, p = 0.8, list = FALSE)

train\_data <- kc\_data[trainIndex, ]

test\_data <- kc\_data[-trainIndex, ]

## **6. Exploratory Data Analysis (EDA) and Visualizations**

### **6.1 House Price Distribution (Before & After Log Transformation)**

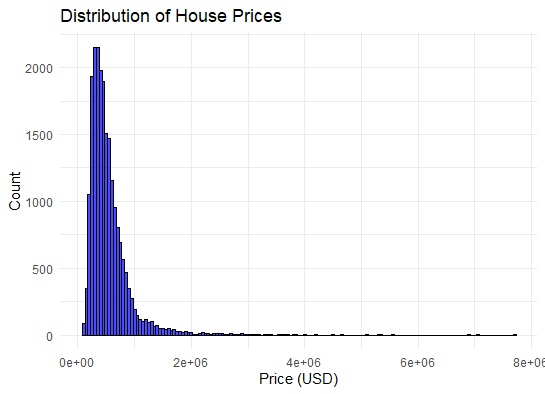
ggplot(kc\_data, aes(x = log\_price)) +

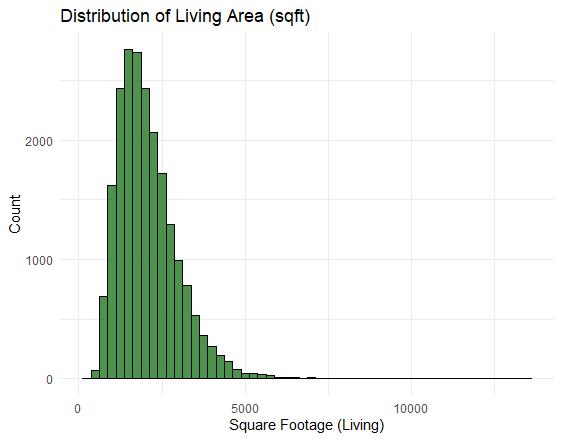
geom\_histogram(binwidth = 0.1, fill = "purple", color = "black", alpha = 0.7) +

labs(title = "Log-Transformed Price Distribution", x = "Log Price", y = "Count") +

theme\_minimal()

*(See Figure 1: Log-Transformed Price Distribution)*

**

**

### **6.2 Scatter Plot: Sqft Living vs. Price**

ggplot(kc\_data, aes(x = sqft\_living, y = price)) +

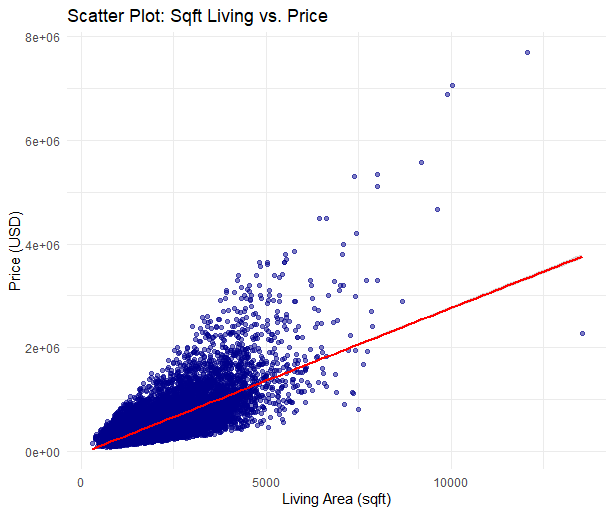
geom\_point(color = "darkblue", alpha = 0.5) +

geom\_smooth(method = "lm", color = "red") +

labs(title = "Scatter Plot: Sqft Living vs. Price", x = "Living Area (sqft)", y = "Price (USD)") +

theme\_minimal()

(See Figure 2: Scatter Plot - Sqft Living vs. Price)

**

### **6.3 Seasonality Effects on House Prices**

library(lubridate)

# Convert Date Column to Proper Date Format

kc\_data$date <- as.Date(kc\_data$date, format="%Y%m%d")

# Extract Month and Year for Seasonality Analysis

kc\_data$month <- month(kc\_data$date, label = TRUE) # Month as Factor (Jan, Feb, etc.)

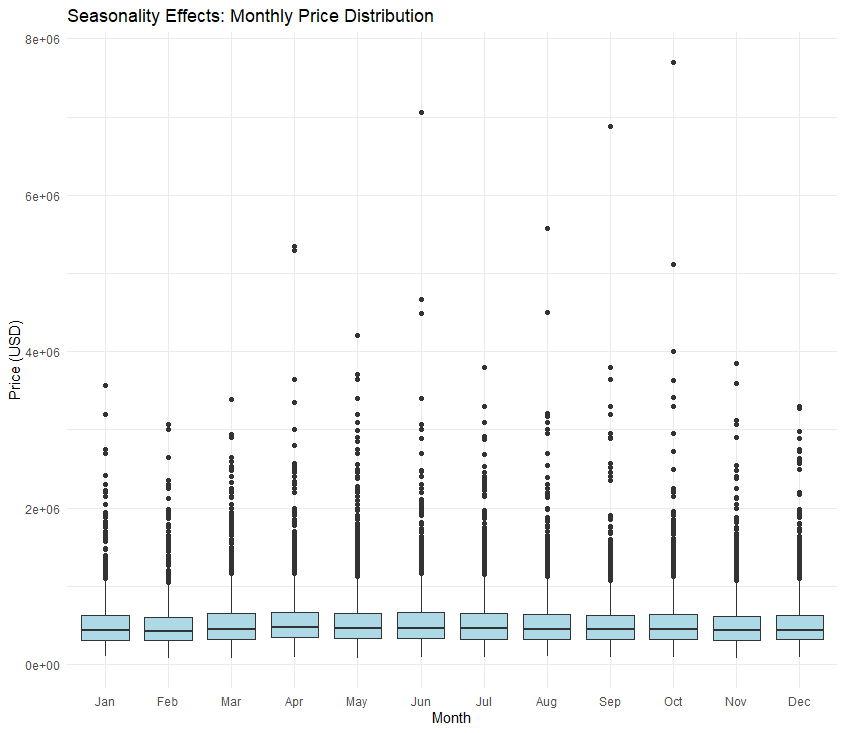
kc\_data$year <- year(kc\_data$date)

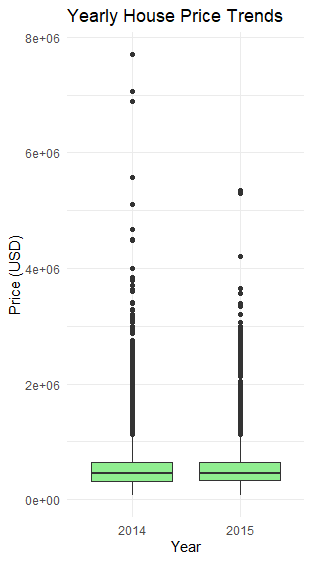
ggplot(kc\_data, aes(x = month, y = price)) +

geom\_boxplot(fill = "lightblue") +

labs(title = "Seasonality Effects: Monthly Price Distribution", x = "Month", y = "Price (USD)") +

theme\_minimal()





**Figure 8** → Yearly House Price Trends - Seasonality Effects

ggplot(kc\_data, aes(x = factor(year), y = price)) +

geom\_boxplot(fill = "lightgreen") +

labs(title = "Yearly House Price Trends", x = "Year", y = "Price (USD)") +

theme\_minimal()

### 

### 

### 

### 

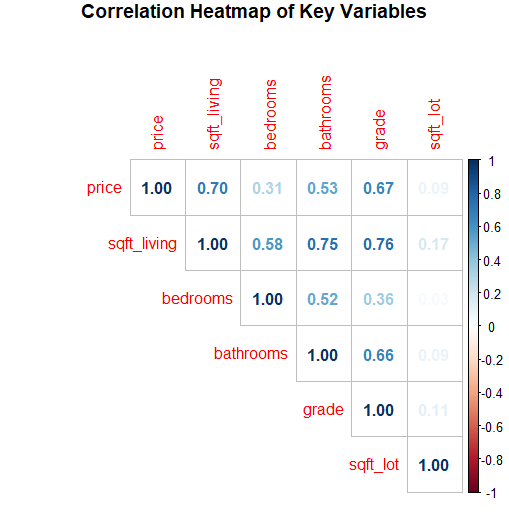
### **6.4 Correlation Heatmap of Key Variables**

cor\_matrix <- cor(kc\_data[, c("price", "sqft\_living", "bedrooms", "bathrooms", "grade", "sqft\_lot")])

corrplot(cor\_matrix, method = "number", type = "upper",

title = "Correlation Heatmap of Key Variables", mar=c(0,0,1,0))

*(See Figure 3: Correlation Heatmap of Key Variables)*

**

## **7. Multiple Linear Regression Model**

lm\_model\_zip <- lm(log\_price ~ sqft\_living + bedrooms + bathrooms + grade + zipcode\_cluster, data = train\_data)

summary(lm\_model\_zip)

predictions\_log\_lm\_zip <- predict(lm\_model\_zip, newdata = test\_data)

predictions\_lm\_zip <- exp(predictions\_log\_lm\_zip)

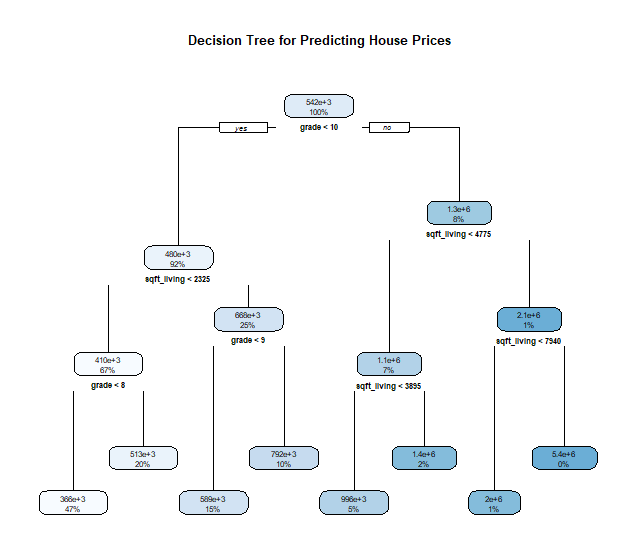
*(See Figure 4: Regression Model Residual Analysis)*

## **8. Decision Tree Model**

tree\_model\_zip <- rpart(log\_price ~ sqft\_living + bedrooms + bathrooms + grade + zipcode\_cluster, data = train\_data, method = "anova")

rpart.plot(tree\_model\_zip, main = "Decision Tree for Predicting House Prices (With Zipcode Clusters)")

*(See Figure 5: Decision Tree Model Output)*

**

## **9. Random Forest Model**

rf\_model\_zip <- randomForest(log\_price ~ sqft\_living + bedrooms + bathrooms + grade + zipcode\_cluster,

data = train\_data, ntree = 500, importance = TRUE)

predictions\_rf\_zip <- predict(rf\_model\_zip, newdata = test\_data)

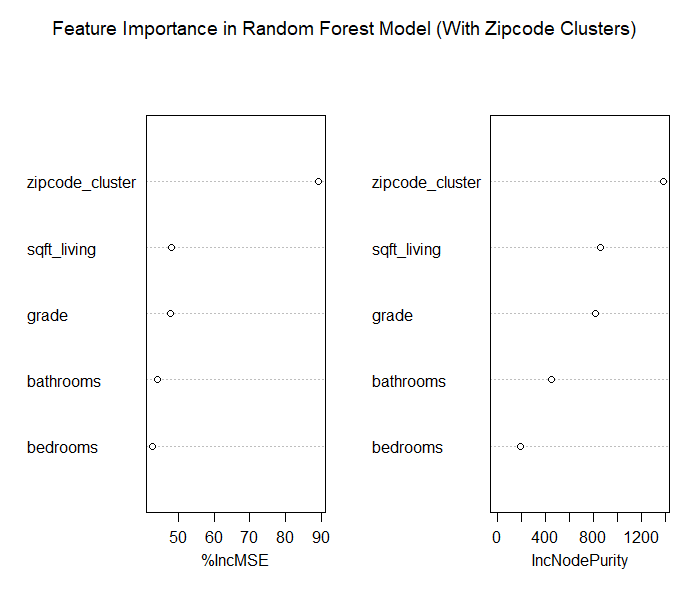
predictions\_rf\_zip <- exp(predictions\_rf\_zip)

## **10. Feature Importance from Random Forest**

importance(rf\_model\_zip)

varImpPlot(rf\_model\_zip, main="Feature Importance in Random Forest Model (With Zipcode Clusters)")

*(See Figure 6: Feature Importance in Random Forest Model)*

**

This appendix provides the **complete R code and visualizations** for the house price prediction project, covering **data preprocessing, EDA, and model training** using **Multiple Linear Regression, Decision Tree, and Random Forest models**.

All major visualizations, including **price distribution, correlation heatmap, scatter plots, decision tree structures, and feature importance plots**, are labeled accordingly. Please refer to the **figures referenced** for further insights.