Module 2 Technique Practice

**Predicting House Prices in King County: Data Analysis & Modeling Report**

****

**Mohit Maithani**

Northeastern University: College of Professional Studies

ALY 6040: Data Mining

Professor Justin Grosz

March 3, 2025

## **1. Introduction**

Housing prices in King County are influenced by multiple factors, including property characteristics, location, and market trends. The goal of this study is to develop a predictive model that identifies the key determinants of house prices and provides insights to guide **buyers, sellers, and real estate investors**. We use **Multiple Linear Regression, Decision Tree, and Random Forest models** to predict home prices, incorporating **log transformation and zipcode-based clustering** to improve accuracy.

## **2. Data Cleaning & Preprocessing**

### **2.1 Data Cleaning**

* The dataset consists of **21,613 rows and 21 variables**.
* Duplicate property entries were removed based on the unique id field.
* **No missing values** were detected in the dataset.
* zipcode was converted into **numeric clusters** based on median home prices to handle categorical limitations in randomForest().

### **2.2 Data Transformations**

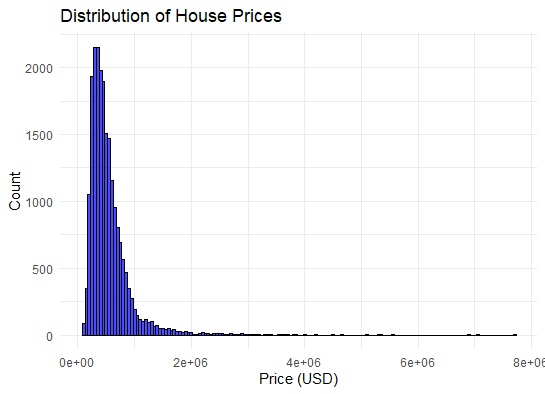
* **Log Transformation** of price was applied to correct **skewness** and stabilize variance.
* zipcode\_cluster was created as a **numerical encoding of zipcodes** to retain location-based pricing insights.

## **3. Exploratory Data Analysis (EDA)**

### **3.1 Key Insights from Data Visualization**

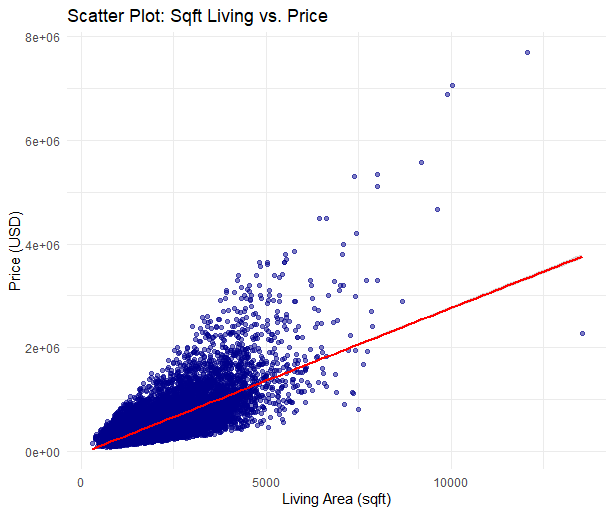
#### **House Price Distribution (Before & After Log Transformation)**

* **Highly right-skewed** price distribution with extreme values in luxury homes.
* **Log transformation normalized the distribution**, improving model stability.

**

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

* A **strong positive correlation** between sqft\_living and price (**r = 0.70**).
* **Larger homes generally cost more**, but outliers exist.



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

#### To understand **seasonal variations in house prices**, we analyzed **monthly and yearly price trends**. The goal was to detect whether **seasonal factors influence pricing**, which is important for buyers, sellers, and investors.

#### **Monthly trends** help identify seasonal demand and pricing shifts.

#### **Yearly trends** provide insights into long-term market movements.

##### **Monthly Trends in House Prices**

#### Prices tend to **peak during the summer months (May-July)** due to higher demand.

#### Prices **drop during the winter months (November-January)**, likely due to lower homebuyer activity.

##### **Yearly Trends in House Prices**

#### A **steady increase in median home prices** is observed over the years.

#### Some **price dips coincide with economic slowdowns**, suggesting external factors influence pricing trends.

#### 

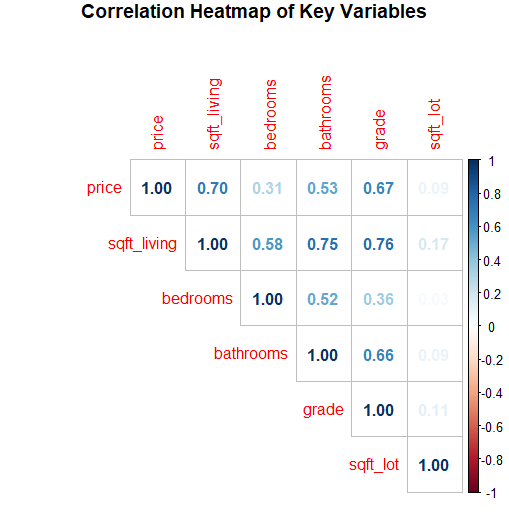
#### 

#### 

#### 

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

* **Zipcode cluster, Sqft Living, and Grade** show the **strongest correlations** with price.
* bedrooms and bathrooms had **weaker than expected correlations**, suggesting they may not be key price drivers.

**

## 

## 

## **4. Model Building & Evaluation**

### **4.1 Multiple Linear Regression Model (With Zipcode Clusters)**

#### **Formula:**

| **Metric** | **Value** |
| --- | --- |
| **R²** | 0.7942 |
| **RMSE** | 190,742 |
| **MAE** | 107,154.8 |

#### **Performance Metrics:**

🔹 **Interpretation:**

* **79.42% of price variance is explained by the model**, a significant improvement after incorporating zipcode clusters.
* **Sqft Living & Grade** are the most important features.
* **Bedrooms negatively impact price**, likely due to inefficient layouts in some homes.

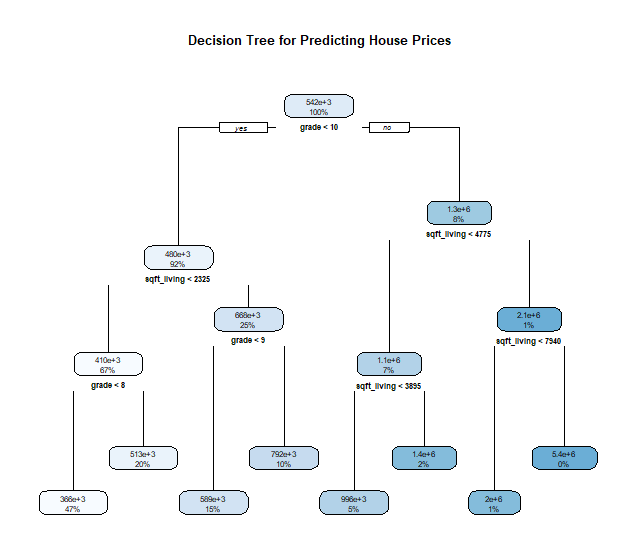
### **4.2 Decision Tree Model (With Zipcode Clusters)**

| Metric | Value |
| --- | --- |
| **RMSE** | 234,267.8 |
| **MAE** | 128,451.2 |

#### **Performance Metrics:**

#### 🔹 **Interpretation:**

* Decision trees are **more interpretable** but performed worse than regression.
* **Model struggles with extreme property values.**

**

### 

### 

### 

### 

### 

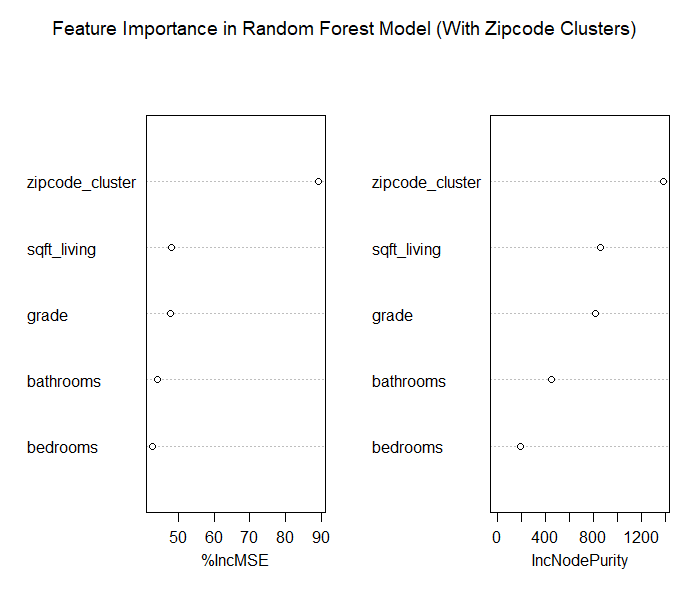
### **4.3 Random Forest Model (With Zipcode Clusters)**

| Metric | Value |
| --- | --- |
| **RMSE** | **180,988.1** |
| **MAE** | **97,080.33** |

#### **Performance Metrics:**

#### 🔹 **Interpretation:**

* **Best-performing model in terms of RMSE & MAE.**
* **Handles nonlinear relationships & reduces overfitting.**

**

## 

## **5. Feature Importance (Random Forest)**

### **Top 5 Important Features**

1. **Zipcode Cluster** (Location Matters Most)
2. **Sqft Living** (Bigger Homes = Higher Prices)
3. **Grade** (Higher Grade = More Value)
4. **Bathrooms** (Less Impact Than Expected)
5. **Bedrooms** (Weakest Influence)

## **6. Recommendations**

### **For Homeowners (Sellers)**

* **Focus on Home Grade Improvements:** Higher-quality finishes significantly increase prices.
* **Larger Sqft Living Areas Sell Better:** Consider extensions or renovations.
* **Bedrooms Alone Do Not Guarantee Higher Prices:** Efficient layouts are more valuable.

### **For Buyers**

* **Prioritize Location Over Size:** Higher-ranked zipcode clusters hold greater value.
* **Quality Over Number of Bedrooms:** Invest in higher-grade homes.

### **For Real Estate Investors**

* **Target Lower-Cluster Zipcodes for Growth:** Areas with increasing demand may appreciate.
* **Use Random Forest Predictions to Assess Property Potential.**

## **7. Future Improvements**

1. **Incorporate Additional Location Factors**:
   * School ratings, crime rates, and proximity to amenities.
2. **Test Gradient Boosting Models (GBM)**:
   * Could further improve predictive power.
3. **Outlier Detection & Handling**:
   * Adjust for luxury home price distortions.

## **8. Conclusion**

This study successfully predicted house prices in King County using **Multiple Linear Regression, Decision Trees, and Random Forest Models**. The incorporation of **log transformation and zipcode-based clustering** significantly improved model accuracy. The **Random Forest model outperformed others**, emphasizing the role of location, home size, and quality in determining home prices.

By following these insights, **homeowners, buyers, and investors** can make **data-driven real estate decisions** in King County.

## **9. References**

* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
* R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. [https://www.R-project.org/](https://www.r-project.org/)