Assignment 2
Predictive Analysis of Housing Prices in King County: A Guide for Real Estate Investors

Course: ALY6040 Data Mining Applications

Professor: Justin Grosz

Submitted by: Kaushal Nagrecha

Abstract

This study examines the primary determinants of house prices in King County, Washington, using exploratory data analysis (EDA) and predictive modeling. From an examination of 21,613 house sales, this report uncovers property characteristics that drive value, such as square footage, location, condition, and upgrades. A Multiple Linear Regression model and a Random Forest model are used to predict house prices, and the findings show Random Forest outperforms Linear Regression in predictive capability. The report aims to provide investment guidance, guiding real estate investors on the most suitable property acquisition and refurbishment opportunities.

Introduction

The housing market is affected by several factors such as property attributes, location, and economic trends (Gyourko et al., 2013). The investors aim at high-return houses while keeping the risks low. This research applies machine learning algorithms to forecast the prices of houses, offering hints on high-potential investments.

Research Objectives:

1. Identify significant property features affecting house prices.
2. Develop a predictive model to predict property prices.
3. Provide investment suggestions based on data-driven analysis.

Data & Methodologies

Dataset Overview

The dataset includes 21,613 property sales in King County, WA. Key variables include:

- Dependent Variable: House price ($).
- Independent Variables: Square footage, bedrooms, bathrooms, floors, waterfront status, view quality, condition, grade, year built, renovation status, location (latitude, longitude, and zip code).

Data Cleaning & Preparation

A robust **data preparation pipeline** ensures model accuracy and removes anomalies.

1. Handling Missing Values

- The dataset **has no missing values**, reducing preprocessing complexity.

2. Removing Outliers & Data Anomalies

- **Zero bedrooms & bathrooms** → Likely errors or misclassified properties. **Removed** these rows.
- **Bedrooms > 10** → Outlier values unlikely for typical residential properties. **Removed** these rows.

3. Transforming Categorical Variables

- **Zip Code Encoding**: Converted zipcode into **dummy variables (one-hot encoding)** for modeling.
- **Date Format**: Transformed date into a **datetime format**.

4. Feature Engineering

- **Age of Property**: Derived from yr_built and yr_renovated to capture renovation effects.
- **Basement Proportion**: sqft_basement / sqft_living to analyze basement impact.

Model Selection

1. **Multiple Linear Regression (MLR)**: Establishes a baseline for price prediction.
2. **Random Forest Regressor (RFR):** Captures non-linear price relationships for better accuracy.

Results & Analysis

Exploratory Data Analysis (EDA)

Price Distribution

- Prices range from **$75,000 to $7.7 million**.
- **Most homes are priced between $300K and $700K**.
- **Median price**: **$450,000**
- **Right-skewed distribution** → Some ultra-luxury properties create a long tail.

Key Price Drivers (Correlation Analysis)

- sqft_living has the **strongest correlation** with price (**r = 0.70**).
- **Bathrooms impact price more than bedrooms.**
- **Renovations significantly increase home values** (up to 30% if renovated after 2000).

Feature	Correlation
sqft_living	(highest corr) 0.70
grade	0.67
sqft_above	0.61
bathroom	0.53
view	0.4

Table 1: Feature Correlation

Predictive Model Performance

Model	R^2	RMSE	MAE	AIC
Multiple Linear Regression	0.81	167018.49	97998.11	78117.62
Random Forest	0.87	133578.17	73376.74	76669.43

Table 2: Comparison of Predictive Models

Random Forest is the superior model, capturing 87% of price variations (compared to 81%), an RSME of ~13K (compared to ~16K), and an MAE of ~73K (compared to ~97K) and an AIC of ~76K (compared to ~78K).

Investment Strategies

High-ROI Property Acquisition

Investors should target:

- **3–5-bedroom homes** in **growing neighborhoods** with moderate price appreciation.
- **Pre-1980 homes with no renovations** (to renovate and increase value).
- **Waterfront properties**, which are valued **80% higher than non-waterfront homes**.

Renovation Strategy for Maximum ROI

- **Focus on bathrooms & kitchens** (highest return on upgrades).
- **Energy-efficient renovations** increase demand (e.g., solar panels, smart home tech).
- **Avoid excessive luxury upgrades**—mid-range finishes provide better ROI.

Buy-and-Hold vs. Fix-and-Flip

Fix & Flip Strategy (Short-Term Gains)

- **Target homes priced below $500K**, needing cosmetic updates.
- **Renovate kitchens and bathrooms** to increase market value.
- **Sell within 12-18 months** to maximize capital gains.

Buy-and-Hold Strategy (Long-Term Rental Income)

- **Invest in 3–4-bedroom homes** in **high-rental demand zip codes** (e.g., 98133, 98042).
- **Waterfront rentals attract premium tenants** and yield **higher cash flow**.
- **Multi-family units offer stable income streams** (consider duplex/triplex investments).

Market Risks & External Factors

While predictive modeling improves investment decisions, external factors should also be considered:

- **Interest Rates:** Higher mortgage rates reduce buyer affordability (Glaeser et al., 2014).
- **Neighborhood Growth:** Proximity to **tech hubs, schools, and transit** boosts long-term property values.
- **Regulatory Risks:** Rental restrictions and zoning laws impact investment returns.

Conclusion & Recommendations

Key Takeaways for Investors:

- **Invest in undervalued homes (pre-1980, no renovations) in emerging zip codes.**
- **Prioritize living space & house quality (grade) over extra bedrooms.**
- **Use Random Forest modeling to assess property value trends.**
- **Consider market factors like interest rates & neighborhood growth.**

Final Investment Strategy

For **long-term appreciation**, buy **mid-range homes ($400K–$700K) in growing areas** and renovate strategically. For **short-term flips**, focus on **bathroom & kitchen upgrades in underpriced properties**.

References

Glaeser, E. L., Gottlieb, J. D., & Gyourko, J. (2014). **The Economics of Housing Markets**. *Journal of Economic Perspectives, 28*(3), 3-26.

Gyourko, J., Mayer, C., & Sinai, T. (2013). **Superstar Cities**. *American Economic Journal: Economic Policy, 5*(4), 167-199.

Appendix

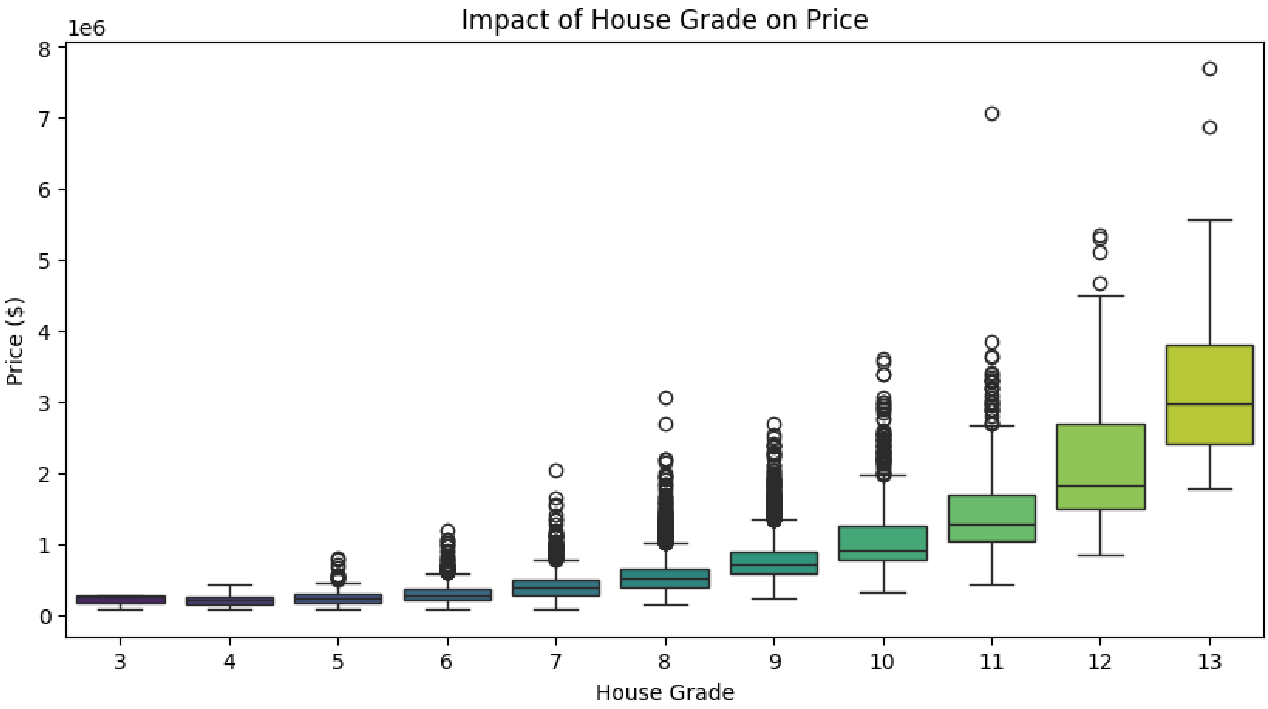

Figure 1: Impact of House Grade on Price

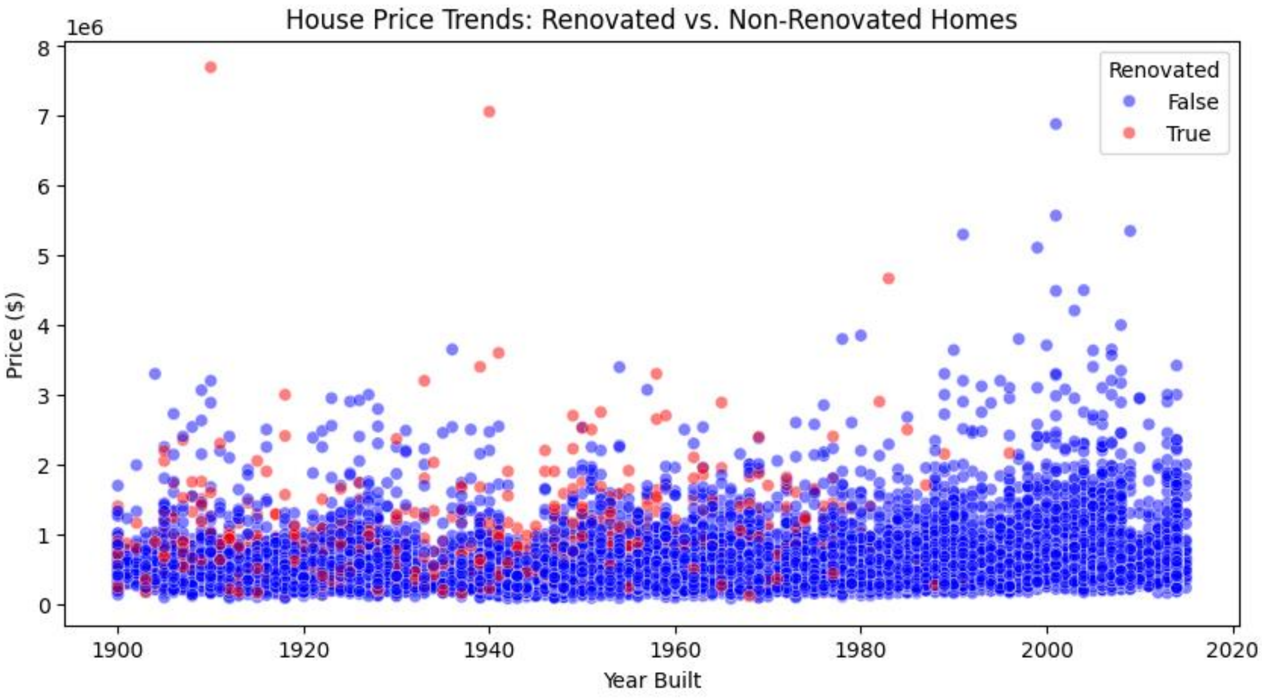

Figure 2: Scatter Plot showing Price Trends of Renovated Houses vs Non-Renovated Houses

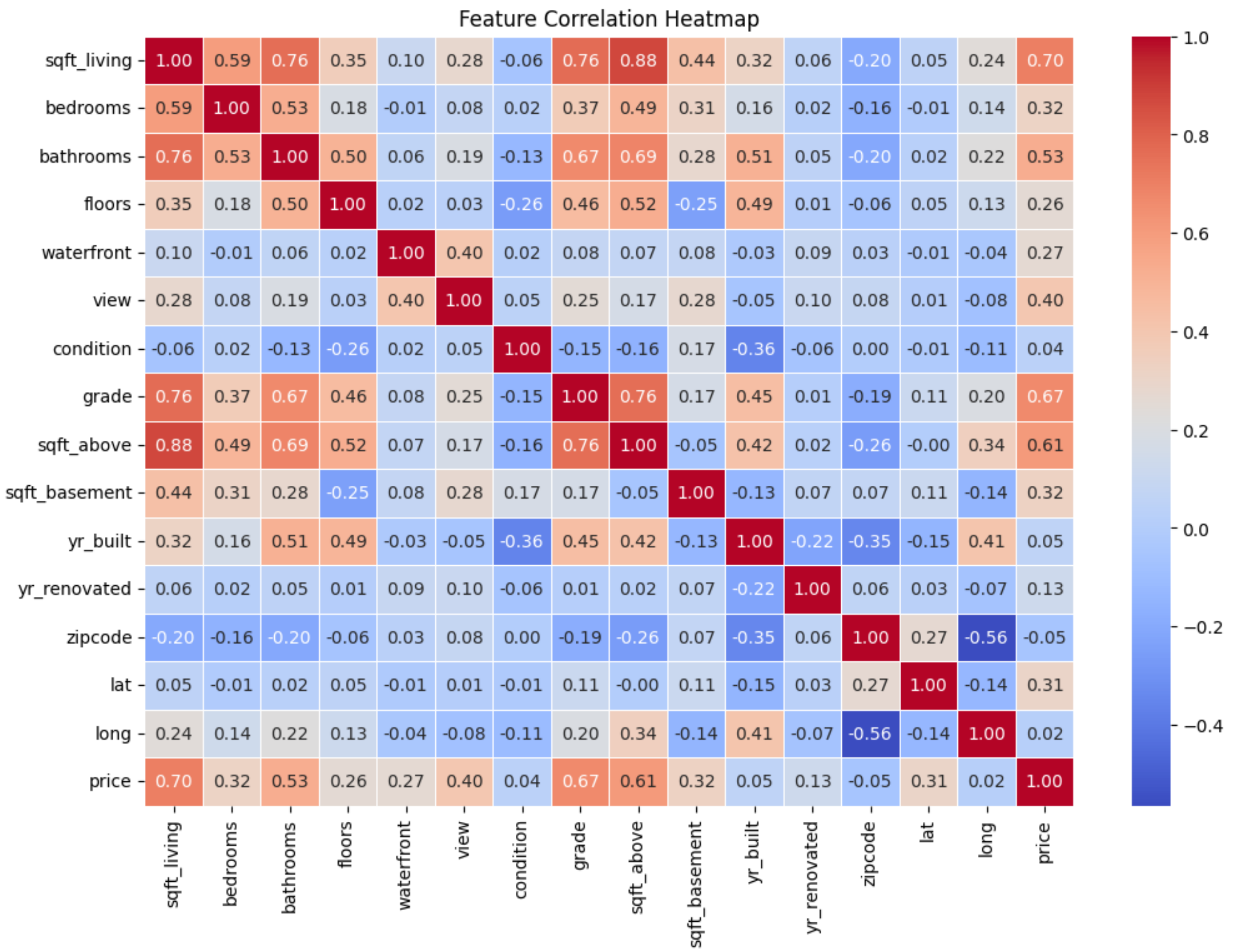

Figure 3: Heatmap showing the Correlation between the Features

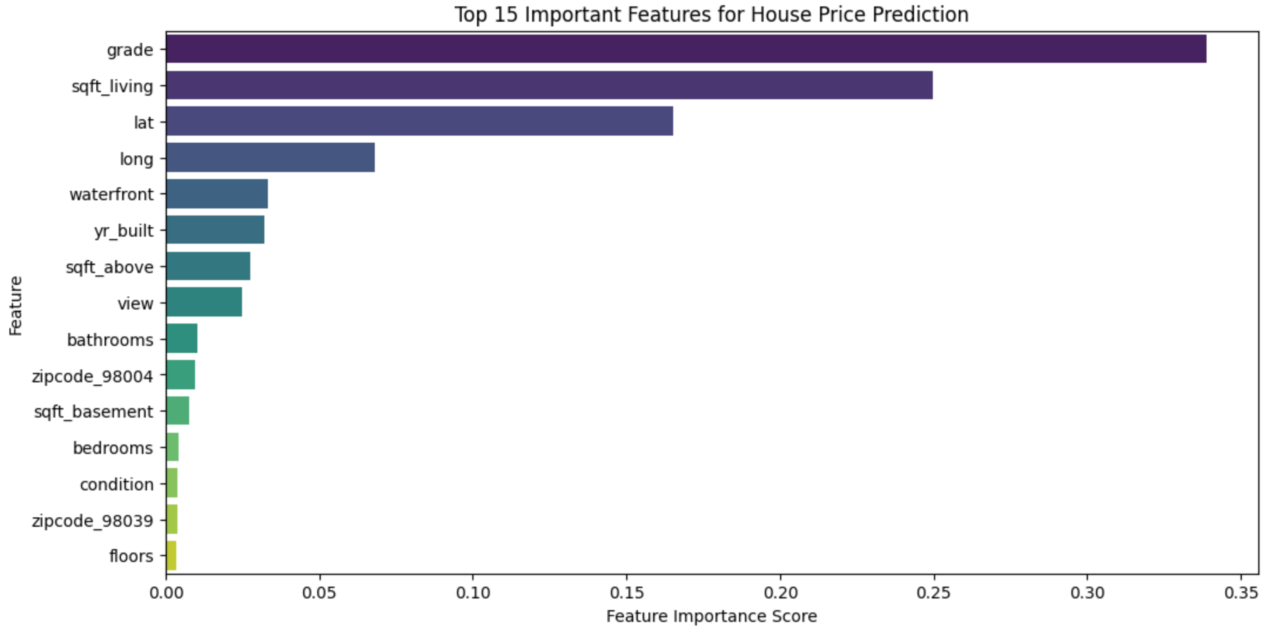

Figure 4: Top 15 Important Features used in the Random Forest Model