Seattle Housing Market Analysis

Justin Mai

Contents

1. Abstract

Keywords: King County Home Sales, Multiple Linear Regression, Prediction Modeling

2. Introduction

Within a highly populated like the Great Seattle and King County area, looking to be a first time homeowner or to purchase a home within the area in general is a tall task. Prices can fluctuate based on several factors making it difficult to make financial decisions. The goal of this report is to develop a model that can support home buyers, investors, and real estate agents navigate the housing market by including the variables that are the most impactful to home prices. Using historical home sales data in King County from 2023-2025, this report aims to use machine learning tools used in multiple linear regression to predict prices. Our goals are **(1) Discover the variables and factors that are significant to forecasting King County home sales using linear regression and ANOVA tools (2) See how well our model predicts prices on our test dataset**

3. Data

3.1 Data Overview

The primary dataset we are using was developed by *Andy Krause* who works as the Director of Valuation and Market Dynamics at Zillow. The dataset he produced was developed with the goal of creating an open access user tool to support in analyzing the housing market. Our dataset consists of 33,333 different observations which represents home sales from 2023 to the start of 2025. Our model will be primarily used to make predictions on `sale_price` and the dataset consists of continuous variables like `land_val`, `sqft`, and `sqft_lot`, and categorical variables like `city`, `zoning`, and `subdivision`. The dataset comes with 45+ predictors for us to use. See https://github.com/andykrause/kingCoData for more details.

3.2 Data Splitting

To avoid data leakage, the influence of training data on testing data, we are splitting our data before applying any type of manipulation. We will be using a standard 80% by 20% split between training and testing data respectively. The data manipulation and transformations will be applied to the training data before modeling, our optimal model developed by the training data will be used on the testing data.

After applying splitting our data, we have 26667 observations in the training data and 6666 observations in the testing data which are both considered sufficient amounts of data. We have chosen the split randomly to eliminate bias in splitting and are also using `set.seed(101)` for the experiment to be reproducible.

3.3 Data Manipulation

3.3 Data Transformation

4. Model

4.1 Model Assumptions / Diagnostics

4.2 Model Selection

5. Results

6. Discussion / Conclusion