DATA WRANGLING, VISUALIZATION, AND MODELING:

King County, USA Housing Price Prediction

Name: Norden Sherpa

Word Count: 1249

Page Count: 13

Table of Content:

King County, USA Housing Price Prediction.......... 1

Abstract.......... 2

Introduction.......... 2

Dataset Description.......... 3

Data Wrangling and Preprocessing.......... 4

Outlier Detection and Handling.......... 4

Handling Irrelevant Columns.......... 4

Exploratory Data Analysis (EDA).......... 4

Univariate Analysis.......... 4

Bivariate Analysis.......... 7

Correlation Matrix.......... 10

Data Modeling: Linear Regression.......... 11

Modeling Approach.......... 11

Model Performance.......... 11

Key Insights and Interpretations.......... 12

Limitations.......... 12

Future Work and Improvements.......... 12

Conclusion.......... 13

Abstract

This report analyzes housing prices in King County, USA, focusing on identifying factors that influence prices and predicting house values. After cleaning the dataset, we explored key features such as square footage, number of bedrooms, and the quality of construction to see how they relate to house prices. Outliers were removed to avoid skewing the analysis, and a linear regression model was built to predict house prices. The model explaines 64% of the price variation. While the model works reasonably well, there are areas for improvement, which we discuss in the conclusion.

Introduction

In real estate, house prices depend on many factors, such as the size of the house, location, and overall condition. For this project, we explored a dataset from King County, USA, which includes houses sold between May 2014 and May 2015. The goal is to better understand what drives house prices and build a model to predict those prices based on the available data.

This project involves the following key steps:

1. **Data Cleaning**: Preparing the dataset by handling irrelevant data and outliers.
2. **Exploratory Data Analysis (EDA)**: Investigating relationships between features like square footage and price.
3. **Modeling**: Building and evaluating a linear regression model to predict house prices.
4. **Insights and Recommendations**: Presenting the findings and offering suggestions for future improvements.

The data is obtained from Kaggle and contains several variables related to house features.

Link: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/data

Dataset Description

The dataset consists of 21,613 records and 21 variables. The most important variables for this analysis are:

- **price**: The sale price of the house.
- **bedrooms**: Number of bedrooms.
- **bathrooms**: Number of bathrooms.
- **sqft_living**: Square footage of the living space.
- **sqft_lot**: Square footage of the land.
- **grade**: Quality of the house based on construction and design.
- **condition**: The condition of the house at the time of sale.
- **yr_built**: The year the house was built.
- **yr_renovated**: The year the house was last renovated.
- **lat** and **long**: Latitude and longitude, indicating the house's location.

Before analysis, we removed some columns, like id, which only served as unique identifiers and didn't contribute to understanding house prices.

Data Wrangling and Preprocessing

Outlier Detection and Handling

Outliers, or extremely high or low values, can distort the results of an analysis. We used two methods to detect and remove outliers:

1. **Interquartile Range (IQR)**: This method identified outliers as prices far above or below the typical range. Using this method, we removed 1,146 outliers.
2. **Z-Score Method**: This method flagged any price that was unusually far from the average price. A total of 406 outliers were removed using this method.
3. **Box-Plot:** Boxplot is used to identify the outlier and boxplot for before and after handling the outliers is shown respectively.

Since the IQR method removed more outliers, we used that cleaned dataset for the rest of the analysis.

Handling Irrelevant Columns

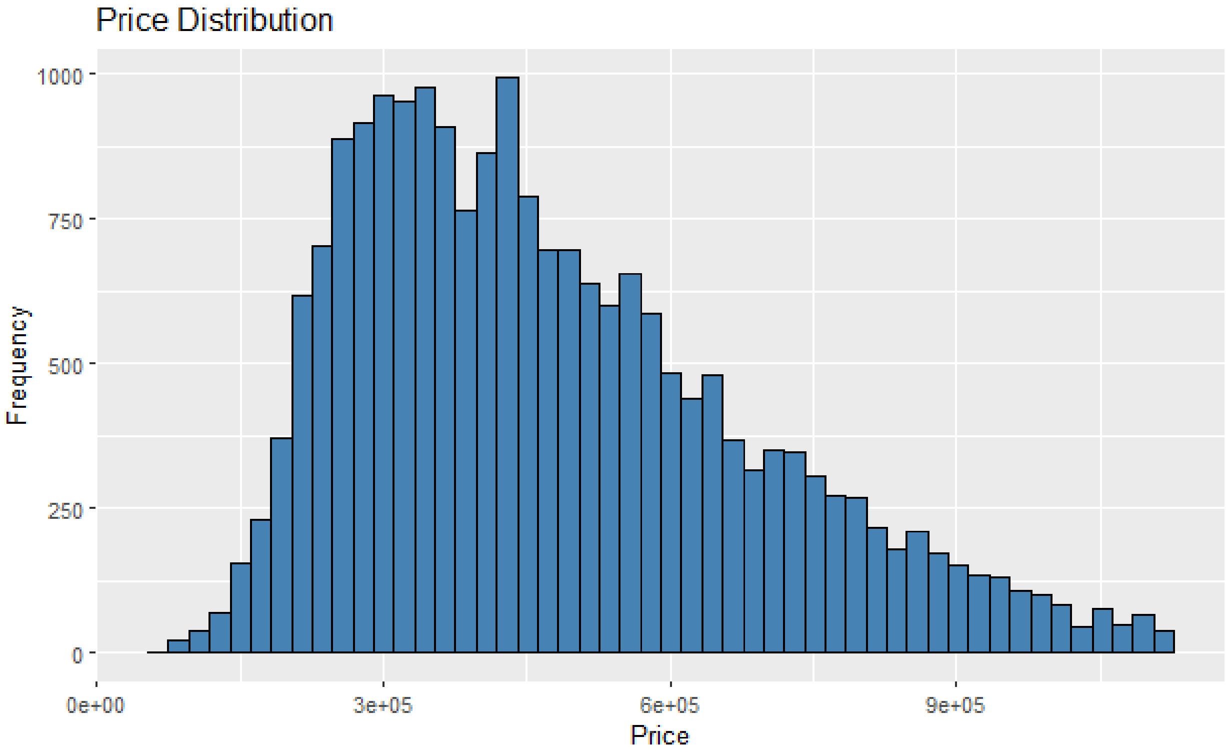
We also removed columns that did not contribute to predicting house prices. For example, the id column was removed because it's just a unique identifier for each house and doesn't provide useful information for predicting prices. Other features, such as waterfront, yr_renovated, sqft_lot, sqft_lot15, yr_built, condition, and zipcode were also removed due to their low correlation with the target variable (price).

Exploratory Data Analysis (EDA)

To better understand the data, we performed EDA to explore the relationships between house features and price.

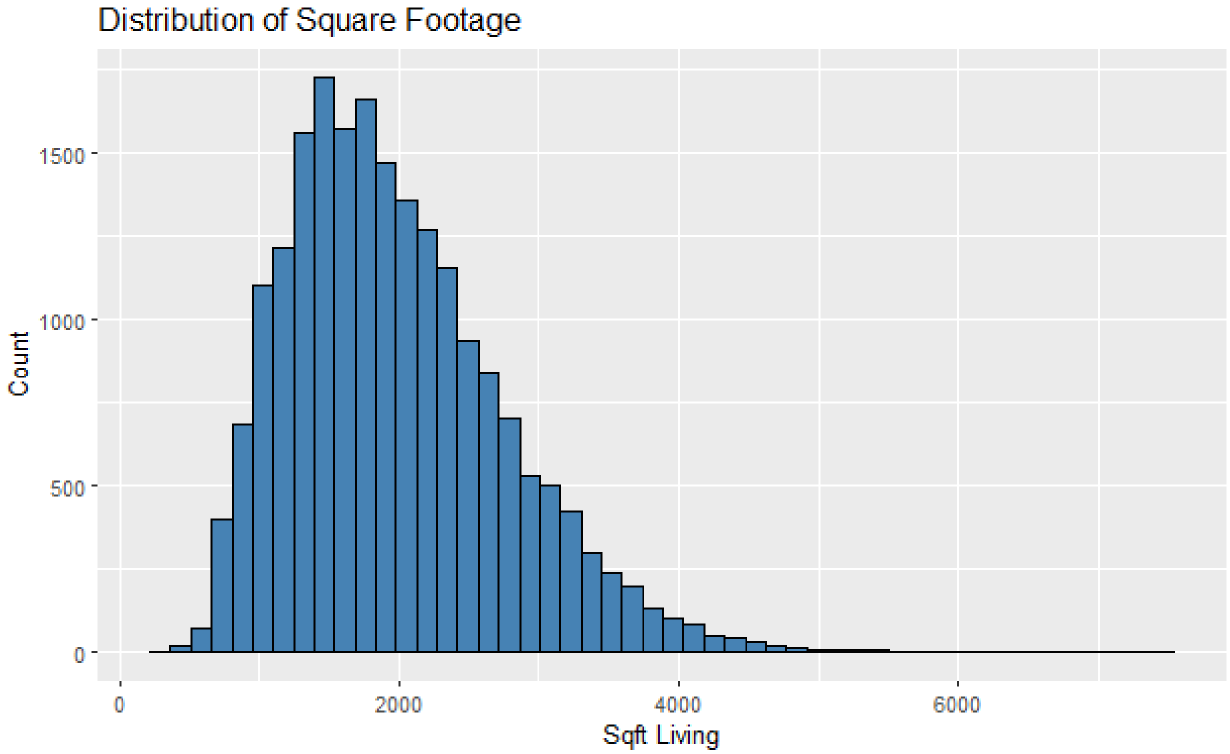
Univariate Analysis

- **Price Distribution**:



Most houses are priced below $500,000, but there are a few very expensive homes that extend the price distribution.

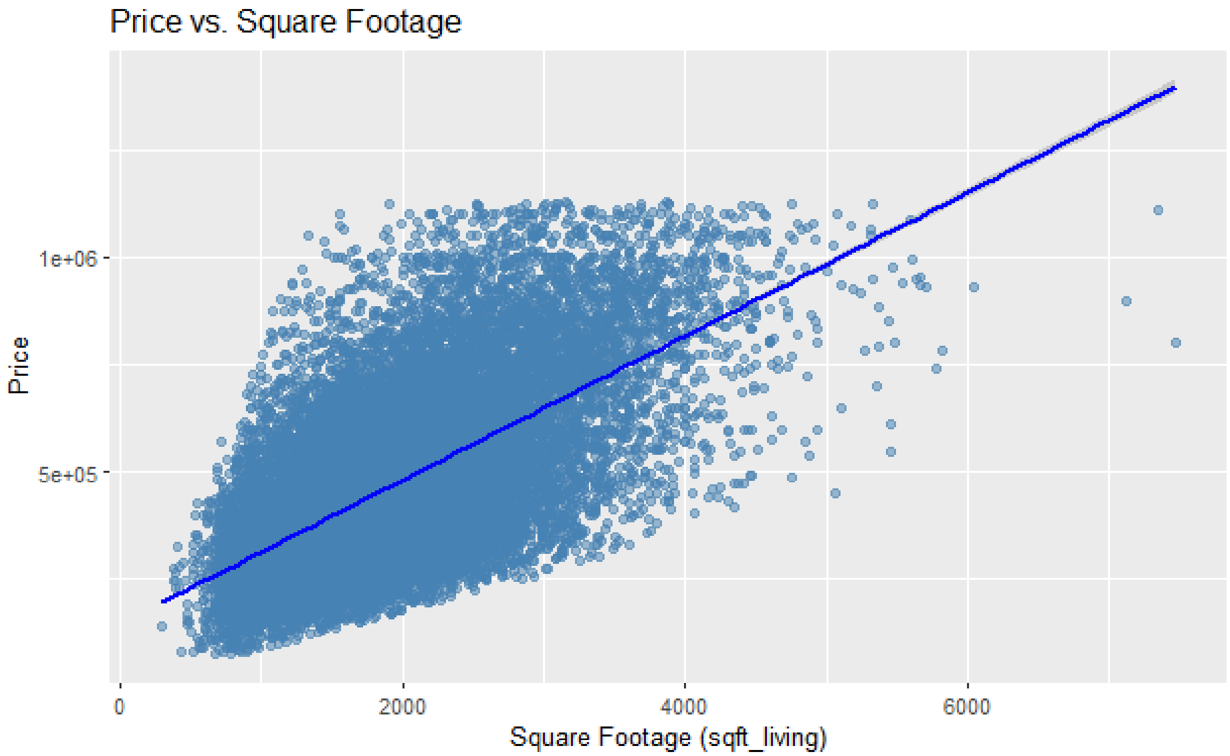
- **Square Footage**:



The majority of homes have between 1,000 and 3,000 square feet of living space, with only a small number of very large homes.

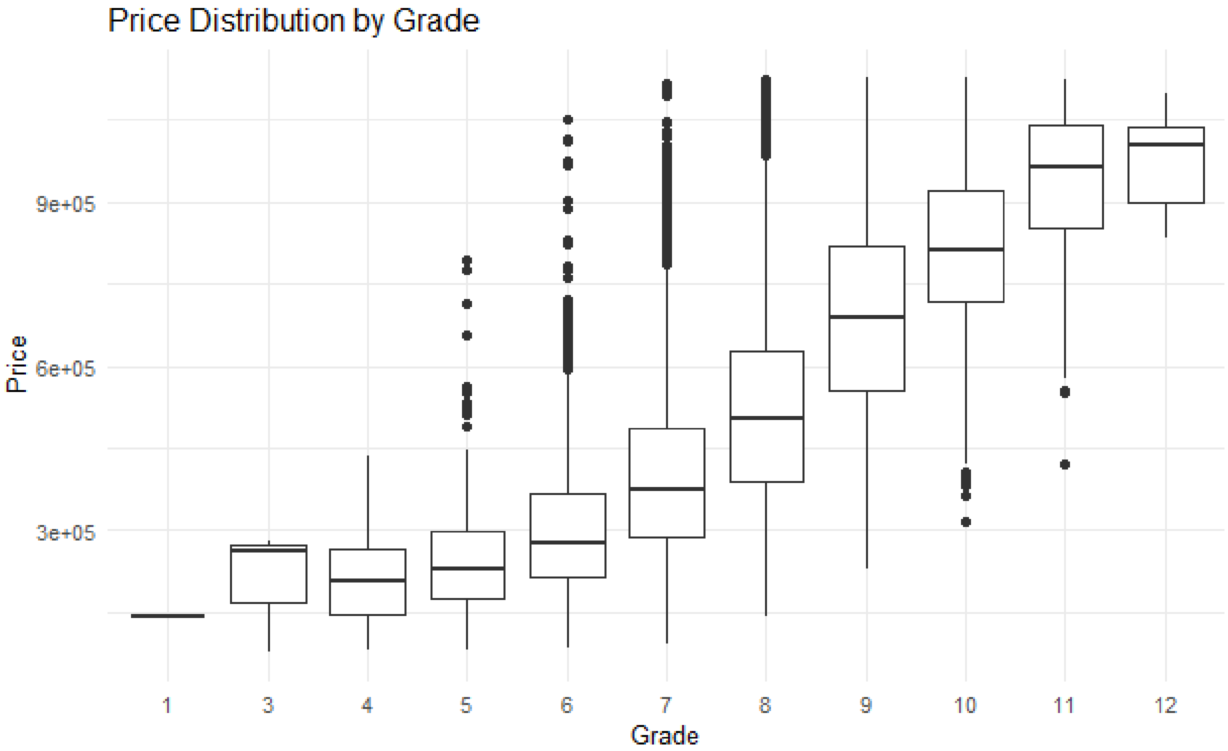
Bivariate Analysis

- **Price vs. Square Footage**:



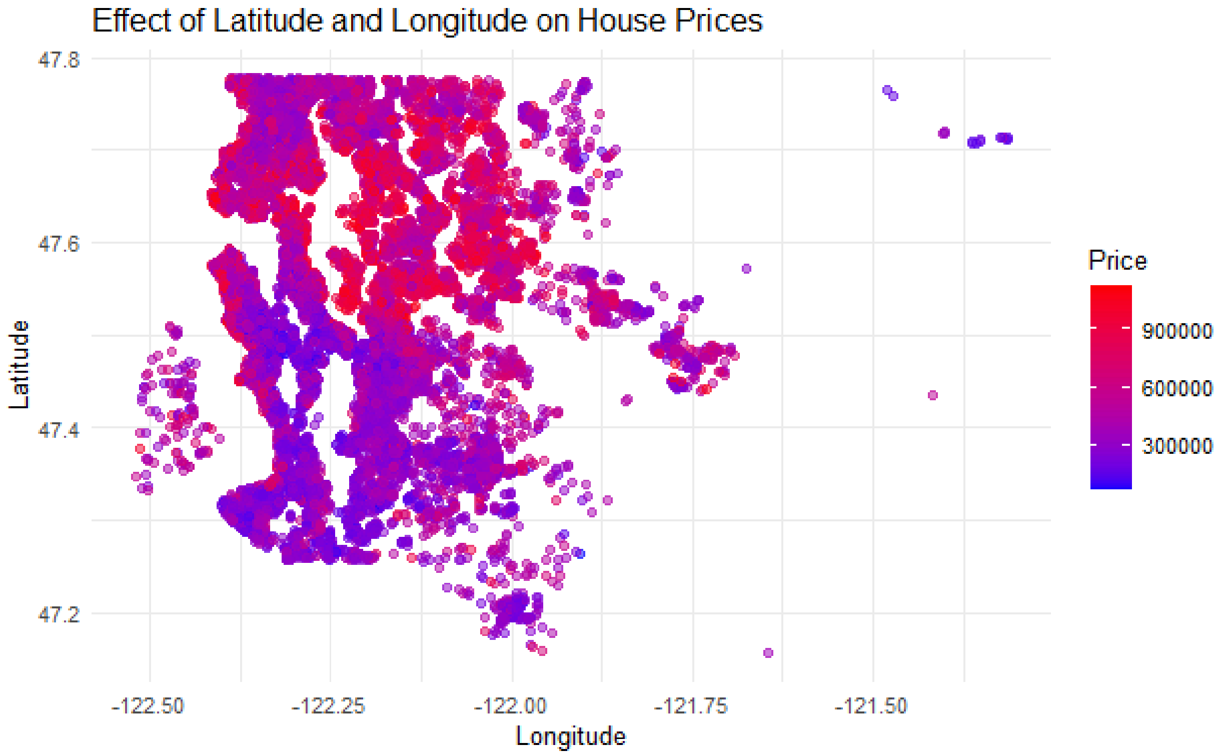
There is a clear positive relationship between square footage and price. Larger homes tend to sell for higher prices.

- **Price vs. Grade**:



Higher-quality houses (as measured by grade) are sold for higher prices.

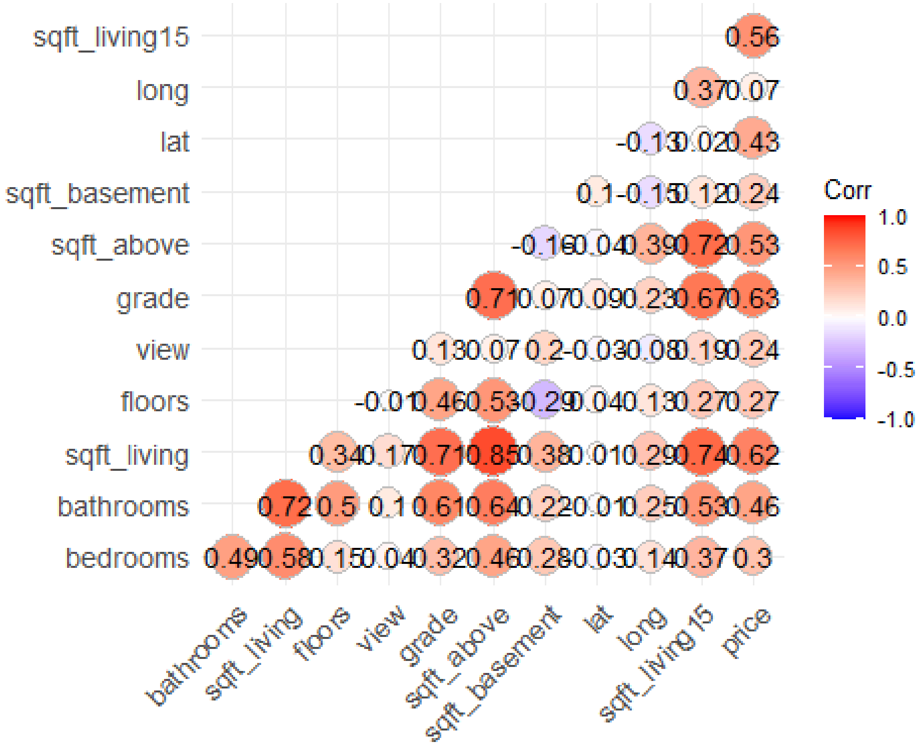
- **Price vs. Latitude/Longitude**:



Homes in certain areas, especially those near the water or city center, tend to be more expensive.

Correlation Matrix

A correlation matrix was created to measure how strongly different variables are related to price.



The most important variables were:

- **sqft_living** (square footage): Correlation of 0.70 with price.
- **grade** (house quality): Correlation of 0.67 with price.
- **sqft_above** (above-ground square footage): Correlation of 0.60 with price.

These features showed the strongest relationship with house prices.

Data Modeling: Linear Regression

After cleaning the data and understanding the relationships between variables, we built a linear regression model to predict house prices.

Modeling Approach

We used the following features to predict house prices:

- **sqft_living**
- **bedrooms**
- **bathrooms**
- **floors**
- **grade**
- **condition**
- **lat** and **long**

Model Performance

The model was trained on 80% of the data and tested on the remaining 20%. Here's how the model performed:

- **R-squared**: 0.64, meaning that 64% of the variation in house prices is explained by the model.
- **Root Mean Squared Error (RMSE)**: $119,630. This means that, on average, the model's predictions are off by about $119,630.
- **Mean Absolute Error (MAE)**: $90,805. This indicates that, on average, the predictions differ from actual house prices by about $90,805.

While the model works reasonably well, the error values suggest that there is room for improvement.

Key Insights and Interpretations

1. **Square Footage and Price**: Larger homes with more living space tend to have higher prices. This is one of the strongest factors affecting house prices.
2. **House Quality (Grade) and Price**: Homes with a higher quality grade are generally more expensive. Buyers are willing to pay a premium for better-built homes.
3. **Location (Latitude and Longitude)**: The closer a home is to city centers or desirable locations (e.g., waterfronts), the higher its price tends to be.
4. **Number of Bedrooms and Bathrooms**: While these features do impact price, their effect is smaller compared to the size and quality of the home.

Limitations

While this project provided valuable insights, there are some limitations to consider:

1. **Limited Features**: The dataset does not include information on factors like proximity to schools, crime rates, or neighborhood amenities, all of which could affect house prices.
2. **Model Simplicity**: The linear regression model assumes that the relationship between house prices and features is linear. More complex models could capture non-linear relationships better.
3. **Right-Skewed Price Distribution**: The dataset contains more affordable homes than luxury properties, which might affect the model’s ability to predict high-end house prices accurately.

Future Work and Improvements

1. **Advanced Modeling**: Future work could explore more advanced models such as Random Forests or Neural Networks. These models might provide better accuracy by capturing more complex patterns in the data.
2. **Additional Features**: Including more features, like distance to schools or public transport, could improve the model’s predictions.
3. **Time-Series Analysis**: Although the time series analysis on how price is changing over the course of year is shown in the analysis, but the prices fluctuate heavily from day to day, making it difficult to identify any obvious long-term trend in the data. Analyzing how house prices have changed over time could provide valuable insights into market trends.

Conclusion

In this project, we explored the factors that influence house prices in King County and developed a linear regression model to predict house prices based on several important features. The model performed well, explaining 64% of the price variation, with square footage and house quality being the most important predictors. However, the model could be improved with additional features and more advanced modeling techniques.

By understanding these key factors, buyers, sellers, and investors can make more informed decisions about property values in the King County housing market.