STA 210 Final Project

Jacob You

Introduction

The Melbourne housing market has exhibited significant volatility over recent years, characterized by fluctuating demand and varying price trends across different suburbs. Understanding these dynamics is crucial for potential homeowners, investors, policymakers, and those looking to buy a home. This project seeks to delve into the factors influencing property prices in Melbourne, utilizing a dataset that captures a wide range of property characteristics.

Research Question

The primary research question this project seeks to answer is: "What are the key factors influencing property prices in Melbourne, and to what extent do they affect these prices?"

Data

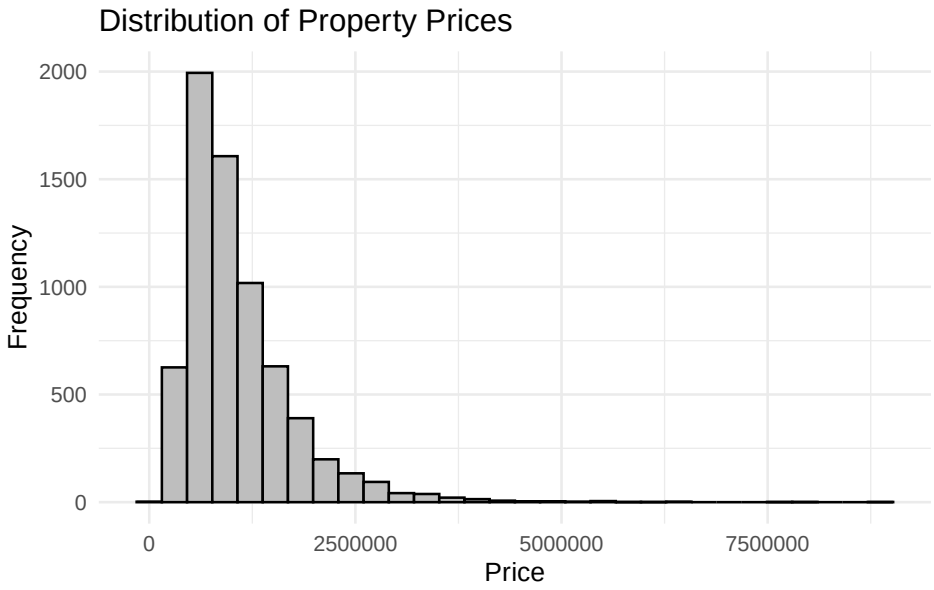
The dataset was taken from the website Kaggle, a site that holds databases for research and machine learning. The dataset "Melbourne Housing Snapshot" was sourced from public real estate records, comprising of 13,580 properties sold in Melbourne during a recent period. Each entry in the dataset includes quantitative details about the home including the price, the number of rooms, bedrooms, and bathrooms, the year built, and building area. The dataset also has categorical variables such as the general region in Melbourne, the suburb, and the type of property between houses, units, and townhouses.

To clean the data, all extreme outliers and invalid data was removed. Five buildings with building areas greater than 3000 square meters were removed, as such large outliers are not representative of the average house, and may cause errors in the model. Additionally, one building built before 1200 was removed, as the extreme outlier may also cause errors in the model. Finally, houses with a building area of 0 were treated as invalid and were removed.

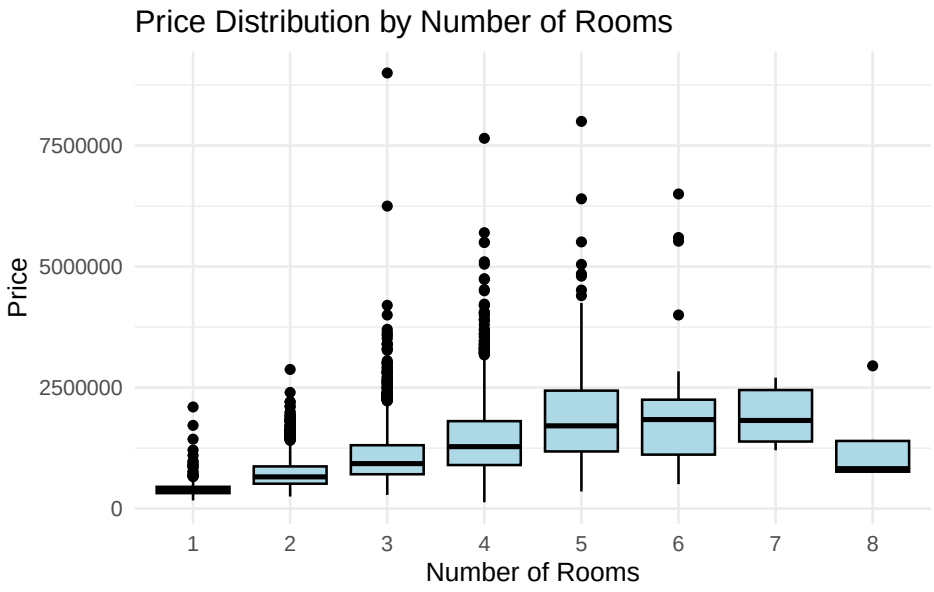
Exploratory Data Analysis

In order to build an accurate model, we must first take a look at the relationship between price and other variables. Through this analysis, we find that the number of rooms and building size seem to have a strong correlation with price. Price also seems to have a correlation with the number of bedrooms and bathrooms, general region, the type of property, latitude and longitude, and the year built, which cannot be included here due to space constraints.

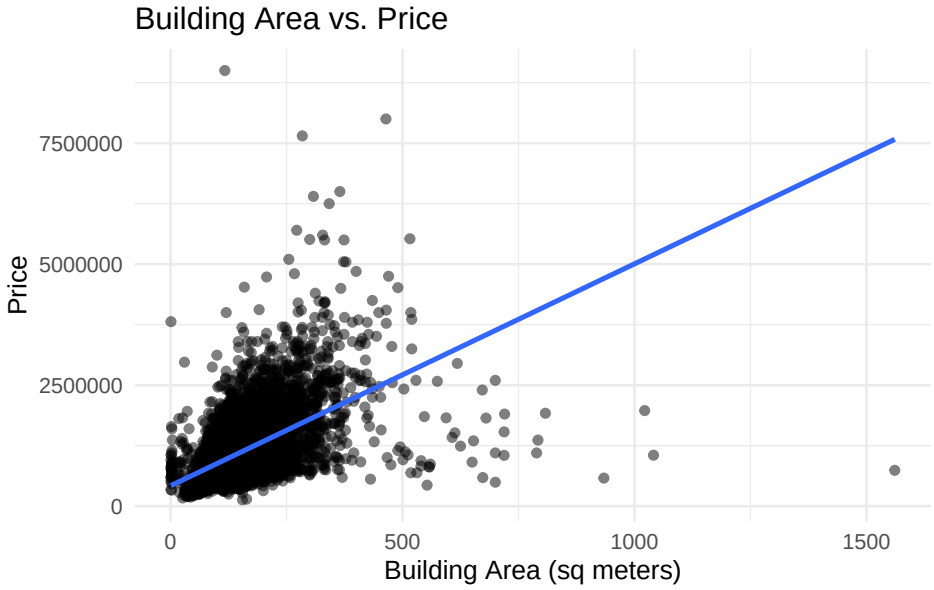
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 131000  630000  891000 1077785 1335000 9000000
```



The median of the data was around 0.89 million dollars, with the mean being slightly higher at 1.08 million dollars. The histogram of property prices illustrates the distribution across the dataset. The distribution is right-skewed, suggesting that while affordable properties dominate the market, a smaller number of high-priced properties significantly exceed the general price range, potentially due to rare features or desirable locations.



The boxplots show a positive correlation between the number of rooms and the overall price. As the number of rooms increases, both the median price and the range of prices tend to increase, suggesting that properties with more rooms typically have higher prices, likely reflecting larger living spaces.



The scatterplot shows an overall positive correlation, suggesting that, on average, properties with larger building areas tend to have higher prices. This also correlates with the positive

correlation between number of rooms and price, as the more rooms there are, the larger the building area.