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

Because there are so many factors that go into the price of a home, my research question is: What factors have the greatest influence on the price of a home in Melbourne, and how do these factors impact the price?

Dataset Source

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 the property was built, and total area of the building. 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.

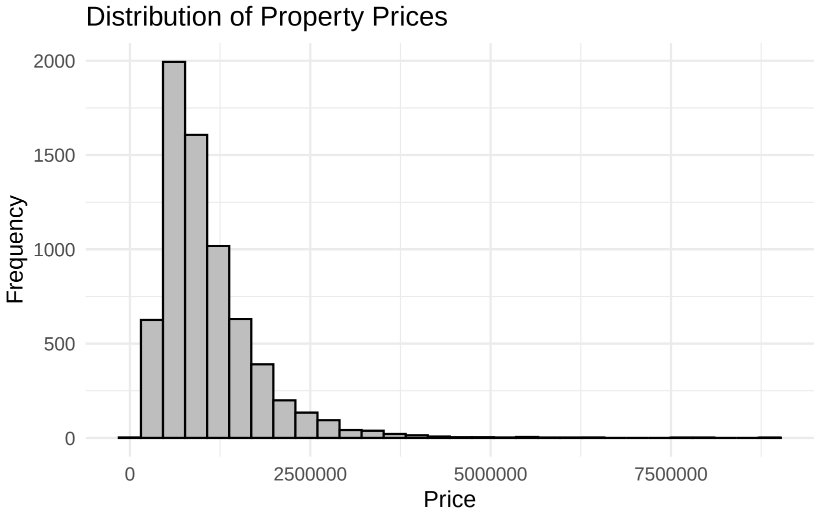
Cleaning

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

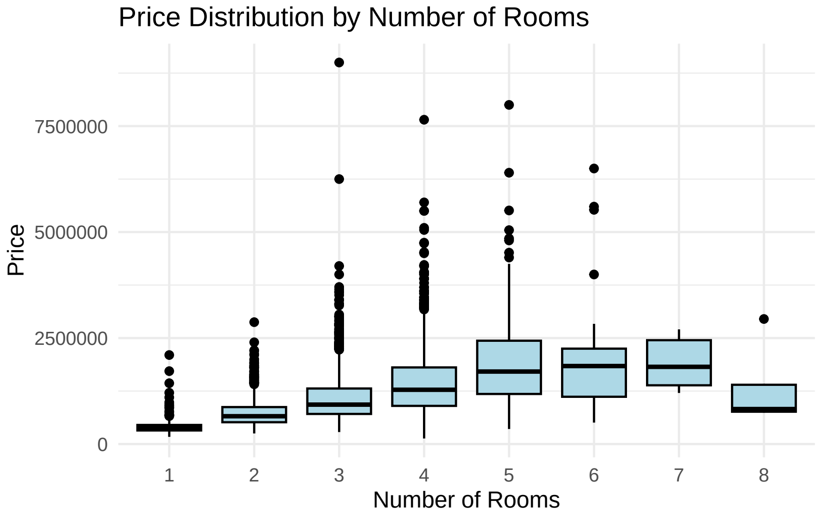
In order to build an accurate model, we must first take a look at the relationship between price and other variables. Through this analysis, we find that the number of rooms and building size seem to have a strong correlation with price. Price also seems to have a correlation with the number of bedrooms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

Price



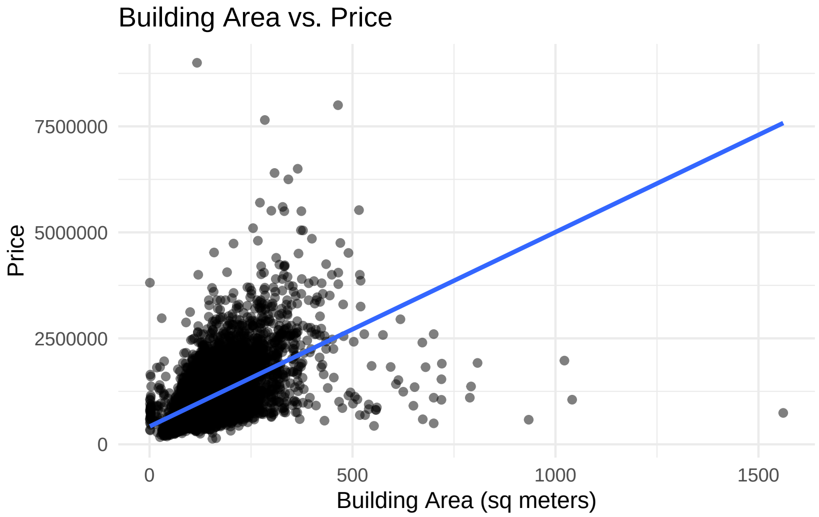
The median of the data was around 0.89 million dollars, with the mean being slightly higher at 1.08 million dollars. The histogram of property prices illustrates the distribution across the dataset. The distribution is right-skewed, suggesting that while affordable properties dominate the market, a smaller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

Number of Rooms



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.

Building Area



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.

Methodology

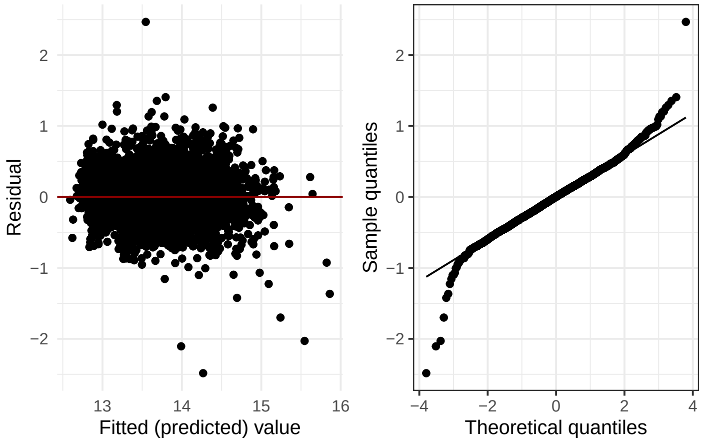
To achieve my goal of finding the factors with the greatest impact on price and understanding how they affect the price, I decided on implementing a log-linear regression model. The prices exhibited a right-skewed distribution, which led to problems for an ordinary linear regression. For a normal multiple predictor model, constant variance and normality were violated. However, applying a logarithmic transformation to price helped normalize the distribution, allowing us to assume homoscedasticity and normality.

Predictor Variables

For the model, I included all of the variables I found to have a noticable correlation with price. The number of rooms, bedrooms, bathroom, car spaces, and building area all related to the overall size of the house and seemed to have a correlation to price. Furthermore, the year the house was built in, the type of the house, and the region of Melbourne the house is in all showed correlation when conducting EDA.

Although longitude and latitude were significant, using them as predictors for a linear model would likely harm the model, as an increase or decrease in latitude could have drastically different effects depending on the region and surroundings. Furthermore, the irregular shape of Melbourne would cause further complications in including coordinates.

Assumptions



The assumption of linearity is satisfied, as the residuals are symmetrically distributed about the horizontal axis. After the log transformation, the variance of the residuals is much more even, and we can assume that the constant variance assumption is satisfied. In the Q-Q plot, there is some degree of deviation in the tails. However, this deviation is quite small, and only on the very far ends of the Q-Q plot, so normality seems to be satisfied.

Independence however, may not be satisfied. In the real world, complex markets like the housing market are rarely ever independent. Properties built during similar times may have been built in similar economic conditions and similar market cycles. Changes in building standards and styles may have influenced property values systematically. Furthermore, houses in the same neighborhood or area might have similar determinants from shared amenities, crime rates, or more. I have incorporated the general region of the property into the model to mitigate the effect of properties being nearby each other, but due to the incredibly complex nature of the housing market, there is not much more I can do to satisfy independence within the contents of the course.

Model Output

```r
tidy(housing_model)
```

```
# A tibble: 15 x 5
   term                                 estimate std.error statistic   p.value
   <chr>                                   <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)                          21.8     0.237         92.0  0
 2 Rooms                                 0.0884  0.00631       14.0  4.40e- 44
 3 RegionnameEastern Victoria           -0.470   0.0593        -7.93 2.51e- 15
 4 RegionnameNorthern Metropolitan      -0.0918  0.0139        -6.60 4.51e- 11
 5 RegionnameNorthern Victoria          -0.625   0.0614       -10.2  4.08e- 24
 6 RegionnameSouth-Eastern Metropolitan -0.158   0.0241        -6.56 5.85e- 11
 7 RegionnameSouthern Metropolitan       0.222   0.0138        16.1  1.97e- 57
 8 RegionnameWestern Metropolitan       -0.182   0.0142       -12.8  4.53e- 37
 9 RegionnameWestern Victoria           -0.878   0.0698       -12.6  6.66e- 36
10 BuildingArea                          0.00143 0.0000642     22.3  8.74e-107
11 YearBuilt                            -0.00446 0.000122     -36.5  6.85e-267
12 Bathroom                              0.149   0.00730       20.4  5.49e- 90
13 Typet                                -0.0367  0.0146        -2.51 1.22e-  2
14 Typeu                                -0.337   0.0124       -27.2  2.52e-154
15 Car                                   0.0201  0.00447        4.50 6.81e-  6
```