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, and policymakers. This project seeks to delve into the factors influencing property prices in Melbourne in order to better understand the details of the housing market, as well as to provide accurate estimates of prices.

Research Question

Because there are so many factors that go into the price of a home, my research question is: What factors significantly 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 in Australian dollars, 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, which is categorized by houses, units, and townhouses.

Note: Although this dataset has time variables, I decided to use it as it was one of the best I found, and I find this topic extremely interesting. I address the autocorrelation from time later on, along with methods to further improve independence. As my research question is primarily on getting broad conclusions about the data and not about getting precise predictions and accuracy, I believe this dataset is suitable to use here.

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 property, and might cause inaccuracies in the model. Additionally, one building built before 1200 was removed, as the extreme outlier may also cause the model to be less accurate. Finally, properties with a building area of 0 were treated as invalid and were removed.

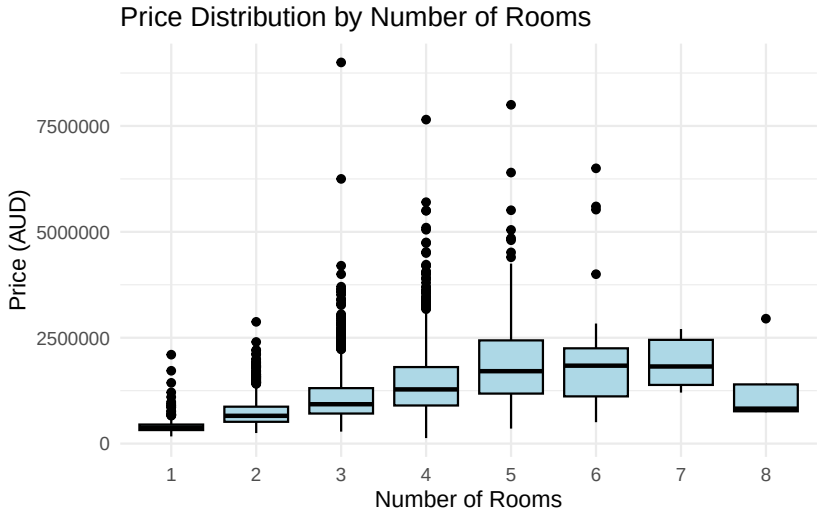
Exploratory Data Analysis

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 bathrooms, the general region, the type of property, the latitude and longitude, and the year built, which could not be included here due to space constraints.

Price

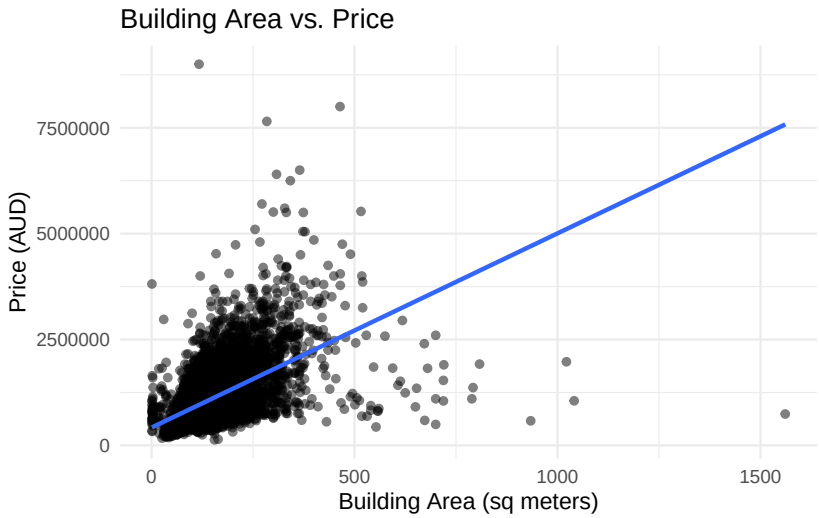
The median of the data was around 0.89 million dollars, with the mean being slightly higher at 1.08 million dollars. The histogram of property prices illustrates the distribution across the dataset. The distribution is right-skewed, suggesting that while affordable properties dominate the market, a smaller 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

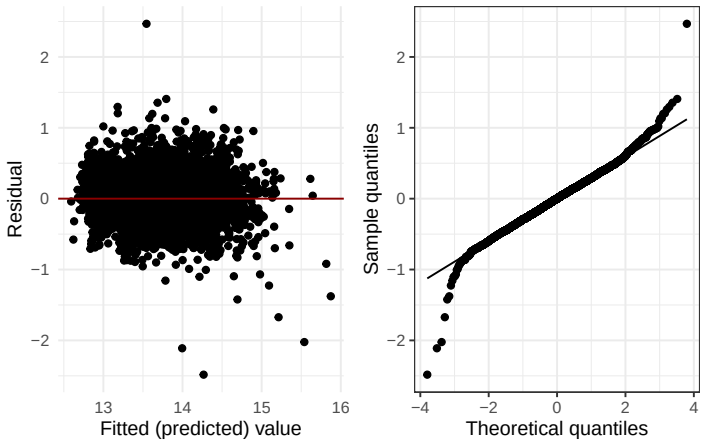
To achieve my goal of finding the factors with the greatest impact on price and understanding how they affect the price, I decided on implementing a log-linear regression model. The prices exhibited a right-skewed distribution, which would cause problems for an ordinary linear regression. When using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

Predictor Variables

For the model, I included all of the variables I thought would logically have a correlation with price according to the EDA. I felt that the number of rooms, bedrooms, bathroom, car spaces, and building area would all be related to price, as larger properties with more space would most logically be more expensive. Furthermore, the year the property was built in, the type of the property, and the region of Melbourne the property is in all showed a correlation with price when conducting the EDA.

Although longitude and latitude were significant, using them as predictors for a linear model would likely harm the model. A property's value doesn't uniformly increase or decrease when moving north, south, east, or west. These variables are more tied to geographic features like proximity to city centers, bodies of water, mountains, and economic activity zones, which do not change linearly across space.

Assumptions



The assumption of linearity is satisfied, as the residuals are symmetrically distributed about the horizontal axis. After the log transformation, the vertical 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.

However, the data is not perfectly independent. In the real world, complex markets like the housing market are rarely completely independent. Properties built during similar times may have been built in similar economic conditions and similar market cycles. Furthermore, buildings in the same neighborhood or area might have similar determinants from shared resources, crime rates, or more. I accounted for some of this dependence by incorporating the region variable, as certain regions would have similar ameneties nearby Finally, some of the dependence is incorporated by including the year the property was built.

Results

```
# A tibble: 16 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                    0.0939  0.0135        6.98  3.26e- 12
 3 BuildingArea             0.00143 0.0000642    22.3   8.63e-107
 4 Bedroom2                -0.00608 0.0132       -0.460 6.46e-  1
 5 YearBuilt               -0.00446 0.000122    -36.5   8.83e-267
 6 Bathroom                 0.150   0.00736      20.3   4.22e- 89
 7 Typet                   -0.0369  0.0146       -2.52  1.16e-  2
```

```
 8 Typeu                                -0.337   0.0124      -27.1   4.87e-154
 9 Car                                   0.0203  0.00448       4.52  6.23e-  6
10 RegionnameEastern Victoria           -0.470   0.0593       -7.93  2.54e- 15
11 RegionnameNorthern Metropolitan      -0.0919  0.0139       -6.60  4.28e- 11
12 RegionnameNorthern Victoria          -0.625   0.0614      -10.2   4.16e- 24
13 RegionnameSouth-Eastern Metropolitan -0.158   0.0241       -6.56  5.95e- 11
14 RegionnameSouthern Metropolitan       0.222   0.0138       16.1   3.26e- 57
15 RegionnameWestern Metropolitan       -0.182   0.0142      -12.8   4.16e- 37
16 RegionnameWestern Victoria           -0.878   0.0698      -12.6   7.09e- 36
```

Every one of the statistics are statistically significant (p-value < 0.05) except for the statistic for the number of bedrooms.

For each additional room in the building, the model predicts that the price is multiplied by approximately 1.0984, while controlling for all other variables. The number of bathrooms has a similar effect, as for each additional bathroom, the model predicts the price to be multiplied by approximately 1.161, controlling for other variables. The number of car slots also has a positive effect, but the effect is weak. However, due to the high p-value for bedrooms (0.64) we fail to reject the null hypothesis that bedrooms has an effect on the price of a property.

Interestingly, there seems to be a slight negative correlation with the year the property was built. For every additional year, the price is predicted to be multiplied by approximately 0.996, while controlling for all other variables. For buildings built many years apart, this effect can significantly impact the predicted price. Similarly, for each additional square meter in building area, the model predicts a 1.0014 times increase in price, holding other variables constant. Because properties can vary by thousands of square meters in size, this effect can also be substantial.

Furthermore, the type of property seems to have a very strong impact on price. The model predicts that for a unit type property, the price is approximately 0.714 times that of a house type property, holding other variables constant. Townhouses also have a smaller negative effect on price. Finally, the region of the property also has a large impact, with properties in the Southern Metropolitan region predicted to have 1.249 times the price of properties in the Eastern Metropolitan region, holding other variables constant. All the other regions seem to be less expensive than the Eastern Metropolitan region, with the Western Victora region being the least expensive, with the price of a property in the region being predicted to be approximately 0.416 times that of a property in the Eastern Metropolitan region, holding all other variables constant.

Discussion

From my model, I found that the number of rooms, the number of bathrooms, the total area of the building, the year the building was built in, the type of property, and the region in

Melbourne that the property is in are factors that have a large impact on the price of a property in Melbourne. These factors are statistically significant, and can have a significant impact on the price of a building. The model also found that buildings with more rooms, bathrooms and a larger area are higher price. I found that older properties are more expensive, and houses are more expensive than townhouses, which are more expensive than units. This may be because older properties are valued for their heritage, or possibly due to being in well-established neighborhoods. Finally, certain regions are more expensive than others, with the Eastern Metropolitan region being the most expensive, and the Western Victoria region being the least expensive. This is likely because the Southern Metropolitan region being close to the central business district and other amenities. Meanwhile, the Western Victoria region is largely rural, and is quite far away from any large developments.

Limitations

One of the major limitations of this analysis is the fact that independence is not perfectly satisfied. Because of how complex the housing market is, there are many things that determine the price of a house such as the current status of the market. Furthermore, there are possible variables that a log-linear regression model can't interpret such as the latitude and longitude due to the non-linear nature of the data. There are also some variables not included in the dataset such as the climate of the region, and nearby amenities. Most importantly, there appears to be some autocorrelation due to time.

To address these concerns, we could implement different models that handle some of these variables better. For example, for the latitude and longitude variables, we could incorporate spatial regression models. For the data on the year the property was built as well as the year the property was sold, we could do a time series analysis to understand how the market changes over time, some of the dependencies between entries. We could also include lagged versions of the price, add specific indicators for key economic time periods, and implement differencing to the data series, although these methods are beyond the scope of the course.

Future Work

In the future, more variables could be included such as the sizes of backyards and extra structures like pools, or variables such as crime rates or school quality scores of the area. There are countless factors that can influence the price of a property, and although incorporating all of them would be incredibly difficult, adding more predictors may lead to more insights and better predictions. This also might mitigate some of the ommited variable biases and dependencies in the current data. Furthermore, a longitudinal study tracking the changes in property prices over time could lead to more insights on how the Melbourne market has changed.