Melbourne Housing Prices

Trung Pham

1. Introduction

1.1 Background

In the last 10 years, the city of Melbourne, Australia has seen its population increasing almost consistently, from just above 4 million in 2011 to the current figure of more than 5 million citizens, representing a staggering 25 per cent growth rate. The city is expected to surpass Sydney to become the largest in Australia before 2020. A significant proportion is accounted by overseas immigrants, which raises the challenge of meeting accommodation needs for this ever growing populace. Housing prices have indeed been rising in response to the burgeoning demand, but trying to forecast the growth is not a simple endeavour, especially after 2008 when the housing bubble burst in the US causing ripple effects throughout the world. Therefore, having a logical, data-driven analysis of the monetary value of a real estate in Melbourne before deciding would be extremely favourable for future owners who would like to offset the risk of making decision blindfoldedly. Other audiences that might have interest in this analysis could be real estate agents or investors trying to gauge the value of particular properties.

1.2 Problem

This project aims to leverage historical data of the characteristics and prices of 34,857 properties in Melbourne area from January 2016, then seeks to reveal certain correlation of different factors such as land size, number of rooms, suburbs, distance to centre, etc. with the final prices. From there, the author endeavours to build a reliable model to predict the price of a given estate or even one that is to be constructed in the near future.

2. Data collection and preprocessing

The author obtains the dataset from Kaggle, at this address. Details of the columns in the raw dataset can be found below.

- Suburb: Suburb
- Address: Address
- Rooms: Number of rooms
- Price: Price in Australian dollars
- Method: S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to

auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.
- Type: br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.
- SellerG: Real Estate Agent
- Date: Date sold
- Distance: Distance from CBD in Kilometres
- Regionname: General Region (West, North West, North, North east ...etc)
- Propertycount: Number of properties that exist in the suburbs.
- Bedroom2 : Scraped # of Bedrooms (from different sources)
- Bathroom: Number of Bathrooms
- Car: Number of carspots
- Landsize: Land Size in Metres
- BuildingArea: Building Size in Metres
- YearBuilt: Year the house was built
- CouncilArea: Governing council for the area
- Lattitude: Self explanatory
- Longtitude: Self explanatory

The set, however, is a mix of numerical and categorical data, therefore splitting that into separate tabular forms is necessary for a more in-depth analysis. Also, there are a significant amount of missing values in almost all columns, thus entries missing any data will be removed from the dataset to ensure pristine examination. Another duplicates are observed in columns ‘Rooms’ and ‘Bedroom2’. The variance between these two is minimal and the author decides to drop the ‘Bedroom2’ variable and rename ‘Rooms’ to represent both columns.

The ‘YearBuilt’ column in the dataset informs the year the house was built. Its main meaning is to probe the age of the house, as such the age can be expressed in terms of historic (greater than 50 years old) vs modern (less than 50 years old) to get the essence of this information in a more condensed way, allowing for better analysis and visualisation.

Besides, the observations with a 'zero' in the feature ‘BuildingArea’ will be removed because it is not possible for a house to have a size of zero. Also, this observation is priced remarkably high at AUD 8.4m, which is an outlier in the dataset, signaling a possible error in the data point. Thus, this observation will be removed.

After cleansing, the final data set consists of 22 columns and 8,842 rows excluding the headers.

Foursquare API provides an exceedingly handy platform for location data to visualise the clustering addresses and surrounding venues for the listed properties. Such factors might have telling influence on the value of one estate. The limit is set at 50 venues at radius 200 for each suburb from its coordinates.

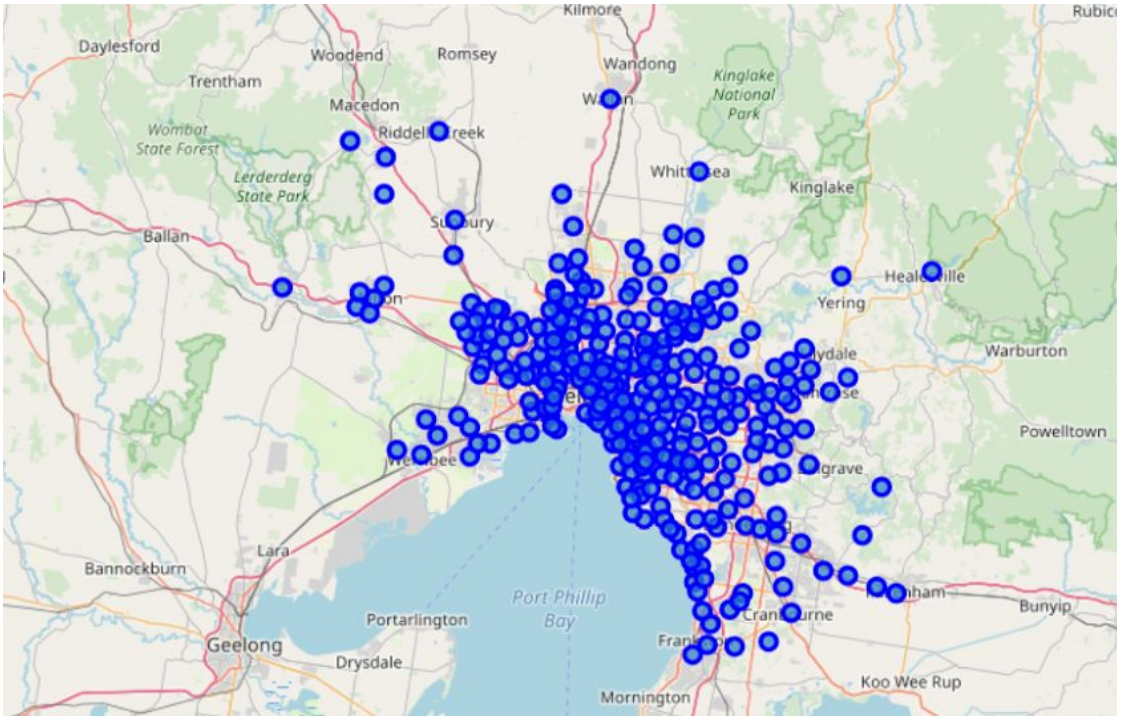

Figure 1. Visualisation of house clusters in Melbourne

3. Exploratory data analysis

Putting ‘Price’ in a vacuum, the graph shows this feature to be normally distributed and skewed to the right (skewness = 2.42, kurtosis = 11.11). That is expected because luxury houses are generally smaller in number compared to the affordable ones. The majority of houses’ price clustered in the range from 200 thousand to AUD 1 million.

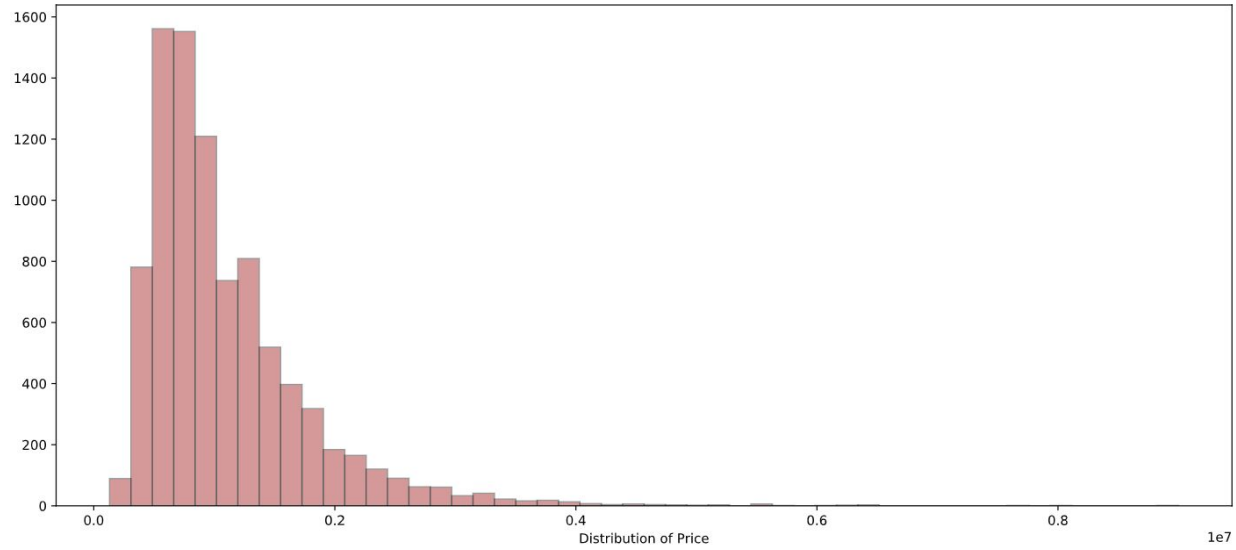

Figure 2. Distribution of price

3.1 Categorical features

Categorical features in the dataset consists of house type, selling method, region name, and the age of the estate. Expanding the scope of the analysis and taking into account them against the price, a few insights can be made:

- Houses are the most expensive type of estate, with the median recorded at 1 million, followed by townhomes at 850 thousand and the least are units, at just under AUD 500 thousand.

- Selling method has minimal impact on house prices, as all figures hover around the AUD 1 million mark.

- Regionally speaking, Metro Area has higher house prices than others in the state of Victoria. Within the Melbourne Metro Area, the South Area reported the highest selling price at AUD 1.5 million. North and West Metro figures are almost identical, at around 800 - 900 thousand for a house.

- With an average price of AUD 1 million, historic properties (older than 50 years old) are valued much higher than newer homes in the area, however they have more variation in price.

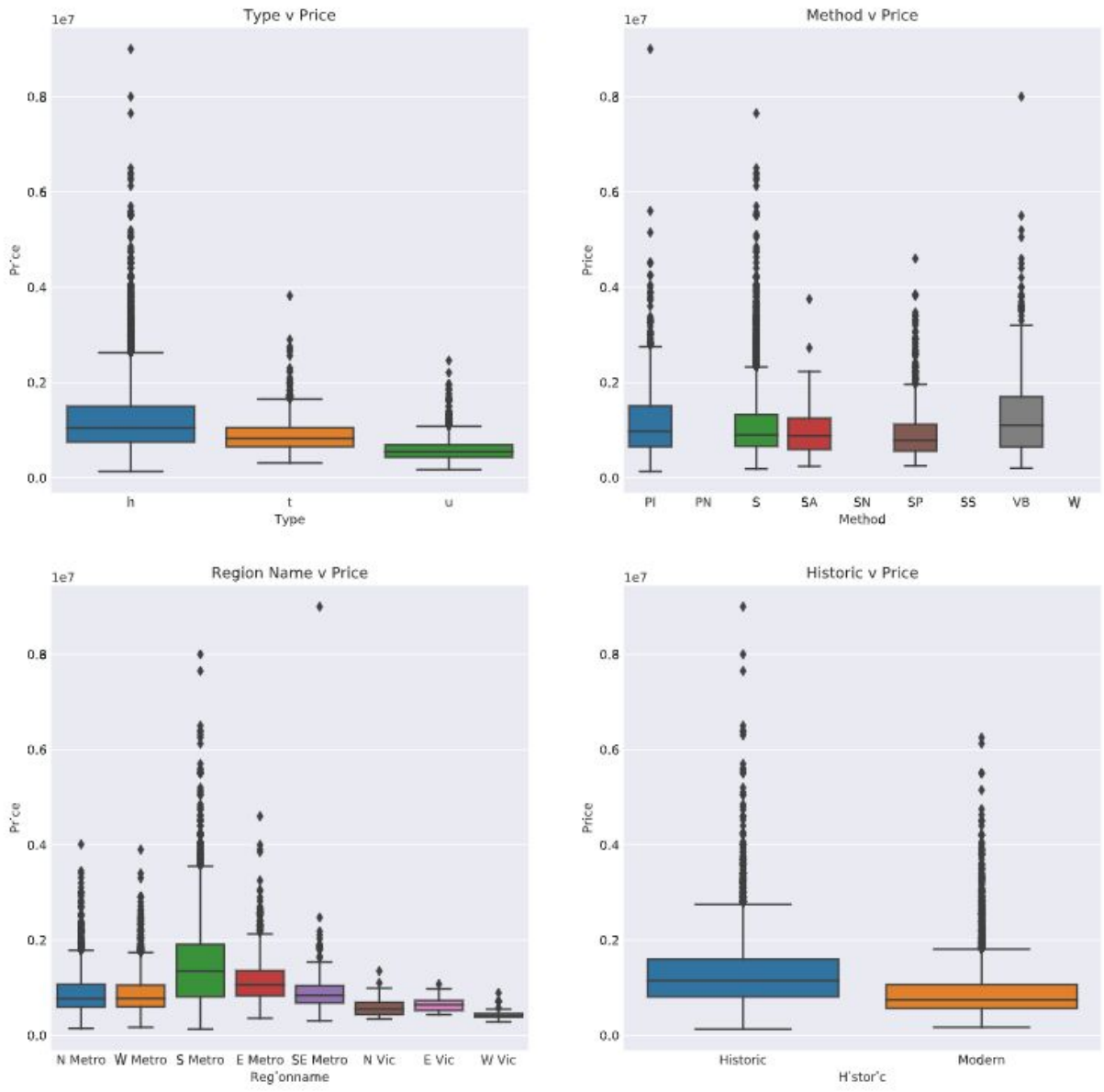

Figure 3. Box plots of categorical variables against prices

3.2 Numerical features

Regarding numerical variables, the dataset contains 'Rooms', 'Price', 'Distance', 'Bathroom', 'Car', 'Landsize', 'BuildingArea', 'YearBuilt', 'Propertycount', columns. Plotting them against price using scatterplots, several conclusions can be drawn:

- The majority of houses have 3-5 rooms with 1 to 3 bathrooms

- The price has a negative correlation with distance from Melbourne Centre of the Business District. Most expensive houses typically cluster within 20 kilometres from this centre.
- Land size and building area do not have very significant impact on the final prices, it can be explained as very large lands are only available in remote areas, thus limiting the price one owner can ask for the estate.

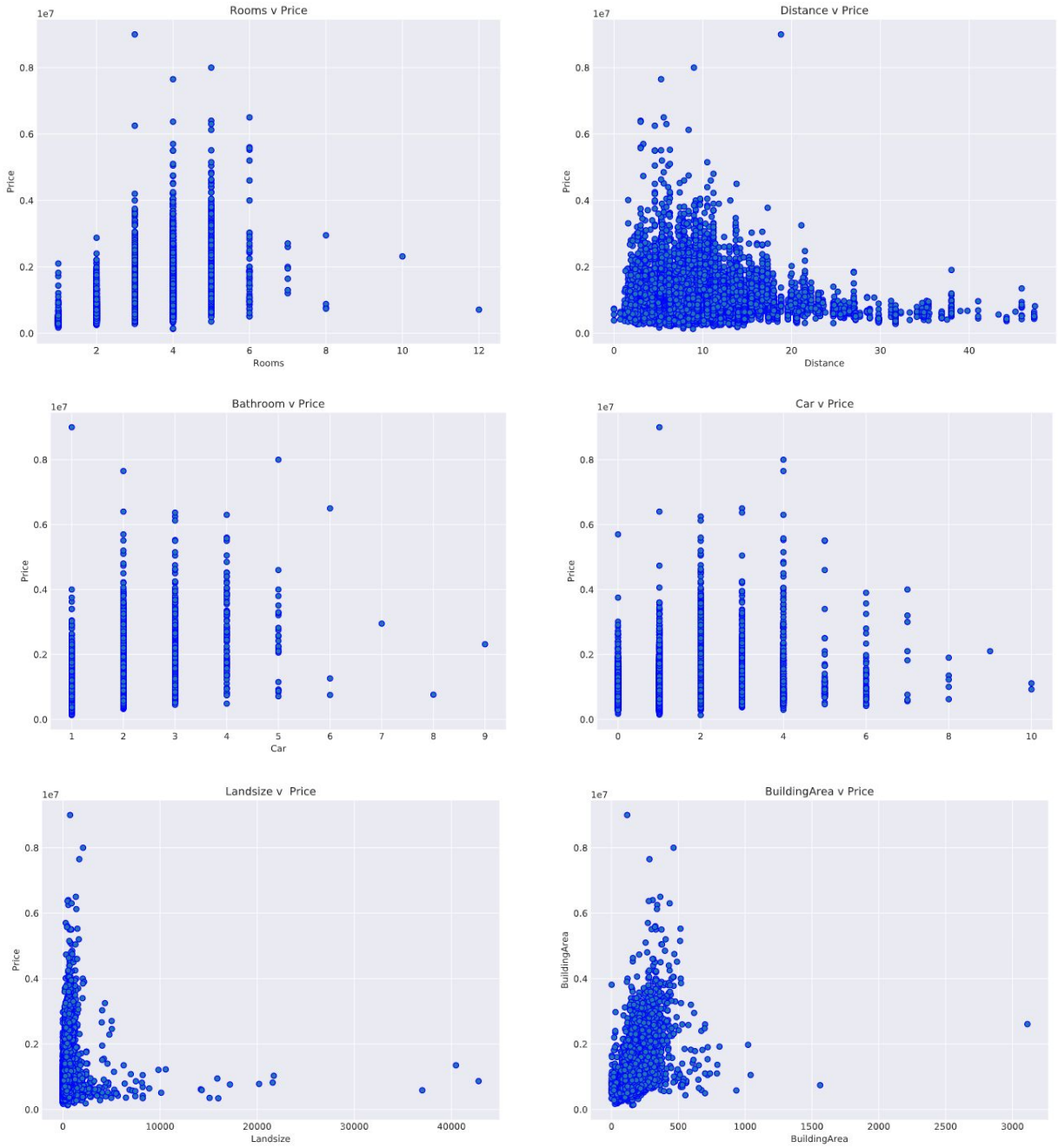

Figure 4. Scatterplots of numeric variables against prices

3.2 Correlation

Between variables, the heatmap confirms the hypothesis proposed earlier about age of a house and price (corr = 0.3). Although on average historic houses can be sold for higher prices, the variance is noticeable, as such, it is not definitive for a house's price to increase as it ages along.

Stronger indicators for prices can be observed in features 'Rooms', 'Bathrooms' and 'Building Area'. However, all three are moderately correlated to one another naturally, as bigger houses tend to have more rooms / bathrooms and vice versa.

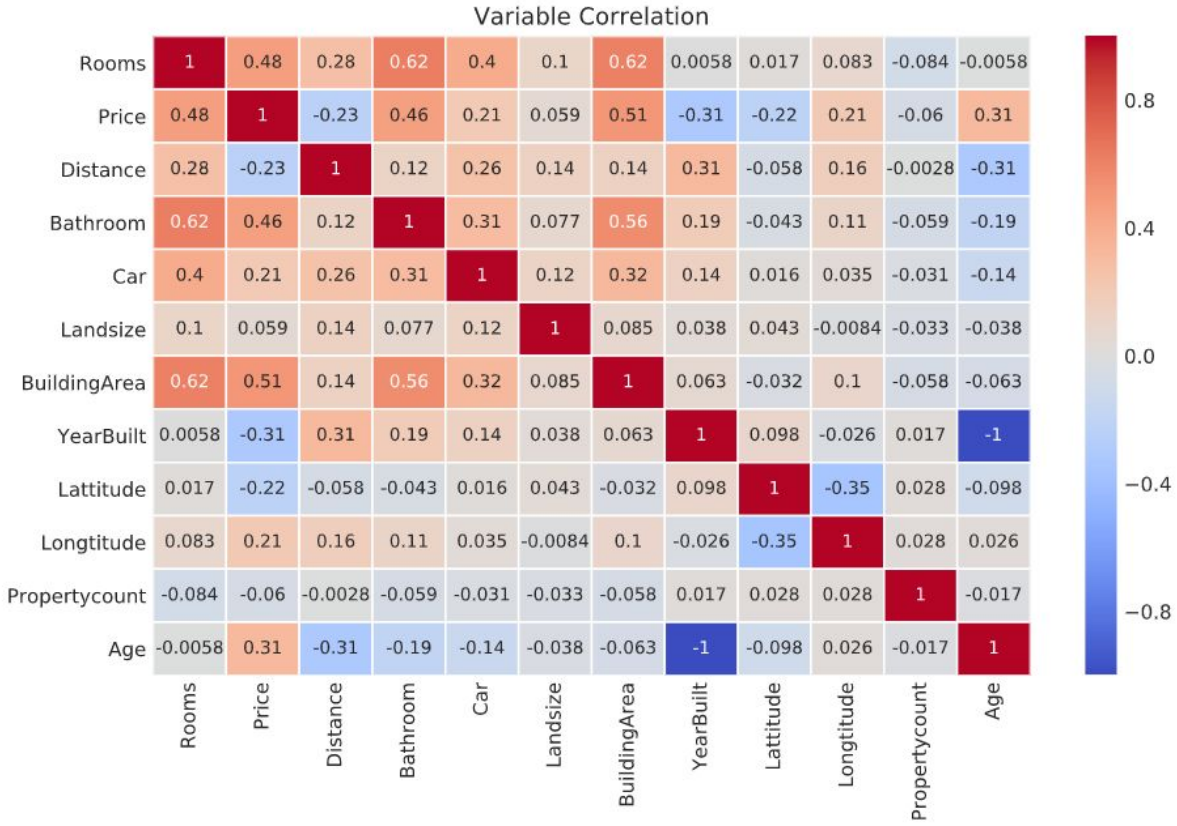

Figure 5. Heatmap showing relationships between variables with one another

4. Predictive modeling

The chosen model for this study is multiple linear regression, which can provide information of the dependency of house's prices on other features such as number of rooms, land size and house age. The model returns good scores for main accuracy measurements, summarised in the table below.

Mean absolute error	Root mean squared error	R squared	Cross validation score
312564.47	470194.87	0.52	0.5

Besides, a few observations can be drawn from the coefficients table:

- A unit increase in ‘Rooms’ would lead to an **increase** in ‘Price’ by AUD 130,782
- A unit increase in ‘Distance’ would lead to a **decrease** in ‘Price’ by AUD 28,481
- A unit increase in ‘Bathroom’ would lead to an **increase** in ‘Price’ by AUD 255,950
- A unit increase in ‘Car’ would lead to an **increase** in ‘Price’ by AUD 49,936
- A unit increase in ‘Landsize’ would lead to an **increase** in ‘Price’ by AUD 24.12
- A unit increase in ‘BuildingArea’ would lead to an **increase** in ‘Price’ by AUD 2,160
- A unit increase in ‘Propertycount’ would lead to a **decrease** in ‘Price’ by AUD 1.23
- A unit increase in ‘Age’ would lead to an **increase** in ‘Price’ by AUD 5,491

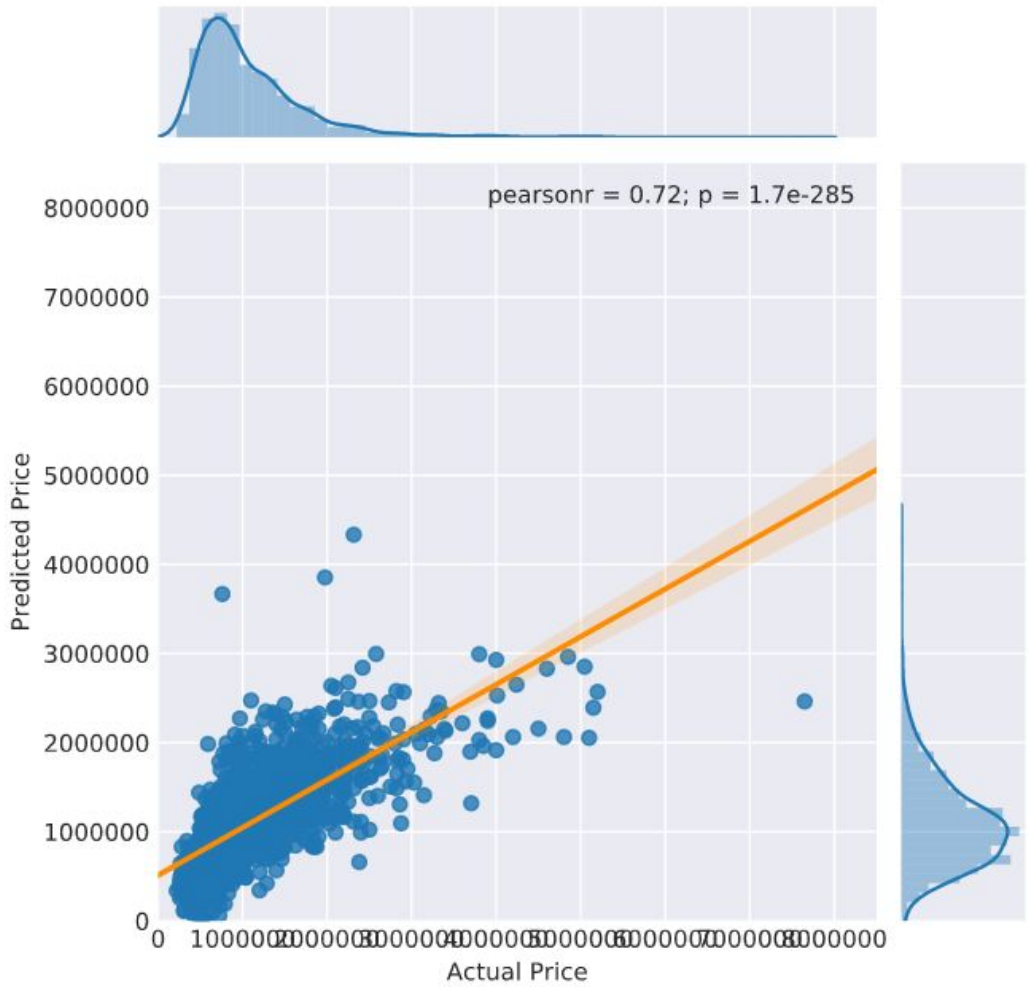

Figure 6. Linear regression of numerical features against prices

	Coefficients
Rooms	130781.994615
Distance	-28481.473952
Bathroom	255950.485267
Car	49936.398377
Landsize	24.128532
BuildingArea	2160.118289
Propertycount	-1.229619
Age	5491.481614

Figure 7. Coefficients of numerical features

Figure 8. Comparison between actual and fitted houses' values

5. Conclusions

In this study the author aims to suggest a model for houses' prices prediction in Melbourne Area, Australia. A multiple linear regression model was made to explore the scalar response of prices to changes in other features such as 'Rooms', 'Distance', 'Bathroom', 'Car', 'Landsize', 'BuildingArea', 'Propertycount' and 'Age'. The feature with the biggest impact on property's price is the number of rooms, while land size has the least influence. This study can provide a handy reference for Melbourne citizens who are looking for a new home or a real estate agent who wants to evaluate the value of a given property. Further studies can also use this model as a building block to draw more comprehensive recommendations.