Hedonic Analysis of Melbourne Real estate Prices

Aslan Bakirov

May 2021

Introduction.

Since 2010 various officials and real-estate analysts have claimed that Australia is in a property bubble. Pandemic only reviled those characteristics of the real estate market, since the start of 2021 number of house loans, property prices and the number of auctions has increased significantly. There are multiple reasons behind that such as low interest rates, the government program called *"First Time Buyers"*, and apparently, the disruptions brought about by COVID-19. All those market-specific characteristics make it compelling to work with Australian real estate data.

The most hectic part of the decade started is the period from 2016 to 2018. It is during these years that real estate prices in large cities, have increased substantially.

The data indicates that 2016-2017 is the period when real-estate prices boomed in the biggest cities of Australia. Later, since 2018 Q2 it was believed that the housing bubble deflated at that time. Yet, given the current trends of the market, as with the pandemic the prices are on the rise again, it is hard to conclude what is the driver and the mechanism behind the housing prices in Australia.

A case in point is Melbourne, where the house prices rose in the given period. Apart from the reasons attributable to Melbourne's mobile population, in this paper I will try to explain how the housing prices were determined in the city. The method of hedonic regression is applied to derive the insights I am after, and finally an own price index is calculated and visualized.

Data Exploration

The dataset used for the analysis is a snapshot of the Kaggle dataset, that was scraped from the publicly available results from the Australian Ibsite Domain.com.au. The original dataset contains a larger number of observations over a bit longer period of time (2016-2018). The snapshot that I use is different in the sense that it only captures the period when the housing prices skyrocketed, i.e. 2016 to 2017.

It contains multiple variables such as **Address, Rooms, Price, Type of the property** *(Type: br - bedroom(s); h - house; u - unit, duplex; t - townhouse; dev site - development site; ores - other residential)*, **Date** when the property was sold, **Distance from CBD, Number of Bathrooms, Number of Parking slots, Number of Bedrooms**.

Initially, the dataset contained 13580 observations. During the exploratory data analysis, I revealed that most of the numeric variables needed outlier handling. As such, real estate slots with more bedrooms than rooms, where number of rooms and bathrooms exceed 5 and 4 respectively were not included in the final version. Additionally, properties which were more expensive than 2.2 millions of Australian dollars were also considered as outliers. This decision can be questioned and countered with argument that conversion to log-scale eliminates this problem. Although that is correct, the price variable did not lose much from this capping, whereas if I were to save, it would save only 200 houses and added outliers from both ends. There are more outliers and non-available values for the year a house was built. As a result, only the buildings that were built after 1875 are involved in the analysis.

Houses where the number of bedrooms was greater than the number of total rooms were excluded. Homes with more than 5 bathrooms and 5 bedrooms also were left out as outliers. Below we can check the descriptive statistics of numeric variables the final dataset.

Table 1: Descriptive statistics

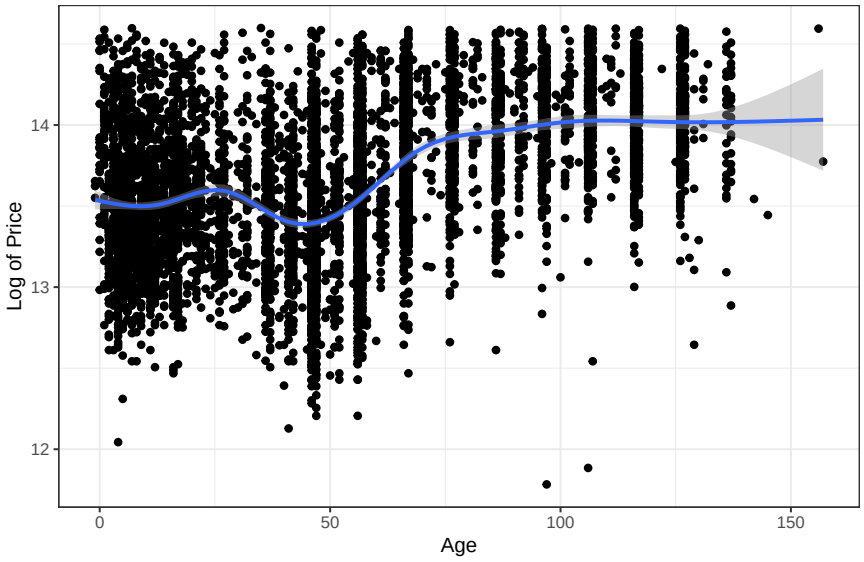
Statistic	N	Mean	St. Dev.	Min	Max
Rooms	6,073	2.846	0.878	1	5
Price	6,073	934,275.600	424,616.200	131,000	2,188,000
Distance	6,073	10.163	6.093	0.000	47.400
Postcode	6,073	3,101.473	90.940	3,000	3,977
Bedroom2	6,073	2.808	0.873	0	5
Bathroom	6,073	1.487	0.601	1	4
Car	6,048	1.536	0.906	0	10
Landsize	6,073	465.692	959.711	0	37,000
BuildingArea	6,073	125.977	50.731	2.000	274.000
YearBuilt	6,073	1,965.240	35.645	1,860	2,017
Lattitude	6,073	−37.806	0.081	−38.165	−37.409
Longtitude	6,073	144.987	0.106	144.542	145.526
Propertycount	6,073	7,474.966	4,425.332	389	21,650
price_log	6,073	13.643	0.464	11.783	14.598
unit_price	6,073	8,934.829	18,637.540	845.161	518,500.000

Before moving on, let's take a step back and familiarize ourselves with how the house market in Australia operates. In large cities, akin Melbourne, houses are sold either privately, directly by the owners, or by the real estate agents. The latter, in turn, can sell the property either privately or on the auction. This dataset that I work on contains houses only traded publicly on auctions, and has several specifications on the type of the trade. For instance, Method how property sold (S - property sold; SP - property sold prior; PI - property passed in; VB - vendor bid; SA - sold after auction). The semantics are as following: If the property is sold prior to auction, means there has been an offer to the vendor that overthrew the benefits from the auction; If the vendor bids on the property, it signals a property of a greater quality; when the public bid does not reach the reservation price of the vendor, the right to negotiate the purchase is passed in to the highest bidder; if the highest bidder and the subsequent higher bids fail to succeed in bargaining for the house, it is sold after the auction; if the bid for property reaches the reservation price of the vendor, it is simply sold.

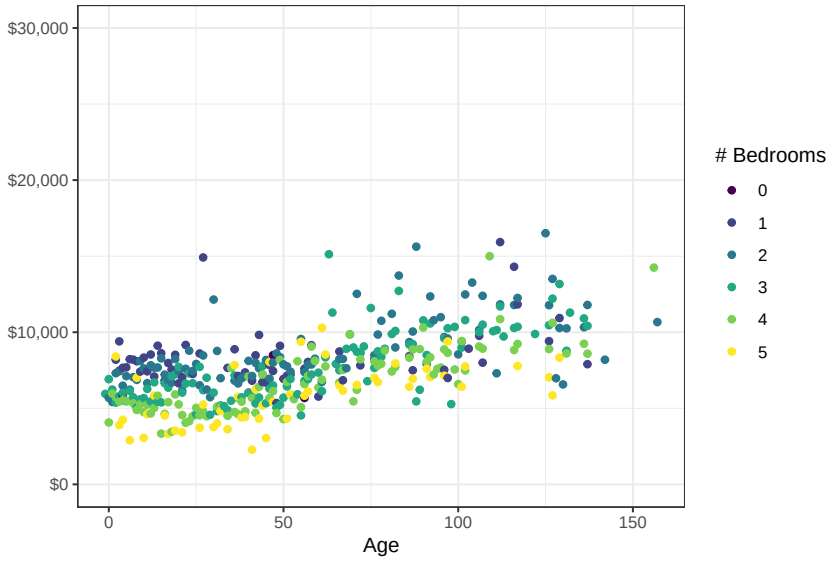
Bearing this in mind, we might hypothesize that properties sold prior to the auction might be overpriced, since this offer makes the vendor to forego the possible highest bids on the auction, where he/she plans to sell for the reservation price.

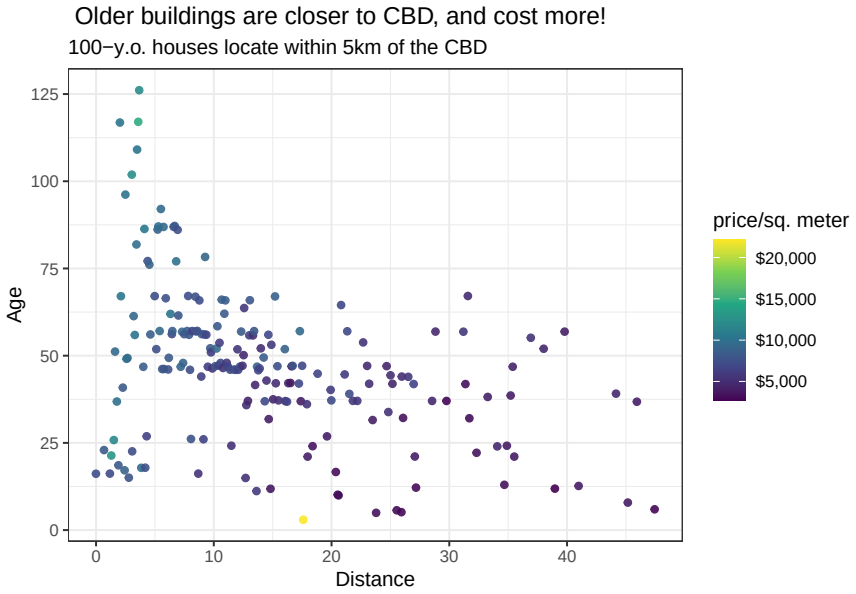
Furthermore, I have created the age variable, measuring how old the building was when sold. Plotting against the price it reveals somewhat quadratic relationship:

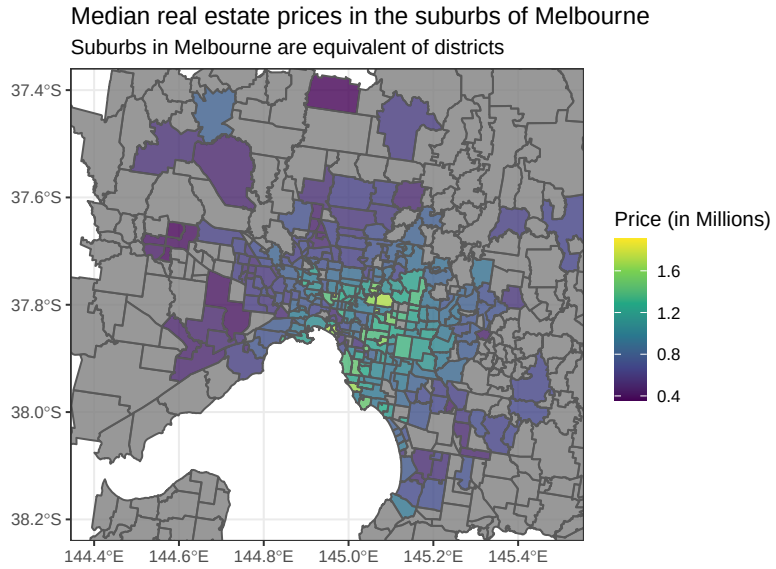
Age of the house vs Price



Older buildings have less bedrooms, but still cost more?!







Modeling

Having become familiar with the dataset, I proceed with modeling. In this section, I construct a set of models, all hedonic regressions, trying to explain the variations in real estate prices in Melbourne. First, I include intrinsic characteristics of a building such as the number of bedrooms, bathrooms, the total number of rooms land and building size in square meters as predictors, and log of price as the dependent variable. The results show that land size is not a significant predictor of the price.

The first model only reflects the internal characteristics' impact on the price, yet even that is hard to state when I left out a couple of extremely important factors. Thus, In the second model, I add the distance from the City Business District (CBD) as a proxy for location. Covariates controlling for the car spots and the age of the building are also added. The model includes polynomials of the second degree for age and bedroom since I have observed a quadratic-like relationship between price and these in the graphs. Finally,

to control for the time trend and possible seasonality that comes with it, I make use of monthly dummies for each year. This decision stems from the fact that our data is only for two years, and contains patterns occurring more often than a year. Monthly dummies are extracted from the Date variable, indicating when a building has been sold, yet there were missing months in the data, namely March 2016. February 2016 is set as the base month.

This also allows us to calculate the price index changing over time, not worrying about the assumption that other characteristics of the buildings hold constant over time. Since I included all the confounders, I believe may affect the price, the regression estimates of the dummies contain only the time change in the prices.

Results

Below, I present the hedonic regression model which explains the log-price using intrinsic characteristics of the house, how it was sold, and distance from CBD.

Table 2: Hedonic regression

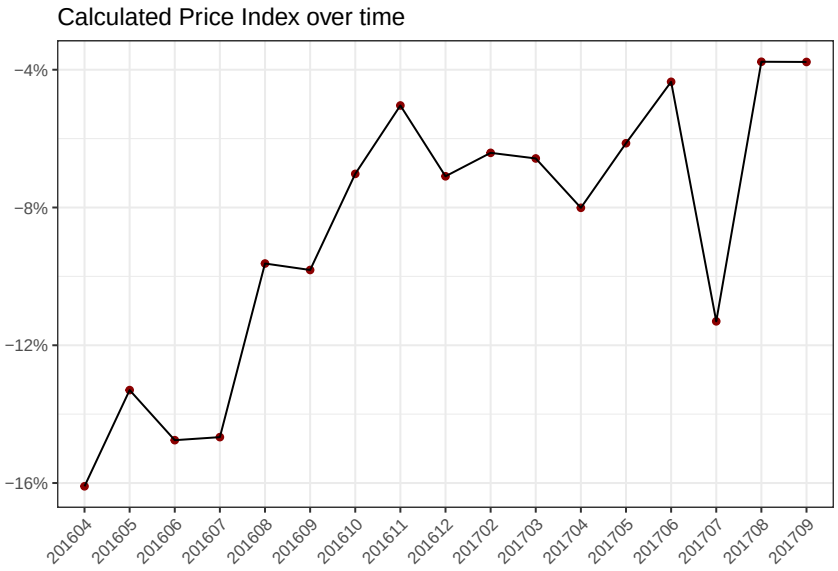
term	estimate	std.error	statistic	p.value
(Intercept)	12.644	0.051	248.030	0.000
Rooms	0.103	0.018	5.885	0.000
Distance	-0.022	0.001	-32.177	0.000
poly(Bedroom2, 2)2	-6.475	0.286	-22.642	0.000
Bathroom	0.109	0.008	13.034	0.000
Car	0.032	0.005	6.966	0.000
BuildingArea	0.004	0.000	30.473	0.000
age	0.004	0.000	9.884	0.000
I(age^2)	0.000	0.000	2.472	0.013
factor(Method)S	0.079	0.012	6.799	0.000

It reveals that, the buildings lose 2.2% in their price for each kilometer they are farther from the CBD. This is in line with the "three-locations rule", albeit the distance from CBD is only a rough proxy for location. For better development of this question in Melbourne, a proper spatial analysis is required. Particularly interest in this approach should be devoted to the metropolitan migration, since Melbourne is one of the fastest growing cities in OECD.

Apparently, the largest predictor of the prices is the number of rooms, as such, of two houses with the same amount of bedrooms, bathrooms, car spots, and in the distance same from the center, the one with an additional room would cost 11% more. This is also the case if they differed only with one extra bathroom, ceteris paribus.

The polynomial Age variable shows a 4% growth in price with each year, after the house turns 50. Among the methods of merchandise, the only significant one is *Sold*, meaning that properties Sold in the Auction cost 8% more than those passed in. This makes sense if the bid does not reach the reservation price of the vendor

Price index



Finally, I added the monthly dummies for each month that I have in the dataset to the previous model, and came up with own house price index. This index is also good since it does not depend on the assumption that the price change is not driven by intrinsic characteristics of the property since they are time-invariant. I can relax this assumption since I include them in our hedonistic regression. According to this, I determine that price boom started from the April of 2016, to the November of the 2016. Prices rose for nearly 11%, base month is February 2016.