Coursera Capstone - A Comparison of a Suburb's Median House Price in Four Major Australian Cities

Stephen Moore

1 Introduction

In this assignment we will study the median house prices in the different suburbs of four major Australian cities, namely Brisbane, Melbourne, Sydney, and Perth. The goal will be to investigate the relatively simple questions of whether the median house price correlates with the distance to the central business district (CBD), or whether it can be better correlated to the types of venues that are found within the suburb. Furthermore, the differences in prices and will be compared between the four cities. The value for such an analysis could be realised by any business involved in real-estate where a detailed knowledge of house prices in different cities and suburbs could add value to potential clients. One particular example could be in an online tool recommending different suburbs in different cities to people interested in immigrating to Australia, based on the financial situation, for example. Furthermore, the analysis could be used to provide information about the types and prevalence of different venues within a suburb of interest.

2 Data

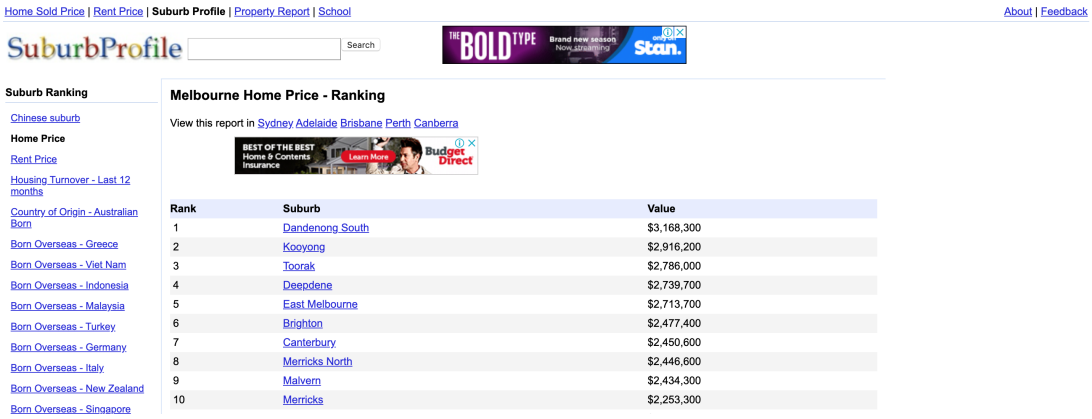


Figure 1: Screenshot of the Suburb Profile website.

The primary data source for this assignment is the 'Suburb Profile' website, that contains tables of median house prices for the different suburbs, as depicted in Figure 1. This data will be scraped and assembled into a data frame using the *Beautiful Soup* library. In addition, the geographic coordinates of the city CBD and suburbs will be

gathered via the *Nominatum* geocoder API and the categories of nearby venues in each suburb will be gathered via the *Foursquare* API.

3 Methodology

In order to assemble the initial data frame from the Suburb Profile website it is important to note that only fifty suburbs are listed per page, with a 'next' link to manually move from page to page. In order to automate the process of web scraping process an entity '`&page=0`' is appended to the URL such that the HTML can be requested inside a for loop, incrementing the page counter. As can be observed in Figure 1, the data is contained in a table, which can be easily found with Beautiful Soup by finding all table rows (with the `tr` tag). Within these rows a regular expression can be used to find standard table cells (with the `td` tag) that have a dollar sign in them, followed by a number. The suburb and price can then be obtained by getting the first and second siblings of the table row.

Having compiled the suburbs and prices for each city, the next step is to find the latitude and longitude for each suburb. The most efficient way to achieve this for a large number of locations is to use the 'RateLimiter' function from the geopy library, passing it the data frame of suburb names. As a side note, it was observed that in order to successfully locate the suburbs in their respective cities, search terms that concatenate the suburb name with the city, state, and country names were added to the data frame so that, for example, Nominatum wouldn't return locations in Europe or North America in the cases where suburbs with the same names exist there. Once the geographic coordinates of each suburb were found, the distance to the CBD was computed using the 'distance' function from the geopy library. This function takes as input two sets of (lat,lon) coordinates and can output a distance in kilometers. To limit the number of suburbs to those which could reasonably be considered part of the city a cutoff value of 40km was used to which any suburb farther from the CBD was dropped from the resulting data frame.

The final step in compiling the dataset involved getting nearby venues within each suburb. Using the Foursquare API to explore each suburb location with radius of 1km and limiting the search to 100 results another data frame was created for each city.

The subsequent analysis involved initially visualising the distributions of house prices for the four cities and then creating Leaflet maps to visualise the house prices within each suburb. Using the nearby venues data frames, some simple k-means clustering was performed for each city using the 'Elbow method' to determine the most appropriate number of clusters and then creating Leaflet maps to display the clusters, such that a visual comparison to the house prices can be made.

Using the distance to the CBD as a simple metric a simple polynomial regression analysis was performed for each city, exploring how reasonable the notion of using just the distance to the CBD as a single feature in predicting a suburbs' median house is. Following this analysis a simple regression was performed using a support vector regressor (SVR) instead using one-hot encoded venue categories, in order to explore whether there is any utility in attempting to use the types of venues found within a suburb to predict its

median house price. While intuitively one would not necessarily think that this would be an effective way to predict house price, a large parameter sweep using the `GridSearchCV` function was performed in order to determine the best possible hyper-parameters with which to make the prediction.

4 Results

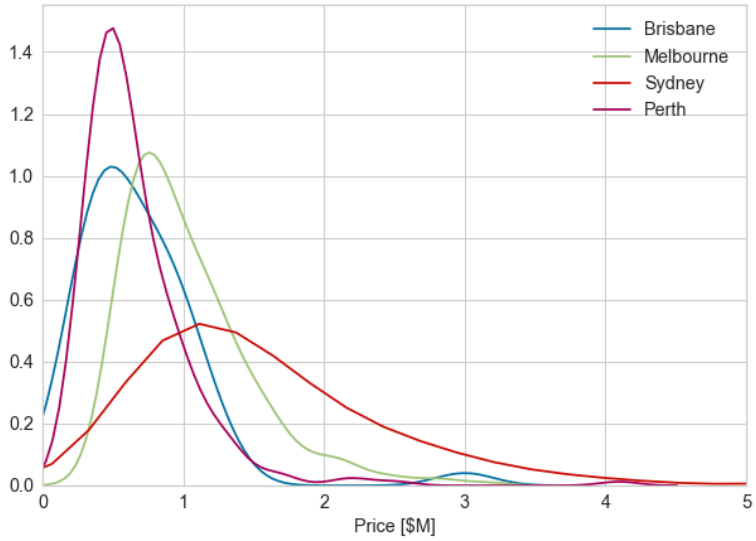
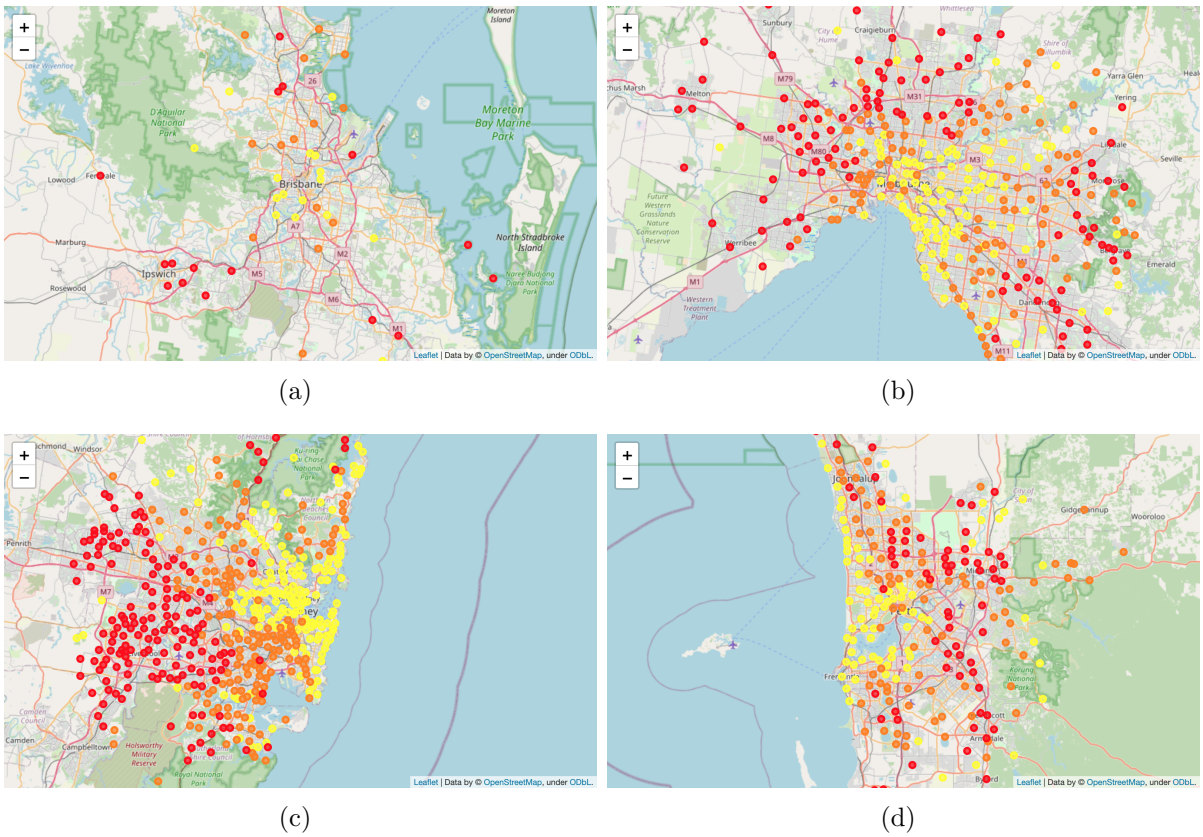


Figure 2: Distributions of house prices in the different suburbs of the four cities considered.

Figure 2 presents the distributions of the median house prices by suburb for the four cities considered. As can be observed for Brisbane, Melbourne, and Perth, the majority of house prices tend to lie in a narrow range between around $0.5 - 0.8 million. Sydney appears as distinctly different with a much broader range of house prices. While the peak of the distribution is centred around $1.2 million, the distribution (unlike the other cities) includes house prices all the way up to $23 million.

Figures 3(a)-3(d) present Leaflet maps generated with the *Folium* library, using markers to locate and colour the suburbs by price. For each city the suburb median house prices were grouped into three bins using the *Pandas* 'qcut' function. Furthermore, for comparison with the distance to the CBD, Figures 4(a)-4(d) present the same Leaflet maps, but colouring the markers by the distance to the CBD. One observation that can be made is that, as one would intuitively believe, the most expensive suburbs tend to be those that are in close proximity to the CBD with a gradual and almost radial reduction in price as one moves further away from the CBD. An interesting exception here is with

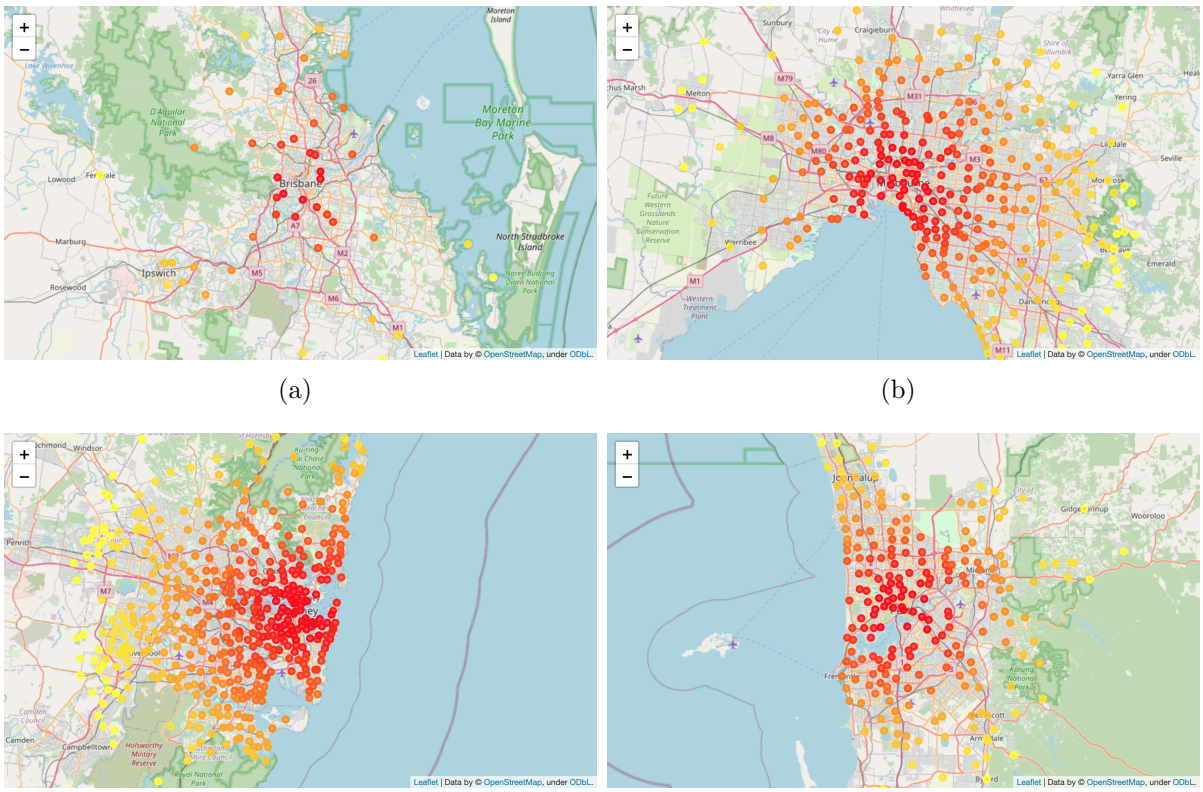


(a) (b)

(c) (d)

Figure 3: City suburb prices grouped into 3 bins (a) Brisbane (b) Melbourne (c) Sydney (d) Perth. Note that with the 'autumn' colourmap used, yellow markers indicate the highest prices and red indicate the lowest prices.

Sydney and Perth, which are located near a long coastline. In these cases the house prices remain high along the coast, even as the distance to the CBD increases. Suburbs that spread inland however, tend to show the more expected decrease in house price with increasing distance from the CBD.



(a) (b)

(c) (d)

Figure 4: City suburb distances to CBD grouped into 10 bins (a) Brisbane (b) Melbourne (c) Sydney (d) Perth. Note that with the 'autumn' colourmap used, yellow markers indicate the suburbs furthest from the CBD and red indicate the closest suburbs to the CBD.

Using the data frame containing nearby values for each suburb, Figures 5(a)-5(d) present the top ten most common venues for the two most expensive and two least expensive suburb for each city. One interesting observation that can be made is that the most common venues tend to include cafes, supermarkets, convenience/liquor stores, various restaurants, parks, gyms, and sports clubs. One perhaps unexpected result is that there does not appear to be any obvious difference in the most common venues when considering the most expensive or least expensive suburbs, hinting at the likelihood that predicting the house price for a suburb, based on the nearby venues is unlikely to be successful.

	Suburb	Distance to CBD [km]	Price [$M]	Venue #1	Venue #2	Venue #3	Venue #4	Venue #5	Venue #6	Venue #7	Venue #8	Venue #9	Venue #10
0	Burbank	16.6	1.3	Gym	Asian Restaurant	French Restaurant	Coffee Shop	Cafe	Bowling Green	Convenience Store	Hungarian Restaurant	Golf Course	Fabric Shop
1	Bulimba	4.2	1.2	Fast Food Restaurant	Diner	Sandwich Place	Furniture / Home Store	Park	Home Service	Dessert Shop	Dim Sum Restaurant	Department Store	Farm
43	Basin Pocket	28.9	0.2	Gym	Cafe	Liquor Store	Park	Supermarket	Pharmacy	Pizza Place	Bus Station	Pub	Shopping Mall
44	Lamb Island	39.2	0.1	Convenience Store	Park	Train Station	Electronics Store	Coffee Shop	Dance Studio	Pharmacy	Pizza Place	Cafe	Playground

(a)

	Suburb	Distance to CBD [km]	Price [$M]	Venue #1	Venue #2	Venue #3	Venue #4	Venue #5	Venue #6	Venue #7	Venue #8	Venue #9	Venue #10
0	Dandenong South	32.8	3.2	Supermarket	Fast Food Restaurant	Grocery Store	Restaurant	Coffee Shop	Spa	Cafe	Big Box Store	Men's Store	Shopping Mall
1	Kooyong	7.1	2.9	Supermarket	Fast Food Restaurant	Bakery	Portuguese Restaurant	Big Box Store	Market	Discount Store	Breakfast Spot	Department Store	Farm
317	Melton	38.7	0.4	Cafe	Zoo Exhibit	Ethiopian Restaurant	Event Space	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant
318	Stony Creek	8.9	0.3	Cafe	Park	Coffee Shop	Food Truck	Grocery Store	Thai Restaurant	Pizza Place	Skate Park	Gym	Bar

(b)

	Suburb	Distance to CBD [km]	Price [$M]	Venue #1	Venue #2	Venue #3	Venue #4	Venue #5	Venue #6	Venue #7	Venue #8	Venue #9	Venue #10
0	Point Piper	3.5	24.0	Farm	Convenience Store	Athletics & Sports	Park	Preschool	Frozen Yogurt Shop	Fruit & Vegetable Store	Field	Filipino Restaurant	Financial or Legal Service
1	Elizabeth Bay	2.2	8.9	Cafe	Thai Restaurant	Italian Restaurant	Pizza Place	Liquor Store	Australian Restaurant	Grocery Store	Park	Pier	Sports Club
499	Hebersham	38.2	0.5	Food & Drink Shop	Home Service	Cafe	Liquor Store	Snack Place	Food	Filipino Restaurant	Financial or Legal Service	Fish & Chips Shop	Fish Market
500	Dharruk	39.0	0.5	Cafe	Park	Furniture / Home Store	Thai Restaurant	Coffee Shop	Pizza Place	Gym	Bakery	Pub	Korean Restaurant

(c)

	Suburb	Distance to CBD [km]	Price [$M]	Venue #1	Venue #2	Venue #3	Venue #4	Venue #5	Venue #6	Venue #7	Venue #8	Venue #9	Venue #10
0	Peppermint Grove	10.2	4.1	Big Box Store	Thai Restaurant	Park	Mobile Phone Shop	Fish Market	Fish & Chips Shop	Field	Fast Food Restaurant	Farmers Market	Eastern European Restaurant
1	Dalkeith	7.4	2.5	Park	Sports Club	Pizza Place	Thai Restaurant	Grocery Store	Pharmacy	Gas Station	Indian Restaurant	Fast Food Restaurant	Farm
223	Koongamia	17.9	0.2	Park	Hockey Rink	Grocery Store	Fast Food Restaurant	Falafel Restaurant	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Event Service	Event Space
224	Medina	31.6	0.2	Pub	Italian Restaurant	Golf Course	Shopping Mall	Coffee Shop	Bookstore	Liquor Store	Supermarket	Home Service	Farm

(d)

Figure 5: The top ten most common venues in the two most expensive and two least expensive suburbs in (a) Brisbane (b) Melbourne (c) Sydney (d) Perth.

In order to perform the k-means clustering, the optimal number of clusters for each city was obtained by using the 'KElbowVisualizer' from the *yellowbrick* library. Figure 4 presents an example result for the city of Perth, illustrating that the elbow is located at ten clusters and hence, this was the the number used for the k-means analysis. The same approach was used for the other cities using 14 clusters for Brisbane, 14 clusters for Melbourne, and 18 clusters for Sydney.

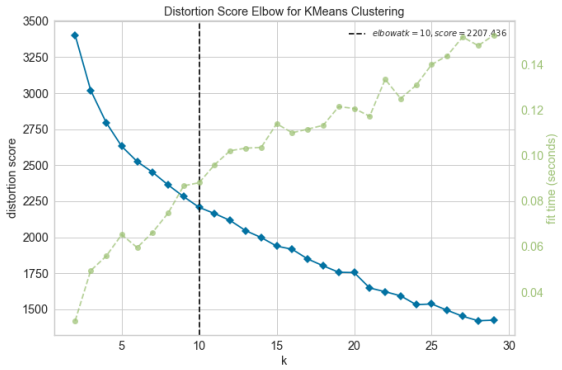
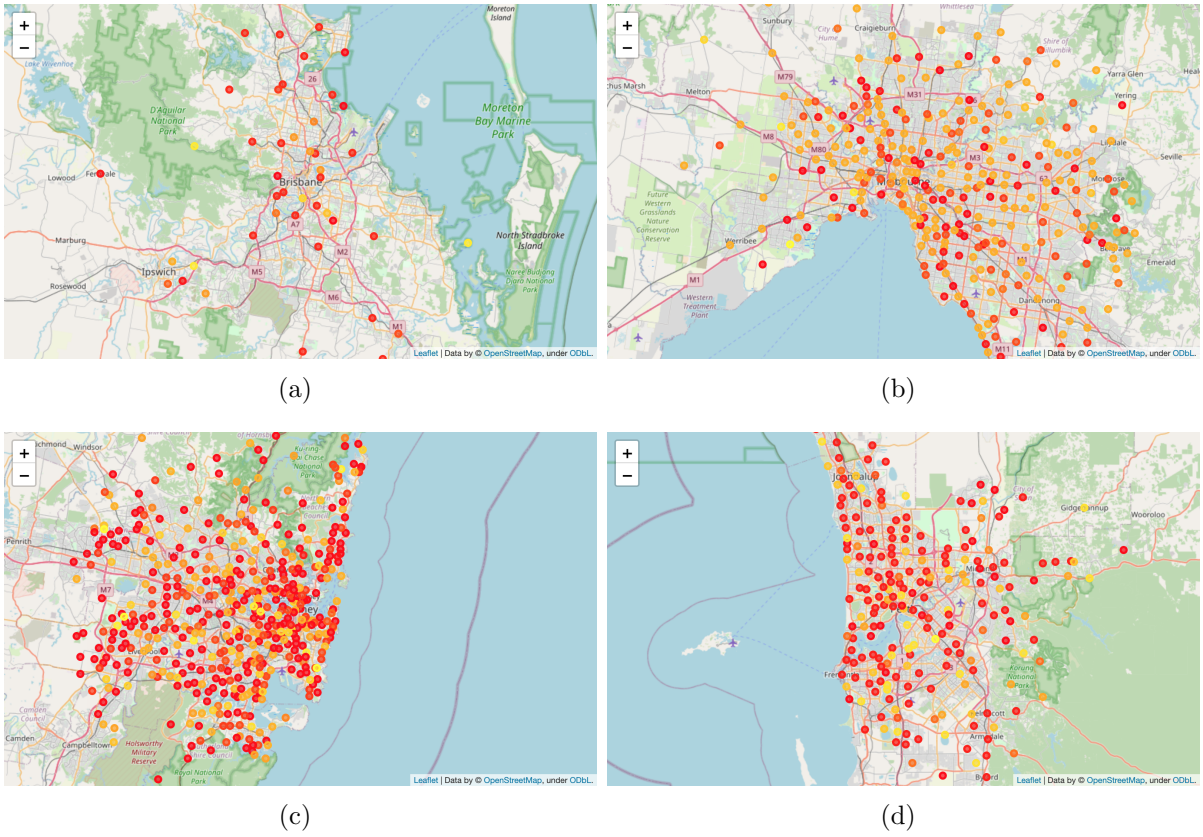


Figure 6: An example plot of the distortion score used to find the elbow and hence the optimal number of clusters for the city of Perth.

Figures 7(a)-7(d) present Leaflet maps for the four cities, using markers to locate and colour the suburbs by the resulting k-means assigned cluster label. As can be observed, there is no discernible pattern in terms of the locations of the clusters, with different clusters being scattered almost randomly throughout the city. The only exception to this trend is with Melbourne, where some unique clusters are found on the coast to the south west.



(a) (b)

(c) (d)

Figure 7: City suburb distances to CBD grouped into 10 bins (a) Brisbane (b) Melbourne (c) Sydney (d) Perth.

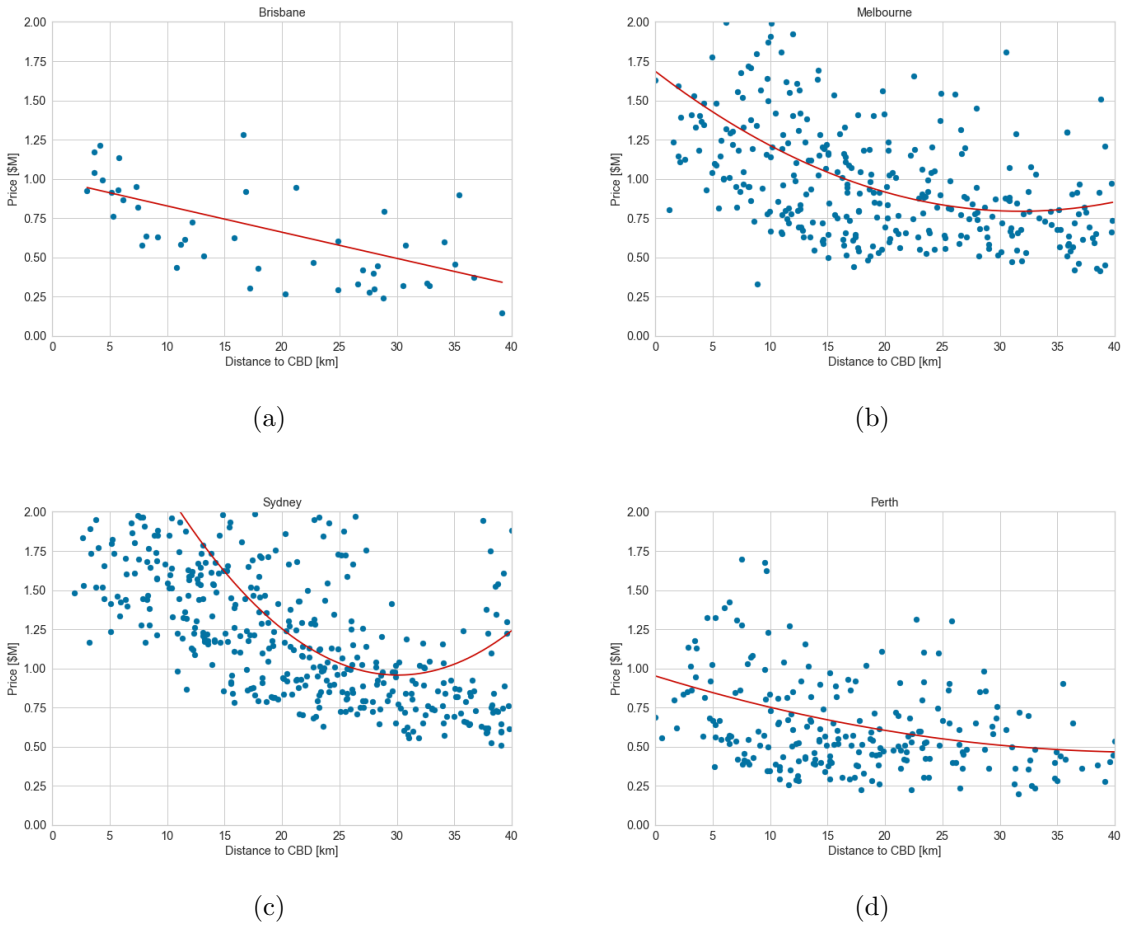


Figure 8: Scatter plots of suburb prices with a second order polynomial fit (a) Brisbane (b) Melbourne (c) Sydney (d) Perth.

In order to test the hypothesis that the distance to the CBD could be used to predict the median house price of a suburb, a simple polynomial regression analysis was performed for each city. Figures 8(a)-8(d) present scatter plots for each city, showing the resulting second order polynomial, fit with the 'LinearRegression' model using 'PolynomialFeatures' from the *sklearn* library. As can be observed, while there is a lot of variance about the regression curves, there also appears to be a general second order decrease in house price with distance from the CBD, but for Melbourne and Sydney the prices increase again around the 40km cutoff value. When looking at the prices as depicted in the Leaflet maps of Figures 3(a)-3(d) it can be observed that there are in fact a number of expensive suburbs far away from the CBD more rural areas, in an almost radial pattern. This result is partly a consequence of the 40km cutoff and if it had been reduced to around 30km or so, this effect would most likely not have been observed.

Finally, in order to test the hypothesis that a suburb's nearby venues could be used

City	Suburb Distance to CBD [km] 2^{nd} Order Polynomial MSE	Suburb Nearby Venues SVR MSE
Brisbane	0.1051	0.0973
Melbourne	0.0626	0.2570
Sydney	1.0823	1.0243
Perth	0.1005	0.2720

Table 1: Comparison of the mean squared errors (MSE) for two different approaches to predicting a suburb's median house price.

as a means to predict the median house price, a support vector regressor (SVR) from the *sklearn* library was used with the suburb's one hot encoded set of nearby venues. Although it would seem unlikely that this approach would be successful, given the almost random clustering of suburbs, an extensive sweep over the SVR parameters was performed with the 'GridSearchCV class' using a five fold cross validation, to obtain the optimal hyper-parameters. The parameters swept over the 'kernel', testing linear, polynomial, and radial basis functions (and for the polynomial kernel swept over polynomial degrees 1-6), and swept over a regularisation parameter in the range of $10^{-6} - 1$. For all cities, the polynomial kernel produced the best score, with degrees 5 for Brisbane and Melbourne, degrees 2 for Sydney, and degree 1 for Perth. The best regularisation parameters were in the range of $10^{-4} - 10^{-3}$. Once the optimal hyper-parameters for each city had been found, the SVR was finally retrained on 70% of the suburb data with the mean squared error (MSE) computed based on the remaining 30%. Table 4 presents a comparison of the mean squared errors for the two approaches. As can be observed, the second order polynomial regression using a suburb's distance to the CBD tended to produce the lowest MSEs. Surprisingly however, the SVR produced MSEs that were in a similar range (at least for Brisbane and Sydney). As a final test the SVR was retrained using *both* the distance to the CBD and nearby venues as features using the same grid search approach, but the MSEs showed only a minor improvement over using just the nearby venues and are hence not presented.

5 Discussion

The analyses performed on the median house prices of each suburb of Brisbane, Melbourne, Sydney, and Perth show that in general, the price follows a quadratic decrease with distance from the CBD. This is intuitively what one might expect, but also useful to see this effect presented in a quantitative manner. When comparing the four cities it is apparent that Sydney has the broadest distribution in house price, with many expensive suburbs having median house prices well over $1 million. When comparing nearby venues, it is apparent that both the most expensive and least expensive suburbs appear to contain similar venues such as cafes, supermarkets, parks, and sports venues. It is interesting to note that for cities like Melbourne, where a suburb's median house price

correlates reasonably well with distance to the CBD, the polynomial regressor using this feature performed significantly better than the SVR using the suburbs nearby venues. For cities like Sydney where expensive suburbs along the coast violate this trend, using the nearby venues proves as useful a predictor as the nearby venues. Perhaps a more detailed analysis of the individual suburbs focussing on specific venues such as beaches, highly ranked schools, or proximity to public transport, or conversely industrial areas, airports, etc, might result in a clearer separation between the expensive suburbs, but this remains future work.

6 Conclusion

This report performed an analysis of the median house prices of four major Australian cities, assembling a dataset of location, price, proximity to the CBD, and nearby venues. The analysis included Leaflet maps visualising the suburbs and their prices, performed both clustering and regression in an attempt to see if the types of nearby venues correlate with the median house price, or could be used to predict the median house price. The results indicated that both the most and least expensive suburbs are comprised of similar venues and that while the proximity to the CBD serves as a 'reasonable' feature with which to predict the house price, the types of venues do not. A more detailed analysis is likely required in order to build a more predictive model of a suburbs median house price.