Project work

MS-E2112
Multivariate Statistical Analysis

Matti Hyyppä

Contents

1 Introduction

Location is often considered as one the most important factors when determining the price of a property. Therefore, the main goal of this study was to find out how much the price of a property depends on the property's location. The goal of this study was also to find out what other variables have the highest dependency with the price. The multivariate analysis of the data set was performed with multiple correspondence analysis (MCA). MCA is similar to principal component analysis (PCA) but MCA is only suitable for categorical variables whereas PCA is appropriate only for continuous variables.

For this project, I used a data set of Melbourne's housing market prices that can be found on Kaggle. The Kaggle page contains two different data sets that have different number of variables. The file named Melbourne_housing_FULL.csv was used for this study. The data set contains 21 variables for the 34,857 properties each of which has been sale on the market between 2016 and 2018. Because the time interval is quite short, the distributions of the variables are considered to be the same during the time interval.

2 Univariate analysis

The variables of the data set are explained in table 1. Some of the variable names in the original data set have typos which have been corrected. In addition, some variable names have been shortened to make the plots in section 4 to be less cluttered. For the same reason, the values for the `Region` variable have been shortened. Values containing the string "Metropolitan" were shortened based on the cardinal and intercardinal directions. For example, "Northern Metropolitan" was shortened to "N" and "South-Eastern Metropolitan" to "S-E". Values containing the string "Victoria" were also shortened based on the cardinal and intercardinal directions but also appended with the string "Vic" to distinguish from the Metropolitan areas. The table contains a written description for each variable. Also, the minimum and maximum values and the quartiles of the values have been included for numerical variables.

Variable	Description	Min	Q_1	Q_2	Q_3	Max
`Suburb`	Suburb of the property	-	-	-	-	-
`Address`	Address of the property	-	-	-	-	-
`Rooms`	Number of rooms	1	2	3	4	16
`Price`	Price in Australian dollars	85 k	635 k	870 k	1.295 M	11.2 M
`Method`	e.g. property sold, withdrawn prior to auction...	-	-	-	-	-
`Type`	e.g. house, townhouse...	-	-	-	-	-
`SellerG`	Real estate agent	-	-	-	-	-
`Date`	Date sold	-	-	-	-	-
`Dist`	Distance from Central Business District (km)	0.0	6.4	10.3	14.0	48.1
`Region`	General Region (West, North West, North, North east, etc)	-	-	-	-	-
`PropertyCount`	Number of properties that exist in the suburb	83	4385	6763	10412	21650
`Bed`	Number of bedrooms	0	2	3	4	30
`Bath`	Number of bathrooms	0	1	2	2	12
`Car`	Number of car spots	0	1	2	2	26
`Land`	Land size (m^2)	0	224	521	670	433,014
`Area`	Building area in square meters	0	102	136	188	44515
`Year`	Year the house was built	1196	1940	1970	2000	2106
`CouncilArea`	Governing council for the area	-	-	-	-	-
`Lat`	Latitude in degrees	-38.19	-37.86	-37.81	-37.75	-37.39
`Lon`	Longitude in degrees	144.4	144.9	145.0	145.1	145.5

Table 1: The variables in the data set. Some variable names have been shortened.

From table 1, it can be seen that there are multiple variables that describe the location of a property. The categorical location-related variables have very different number of categories. `Suburb` has 351 modalities, `Address` has 34,009, `CouncilArea` has 34 and `Region` has 9. Without

more expertise in the real estate market of Melbourne, it is quite difficult to work with the variables `Suburb`, `Address` and `CouncilArea` because of the large number of different categories and the challenges in trying to regroup them into smaller ones. Therefore, `Region` was chosen as one of the location-related variables that will be used for determining the dependency between the location and price of a property.

However, the region of a property might not be sufficient when describing the location of a property. Especially, if the regions are quite large, there can be a lot of differences in the locations of the properties that belong to the same region. Therefore, another variable called `Dist` was included in the set of chosen variables. `Dist` describes the distance between the property and Melbourne's central business district in kilometers.

From table 1, it can be seen that real estate prices are very high in Melbourne. The median property price is 870,000 Australian dollars. In addition, 75 % of the properties that are included in the data set have a price of $635,000 or more. The property prices are visualized more closely with a histogram in figure 1. The figure contains histograms for `Rooms`, `Bath` and `Dist` variables, too.

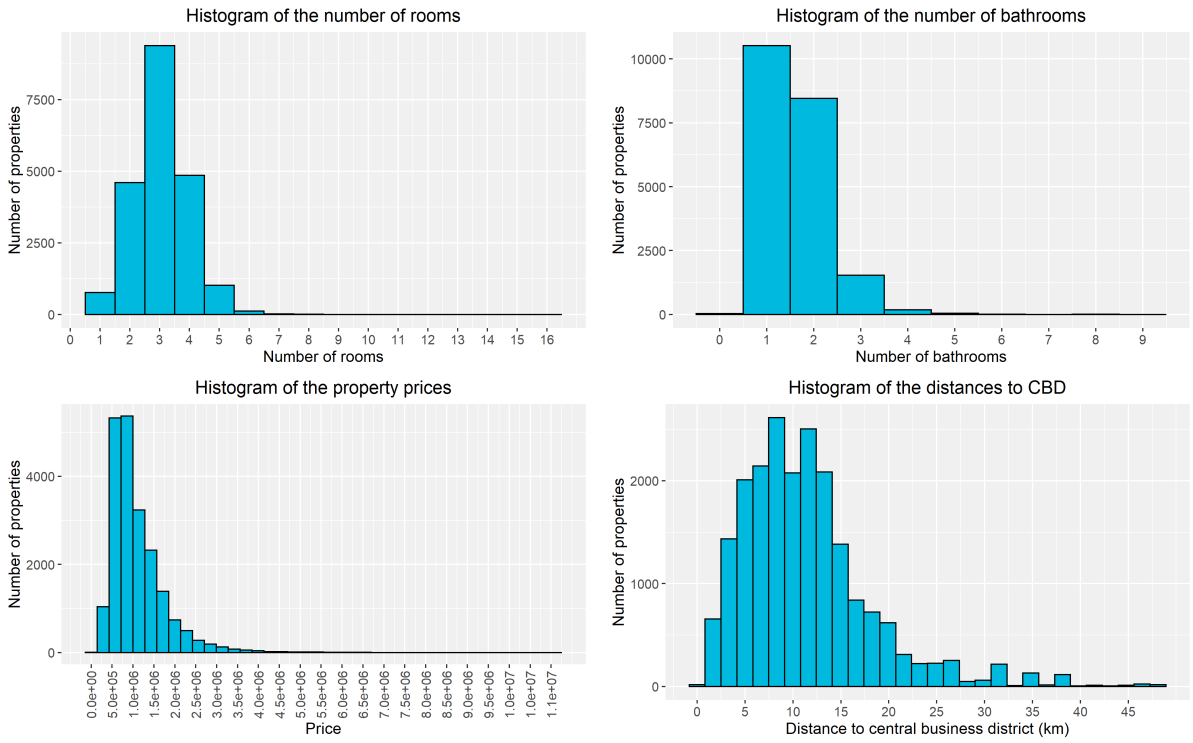

Figure 1: Histograms of some of the numeric variables.

3 Bivariate analysis

To understand which variables seem to correlate with each other, a correlation matrix was calculated. Figure 2 contains a heat map visualized from the correlation matrix. `Rooms`, `Bed` and `Bath` have the highest positive correlation with the property price. The area of the property doesn't seem to correlate with the price as much as the number of rooms, bedrooms or bathrooms. This means that developers trying to maximize their profits should consider building properties with a larger number of rooms even if the rooms were a bit smaller - at least to some extent.

Interestingly, the year of construction is the variable with the largest negative correlation with the price of a property. Usually, newer properties tend to sell for higher prices. This interesting

finding could result from the fact that when cities grow larger and larger, the best building lots are often already constructed. Therefore, the properties built on the most valuable lots could be older compared to properties built on less expensive building lots. The heat map in figure 2 seems to agree with this explanation. According to the heat map, `Dist` and `Year` correlate positively which implies that newer properties tend to be further away from the central business district where the building lots tend to be less expensive.

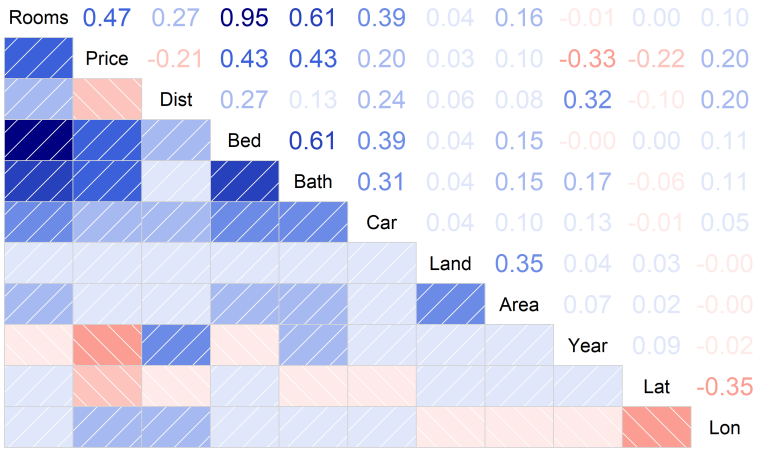

Figure 2: Correlation matrix of the numerical variables.

The distance to central business district has a negative correlation with the price, too, which means that properties closer to the central business district tend to sell for higher prices. However, the correlation is a bit weaker (-0.21) that what one could have imagined before the analysis. This implies that `Dist` doesn't seem to explain `Price` well enough when used as the only location-related variable.

It is also interesting that the size of a building lot doesn't correlate with the prices of the properties included in the data set. Usually, the larger a building lot is, the higher the price of a property is given that the properties are located in the same region. However, because the correlation matrix has been calculated from all the properties in all different regions, the size of a building lot doesn't explain the property price well on its own. In addition, when some regions become more popular and expensive, the building lots are usually divided into smaller ones which means that smaller lots can also be expensive. It is also worth noting that none of the variables seem to correlate very strongly with the price of a property. The absolute values of all the correlation coefficients are smaller than 0.5 when `Price` is paired with the other variables.

When analysing other pairs of variables that don't include `Price`, results that are perhaps more obvious can be obtained. The three pairs with the largest Pearson correlation coefficient are (`Rooms`, `Bed`), (`Rooms`, `Bath`) and (`Bed`, `Bath`). It seems quite obvious that properties with more rooms tend to have more bedrooms and bathrooms. In addition, the area of a property correlates positively with the size of the building lot. This makes sense because building larger properties require larger building lots.

4 Multivariate analysis

4.1 Method selection

Multiple correspondence analysis is a PCA type data analysis method appropriate for analysing categorical variables. Many of the variables in this data set that are of interest when trying to understand the property prices are categorical. This suggests that MCA would be a good method for analysing the data set. MCA is also suitable considering the research question about finding variables with the most dependency with the price. Therefore, multiple correspondence analysis was chosen as the main method of this project work.

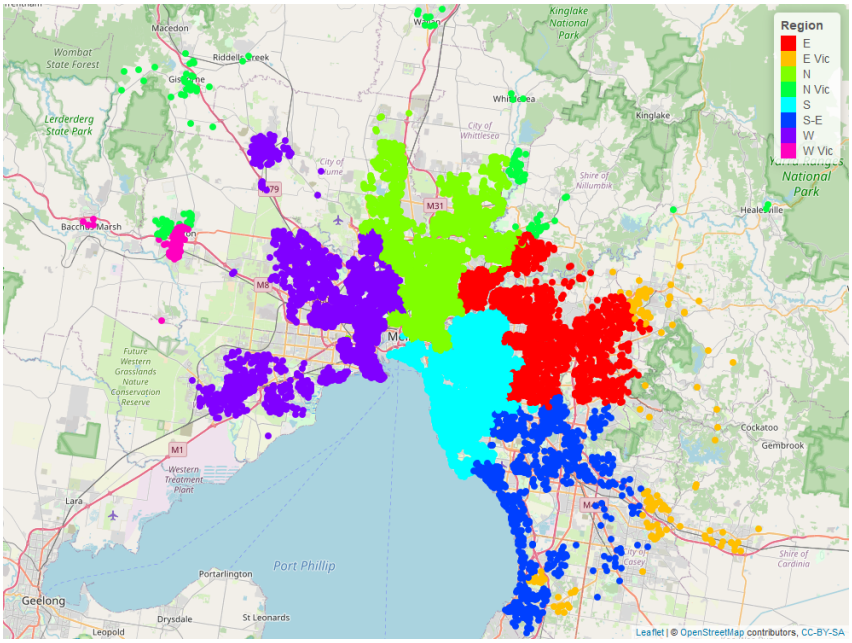

Figure 3: Properties plotted on a map.

Figure 3 contains a map on which the properties of the data set have been plotted. The color of each marker is determined by the region the property belongs to. According to the figure, some regions seem to have a lot fewer properties than others. Particularly, the regions `E Vic`, `N Vic` and `W Vic` seem to have very few properties compared to the other regions. Multiple correspondence analysis is sensitive to having categories with very different number of elements. This makes MCA a non-robust method and to avoid having bad results the smaller regions can be clustered to the larger ones with some clustering method. When regrouping some of the data points, it is important to notice that most properties have already been labeled according to the new categories. This information can be used for trying to relabel the other data points. k-nearest neighbors algorithm seems to be a good choice for this kind of a task. k-NN is a supervised machine learning method meaning that it requires the labels of the training data to be known. In the setting of regrouping some of the regions, the idea in k-NN is to perform the relabeling for each property that belong to the regions that want to be clustered to the larger regions. For each such property, k nearest properties belonging to the larger regions are found out. The new label of a property is determined by a majority vote of its k nearest neighbors.

4.2 Technical implementation

4.2.1 Choosing variables

One of the first parts of the technical implementation is to find out which variables should be included in the multiple correspondence analysis. In section 2, it was already discussed that

`Region` and `Dist` would be included in the set of chosen variables because together they determine the location of a property quite well. Using the two location-related variables helps determining how far the property is from the central business district of Melbourne and to which cardinal (or intercardinal) direction from the city centre.

When choosing between the other variables, it is also worth noting that quite a few properties have missing values for at least some of the variables. In MCA, each individual data point should belong to exactly one category per variable. Therefore, either the properties with missing values would need to be ignored completely or the missing values could be filled in with the category of similar properties. Filling in the missing values can be a very challenging task and requires expertise in the real estate market of Melbourne. Therefore, it was first analysed how many missing values each variable has to better understand the trade-off between the number of suitable variables and the number of data points that would need to be discarded.

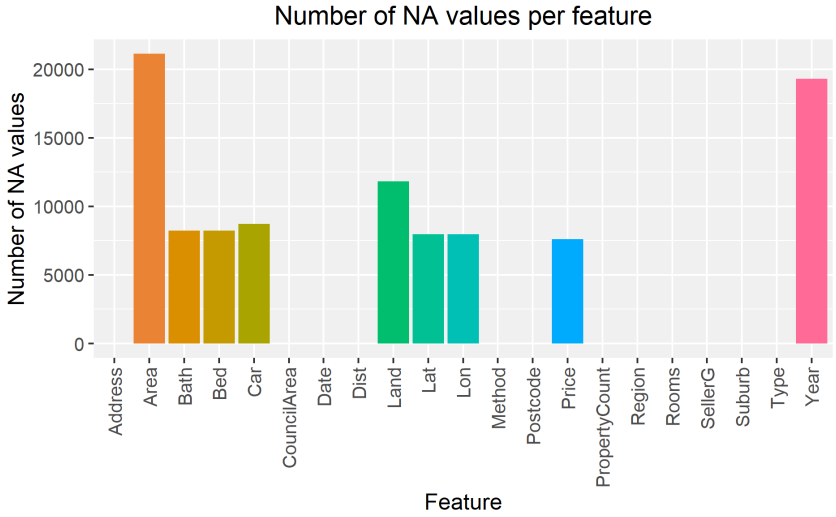

Figure 4: The number of missing values for each variable.

Figure 4 contains a bar plot of the number of missing values for each variable. Interestingly, 7,610 properties did not have a price mentioned in the data set which accounts for more than 20% of the properties. From the description of the data set on Kaggle, it can be seen that not all properties of the data set have necessarily been sold. The variable `Method` has different categories including `NB` (no bid) and `W` (withdrawn prior to auction). For some properties, the price was not disclosed. For simplicity, the properties without a price were simply ignored in the analysis because filling in the missing values would have needed much more knowledge about the real estate market of Melbourne.

Next, it can be seen from figure 4 that the variables without any missing values are mostly related to the location of the property. Of those location-related variables, `Region` and `Dist` were already chosen as the most suitable for MCA. The variable `Rooms` also contains zero missing values and it was included in the set of chosen variables, too. However, these variables are not enough for preprocessing the data set with k-NN algorithm. Therefore, the latitude and longitude are also needed. There are 7,976 properties without values for `Lat` and `Lon` in the data set. Unfortunately, there is not much overlap between these properties and the ones without a price, though. The variable `Bath` can be included in the analysis by reducing the number of data points by only 276 considering that `Lat` and `Lon` are needed in the analysis. All in all, the complete list of chosen variables is `Price`, `Region`, `Dist`, `Rooms`, `Lat`, `Lon` and `Bath`. All the properties with missing values for these variables were ignored in the analysis. The total number of remaining data points was 20,778. `Lat` and `Lon` were only used in the k-NN algorithm but dropped before MCA.

4.2.2 Preprocessing the data

The k-NN algorithm was performed on the data set by using the `knn` function in the R package called `class`. The properties belonging to the larger regions (`E`, `N`, `S`, `S-E` and `W`) were considered as training data and the properties belonging to the other regions as test data. k-NN requires that the parameter k is given to the algorithm and in a more general case, the best parameter k can be obtained by trying many different values k and choosing the one with the smallest test error. In general, a small value for k can result in overfitting and a large value for k in underfitting. However, as can be seen from figure 3, the clusters of the larger regions are quite well separated from each other and the properties needing relabeling are often close to only one or two of the larger regions. Therefore, there doesn't seem to be need for trying out large values for k. As can be seen from figure 5, the results for $k = 3$ are exactly what was wanted.

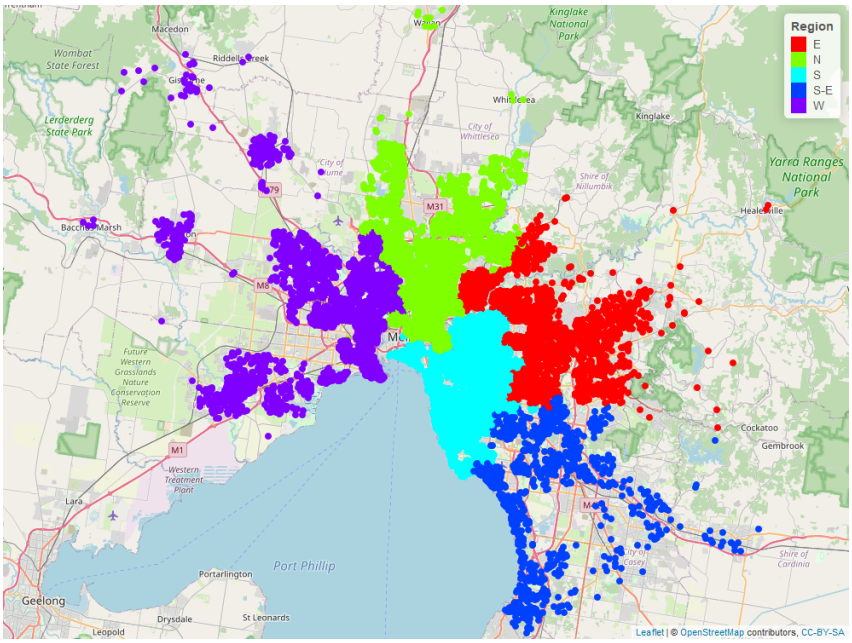

Figure 5: Properties plotted on a map after performing k-NN.

The region `S-E` contains a bit fewer properties than the other regions but it was still kept as a separate category. Some properties in `S-E` are located quite far from the other regions which means that having fewer regions would result in having too many properties located in very different locations but would still be considered belonging to the same region in the analysis.

Variable	The new categories and their ranges				
`Rooms`	1-2	3	>=4	-	-
	1 or 2	3	4 or more	-	-
`Price`	1	2	3	4	-
	< 650,000	650,000..900,000	900,000..1,350,000	>= 1,350,000	-
`Dist`	close	quite close	quite far	far	-
	< 5	5..10	10..15	>= 15	-
`Region`	E	N	S	S-E	W
	-	-	-	-	-
`Bath`	<=1	>=2	-	-	-
	at most 1	at least 2	-	-	-

Table 2: The new categories.

The other variables chosen in the analysis also need regrouping. MCA is sensitive to having variables with very different number of categories. Table 2 presents the new categories for the chosen variables. Each variable has two rows in the table: the first row describes the names of the

new categories and the second row describes the range of the variable from which all properties belong to the corresponding category. For example, all properties that are at least 5km but at most 10km away from the city centre belong to the category called `quite close`. The summary statistics presented in table 1 were utilized when regrouping the variables. The summary statistics were rounded to some sensible numbers when determining the ranges for each category.

4.2.3 Multiple correspondence analysis

After preprocessing the data, multiple correspondence analysis was performed using the `mjca` function from the R package called ca. The analysis was performed by applying correspondence analysis on the indicator matrix and, therefore, the parameter `lambda = 'indicator'` was used. The parameter `nd` was set to 6 so that the summary output contains the quality of representations for the first six dimensions.

4.3 Results

The representation of multiple correspondence analysis contains 13 dimensions in total. The first pair of dimensions explain 29.1% of the total variance in the data set. Together the dimensions 3 and 4 explain 19% and the dimensions 5 and 6 explain 16.1% of the total variance. Therefore, the first six dimensions explain 64.2% of the total variance. To explain 85% of the total variance, nine dimensions would be needed.

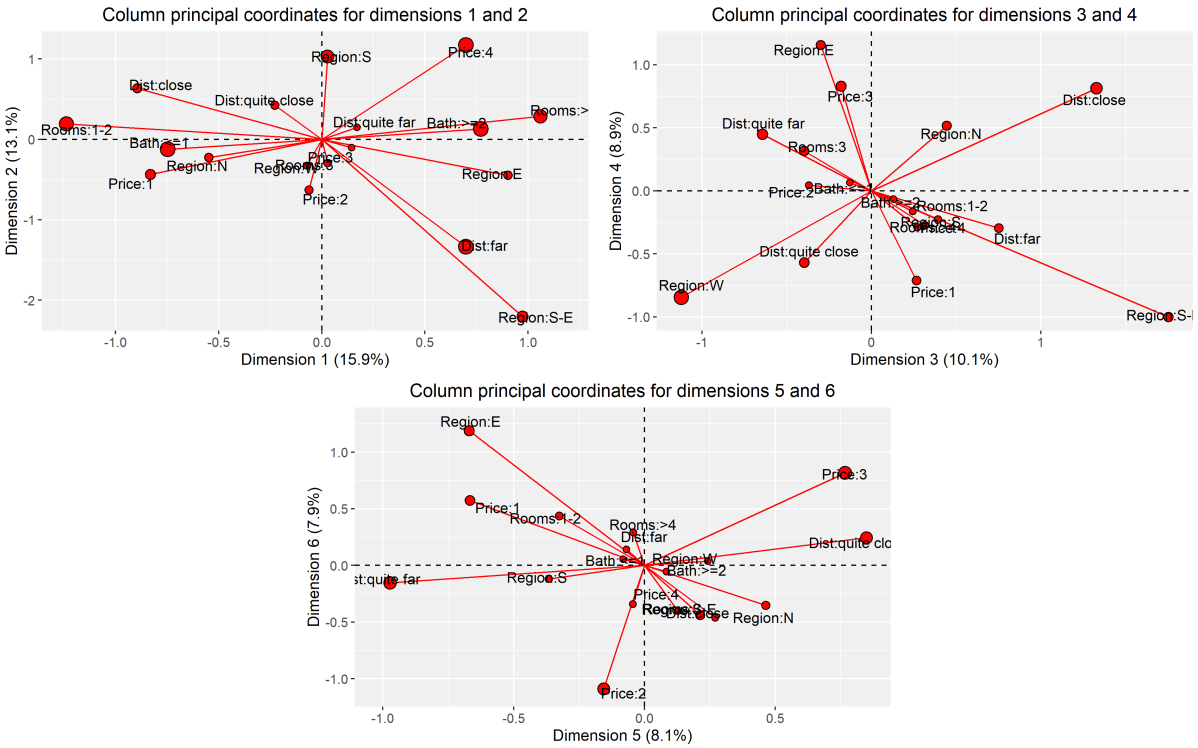

Figure 6: The column principal coordinates plotted pairwise for the first six dimensions.

Figure 6 visualizes the column principal coordinates for the first six dimensions. The size of each circle representing a category depends on the quality of representation of the category in the two dimensions of the plot. The larger the circle, the better the quality of representation is. When trying to make inferences from the plots, it is important to remember that only the categories with high enough a quality of representation in the related dimensions should be included in the analysis. Therefore, only the largest circles of each plot should be taken into consideration.

From the first plot of figure 6, it can be seen that the first dimension of the MCA representation is mostly determined by the number of rooms and bathrooms the property has. The categories `Rooms:1-2` and `Bath:<=1` on the left and `Rooms:>=4` and `Bath:>=2` on the right are all well represented. Because the angle between `Rooms:>=4` and `Bath:<=1` is greater than 90 degrees, the two categories repulse each other. In fact, the angle between the two modalities is almost 180 degrees. The angle between `Rooms:>=4` and `Bath:>=2` is smaller than 90 degrees which implies that the two categories attract each other. In fact, the angle is almost zero and, therefore, the attraction between the two modalities is very strong. A similar inference applies also for the categories `Rooms:1-2` and `Bath:<=1`.

The second dimension in the first plot of figure 6 is determined by both the location and the price of a property. The categories that stretch out the furthest away from the origin along the second dimension are `Region:S` and `Price:4` above the origin and `Dist:far` and `Region:S-E` below the origin. The angle between `Dist:far` and `Region:S-E` is small meaning that the properties in South-Eastern Metropolitan area tend to be quite far away from the city centre. On the other hand, the angle between `Region:S` and `Price:4` is around 45 degrees meaning that the properties in Southern Metropolitan area tend to be very expensive. This inference seems sensible based on the map in figure 5 because the properties in that area are located close to both the ocean and the city centre.

The interpretation of dimensions 3 and 4 is not as clear as in the case of the first two dimensions. Both dimensions seem to be composed of some combination of the location-relation variables `Region` and `Dist`. The angle between categories `Region:E` and `Price:3` is close to zero which implies that there are a lot of properties in Eastern Metropolitan area that belong to the third quartile in terms of price. In addition, the angle between modalities `Price:3` and `Dist:close` is approximately 90 degrees. From this finding, it can be inferred that having a relatively high price point does not depend on whether the property is located within 5km from the city centre or not. Moreover, the angle between the categories `Price:1` and `Dist:close` is approximately 90 degrees. This implies that having a lower price point does not seem to depend on whether the property is located within 5km from the city centre or not, either. These findings are somewhat in accordance with the inferences made based on the heat map in figure 2 in section 3. Based on the heat map, it was inferred that being further away from the city centre correlates with lower prices but the correlation might not have been as evident as one could have imagined. It is worth noting that the categories for `Price` are not as well represented in the second plot of figure 6 as the other categories that have been used for making inferences so far.

According to the third plot in figure 6, the fifth dimension represents the distance to the city centre. The categories of the variable `Dist` stretching out the furthest away from the origin in the direction of the fifth dimension are `Dist:quite far` on the left and `Dist:quite close` on the right. For this reason, the fifth dimension puts more weight on the properties that are located approximately median distance away from the city centre instead of the properties either very close or very far away from the centre. According to the plot, the modalities `Price:3` and `Dist:quite close` attract each other meaning that there is dependency between having a relatively high price and being located between 5 to 10 km away from the city centre. It was previously discussed that the price of a property doesn't seem to depend as much on being located within 5km away from the city centre as one could have imagined. However, the third plot in figure 6 suggests that relatively expensive properties still tend to be located between 5 to 10 km away from the centre. This could result from the following reasoning: there are usually a lot of smaller studio and one-bedroom apartments located very close to the city centre because the price per square meter is high. In addition, there are often a smaller number of very expensive properties located very close to the city centre with a large building area. These two reasons can affect to dependency between the different price categories and the category `Dist:close`. There tends

to be more detached houses a bit further away from the city centre, though, and the properties tend to be larger there. However, the properties belonging to the category `Dist:quite close` are still only between 5 to 10 km away from the centre meaning that people could be willing to pay extra money for a larger house from which it is still easy and fast to commute to work. To verify how correct this analysis is, the property type would needed to be included in the multiple correspondence analysis, though. However, the variable `Type` could not be included in MCA as such because too many different types of properties have been included in the same categories. For example, 69% of properties in the data set have a property type of house but there is no additional information whether the house is a detached, semi-detached or terraced house or a cottage or a villa.

5 Evaluation

The results of multiple correspondence analysis seem good but the fact that some categories had quite a small quality of representation made the analysis a bit more challenging. This could partly result from the fact that dropping quite a few data points from the analysis made it more difficult to explain the variance of the data. Therefore, more data points from a longer time period could be collected. Moreover, some properties did not contain a price in the data set because the price had not been disclosed on the website from which the data had been collected. The author of the data set had collected the data from the website of Domain which is a website that lists properties that are for sale and rent in Australia. The price information could perhaps be collected from a real estate register provided by the Australian authorities. It would also be interesting to study the effect of the property type on the price which would require collecting more detailed information about the property type than in the current data set.