REAL ESTATE DATA ANALYSIS: MELBOURNE
report

Andrea Nappi andi.nappi@gmail.com
08-01-2024

SCOPE OF THE PROJECT:

The scope of this study encompasses a comprehensive analysis of the Melbourne real estate market, as if it were a hypothetical consulting commission coming from a real estate business, with focus on understanding the primary determinants of property prices, identifying property groups with similar features, and uncovering potential hidden insights. To simulate a real case scenario, it has been chosen a raw, unknown dataset, where the most suitable procedures for the data are unknown. The analysis will particularly emphasize property types price differences and geographical regions based properties distinction.
In doing so, the study aims to achieve the following:

- **Identify Key Features**: Determine the key features that significantly influence property prices, providing valuable insights for buying and selling.
- **Segmentation of Property Groups**: Employ clustering techniques to identify distinct groups of properties that share common characteristics. This segmentation can enhance the understanding of market trends and preferences within specific property categories.
- **Typological Analysis**: Explore the impact of property features on property prices. This information can be crucial for both buyers and sellers seeking to understand the dynamics of different house types.
- **Predictive Modeling**: Develop predictive models for property prices, allowing for a nuanced understanding of how various features contribute to overall valuation. This predictive approach can aid in making informed decisions regarding property investments.
- **Visualization of Results**: Present the findings in visually compelling formats such as charts, graphs, and maps to facilitate a clear understanding of complex patterns and trends.

The data:

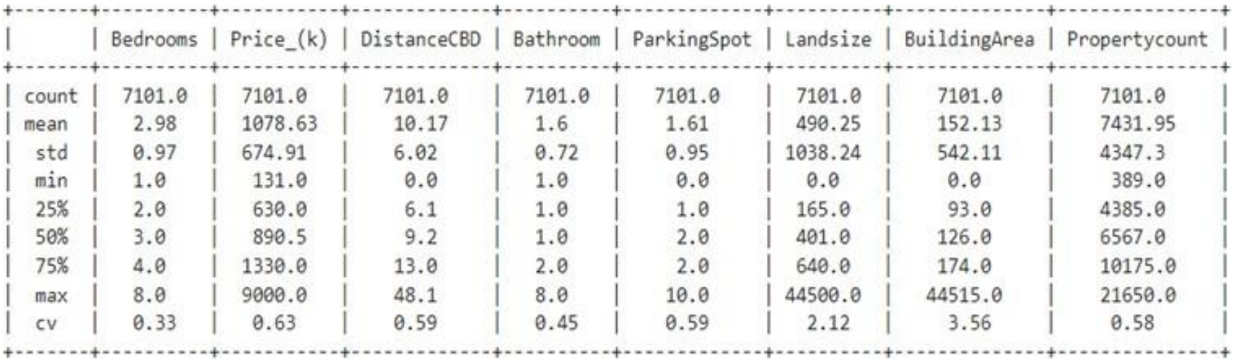
The dataset used is the “Melbourne Housing Snapshot” from Kaggle. It was scraped from publicly available results posted every week on Domain.com.au and contains information about over 14,000 real estate properties in Melbourne.

DATA EXPLORATION AND CLEANING:

The data has been explored and cleaned, checking for potential duplicates that should have been removed, adjusting some house features’ types to make them more appropriate, checking for missing data and dealing with them according to what is needed, removing some features, renaming some and editing others; also, some properties have been taken out of the data because of their extremely atypical characteristics, that would have probably been obstacles for the sake of this elaborate. The processed data, which is the one that has then been used for the project, ended up containing 6008 properties and 16 features for each property, the following:

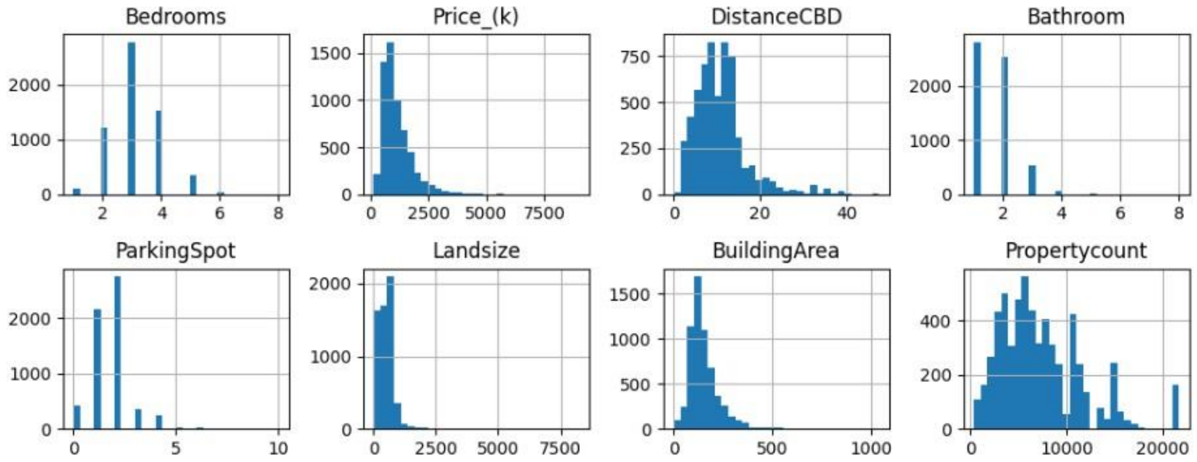
- **Suburb**: Suburb the property belongs to.
- **Address**: Address of the property.
- **Bedrooms**: Number of bedrooms in the property.
- **HouseType**: Type of property: h - house, cottage, villa, semi, terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.
- **Price_(k)**: Price of the property (thousands of $).
- **SaleMethod**: method of sale of the property: 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 before auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.
- **SellerG**: Real estate agent associated to the sale of the property.
- **DateOfSale**: Date of sale of the property.
- **DistanceCBD**: Distance of the property from the CBD (Central Business District).
- **Postcode**: Postcode of the property.
- **Bathroom**: Number of bathrooms in the property.
- **ParkingSpot**: Number of car spots in the property.
- **Landsize**: Land size of the property (m2).
- **BuildingArea**: Building area size of the property (m2).
- **Regionname**: General Region the property is in.
- **Propertycount**: Number of properties that exist in the suburb.

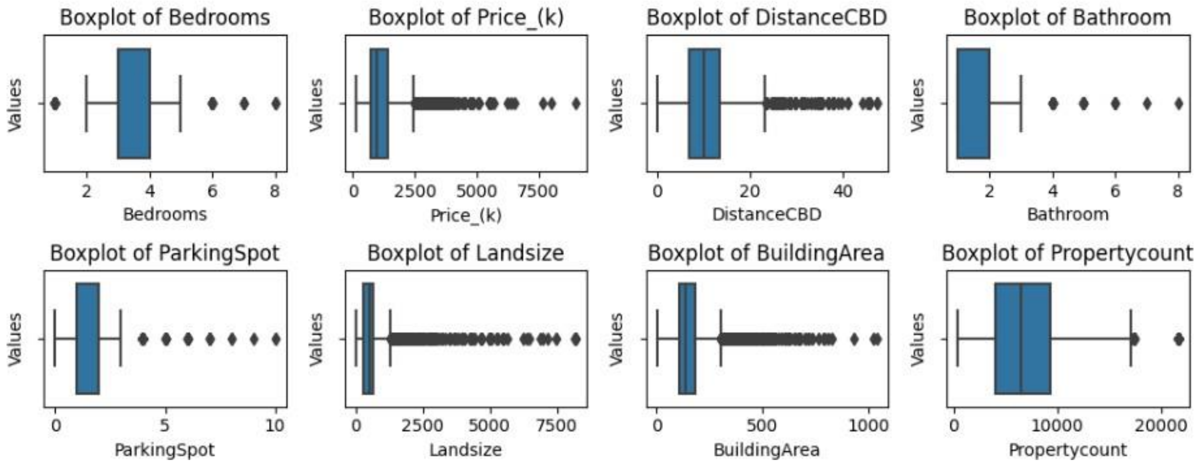
DESCRIPTIVE STATISTICS AND VISUALIZATIONS:



```
+-------+----------+----------+-------------+----------+-------------+----------+--------------+---------------+
|       | Bedrooms | Price_(k) | DistanceCBD | Bathroom | ParkingSpot | Landsize | BuildingArea | Propertycount |
+-------+----------+----------+-------------+----------+-------------+----------+--------------+---------------+
| count |  7101.0  |  7101.0  |    7101.0   |  7101.0  |    7101.0   |  7101.0  |    7101.0    |     7101.0    |
| mean  |   2.98   | 1078.63  |    10.17    |   1.6    |     1.61    |  490.25  |    152.13    |    7431.95    |
|  std  |   0.97   |  674.91  |     6.02    |   0.72   |     0.95    | 1038.24  |    542.11    |     4347.3    |
|  min  |   1.0    |  131.0   |     0.0     |   1.0    |     0.0     |   0.0    |     0.0      |     389.0     |
|  25%  |   2.0    |  630.0   |     6.1     |   1.0    |     1.0     |  165.0   |     93.0     |     4385.0    |
|  50%  |   3.0    |  890.5   |     9.2     |   1.0    |     2.0     |  401.0   |    126.0     |     6567.0    |
|  75%  |   4.0    |  1330.0  |     13.0    |   2.0    |     2.0     |  640.0   |    174.0     |    10175.0    |
|  max  |   8.0    |  9000.0  |     48.1    |   8.0    |     10.0    | 44500.0  |   44515.0    |    21650.0    |
|  cv   |   0.33   |   0.63   |     0.59    |   0.45   |     0.59    |   2.12   |     3.56     |      0.58     |
+-------+----------+----------+-------------+----------+-------------+----------+--------------+---------------+
```

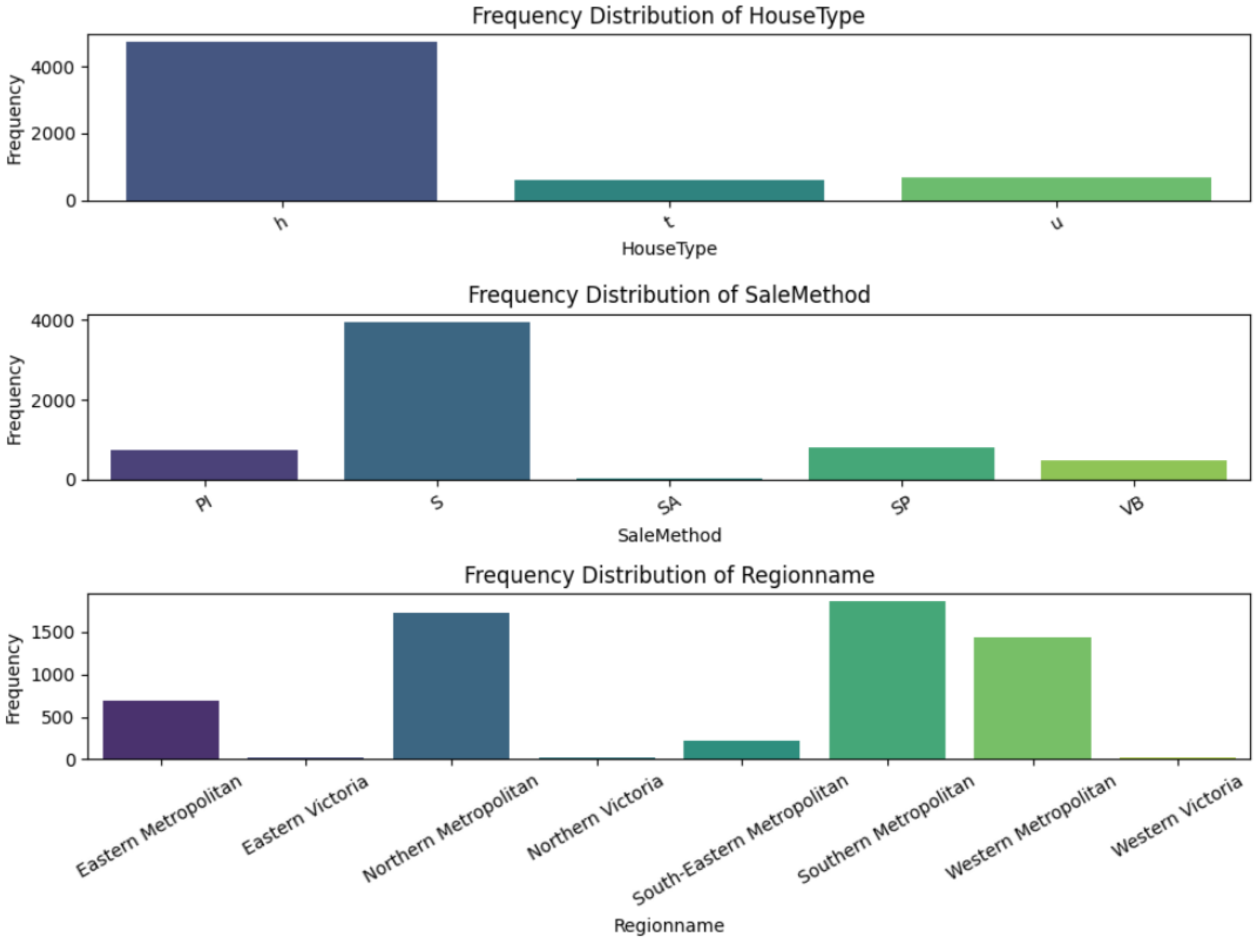
Table 1: *Descriptive statistics for numeric property features*





Picture 1: Histograms of all numeric property features
Picture 2: Boxplots of all numeric property features

As Table 1 indicates and Picture 1 shows, the distributions of the property features exhibit a right-skewed pattern, suggesting that a majority of properties have relatively typical characteristics, but there is a subset of elements presenting atypical features' values, that elevate the mean values in comparison to the median. As Picture 2 confirms, this means there are a few standout properties among a big group of residences with typical characteristics. The price distribution is among the most heavily skewed ones; this might be reflected in additional challenges during the price prediction section.

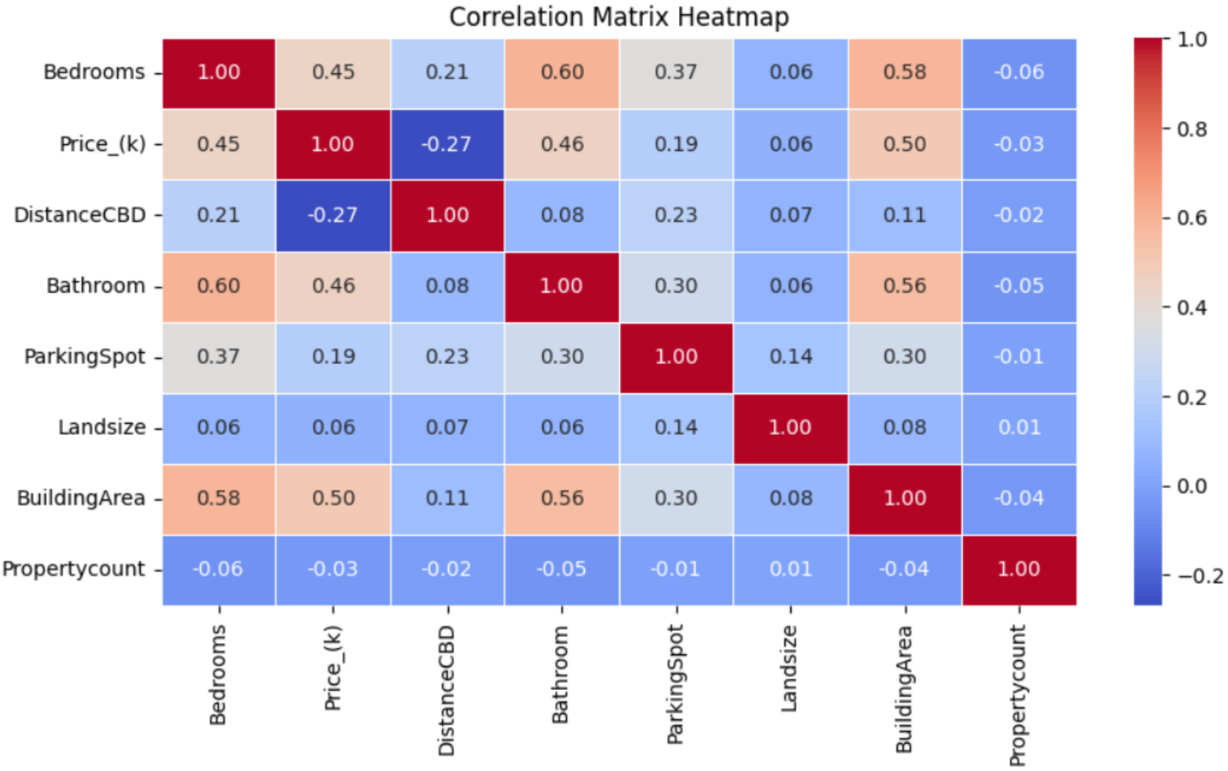


Picture 3: Frequency distribution for categorical property features

Picture 3 shows that the prevalent property type is the house (h), surpassing the combined count of the other two types (townhouses and units).

Regular sales (S) dominate as the most frequent sales type, exceeding the total count of all other sale types.

The Southern Metropolitan and Northern Metropolitan regions stand out as the two most represented in the available data, counting more than 1500 residences each, with the first one having more than the second one. Western Metropolitan counts more than 1000, Eastern and South-Eastern have a few hundred each.
All Victoria regions are heavily under-represented, each one counting less than 40 properties.

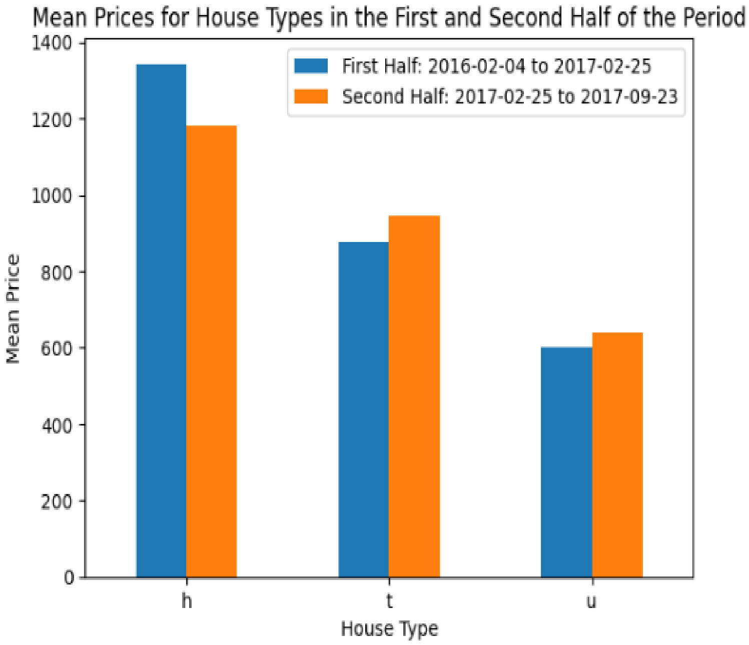


Picture 4: Correlation matrix

The correlation matrix is a tool that indicates what pairs of features tend to change together, and with what intensity from -1 to 1, -1 being when one grows the other one raises and vice-versa, 0 being they do not vary together at all, and 1 being when one grows, the other one grows as well (correlation matrix does not imply causation in any way).

The correlation matrix in Picture 4 indicates no strong negative connections between any variable pairs, the strongest one being the one between the price and the distance from the CBD (-0.27). Notably, features number of bedrooms, building area, and bathrooms show a tendency to higher correlations compared to the others. In particular, the strongest associations are between the number of bedrooms and both the building area (0.58) and the number of bathrooms (0.60). This indicates that larger properties tend to have more bedrooms, and those with more bedrooms often have more bathrooms, which aligns with common expectations.

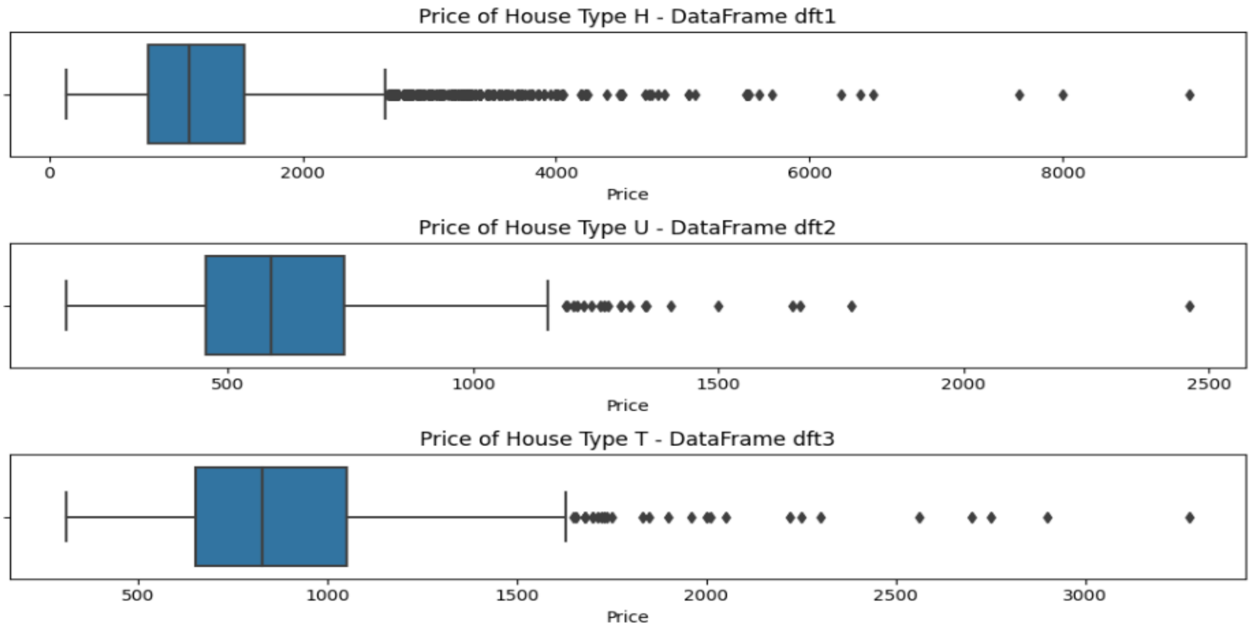
Insights on house types:



Picture 5: Mean price per property type over time

Picture 5 shows how properties classified as "house" (and similar) types exhibit the highest average price over both periods, followed by townhouses (and similar), and then by units and duplexes.
Notably, houses stand out as the only category experiencing a decline in prices during the second half of the sales period, with a mean price drop of over 100.000 $.

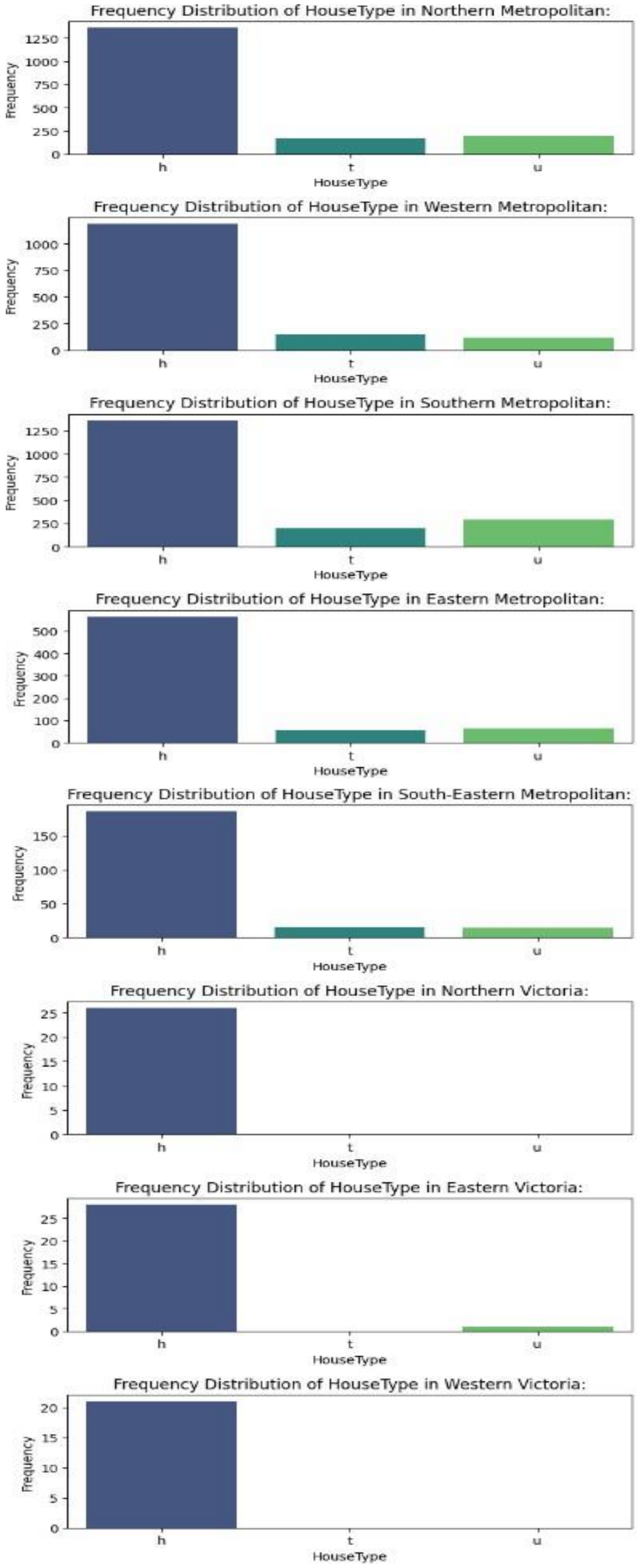
(Note: Halves are determined by the number of properties sold, not by days.)

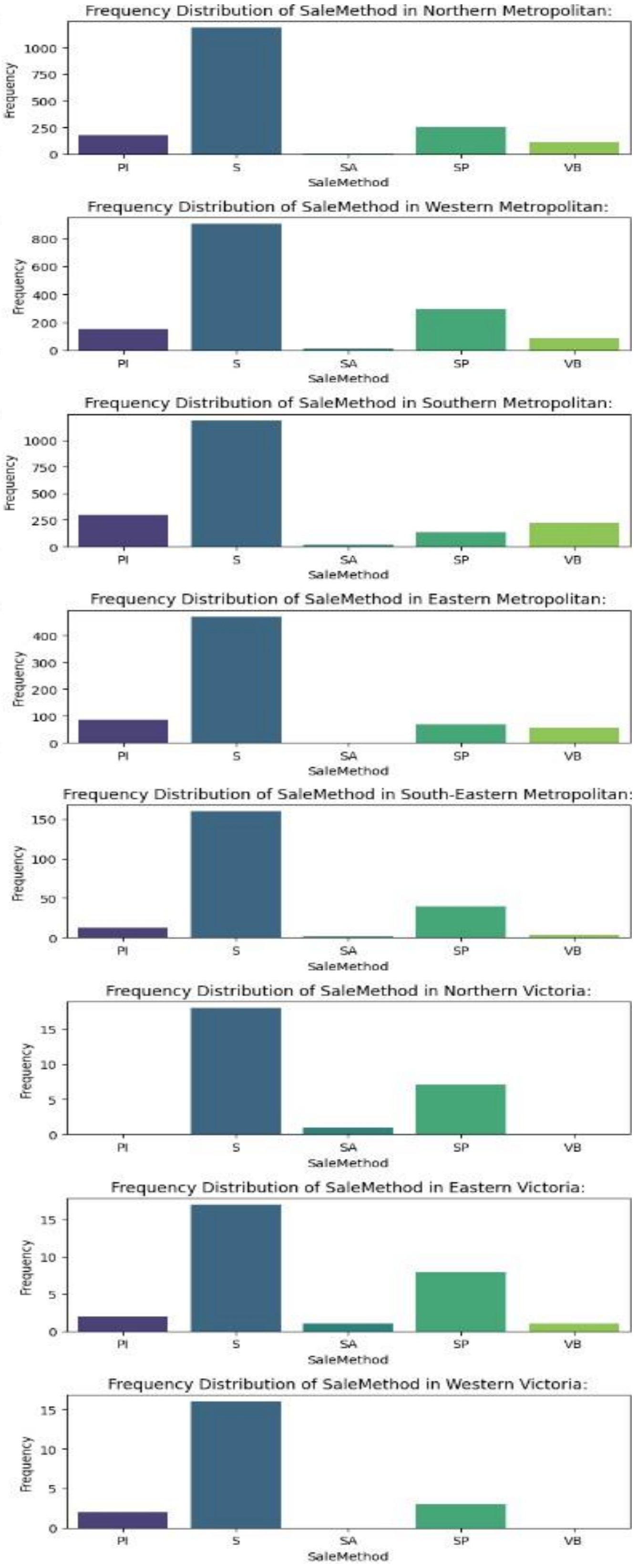


Picture 6: Price boxplot per property type

Picture 6, showing the distributions of property prices based on house types, unveils how the type group presenting the most unusual property is houses, followed by townhouses, and then by units.

Insights on Melbourne regions:



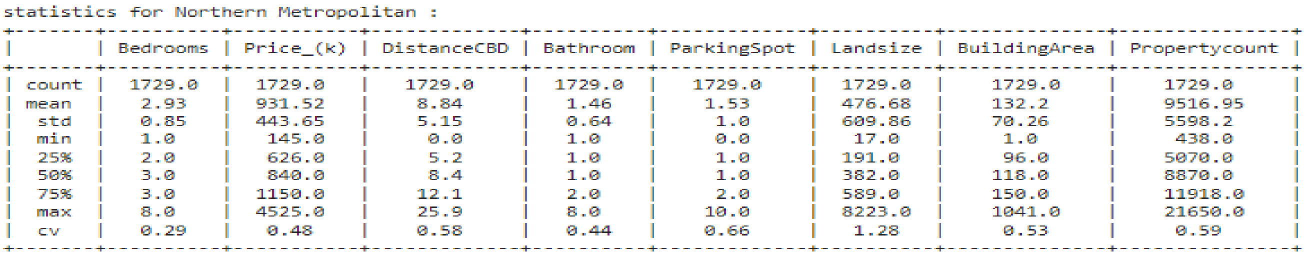


Picture 7: Property type per region

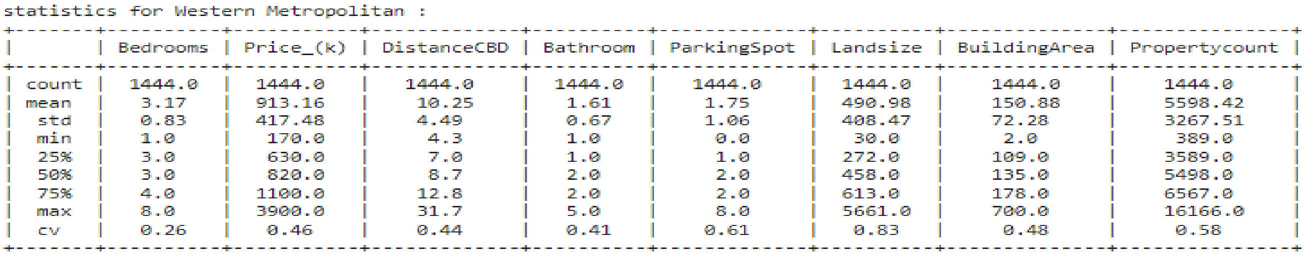
Picture 8: Sale type per region

Picture 7 represents the number of properties per region divided by type, it suggests that all regions have similar kinds of quantitative distributions on this matter.
House type is the most dominant everywhere, and townhouses and units are fairly represented as minor quotas in all regions, with exceptions made for the 3 least represented regions (all Victoria regions count less than 30 properties), where they are virtually absent.

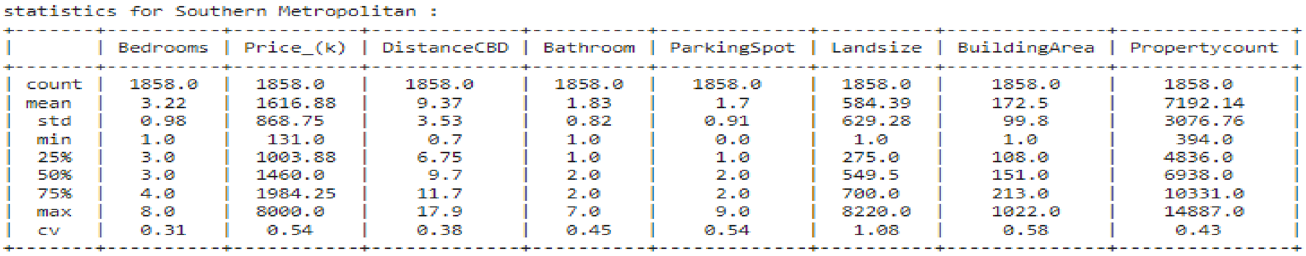
The sale method does not seem to be changing much depending on the region of the property, and it too follows the distribution of the general dataset.



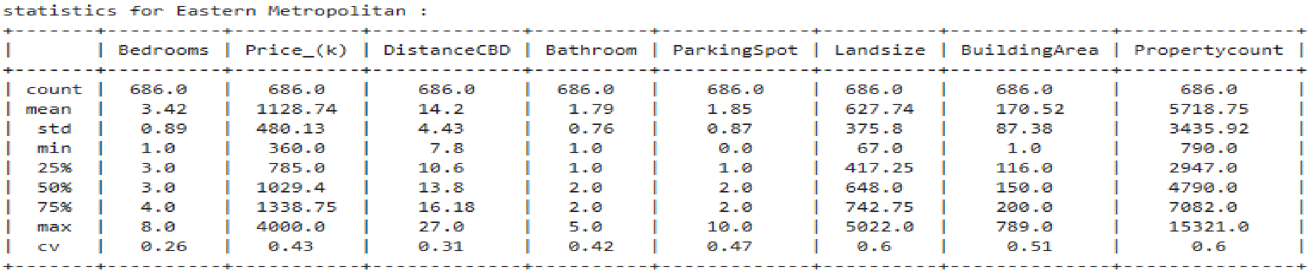
```
statistics for Northern Metropolitan :
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
|       | Bedrooms | Price_(k) | DistanceCBD | Bathroom | ParkingSpot | Landsize | BuildingArea | Propertycount |
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
| count |  1729.0  |  1729.0   |   1729.0    |  1729.0  |   1729.0    |  1729.0  |    1729.0    |    1729.0     |
| mean  |   2.93   |  931.52   |    8.84     |   1.46   |    1.53     |  476.68  |    132.2     |    9516.95    |
|  std  |   0.85   |  443.65   |    5.15     |   0.64   |     1.0     |  609.86  |    70.26     |    5598.2     |
|  min  |   1.0    |   145.0   |     0.0     |   1.0    |     0.0     |   17.0   |     1.0      |     438.0     |
|  25%  |   2.0    |   626.0   |     5.2     |   1.0    |     1.0     |  191.0   |     96.0     |    5070.0     |
|  50%  |   3.0    |   840.0   |     8.4     |   1.0    |     1.0     |  382.0   |    118.0     |    8870.0     |
|  75%  |   3.0    |  1150.0   |    12.1     |   2.0    |     2.0     |  589.0   |    150.0     |    11918.0    |
|  max  |   8.0    |  4525.0   |    25.9     |   8.0    |    10.0     |  8223.0  |    1041.0    |    21650.0    |
|  cv   |   0.29   |   0.48    |    0.58     |   0.44   |    0.66     |   1.28   |     0.53     |     0.59      |
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
```



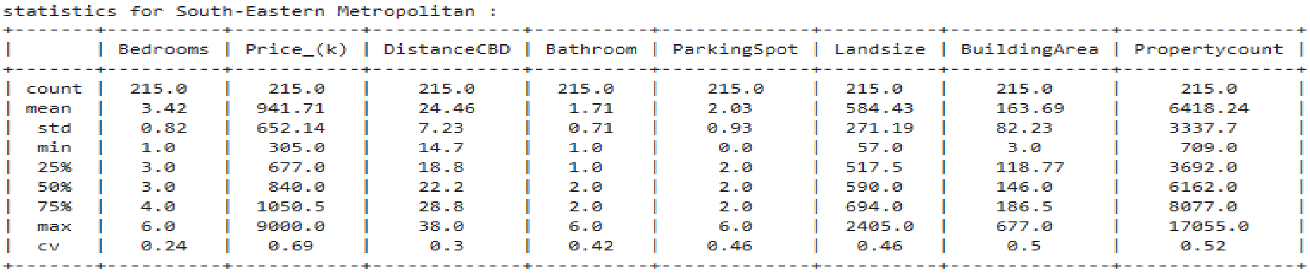
```
statistics for Western Metropolitan :
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
|       | Bedrooms | Price_(k) | DistanceCBD | Bathroom | ParkingSpot | Landsize | BuildingArea | Propertycount |
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
| count |  1444.0  |  1444.0   |   1444.0    |  1444.0  |   1444.0    |  1444.0  |    1444.0    |    1444.0     |
| mean  |   3.17   |  913.16   |    10.25    |   1.61   |    1.75     |  490.98  |    150.88    |    5598.42    |
|  std  |   0.83   |  417.48   |    4.49     |   0.67   |    1.06     |  408.47  |    72.28     |    3267.51    |
|  min  |   1.0    |   170.0   |     4.3     |   1.0    |     0.0     |   30.0   |     2.0      |     389.0     |
|  25%  |   3.0    |   630.0   |     7.0     |   1.0    |     1.0     |  272.0   |    109.0     |    3589.0     |
|  50%  |   3.0    |   820.0   |     8.7     |   2.0    |     2.0     |  458.0   |    135.0     |    5498.0     |
|  75%  |   4.0    |  1100.0   |    12.8     |   2.0    |     2.0     |  613.0   |    178.0     |    6567.0     |
|  max  |   8.0    |  3900.0   |    31.7     |   5.0    |     8.0     |  5661.0  |    700.0     |    16166.0    |
|  cv   |   0.26   |   0.46    |    0.44     |   0.41   |    0.61     |   0.83   |     0.48     |     0.58      |
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
```



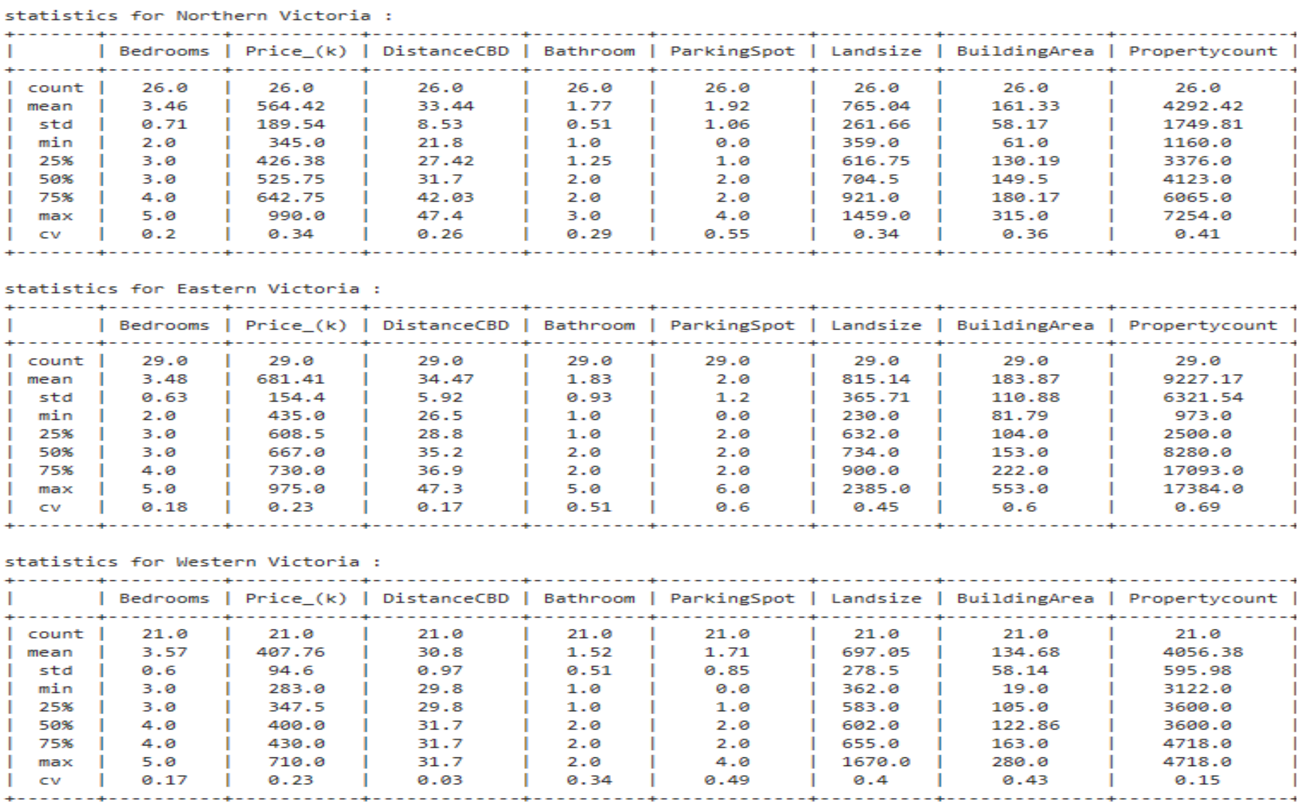
```
statistics for Southern Metropolitan :
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
|       | Bedrooms | Price_(k) | DistanceCBD | Bathroom | ParkingSpot | Landsize | BuildingArea | Propertycount |
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
| count |  1858.0  |  1858.0   |   1858.0    |  1858.0  |   1858.0    |  1858.0  |    1858.0    |    1858.0     |
| mean  |   3.22   |  1616.88  |    9.37     |   1.83   |     1.7     |  584.39  |    172.5     |    7192.14    |
|  std  |   0.98   |  868.75   |    3.53     |   0.82   |    0.91     |  629.28  |     99.8     |    3076.76    |
|  min  |   1.0    |   131.0   |     0.7     |   1.0    |     0.0     |   1.0    |     1.0      |     394.0     |
|  25%  |   3.0    |  1003.88  |    6.75     |   1.0    |     1.0     |  275.0   |    108.0     |    4836.0     |
|  50%  |   3.0    |  1460.0   |     9.7     |   2.0    |     2.0     |  549.5   |    151.0     |    6938.0     |
|  75%  |   4.0    |  1984.25  |    11.7     |   2.0    |     2.0     |  700.0   |    213.0     |    10331.0    |
|  max  |   8.0    |  8000.0   |    17.9     |   7.0    |     9.0     |  8220.0  |    1022.0    |    14887.0    |
|  cv   |   0.31   |   0.54    |    0.38     |   0.45   |    0.54     |   1.08   |     0.58     |     0.43      |
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
```



```
statistics for Eastern Metropolitan :
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
|       | Bedrooms | Price_(k) | DistanceCBD | Bathroom | ParkingSpot | Landsize | BuildingArea | Propertycount |
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
| count |  686.0   |   686.0   |    686.0    |  686.0   |    686.0    |  686.0   |    686.0     |     686.0     |
| mean  |   3.42   |  1128.74  |    14.2     |   1.79   |    1.85     |  627.74  |    170.52    |    5718.75    |
|  std  |   0.89   |  480.13   |    4.43     |   0.76   |    0.87     |  375.8   |    87.38     |    3435.92    |
|  min  |   1.0    |   360.0   |     7.8     |   1.0    |     0.0     |   67.0   |     1.0      |     790.0     |
|  25%  |   3.0    |   785.0   |    10.6     |   1.0    |     1.0     |  417.25  |    116.0     |    2947.0     |
|  50%  |   3.0    |  1029.4   |    13.8     |   2.0    |     2.0     |  648.0   |    150.0     |    4790.0     |
|  75%  |   4.0    |  1338.75  |    16.18    |   2.0    |     2.0     |  742.75  |    200.0     |    7082.0     |
|  max  |   8.0    |  4000.0   |    27.0     |   5.0    |    10.0     |  5022.0  |    789.0     |    15321.0    |
|  cv   |   0.26   |   0.43    |    0.31     |   0.42   |    0.47     |   0.6    |     0.51     |      0.6      |
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
```



```
statistics for South-Eastern Metropolitan :
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
|       | Bedrooms | Price_(k) | DistanceCBD | Bathroom | ParkingSpot | Landsize | BuildingArea | Propertycount |
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
| count |  215.0   |   215.0   |    215.0    |  215.0   |    215.0    |  215.0   |    215.0     |     215.0     |
| mean  |   3.42   |  941.71   |    24.46    |   1.71   |    2.03     |  584.43  |    163.69    |    6418.24    |
|  std  |   0.82   |  652.14   |    7.23     |   0.71   |    0.93     |  271.19  |    82.23     |    3337.7     |
|  min  |   1.0    |   305.0   |    14.7     |   1.0    |     0.0     |   57.0   |     3.0      |     709.0     |
|  25%  |   3.0    |   677.0   |    18.8     |   1.0    |     2.0     |  517.5   |    118.77    |    3692.0     |
|  50%  |   3.0    |   840.0   |    22.2     |   2.0    |     2.0     |  590.0   |    146.0     |    6162.0     |
|  75%  |   4.0    |  1050.5   |    28.8     |   2.0    |     2.0     |  694.0   |    186.5     |    8077.0     |
|  max  |   6.0    |  9000.0   |    38.0     |   6.0    |     6.0     |  2405.0  |    677.0     |    17055.0    |
|  cv   |   0.24   |   0.69    |     0.3     |   0.42   |    0.46     |   0.46   |     0.5      |     0.52      |
+-------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
```



```
statistics for Northern Victoria :
+--------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
|        | Bedrooms | Price_(k) | DistanceCBD | Bathroom | ParkingSpot | Landsize | BuildingArea | Propertycount |
+--------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
| count  |   26.0   |  26.0     |    26.0     |   26.0   |    26.0     |  26.0    |    26.0      |    26.0       |
| mean   |   3.46   |  564.42   |    33.44    |   1.77   |    1.92     |  765.04  |    161.33    |    4292.42    |
|  std   |   0.71   |  189.54   |    8.53     |   0.51   |    1.06     |  261.66  |    58.17     |    1749.81    |
|  min   |   2.0    |  345.0    |    21.8     |   1.0    |    0.0      |  359.0   |    61.0      |    1160.0     |
|  25%   |   3.0    |  426.38   |    27.42    |   1.25   |    1.0      |  616.75  |    130.19    |    3376.0     |
|  50%   |   3.0    |  525.75   |    31.7     |   2.0    |    2.0      |  704.5   |    149.5     |    4123.0     |
|  75%   |   4.0    |  642.75   |    42.03    |   2.0    |    2.0      |  921.0   |    180.17    |    6065.0     |
|  max   |   5.0    |  990.0    |    47.4     |   3.0    |    4.0      |  1459.0  |    315.0     |    7254.0     |
|  cv    |   0.2    |  0.34     |    0.26     |   0.29   |    0.55     |  0.34    |    0.36      |    0.41       |
+--------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+

statistics for Eastern Victoria :
+--------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
|        | Bedrooms | Price_(k) | DistanceCBD | Bathroom | ParkingSpot | Landsize | BuildingArea | Propertycount |
+--------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
| count  |   29.0   |  29.0     |    29.0     |   29.0   |    29.0     |  29.0    |    29.0      |    29.0       |
| mean   |   3.48   |  681.41   |    34.47    |   1.83   |    2.0      |  815.14  |    183.87    |    9227.17    |
|  std   |   0.63   |  154.4    |    5.92     |   0.93   |    1.2      |  365.71  |    110.88    |    6321.54    |
|  min   |   2.0    |  435.0    |    26.5     |   1.0    |    0.0      |  230.0   |    81.79     |    973.0      |
|  25%   |   3.0    |  608.5    |    28.8     |   1.0    |    2.0      |  632.0   |    104.0     |    2500.0     |
|  50%   |   3.0    |  667.0    |    35.2     |   2.0    |    2.0      |  734.0   |    153.0     |    8280.0     |
|  75%   |   4.0    |  730.0    |    36.9     |   2.0    |    2.0      |  900.0   |    222.0     |    17093.0    |
|  max   |   5.0    |  975.0    |    47.3     |   5.0    |    6.0      |  2385.0  |    553.0     |    17384.0    |
|  cv    |   0.18   |  0.23     |    0.17     |   0.51   |    0.6      |  0.45    |    0.6       |    0.69       |
+--------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+

statistics for Western Victoria :
+--------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
|        | Bedrooms | Price_(k) | DistanceCBD | Bathroom | ParkingSpot | Landsize | BuildingArea | Propertycount |
+--------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
| count  |   21.0   |  21.0     |    21.0     |   21.0   |    21.0     |  21.0    |    21.0      |    21.0       |
| mean   |   3.57   |  407.76   |    30.8     |   1.52   |    1.71     |  697.05  |    134.68    |    4056.38    |
|  std   |   0.6    |  94.6     |    0.97     |   0.51   |    0.85     |  278.5   |    58.14     |    595.98     |
|  min   |   3.0    |  283.0    |    29.8     |   1.0    |    0.0      |  362.0   |    19.0      |    3122.0     |
|  25%   |   3.0    |  347.5    |    29.8     |   1.0    |    1.0      |  583.0   |    105.0     |    3600.0     |
|  50%   |   4.0    |  400.0    |    31.7     |   2.0    |    2.0      |  602.0   |    122.86    |    3600.0     |
|  75%   |   4.0    |  430.0    |    31.7     |   2.0    |    2.0      |  655.0   |    163.0     |    4718.0     |
|  max   |   5.0    |  710.0    |    31.7     |   2.0    |    4.0      |  1670.0  |    280.0     |    4718.0     |
|  cv    |   0.17   |  0.23     |    0.03     |   0.34   |    0.49     |  0.4     |    0.43      |    0.15       |
+--------+----------+-----------+-------------+----------+-------------+----------+--------------+---------------+
```

Table 2: Basic statistics per region

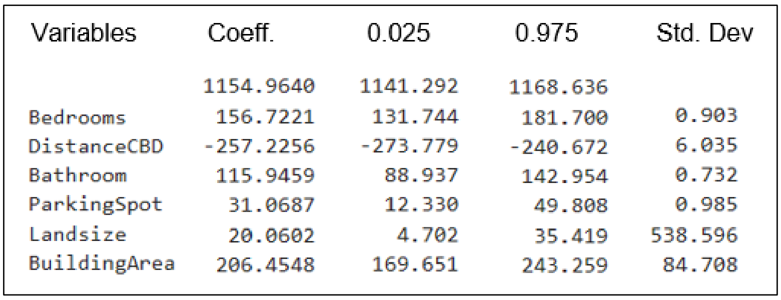
- **Northern Metropolitan** properties have a moderate number of bedrooms and bathrooms, yet with a balanced distribution of land size and building area. The mean price is slightly lower than the overall properties' mean price.
- **Western Metropolitan** region properties stand out for having the highest mean property count, suggesting it may represent properties in more populated areas. The average price is moderate and so is the building size.
- **Southern Metropolitan** properties are higher-end properties with larger land sizes and building areas. The average price is the highest among the regions, indicating a focus on premium real estate.
- **Eastern Metropolitan** properties tend to have relatively high average number of bedrooms, bathrooms, and parking spots and, indeed, larger land size and building areas, suggesting more spacious accommodations. Though, prices are in line with the overall properties, this might be linked with a slightly higher mean distance from the CBD.
- **South-Eastern Metropolitan** region properties have a high average distance to the CBD, indicating suburban or more distant locations. The average price is moderate, yet the properties have a reasonable number of bedrooms and bathrooms.
- **Southern Victoria** region is underrepresented in these data, therefor information coming from this analysis might not be scaled properly to the neighborhood. Properties with larger land sizes and building areas are included. The average distance to the CBD is the highest among the regions, indicating more remote locations.
- **Eastern Victoria** properties have a high average land size and building area, suggesting larger accommodations. The average distance to the CBD is also high. The region is underrepresented.
- **Western Victoria** region properties include houses with a moderate number of bedrooms and bathrooms. The average distance to the CBD is relatively high, similar to Southern Victoria and Eastern Victoria. Again, the number of properties in this region is very small suggesting that the whole Victoria region is heavily underrepresented.

PRICE PREDICTION:

LINEAR REGRESSION PREDICTION:

The initial step in developing a predictive analytical tool for property prices involves employing a multiple linear regression model. This model considers the linear connections among various features to predict property prices based on the values of those specific features.

Overall price prediction model:



Variables	Coeff.	0.025	0.975	Std. Dev
	1154.9640	1141.292	1168.636	
Bedrooms	156.7221	131.744	181.700	0.903
DistanceCBD	-257.2256	-273.779	-240.672	6.035
Bathroom	115.9459	88.937	142.954	0.732
ParkingSpot	31.0687	12.330	49.808	0.985
Landsize	20.0602	4.702	35.419	538.596
BuildingArea	206.4548	169.651	243.259	84.708

Table 3: Overall price prediction model information

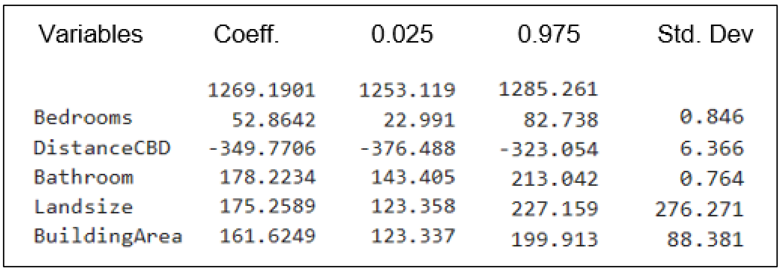
The overall price prediction model elucidates approximately 44% of the variance in Melbourne property prices. While this level of explanatory power might be considered acceptable, it does fall short of being considered great. According to the model, property prices are estimated based on seven values: six out of the seven property features (the eight is the one that we want to predict) and a constant term.

The number of properties in the suburb has been taken out of the model as it turned out to be not very informative (non-significant). Positive correlations exist between the price of a property and the number of bedrooms, bathrooms, parking spots, land size, and building area. This implies that an increase in any of these features is associated with a rise in property price and, of course, the other way around too. Conversely, distance from the CBD demonstrates a negative correlation; hence, as the distance from the CBD increases, the property price tends to decrease and vice versa.

The constant is the first coefficient value in Table 3, and it is an indication of what is the model esteem of the model for a property with a mean value for all features (1155k$ in this case). The following coefficients are those for significant features and can be interpreted as the estimated change of the estimated price (thousands of $) of a property, following a single standard deviation increase/decrease in the considered feature, assuming all other features remain constant. It is crucial to note that such interpretations do not imply causation.

The feature whose one standard deviation change means the greatest impact on price estimation for a property is the only negatively correlated one, distance from the CBD. Holding all other features constant, a property situated 6 km further away from the CBD compared to another one is estimated to be approximately 257k$ cheaper, and with a 95% probability from 240k$ to 273k$ cheaper. Building area emerges as the most influential positive feature standard-deviation wise: For every 85 square meters variation in building area, the estimated property price varies in the same direction by about 206k$, and with a 95% probability from 169k$ to 243k$, assuming all other variables remain constant.

Houses, cottages, villas, semis, and terraces price prediction:



Variables	Coeff.	0.025	0.975	Std. Dev
	1269.1901	1253.119	1285.261	
Bedrooms	52.8642	22.991	82.738	0.846
DistanceCBD	-349.7706	-376.488	-323.054	6.366
Bathroom	178.2234	143.405	213.042	0.764
Landsize	175.2589	123.358	227.159	276.271
BuildingArea	161.6249	123.337	199.913	88.381

Table 4: Houses and similar price prediction model information

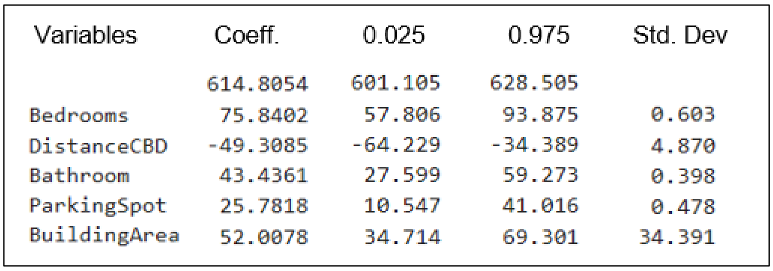
The model manages to pick up approximately 45.4% of the variance in the prices of Melbourne houses, cottages, villas, semis, and terraces, a similar proportion compared to the general model, again decent. According to this model, house prices are estimated based on six values: five house features and a constant. The constant is the first coefficient value in Table 4 and is to be interpreted as the theoretical price (thousands of $) for a house where all other considered features have a value of "0", and in the case of standardized data, which is just the case of all the linear regression models included in this elaborate, is the same as saying “a property with average values in all significant feature”; in this case it would be 1269k$, and to be more accurate, the model is 95% sure that this value is included in the range 1253k$ - 1285k$.

The number of properties in the suburb and the number of parking spots have been excluded from the model due to their lack of informative power in the prediction of house prices. The price is directly correlated with the number of bedrooms, number of bathrooms, land size, and building area: this means that an increase in any of these features is associated with a rise in the price (no causality is implied). Similar to the overall model, the distance from the CBD is the only feature negatively correlated with the price, indicating that the further away a house is from the CBD, the lower the price tends to be.

The coefficients for significant features are to be interpreted as the variation (thousands of $) a single standard-deviation variation in the considered feature would theoretically be matched with in the price prediction of a house. Again, no causality is implied. The feature whose standard deviation swing would be associated with the largest variation in price estimation for a house is, again, the only negatively correlated one: Holding all other features constant, a property 'a' which is 6.3 km further away from the CBD compared to property 'b' will be estimated to be approximately 349k$ cheaper; this is the exact esteem computed by the model, that provides also a range in which it believes the price variation would fall in with a 95% probability (from 323k$ to 376k$ cheaper). The most positively impactful feature (in terms of standard deviation) is the number of bathrooms. Given that all other features are equal, a property 'a' that has 0.76 more bathrooms compared to a property 'b' will be about 178k$ more expensive, and a property ‘c’ that has 0.76 fewer bathrooms compared to property ‘b’ will be about 178k$ cheaper.

Comparisons among models can only be made under a set of strict assumptions. What is possible to say comparing the house's model to the overall properties one is that average houses prices are higher than average overall properties prices. Coefficients comparison can be of hard interpretation, it is advised not to compare the numbers to see what features are more relevant in which model.

Units and duplexes price prediction:



Variables	Coeff.	0.025	0.975	Std. Dev
	614.8054	601.105	628.505	
Bedrooms	75.8402	57.806	93.875	0.603
DistanceCBD	-49.3085	-64.229	-34.389	4.870
Bathroom	43.4361	27.599	59.273	0.398
ParkingSpot	25.7818	10.547	41.016	0.478
BuildingArea	52.0078	34.714	69.301	34.391

Table 5: Units and duplexes price prediction model information

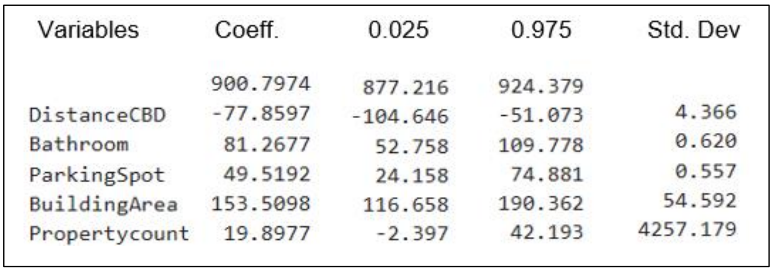
This model explains approximately 41.5% of the variance in the prices of Melbourne units and duplexes, slightly less than the general model and the one for houses.

The number of properties in the suburb and the land size have been excluded from the model as they have been individuated as features of close to null importance (non significant). According to this model reported in Table 5, the price can be estimated through six values: five property features and a constant. The price is directly correlated with the number of bedrooms, number of bathrooms, number of parking spots, and building area, indicating that an increase in any of these features is associated with a price rise. Once again, the distance from the CBD is the only feature negatively correlated with the price, meaning that the further away a unit is from the CBD, the lower the price tends to be.

Due to the adopted procedure of data standardization, coefficients of significant features shall be interpreted as the swing (thousands of $) that a one-standard-deviation variation in the considered feature would theoretically be matched with by the price of a unit; the constant term is to be interpreted as the price the model associates to a unit characterized by average values in all features. It's crucial to note that no causality is implied.
The feature with the potentially highest impact power, in terms of one standard deviation variation, is the number of bedrooms. Given that all other features are kept the same, property "a" with 0.6 more bedrooms compared to property "b" will be estimated to be about 75k$ more expensive than property “b”. The least impactful feature, in terms of standard deviation, is the number of parking spots. Holding all other features constant, a property "c" with 0.47 fewer parking spots compared to a property "b" will be estimated to be 25k$ cheaper, with a 95% probability, according to the model, of being anywhere from 10k$ to 41k$ cheaper.

Comparisons among models can be tricky and, if not properly done, often misleading, so it is advised to not do them in the absence of an expert. What it is possible to say comparing this model to the overall properties one and the houses one is that typical units are predicted to be more affordable than average houses and overall properties. Also, unlike in the houses and overall properties models where the most impactful single standard deviation variation was to be produced by the distance from the CBD, in the units model it is produced by a positively correlated feature: the number of bedrooms, with distance from CBD occupying only the third place, after building area.

Townhouses; development sites and other residentials price prediction:



Variables	Coeff.	0.025	0.975	Std. Dev
	900.7974	877.216	924.379	
DistanceCBD	-77.8597	-104.646	-51.073	4.366
Bathroom	81.2677	52.758	109.778	0.620
ParkingSpot	49.5192	24.158	74.881	0.557
BuildingArea	153.5098	116.658	190.362	54.592
Propertycount	19.8977	-2.397	42.193	4257.179

Table 6: Townhouses and similar price prediction model information

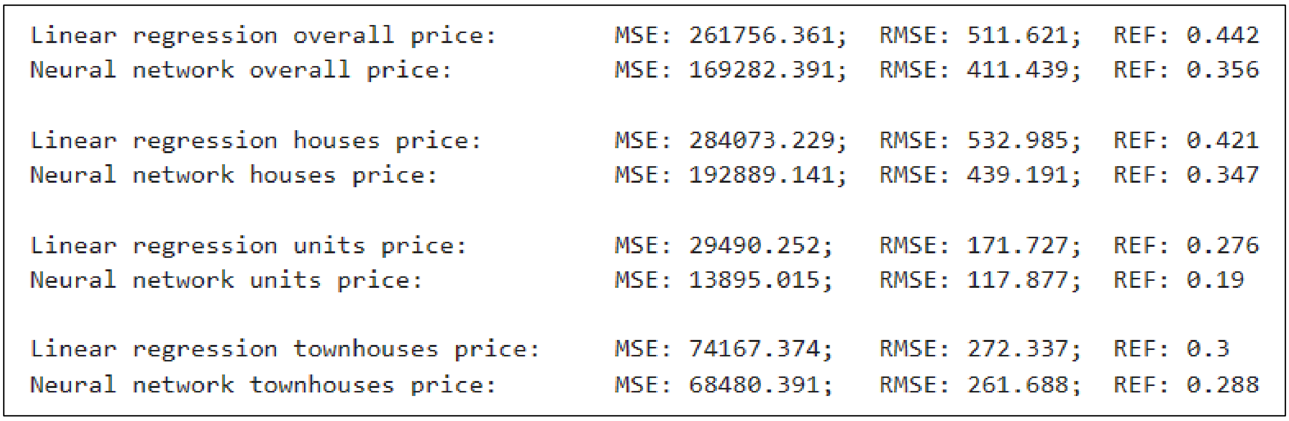
The model manages to explain about 40% of the variance of the price of Melbourne townhouses, which is a slightly smaller amount compared to the general model and the houses one, and in line with the units model.

As it is possible to see in Table 6, number of bedrooms and land size features have been removed from the model as they do not add much information; an argument could be made on the decision to keep the “Propertycount” variable as its test of significance for the estimation of the model was not properly passed, but the reason why it has been kept is that this model explains the least amount of variance among all 4 linear regression models produced, therefore it has been considered appropriate to hold a feature that could explain even a small amount of price variability, especially given the moderate computational demands of a model of such dimension.
According to this model, the price can be esteemed through 6 values: 5 house features and a constant. The price is directly correlated with the number of bathrooms, number of parking spots, building area, and number of properties in the suburb, which means that a variation of one of any of these features is associated with a same direction variation of the price. Again, the distance from the CBD is the only feature that is negatively correlated with the price; this means that the closer to the CBD a townhouse is, the higher the price tends to be.

Coefficients of significative features of the model are to be interpreted as the growth/decrease (thousands of $) a 1 standard deviation increase/decrease of the considered feature would theoretically be matched with by the price of a property (no causality implied). The feature whose 1 standard deviation variation would mean the biggest swing in the price esteem for a townhouse is the building area: given that all other features are equal, a property "a" which is 54 square meters larger compared to property "b" will be about 153k$ more expensive. The least impactful (standard deviation wise) feature is the number of properties in the suburb. Given that all other features are equal, a property "a" which counts 4257 more properties in its suburb compared to a property "b", will be about 19.8k$ more expensive, but the model even says there is the possibility of it being negatively correlated, in the sense that it is 95% confident that such value will fall into the range -2.3k$ to +42.1k$.

Comparisons among models shall be made with extreme attention.
Comparing the townhouses model with the previous 3, it is possible to see that average townhouses are priced (900k$) less than average houses and overall properties, and more than average units. Also, townhouses price is determined by the most unique combination of features, as it is the only model where the number of bedrooms is not considered relevant and in which the number of properties in the suburb has been included.

NEURAL NETWORKS AND RESULTS EVALUATION:



```
Linear regression overall price:       MSE: 261756.361;  RMSE: 511.621;  REF: 0.442
Neural network overall price:          MSE: 169282.391;  RMSE: 411.439;  REF: 0.356

Linear regression houses price:        MSE: 284073.229;  RMSE: 532.985;  REF: 0.421
Neural network houses price:           MSE: 192889.141;  RMSE: 439.191;  REF: 0.347

Linear regression units price:         MSE: 29490.252;   RMSE: 171.727;  REF: 0.276
Neural network units price:            MSE: 13895.015;   RMSE: 117.877;  REF: 0.19

Linear regression townhouses price:    MSE: 74167.374;   RMSE: 272.337;  REF: 0.3
Neural network townhouses price:       MSE: 68480.391;   RMSE: 261.688;  REF: 0.288
```

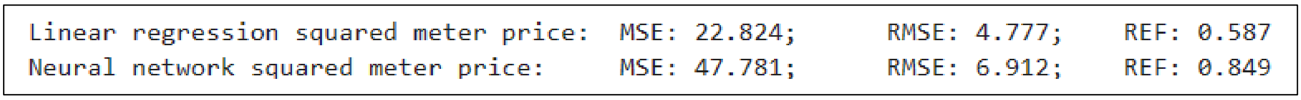
Table 7: Comparison and evaluation of the models

To facilitate a meaningful comparison between linear regression models and neural networks predictions, the same loss function —Mean Squared Error (MSE)— has been employed for both. The Root Mean Squared Error (RMSE) is utilized to express the average prediction error in the same units as the target variable (thousands of $ in this case). In Table 7 it has also been included a REF (Reference) value, which is a normalized metric, calculated as the ratio of RMSE to the mean. This ratio provides insight into the magnitude of prediction errors relative to the average value. In summary, the chosen metrics —MSE, RMSE, and REF— offer a comprehensive evaluation and comparison of regression models by considering both accuracy and scale of errors.

- **Overall price**: The neural network prediction manages to give a more accurate esteem of the price, with a prediction that is about 100.000$ more accurate (from 511k$ off to 411k$ off), and a ref value from 0.44 to 0.35. The neural network is a more complex system, but the advantages brought by its implementation are considered a big enough compensation.
- **Houses price**: Again, the neural network prediction manages to give a more accurate esteem of the price, with a prediction that is over 900.000$ more accurate (from 532k$ off to 439k$ off) and a ref value from 0.42 to 0.34. This prediction tool is quite similar to the overall price one, and this is probably because most properties of the available data fall into this category.
- **Units price**: The improvement brought by the neural network for the prediction of units and duplexes prices is, proportionally, the largest one (from 171000$ to 117000$) with a ref of 0.19, which is the lowest among all predictive tools. One of the reasons behind this esteem's success is probably the "tighter" distribution of the units and duplexes compared to the others, meaning the least outliers, and generally closer to the distribution anyway.
- **Townhouses price**: This is the neural network that brought the least improvement (from 272k$ off to 261k$ off). In this case, depending on the use of these informations, it might be worthy of consideration the use of the linear regression model, which even though is slightly less accurate, is simpler, less computationally demanding, and more interpretable.

SQUARE METER PRICE PREDICTION:

As per demand of the client, it has been made an attempt to predict the values of square meter price for properties. This is an artificially generated feature determined as the ratio of price to the building area. The attempts made did not give the hoped results.



```
Linear regression squared meter price:  MSE: 22.824;     RMSE: 4.777;    REF: 0.587
Neural network squared meter price:     MSE: 47.781;     RMSE: 6.912;    REF: 0.849
```

Table 8: comparison and evaluation of the models for square meter price

From Table 8, it is possible to see that just like the linear regression model, the neural network failed to provide a decent prediction of square meter price. What is concluded is that the numerical features available in this dataset are incapable of providing a proper prediction for the square meter price. Thus, the squared meter price analysis has not been further explored.

PRICE PREDICTION CONCLUSIONS:

The analysis reveals that neural networks generally outperform linear regression in price prediction. However, despite the improvement, there remains a noticeable degree of inaccuracy. This suggests the presence of determinants influencing prices not captured by the available data.

- **Interpretability**: Linear regression, while less accurate, offers a relatively straightforward interpretation. In contrast, neural networks lack a clear explanation of the relationship with the features, as per their "black-box" nature.
- **Limitations**: The observed inaccuracies in predictions, especially with neural networks, highlight the inherent limitations of the dataset. Certain factors influencing prices may be absent or inadequately represented.
- **Consideration of Reference Values**: Evaluation of reference values is context-dependent and should be based on the intended use of predictions. This analysis provides insights, but decisions regarding the suitability of predictions depend on the specific application.

In conclusion, while neural networks enhance prediction accuracy, careful consideration is needed, balancing interpretability, and recognizing potential limitations in the dataset. Knowing the use case is crucial for appropriately assessing the significance of reference values.

CLUSTER ANALYSIS:

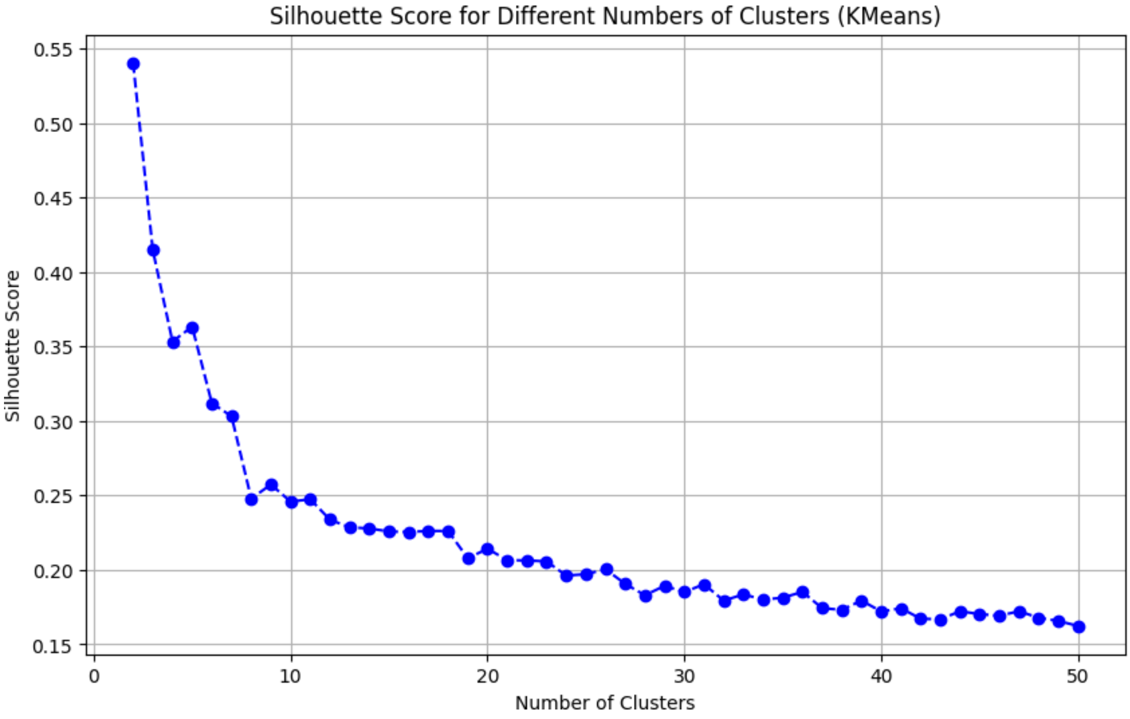
Through cluster analysis, the aim is to group similar properties together, trying to minimize the logical distance among residences in the same group and at the same time maximize the logical distance among groups. The business uses for a real estate agency are multiple:

- Targeted Marketing.
- Customized Marketing Strategies.
- Resource Allocation.
- Pricing Strategies.
- Market Trends and Dynamics.
- Customer Relationship Management.
- Risk Assessment.
- Portfolio Diversification.

K-MEANS CLUSTERING:

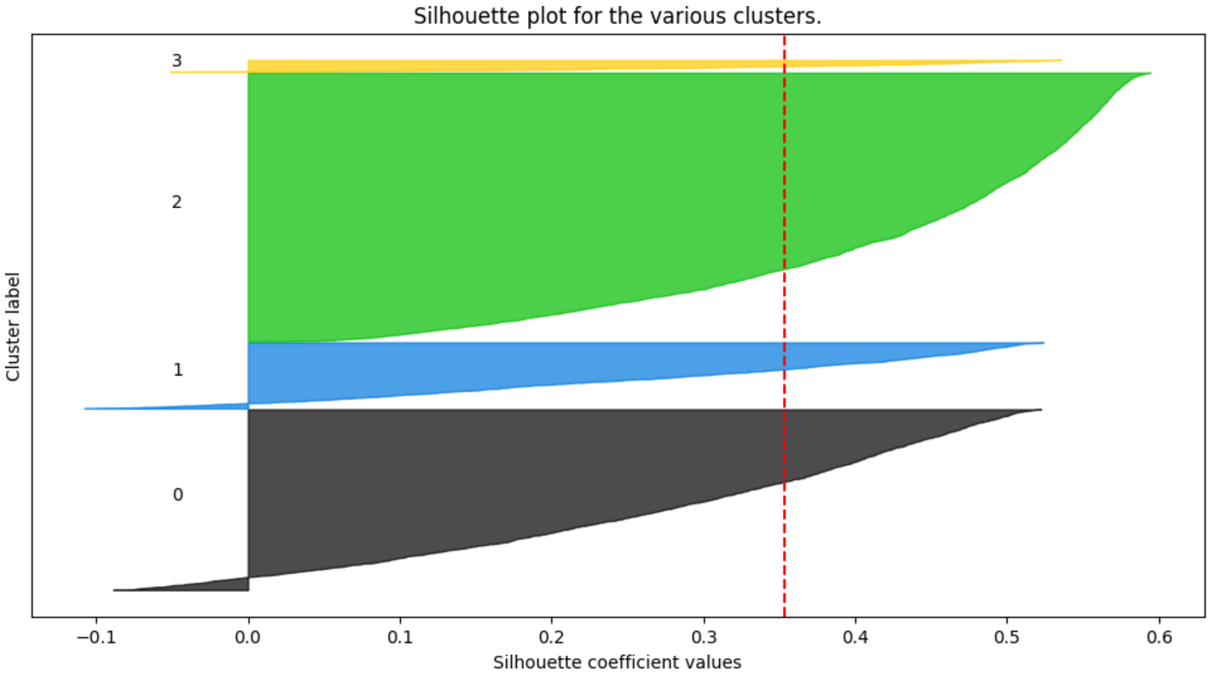
The method used is the k means clustering, which has been checked to be the best performing one in this scenario. Properties have been divided in 4 groups.

About the data, it has been adopted a unique standardization technique, which consists of regular standardization followed by some scaling operations performed ad hoc for some accurately selected features (price*10, building area*5, number of bedrooms*2, and land size*2). This forced the clustering procedure to give more importance to certain features compared to others.



Picture 9: Silhouette score per k-means clustering size

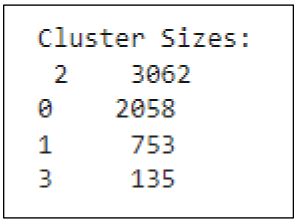
Looking at the silhouette scores (which is a measure of goodness) graph in Picture 9, the most appropriate number of clusters for the sake of the elaborate is either 3 or 4, with a silhouette score of a little over 0.35. Such a score is not great in general, but given the high skewness of the data, and the presence of such a high number of outliers, it can be considered decent. Clustering of 2 and 1 groups have better scores, but it is considered a worthy move to sacrifice a contained amount of accuracy to get a more suitable number of groups to work with. So the partition chosen is the one considering 4 clusters.



Picture 10: Silhouette score cluster

Picture 10 demonstrates how 3 out of the 4 clusters experience a fairly low presence of properties with negative silhouette scores, with most properties that are considered to have discrete silhouette scores, which is a sign of the fact that properties are somewhat "happy" to belong to the cluster they have been assigned to.

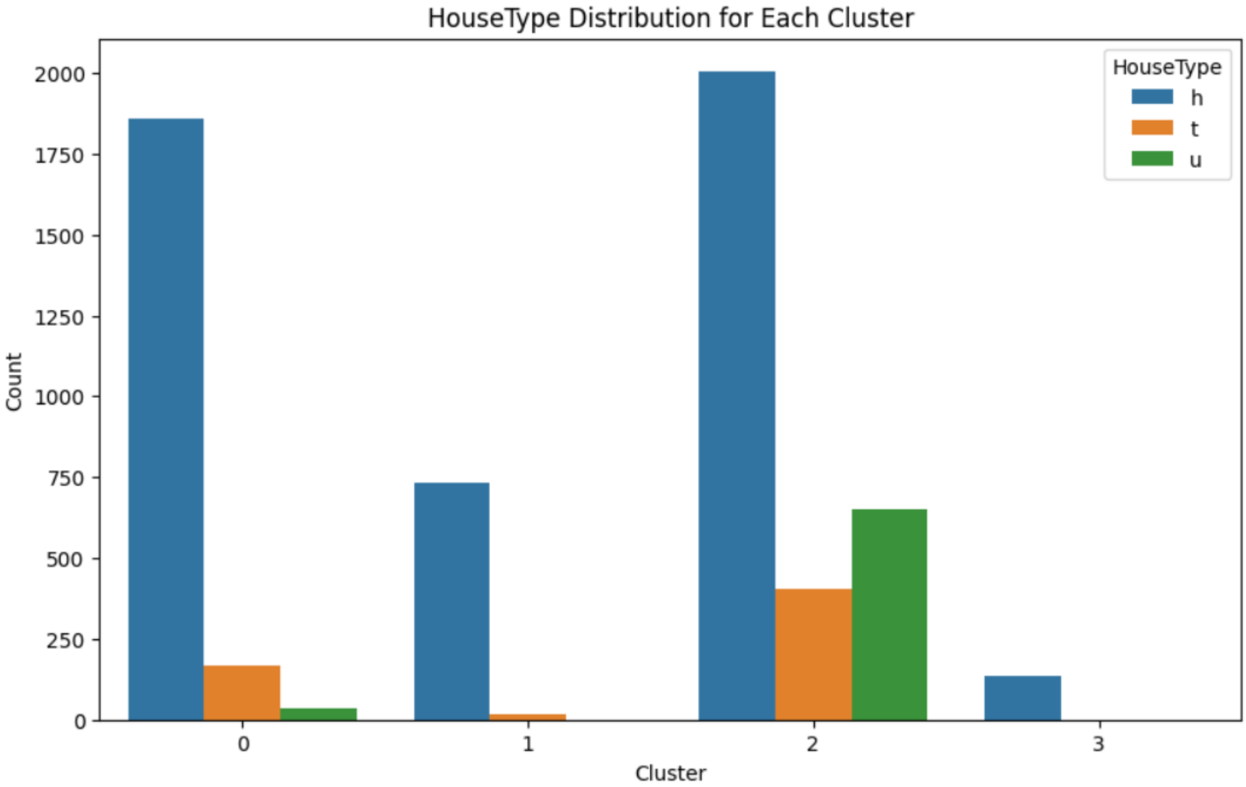
Analysis of the clustering:



```
Cluster Sizes:
 2     3062
0     2058
1      753
3      135
```

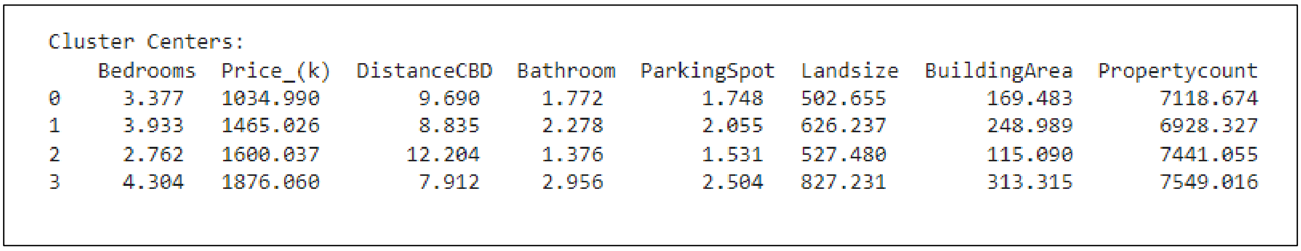
Table 9: Clusters sizes

Table 9 reports the dimensions of the clusters. They vary in size, with the largest one including over 3000 properties and the smallest one counting 135.



Picture 11: House type distribution per cluster

Picture 11 reports the house type distribution per cluster. All clusters show a dominance of the most common house type, which is the house, additionally, cluster 3 presents exclusively house type properties. Cluster 1 completely lacks unit properties, the (almost) totality of which is included in cluster 2.



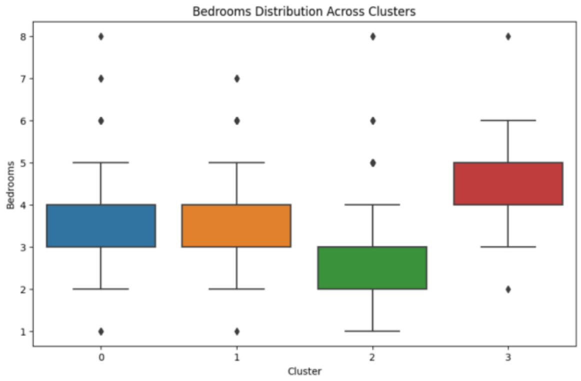
```
Cluster Centers:
    Bedrooms  Price_(k)  DistanceCBD  Bathroom  ParkingSpot  Landsize  BuildingArea  Propertycount
0      3.377   1034.990        9.690     1.772        1.748   502.655       169.483       7118.674
1      3.933   1465.026        8.835     2.278        2.055   626.237       248.989       6928.327
2      2.762   1600.037       12.204     1.376        1.531   527.480       115.090       7441.055
3      4.304   1876.060        7.912     2.956        2.504   827.231       313.315       7549.016
```

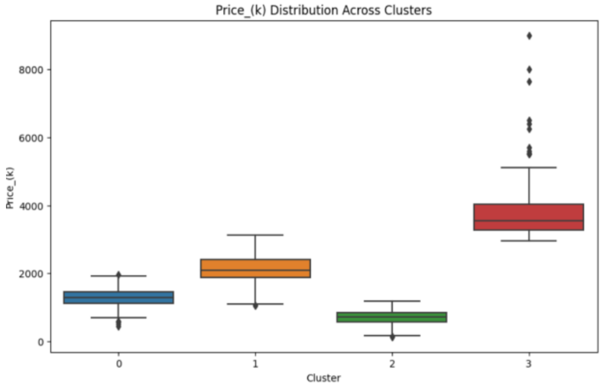
Table 10: Clusters centers

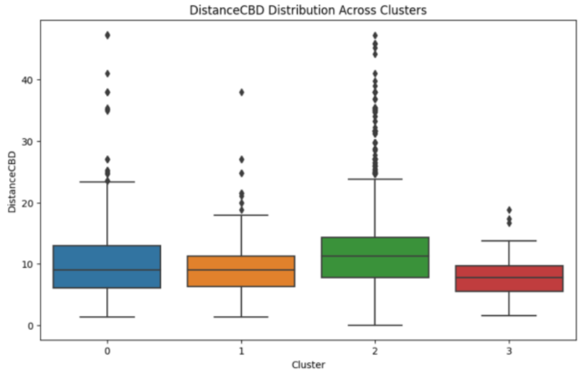
Before discussing the clusters' centers (Table 10), it is worth underlying that the values are brought back to the original scale, but are not to be taken as accurate esteems. Values might be biased, and what it is possible to extract from them are insights about the relationships, and their magnitude, among groups. Also, given the distortions created in the data that gave some features a larger weight, it is appropriate to discuss those features with more energy.

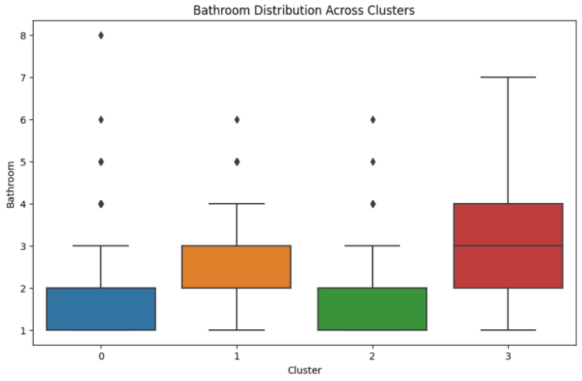
Cluster centers indicate the general tendencies of properties belonging to those clusters.

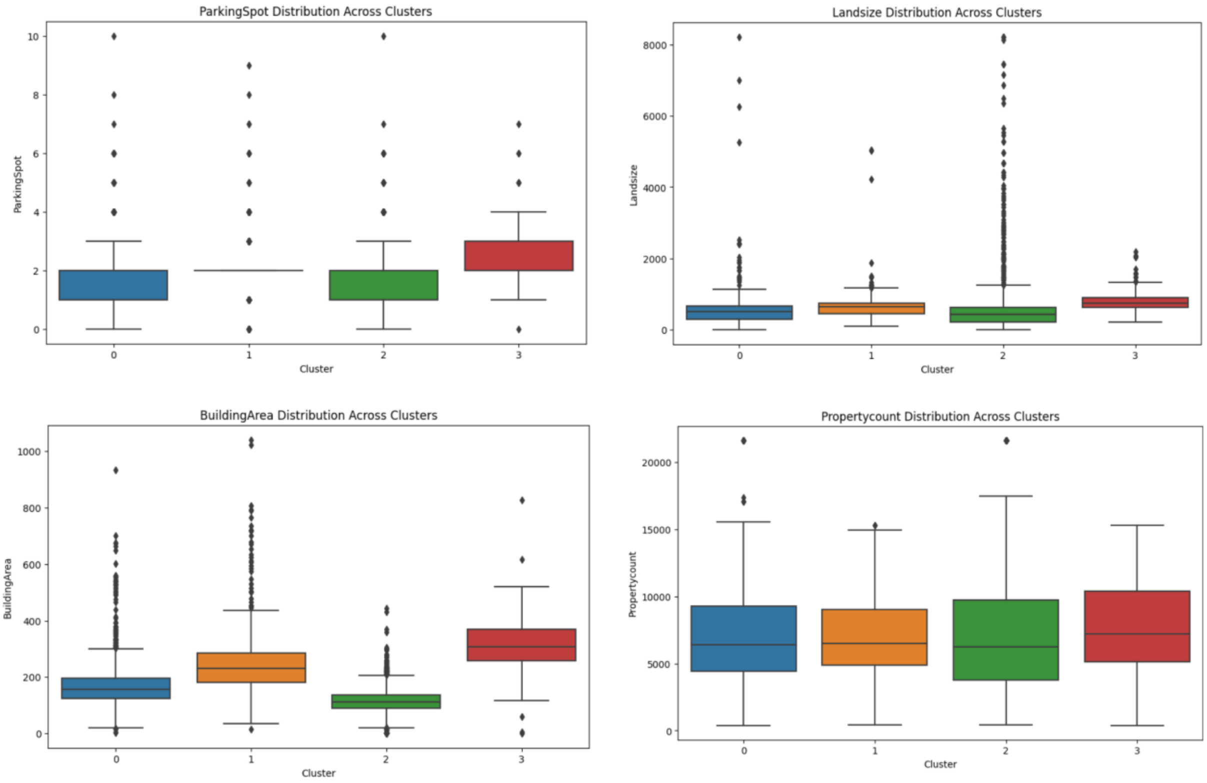
- **Cluster 0** includes the cheapest properties, with the smallest land size.
- **Cluster 1** properties are the closest to the CBD, yet the ones that average the least number of residences in their neighborhoods.
- **Cluster 2** properties are relatively small properties, with the smallest average building area, number of bedrooms, bathrooms, and parking spots. This is also the only group including no residence with a negative silhouette coefficient, other than the largest group
- **Cluster 3** properties are the most expensive, the biggest ones, counting the highest average number of bedrooms, bathrooms, and parking spots. this is the smallest cluster, including only house type properties, probably high-end homes.











Picture 12: Clusters features distributions

Together with the silhouette coefficient, it is possible to take a look at the clusters distributions for all numeric features (Picture 12) to check how cohesive they are (the tighter the better) as it means there are fewer properties different from the others in the same group. All clusters seem to have similar cohesion levels. Boxplots confirm the analysis based on clusters centers, providing the additional possibility to take a look at the distribution and outliers of each group for each feature.

CLUSTERING CONCLUSIONS:

- **Cluster 0** counts 2058 properties, including less expensive, and with smaller land sizes.
- **Cluster 1** counts 753 properties, including no units nor duplexes, in low-densely inhabited neighborhoods.
- **Cluster 2** is the biggest group with 3062 residences belonging to it. It includes smaller properties, that consequentially average fewer bedrooms, bathrooms, parking spots and smaller building areas.
- **Cluster 3** is the smallest group counting only 135 properties. To this cluster are assigned the highest-end, larger, and more expensive properties, including only house type ones.