- **ACM476 TERM PROJECT PHASE 1**

Melbourne Housing

Kaan Akkök, Hasan Emir Çilesiz

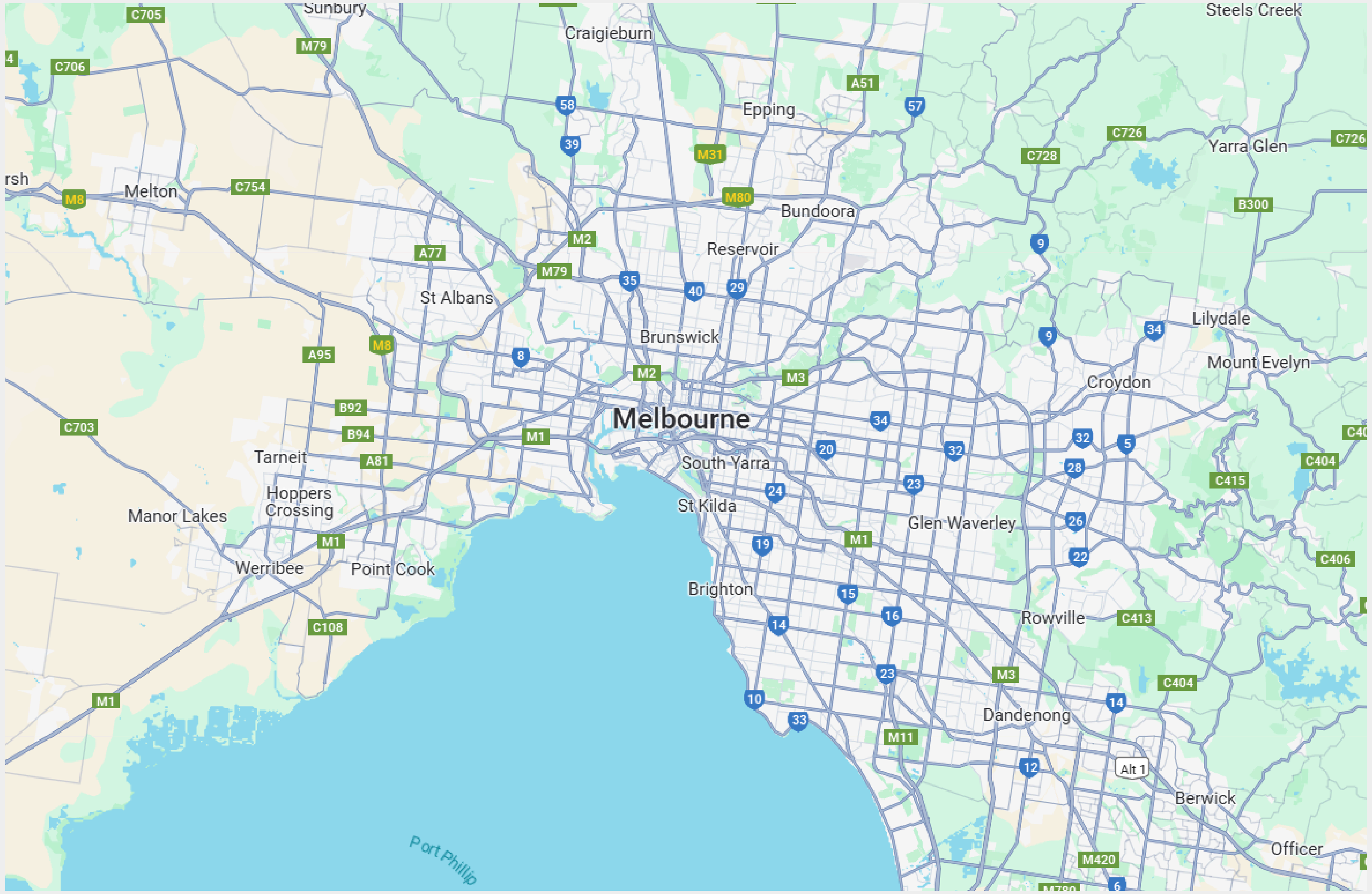
14.05.2025

Content

Introduction:

Target:

The aim of this project is to analyze and predict house prices in Melbourne (Australia) using various data mining techniques, including preprocessing, regression, classification, clustering, feature selection, and dimensionality reduction.



Dataset

Dataset Information:

- Data Source: Kaggle – Melbourne Housing Dataset
- Total Observations: 34,857
- Sample Used: 2,000 rows (random_state=5006)
- Number of Features (columns): 21
- This data was scraped from publicly available results posted every week from Domain.com.au

[o: object, i: integer, f: float]

Suburb (o), Address (o), Rooms (i), Type (o), **Price (f)**, Method (o), SellerG (o),
Date (o), Distance (f), Postcode (f), Bedroom2 (f), Bathroom (f), Car (f), Landsize (f), BuildingArea (f), YearBuilt (f), CouncilArea (o), Lattitude (f), Longtitude (f), Regionname (o), Propertycount (f)

	Suburb	Address	Rooms	Type	Price	Method	SellerG	Da
1	Hillside	105 Community Hub	3	h	781500	S	YPA	2017-09-
2	Brighton East	2/13 Tatong Rd	3	h	1056532	S	Marshall	2016-05-
3	Glen Iris	43 Albion Rd	4	h	1997500	S	Marshall	2018-02-
4	Oak Park	74a Winifred St	3	u	600000	SP	Brad	2016-08-
5	Beaumaris	17A Towers St	4	h	1140000	S	Hodges	2017-10-
6	Kensington	37 Cakebread Mw	4	h	1200000	S	Nelson	2017-02-
7	Mont Albert	1/12 Hotham St	2	u	850000	SP	hockingstuart	2017-09-
8	Coburg	25 Stock St	3	h	1056532	S	Jellis	2017-09-
9	Bulleen	25 William St	4	h	1235000	S	Jellis	2016-09-
10	Bentleigh East	50a Brady Rd	3	u	905000	SP	Woodards	2017-08-

Preprocessing of data:

We identified missing values from the sample dataset in the following columns:

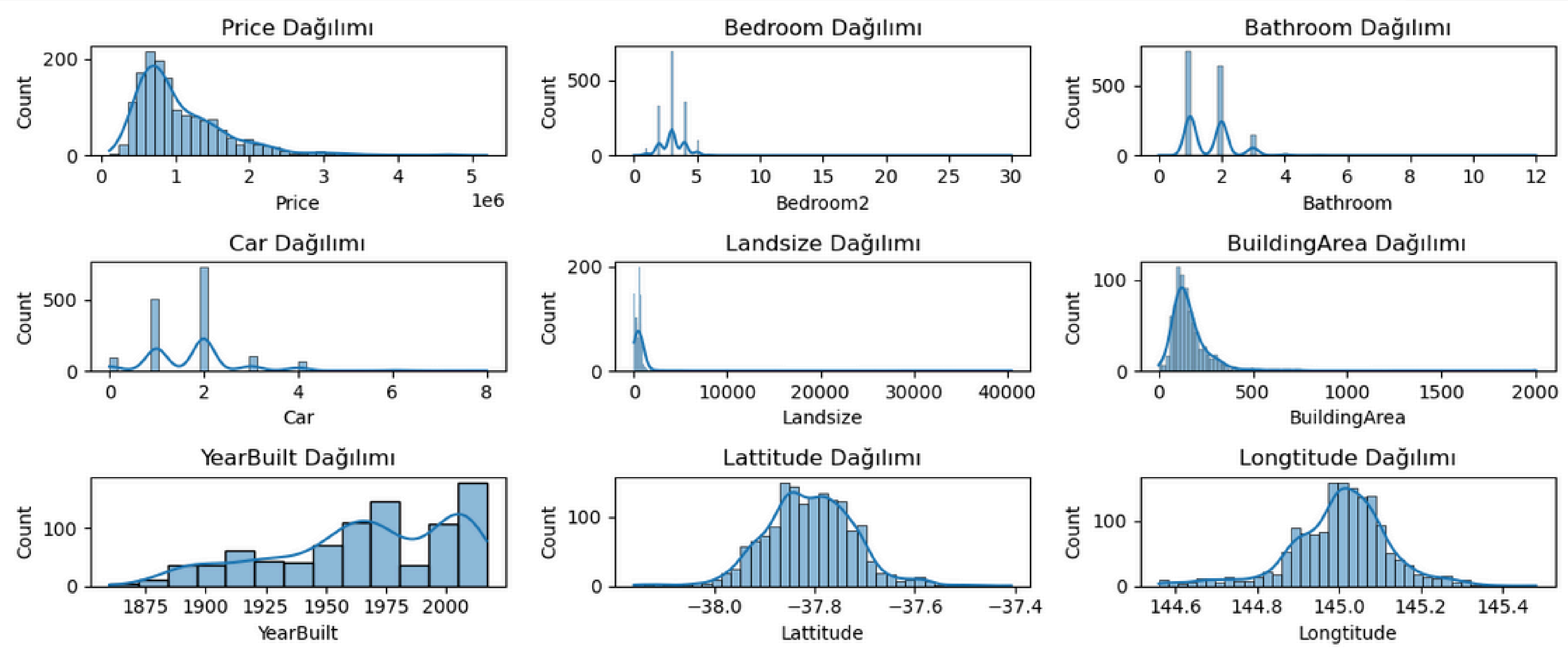
- “Price”: 443
- “Bedroom2”: 453
- “Bathroom”: 453
- “Car”: 482
- “Landsize”: 669
- “BuildingArea”: 1217
- “YearBuilt”: 1123
- “Lattitude”: 437
- “Longtitude”: 437

These are all numerical variables. There is no missing values on categorical variables.

```
Suburb:
        dtype:object null:0
Address:
        dtype:object null:0
Rooms:
        dtype:int64 null:0
Type:
        dtype:object null:0
Price:
        dtype:float64 null:443
Method:
        dtype:object null:0
SellerG:
        dtype:object null:0
Date:
        dtype:object null:0
Distance:
        dtype:float64 null:0
Postcode:
        dtype:float64 null:0
Bedroom2:
        dtype:float64 null:453
Bathroom:
        dtype:float64 null:453
Car:
        dtype:float64 null:482
Landsize:
        dtype:float64 null:669
BuildingArea:
        dtype:float64 null:1217
YearBuilt:
        dtype:float64 null:1123
CouncilArea:
        dtype:object null:0
Lattitude:
        dtype:float64 null:437
Longtitude:
        dtype:float64 null:437
Regionname:
        dtype:object null:0
Propertycount:
        dtype:float64 null:0
```

Preprocessing of data:

We inspected the graph showing the distribution of missing values to determine how to fill them.

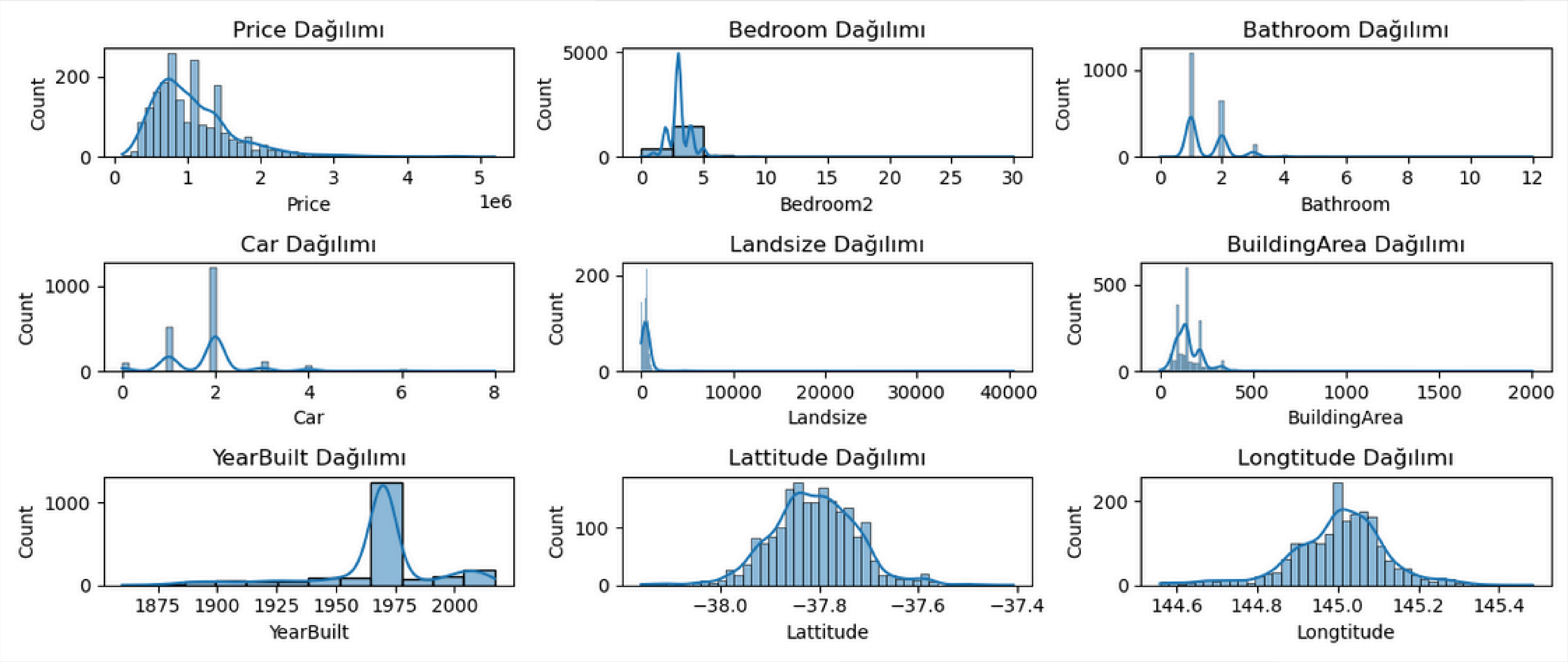


- “Price”: Filled with the average price of properties that have the same number of 'Rooms'.
- “Bedroom2” : (numeric categorical value) Filled missing values with the most frequent value (mode).
- “Bathroom”: (numeric categorical value) Filled missing values with the most frequent value (mode).
- “Car”: (numeric categorical value) Filled missing values with the most frequent value (mode).
- “Landsize”: Filled missing values with the average landsize of properties in the same 'Suburb'.
- “BuildingArea”: Filled missing values with the average BuildingArea of properties in the same 'Rooms'.
- “YearBuilt”: Filled missing values with the median year.
- “Lattitude” and “Longtitude”: Filled missing values with the average Lattitude and Longtitude values of properties in the same 'Suburb'.

Preprocessing of data:

After the handling the missing values, we saved the sample dataset.

```
Suburb           0
Address          0
Rooms            0
Type             0
Price            0
Method           0
SellerG          0
Date             0
Distance         0
Postcode         0
Bedroom2         0
Bathroom         0
Car              0
Landsize         0
BuildingArea     0
YearBuilt        0
CouncilArea      0
Lattitude        0
Longtitude       0
Regionname       0
Propertycount    0
dtype: int64
```



EDA:

We separated numerical and categorical columns:

- We applied detailed statistics for numerical columns
- We found the unique values in the categorical columns and noticed that 'address' has too many unique values. For this reason, we concluded that 'address' is unnecessary.
- We graphed the distribution of numerical values and noticed that “Bedroom2”, “Bathroom” and “Car” values behave like categorical values.
- We graphed the distribution of categorical values.
- We graphed the comparison between “Price” and other numerical variables.
- We graphed the comparison between “Price” and categorical variables and noticed that “Type”, “Method” and “Regionname” are related to “Price”.
- We graphed the distribution of house type based on region name.
- We have created a table showing the which types of houses are sold more frequently in each region and by which method.
- We have created a table showing the number of houses sold by sellers in different areas.

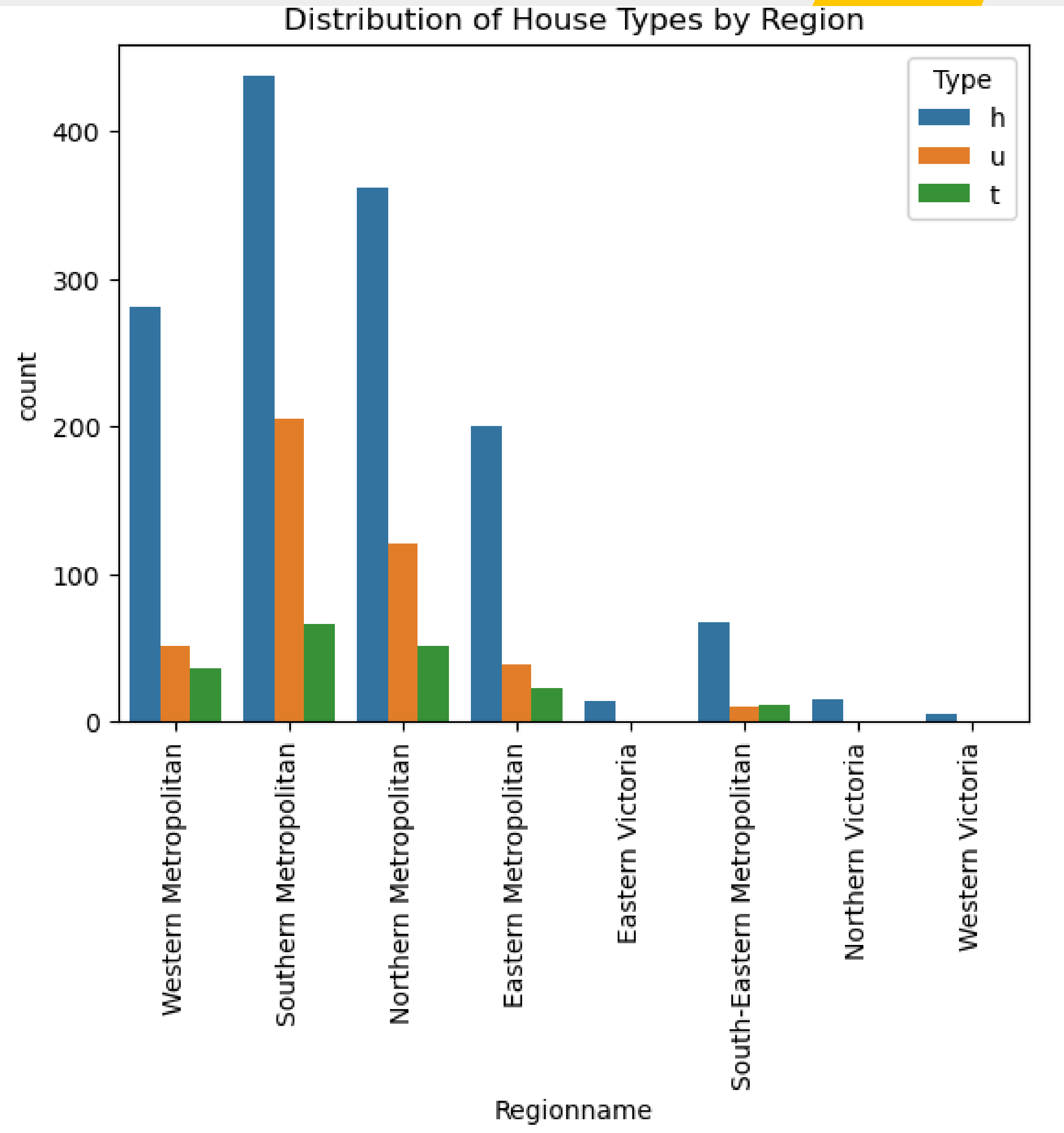
```
Numerical:
        Rooms
        Price
        Distance
        Postcode
        Bedroom2
        Bathroom
        Car
        Landsize
        BuildingArea
        YearBuilt
        Lattitude
        Longtitude
        Propertycount
Categorical:
        Suburb
        Address
        Type
        Method
        SellerG
        Date
        CouncilArea
        Regionname
```

EDA:

```
Suburb unique vals: 274
Address unique vals: 1997
Type unique vals: 3
Method unique vals: 9
SellerG unique vals: 150
Date unique vals: 77
CouncilArea unique vals: 32
Regionname unique vals: 8
```

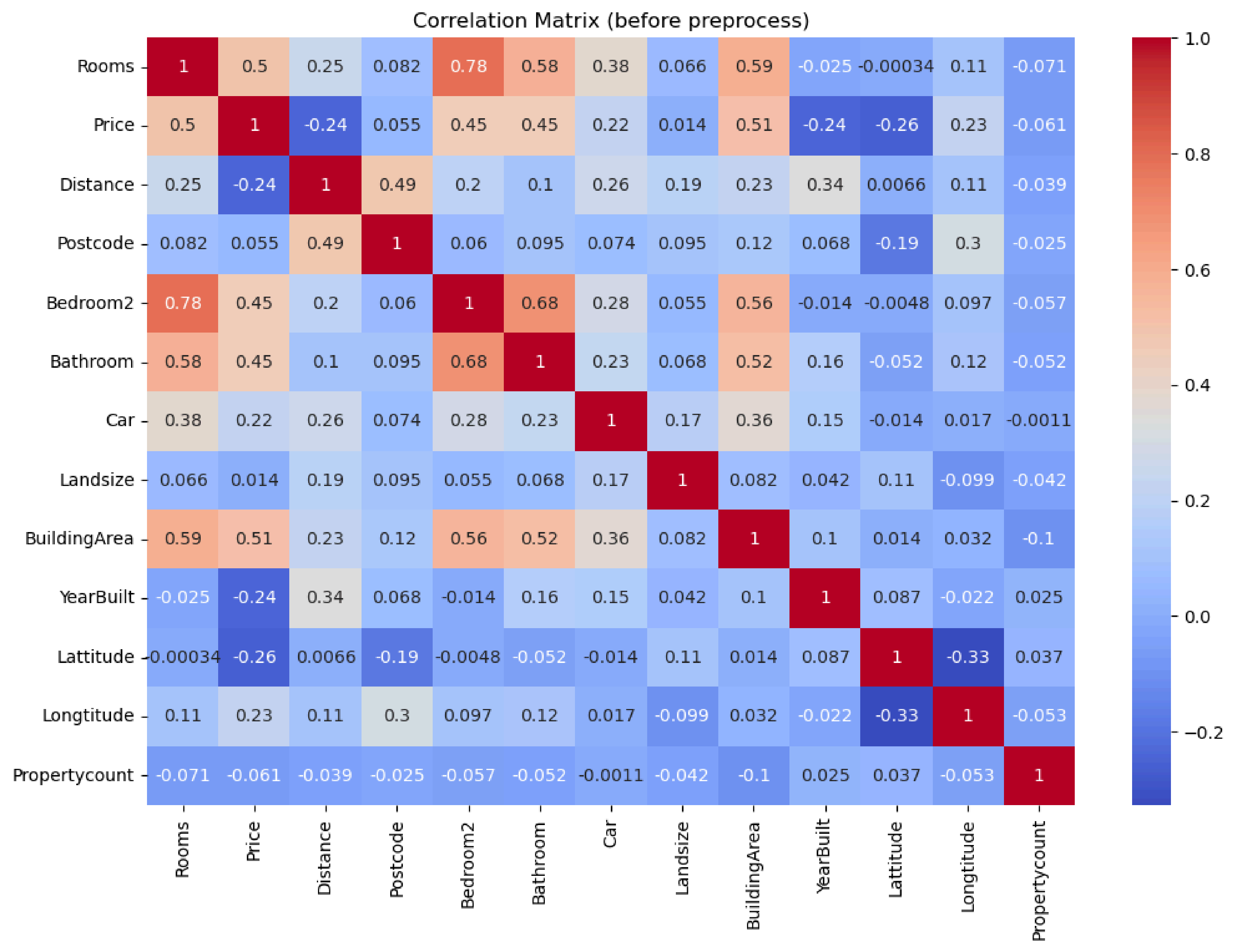
Regionname	SellerG	h	t	u
Eastern Metropolitan	Appleby	1	0	0
	Barry	28	3	3
	Bekdon	2	0	0
	Biggin	2	0	0
	Buxton	6	0	0
...	...	...	...	...
Western Metropolitan	hockingstuart	20	4	4
Western Victoria	Raine	1	0	0
	Reliance	1	0	0
	YPA	1	0	0
	hockingstuart	3	0	0

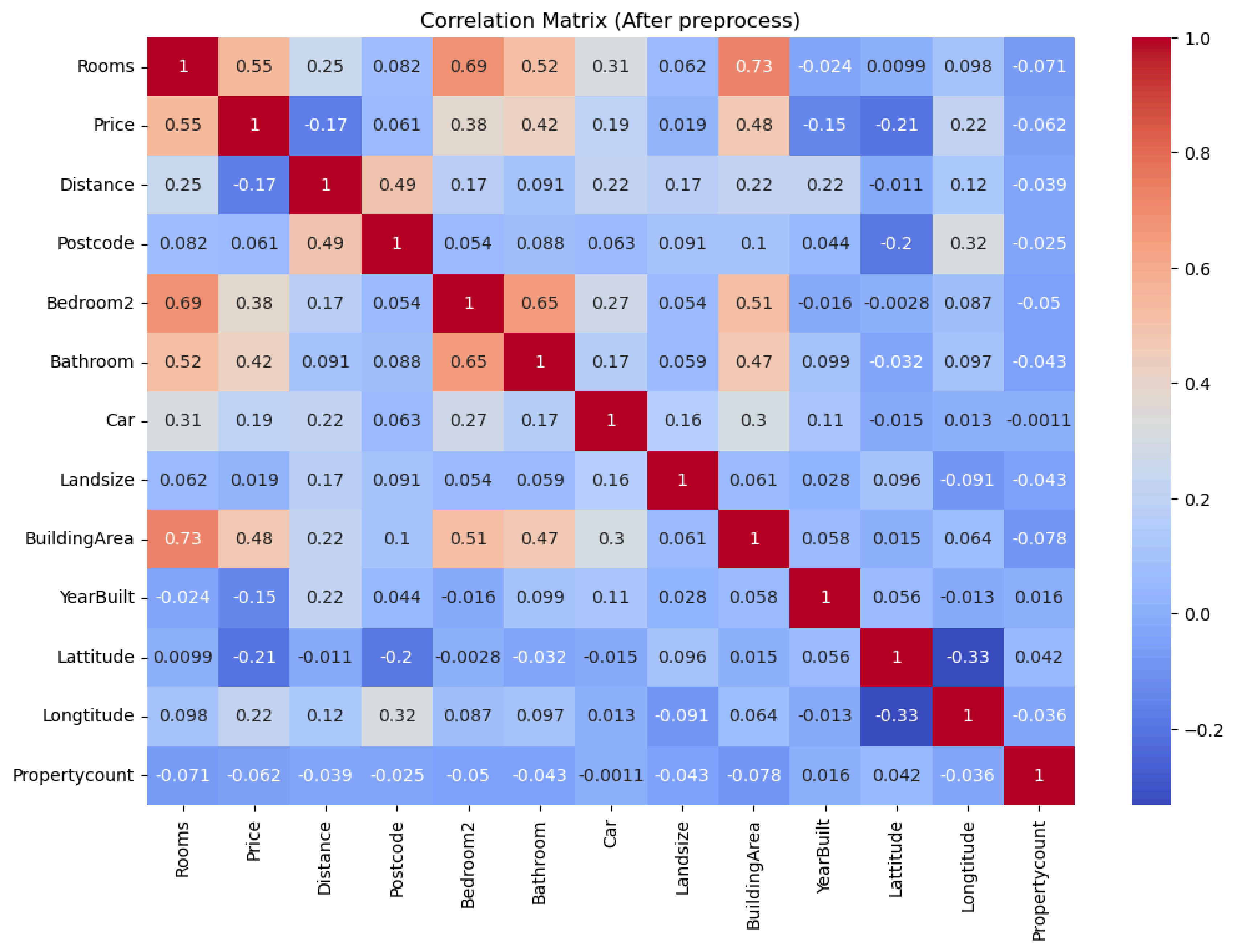
Regionname	Type	PI	PN	S	SA	SN	SP	SS	VB	W
Eastern Metropolitan	h	29	1	115	0	14	26	0	15	0
	t	3	2	7	0	1	6	0	4	0
	u	4	0	27	0	1	5	0	2	0
Eastern Victoria	h	2	0	8	0	2	0	0	2	0
	u	0	0	1	0	0	0	0	0	0
Northern Metropolitan	h	46	2	245	2	6	42	0	17	2
	t	7	0	25	0	1	12	0	6	0
	u	28	2	49	1	1	26	1	10	3
Northern Victoria	h	0	0	8	1	1	4	0	1	0
South-Eastern Metropolitan	h	12	0	32	1	5	9	0	9	0
	t	3	0	6	0	1	1	0	0	1
	u	2	0	3	0	0	3	0	2	0
Southern Metropolitan	h	61	6	253	4	17	44	0	52	0
	t	11	2	32	1	3	10	0	8	0
	u	32	7	111	2	3	33	0	14	3
Western Metropolitan	h	26	3	170	2	12	46	0	19	3
	t	7	0	17	0	0	6	0	6	0
	u	8	0	27	0	1	8	0	8	0
Western Victoria	h	0	0	6	0	0	0	0	0	0



	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt	Lattitude	Longtitude
count	2000.000000	2.000000e+03	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	3.057500	1.090117e+06	11.248200	3117.073500	3.099000	1.506500	1.799000	620.779000	154.697500	1968.137500	-37.810891	145.002227
std	0.977073	5.981216e+05	6.720257	106.567502	1.052974	0.741102	0.857302	1666.571369	87.982194	24.859557	0.087331	0.119269
min	1.000000	1.120000e+05	0.000000	3000.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1860.000000	-38.159690	144.559290
25%	2.000000	6.953750e+05	6.500000	3052.750000	3.000000	1.000000	1.000000	288.000000	92.000000	1970.000000	-37.862927	144.937600
50%	3.000000	9.610000e+05	10.500000	3104.000000	3.000000	1.000000	2.000000	525.500000	139.000000	1970.000000	-37.810891	145.009686
75%	4.000000	1.359154e+06	14.000000	3161.000000	3.000000	2.000000	2.000000	666.000000	201.000000	1970.000000	-37.754595	145.073440
max	10.000000	5.200000e+06	48.100000	3977.000000	30.000000	12.000000	8.000000	40469.000000	2002.000000	2017.000000	-37.407580	145.482460

EDA:


Correlation Matrix (before preprocess)
Rooms 1 0.5 0.25 0.082 0.78 0.58 0.38 0.066 0.59 -0.025 -0.00034 0.11 -0.071
Price 0.5 1 -0.24 0.055 0.45 0.45 0.22 0.014 0.51 -0.24 -0.26 0.23 -0.061
Distance 0.25 -0.24 1 0.49 0.2 0.1 0.26 0.19 0.23 0.34 0.0066 0.11 -0.039
Postcode 0.082 0.055 0.49 1 0.06 0.095 0.074 0.095 0.12 0.068 -0.19 0.3 -0.025
Bedroom2 0.78 0.45 0.2 0.06 1 0.68 0.28 0.055 0.56 -0.014 -0.0048 0.097 -0.057
Bathroom 0.58 0.45 0.1 0.095 0.68 1 0.23 0.068 0.52 0.16 -0.052 0.12 -0.052
Car 0.38 0.22 0.26 0.074 0.28 0.23 1 0.17 0.36 0.15 -0.014 0.017 -0.0011
Landsize 0.066 0.014 0.19 0.095 0.055 0.068 0.17 1 0.082 0.042 0.11 -0.099 -0.042
BuildingArea 0.59 0.51 0.23 0.12 0.56 0.52 0.36 0.082 1 0.1 0.014 0.032 -0.1
YearBuilt -0.025 -0.24 0.34 0.068 -0.014 0.16 0.15 0.042 0.1 1 0.087 -0.022 0.025
Lattitude -0.00034 -0.26 0.0066 -0.19 -0.0048 -0.052 -0.014 0.11 0.014 0.087 1 -0.33 0.037
Longtitude 0.11 0.23 0.11 0.3 0.097 0.12 0.017 -0.099 0.032 -0.022 -0.33 1 -0.053
Propertycount -0.071 -0.061 -0.039 -0.025 -0.057 -0.052 -0.0011 -0.042 -0.1 0.025 0.037 -0.053 1
Rooms Price Distance Postcode Bedroom2 Bathroom Car Landsize BuildingArea YearBuilt Lattitude Longtitude Propertycount
1.0 0.8 0.6 0.4 0.2 0.0 −0.2

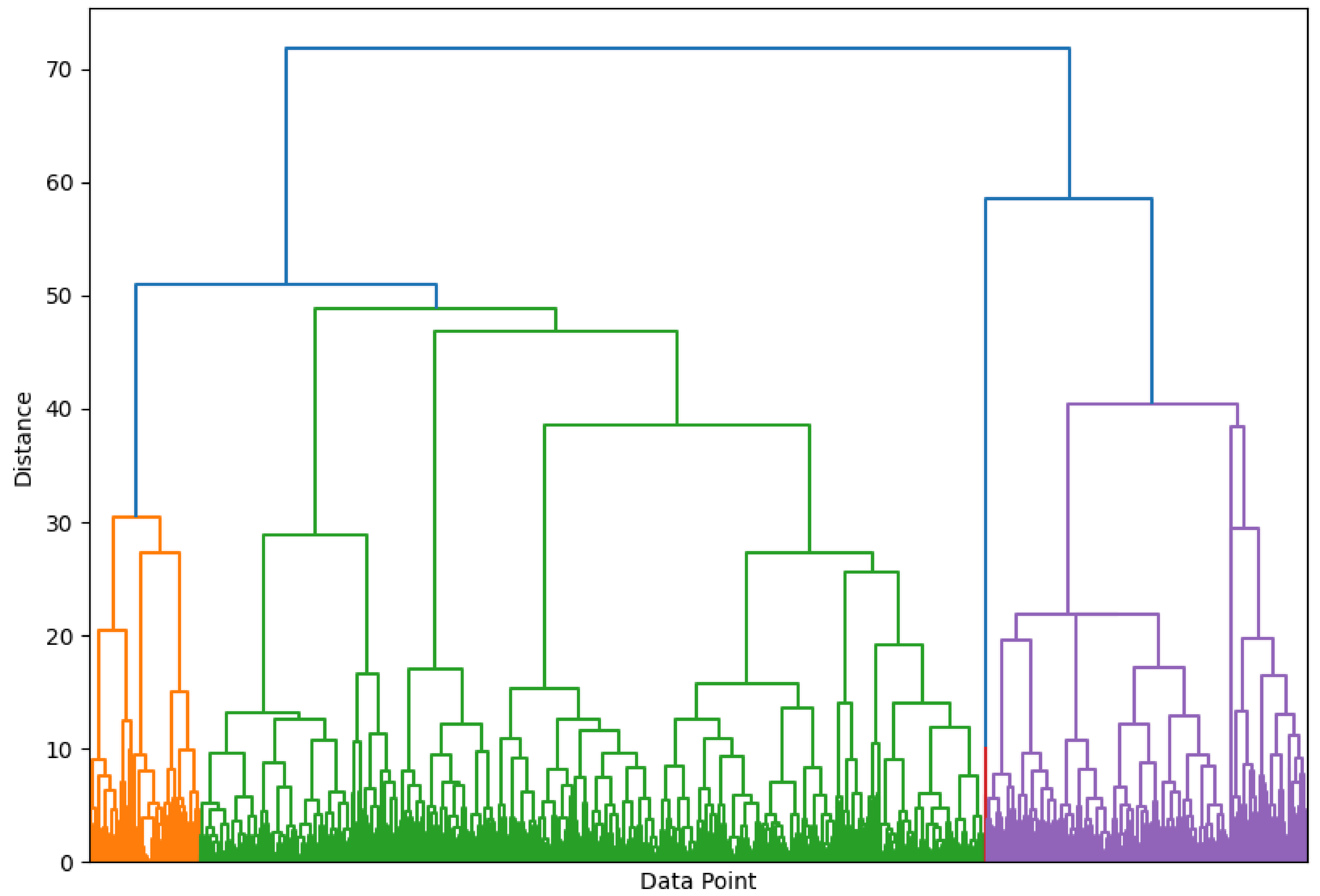

Correlation Matrix (After preprocess)
Rooms 1 0.55 0.25 0.082 0.69 0.52 0.31 0.062 0.73 -0.024 0.0099 0.098 -0.071
Price 0.55 1 -0.17 0.061 0.38 0.42 0.19 0.019 0.48 -0.15 -0.21 0.22 -0.062
Distance 0.25 -0.17 1 0.49 0.17 0.091 0.22 0.17 0.22 0.22 -0.011 0.12 -0.039
Postcode 0.082 0.061 0.49 1 0.054 0.088 0.063 0.091 0.1 0.044 -0.2 0.32 -0.025
Bedroom2 0.69 0.38 0.17 0.054 1 0.65 0.27 0.054 0.51 -0.016 -0.0028 0.087 -0.05
Bathroom 0.52 0.42 0.091 0.088 0.65 1 0.17 0.059 0.47 0.099 -0.032 0.097 -0.043
Car 0.31 0.19 0.22 0.063 0.27 0.17 1 0.16 0.3 0.11 -0.015 0.013 -0.0011
Landsize 0.062 0.019 0.17 0.091 0.054 0.059 0.16 1 0.061 0.028 0.096 -0.091 -0.043
BuildingArea 0.73 0.48 0.22 0.1 0.51 0.47 0.3 0.061 1 0.058 0.015 0.064 -0.078
YearBuilt -0.024 -0.15 0.22 0.044 -0.016 0.099 0.11 0.028 0.058 1 0.056 -0.013 0.016
Lattitude 0.0099 -0.21 -0.011 -0.2 -0.0028 -0.032 -0.015 0.096 0.015 0.056 1 -0.33 0.042
Longtitude 0.098 0.22 0.12 0.32 0.087 0.097 0.013 -0.091 0.064 -0.013 -0.33 1 -0.036
Propertycount -0.071 -0.062 -0.039 -0.025 -0.05 -0.043 -0.0011 -0.043 -0.078 0.016 0.042 -0.036 1
Rooms Price Distance Postcode Bedroom2 Bathroom Car Landsize BuildingArea YearBuilt Lattitude Longtitude Propertycount
1.0 0.8 0.6 0.4 0.2 0.0 −0.2

Clustering

After removing the target class label and non-numerical features, we used the remaining 12 numerical attributes. These features represent the physical and locational characteristics of the properties

ROOMS	INT64
DİSTANCE	FLOAT64
BEDROOM2	INT64
BATHROOM	INT64
CAR	INT64
LANDSİZE	INT64
BUİLDİNGAREA	INT64
YEARBUİLT	INT64
LATTİTUDE	FLOAT64
LONGTİTUDE	FLOAT64
PROPERTYCOUNT	INT64

'Postcode' also removed even it is numerical since it's a categorical location label, not a meaningful numerical feature.



```
🔍 Average characteristics of each cluster:
            Rooms   Distance   Bedroom2   Bathroom        Car      Landsize  \
Cluster
1        2.692726  11.214004  2.800136  1.278722  1.700884    550.957852
2        3.333333  33.500000  3.333333  2.000000  4.333333  39312.333333
3        4.076046  11.216920  3.933460  2.140684  2.058935    595.365019

         BuildingArea    YearBuilt  Lattitude  Longtitude  Propertycount
Cluster
1          125.702243  1965.342624 -37.807155  144.992298    7823.066621
2          184.333333  1976.666667 -37.609163  144.684613    2861.333333
3          235.615970  1975.904943 -37.822490  145.031807    6903.233840
```

— Necessity of the Stardartization

In our dataset, features like 'Landsize' and 'BuildingArea' have much larger scales compared to features like 'Bathroom' or 'Car'.

Without standardization, these large-scale features would dominate the clustering process and distort the grouping structure.

Impact of Stardardization