IMMO ELIZA
DATA ANALYSIS

TEAM 5

Yusra
Zelim
Rasmita
Muntadher

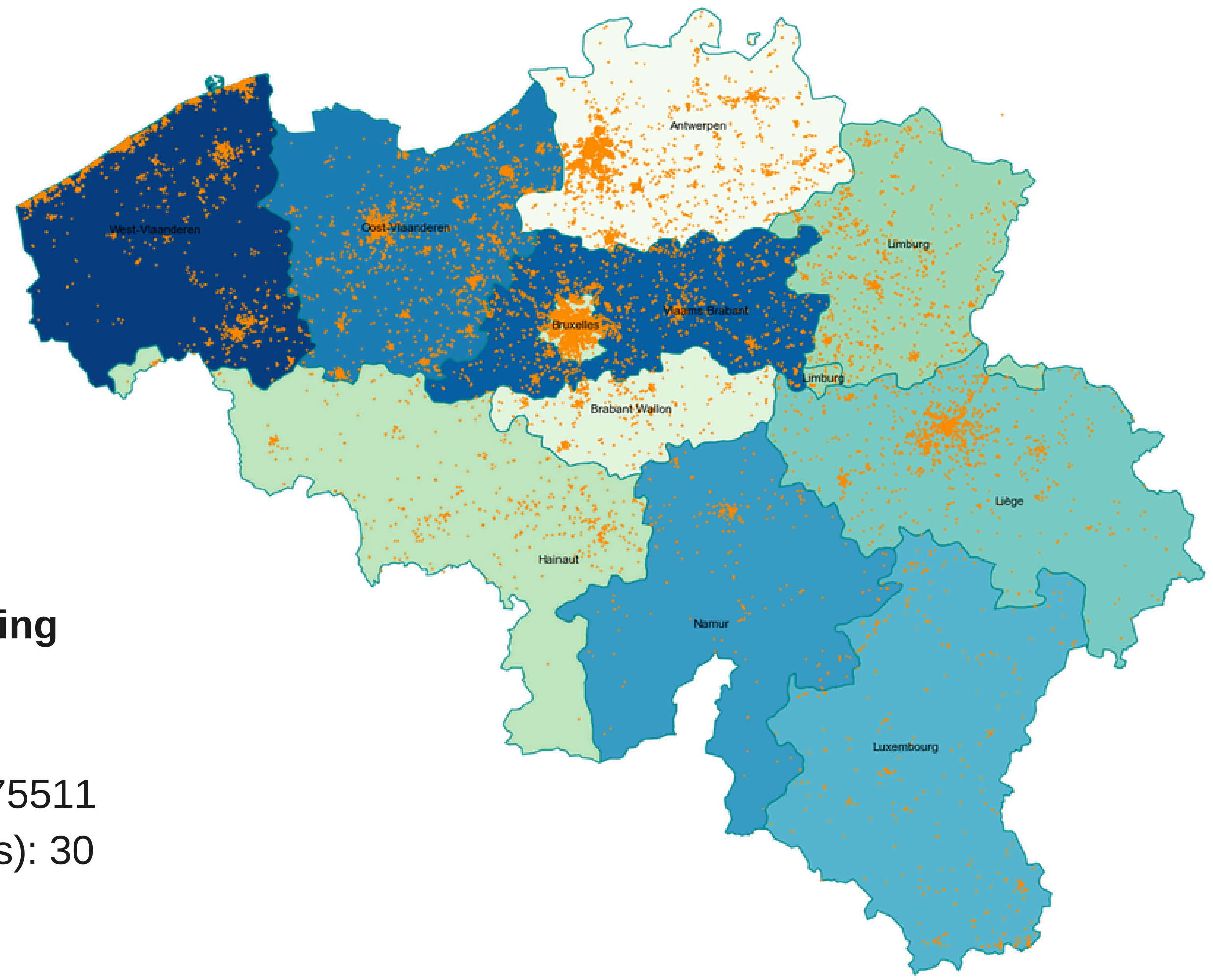
OBJECTIVE

Perform an initial analysis of the scraped data to create visualizations and establish the foundation for the Machine Learning model.

OBSERVATIONS AND FEATURES

Observations: These are the rows in your dataset, representing each individual instance .

Features: These are the columns in your dataset, representing the attributes or characteristics of each observation. For example, features in a real estate dataset could include "location," "price," "number of bedrooms.



Before Data Cleaning

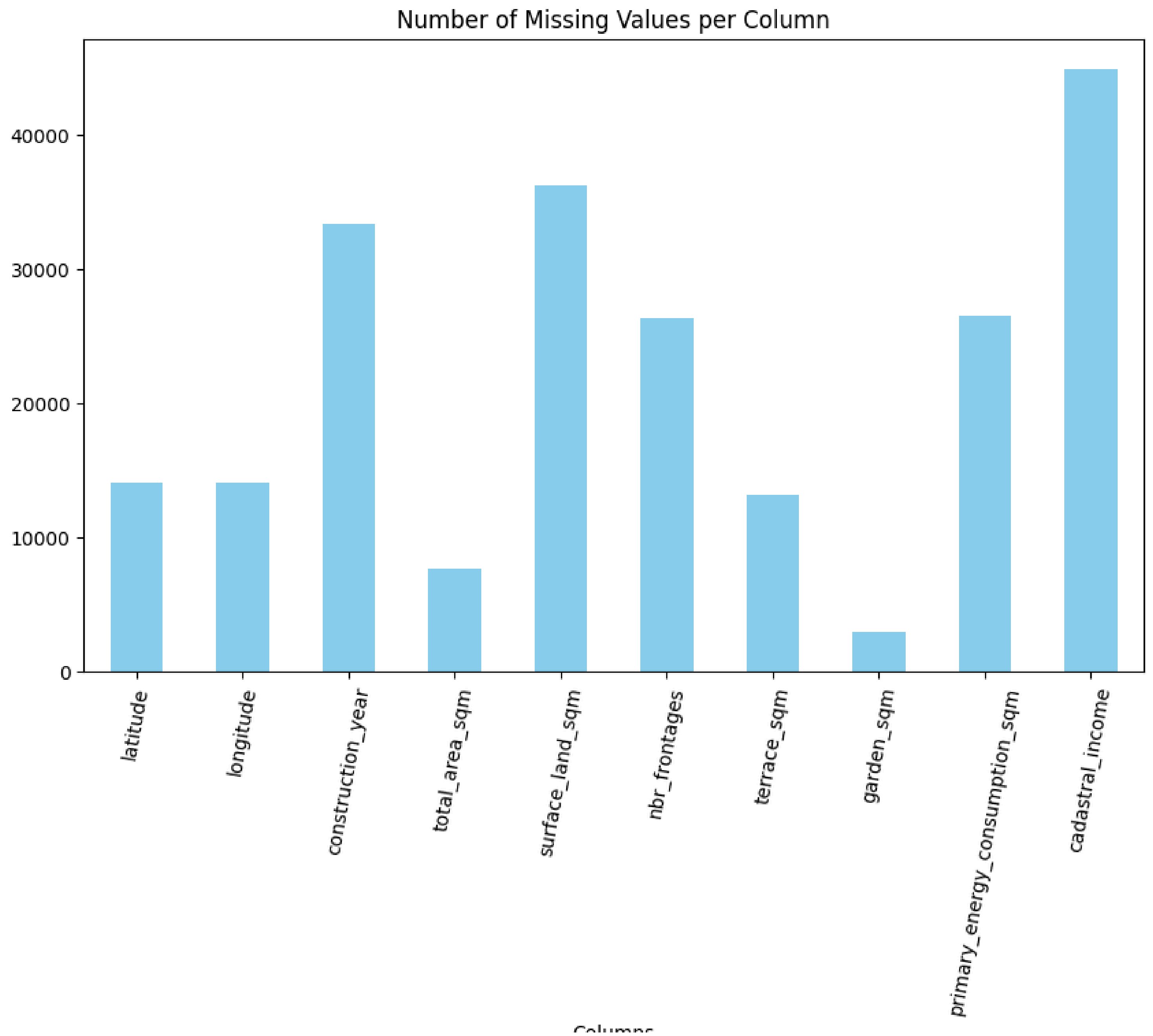
Number of (rows): 75511
Number of (columns): 30

After Data Cleaning

Number of (rows): 17258
Number of (columns): 23

THE PROPORTION OF MISSING VALUES PER COLUMN.

- Construction year 44%
- Total area sqm 10.08%
- Surface land sqm 48.01%
- Frontages 34.89%
- Terrace sqm 17.40%
- Garden sqm 3.89%
- Latitude 18.67%
- Longitude 18.67%
- Primary Energy Consumption sqm 35.18%
- Cadastral Income 59.55%



VARIABLES YOU WOULD DELETE .

Variable Deletion in Real Estate Dataset for Immo Eliza

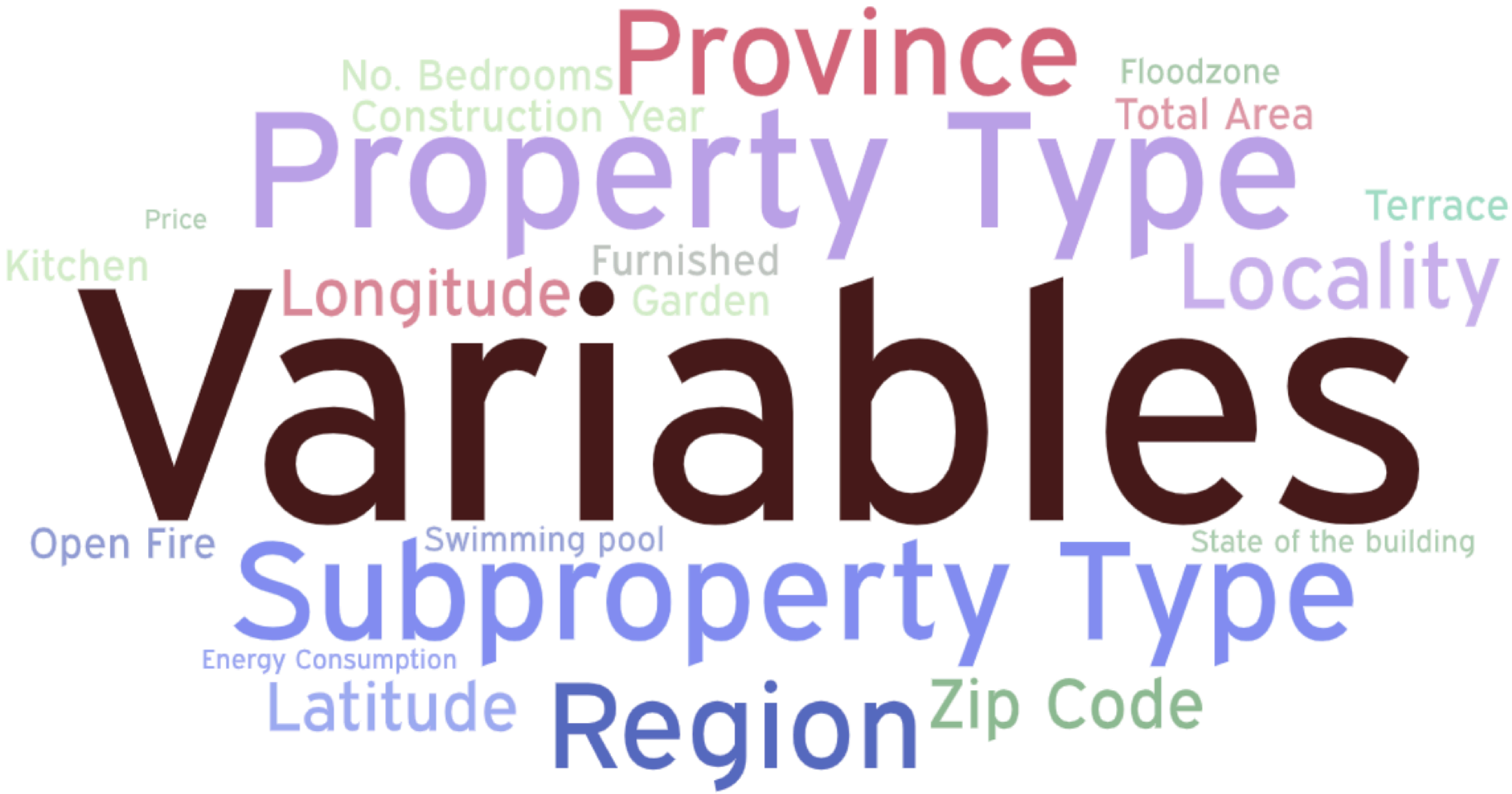
During this review, certain variables were identified as redundant or irrelevant for our specific objectives and were therefore removed. Also variables with a high percentage of missing values (e.g., more than 40%).

1. Surface land sqm
2. Frontages
3. Cadastral Home
4. Terrace FL
5. Garden FL
6. Primary Energy Consumption
7. Heating type

REMOVING THESE VARIABLES IS EXPECTED TO

- **Improve Model Simplicity:** Fewer variables reduce model complexity, which can lead to better interpretability and faster computation.

- **Focus on Key Features:** Concentrating on variables that have a more direct impact on property prices will help refine the accuracy of predictions.


Province
No. Bedrooms
Floodzone
Construction Year
Total Area
Property Type
Price
Terrace
Kitchen
Furnished
Locality
Longitude
Garden
Variables
Open Fire
Swimming pool
State of the building
Subproperty Type
Energy Consumption
Latitude
Region
Zip Code

QUANTITATIVE

- Price
- Total Area
- No. of bedrooms
- Terrace(sqm)
- Garden(sqm)
- Primary Energy Consumption per sqm

QUALITATIVE

Nominal

- ID
- Property Type
- Sub property type
- Location data

Ordinal

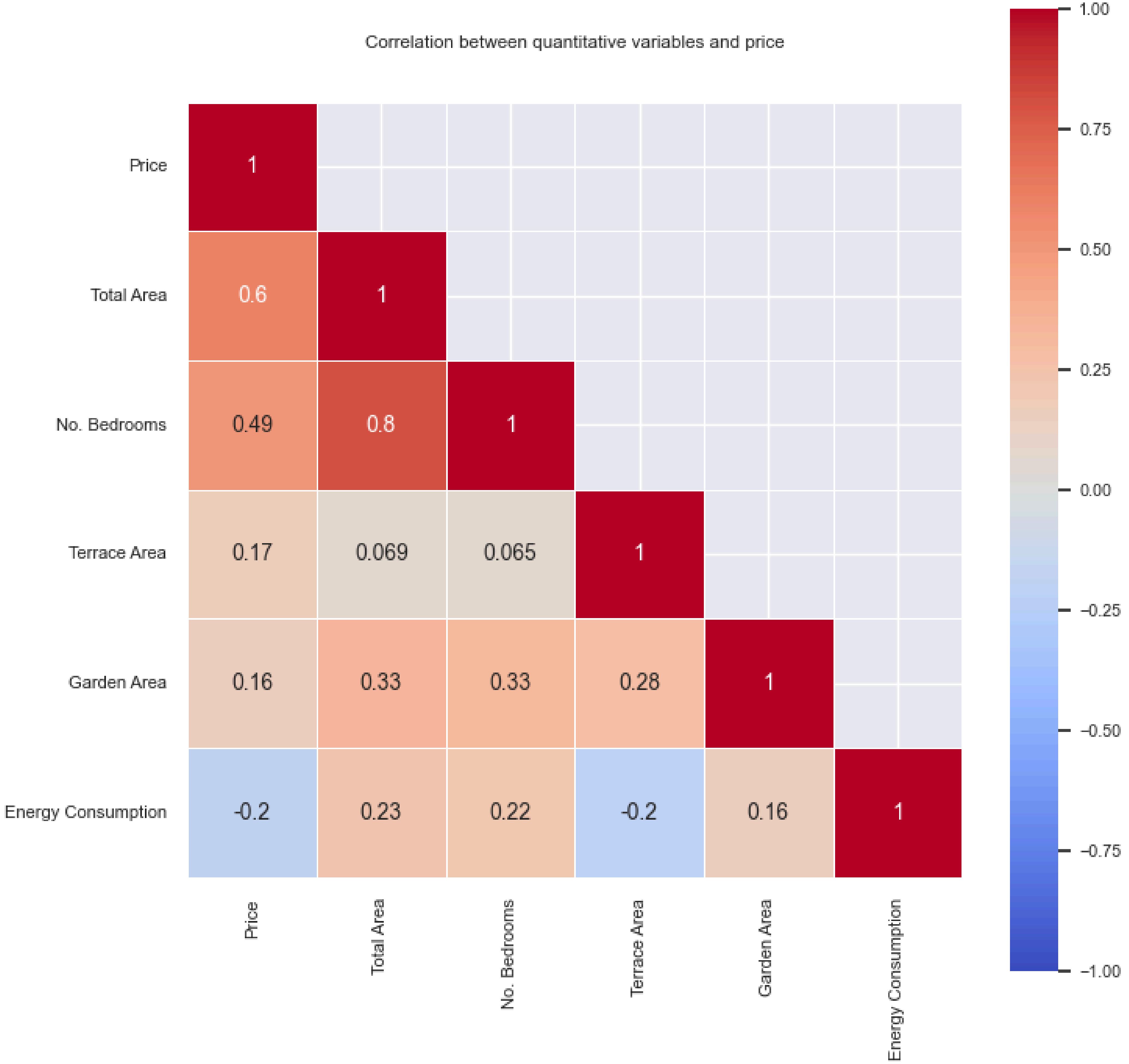
- Construction year
- State of building
- Kitchen level

Binary

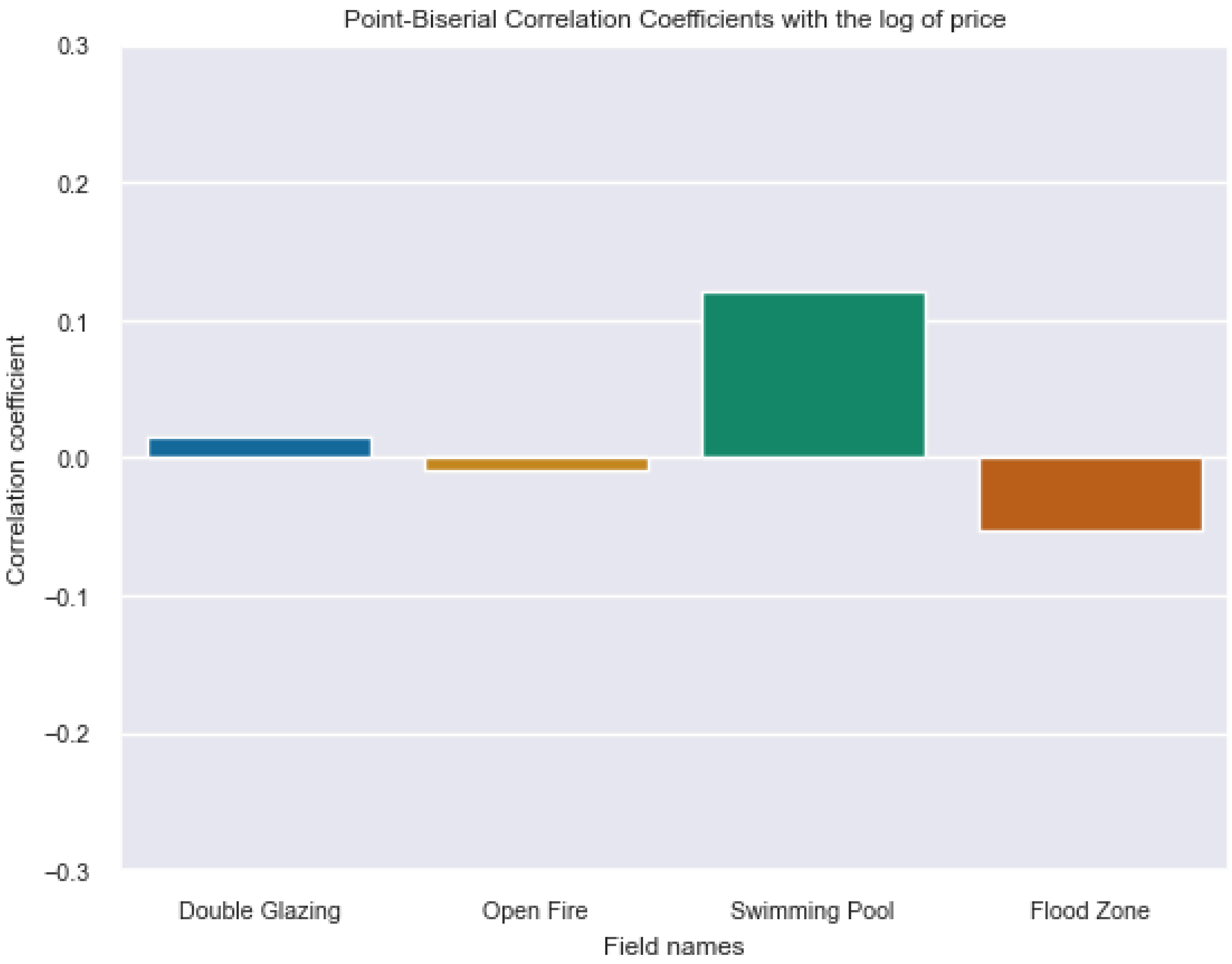
- Swimming pool
- Double glazing
- Furnished
- Open Fire
- Flood zone

CONVERSION METHODS

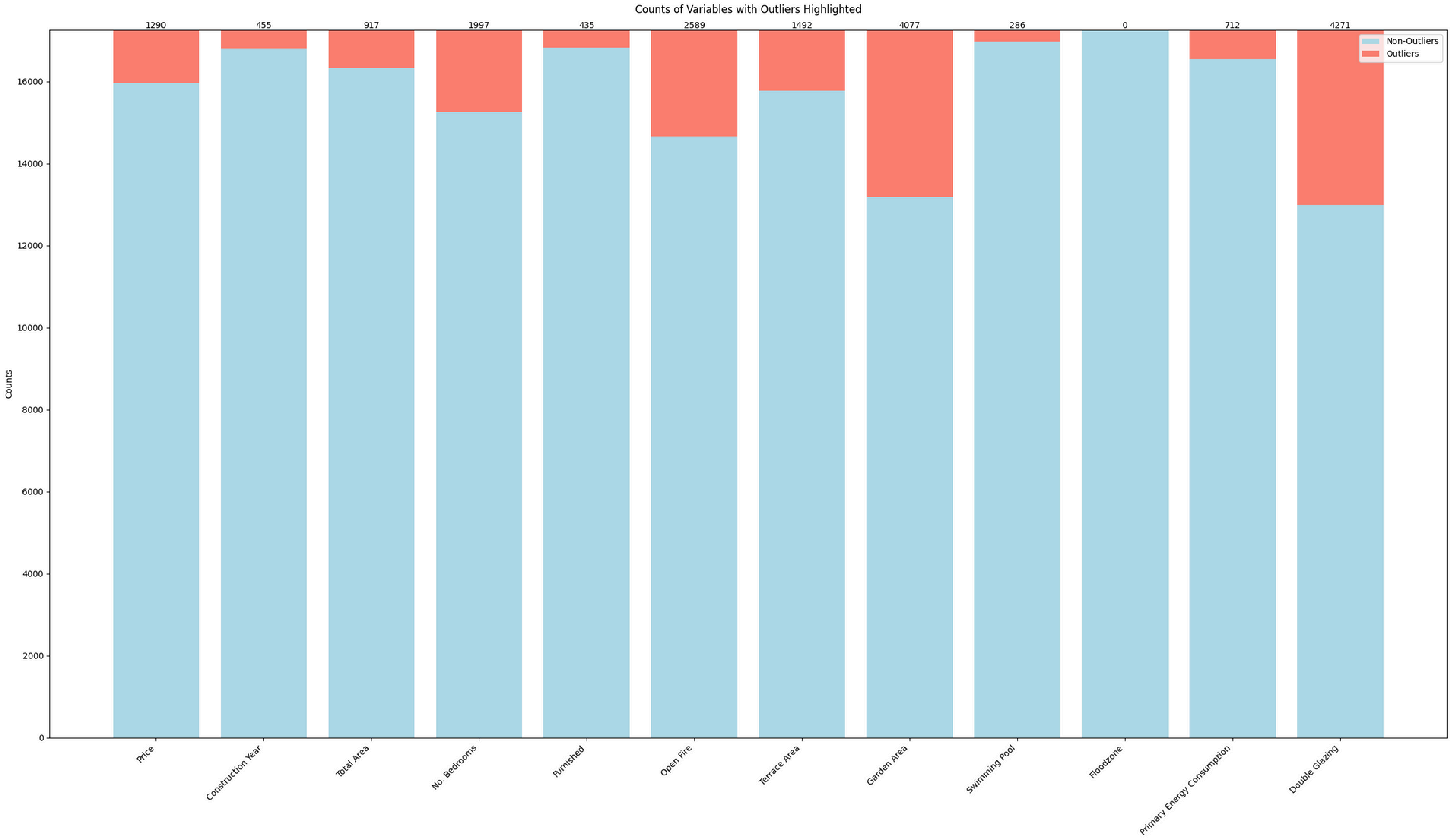
- Numeric Encoding
- Label Encoding
- One hot encoding
- Frequency encoding

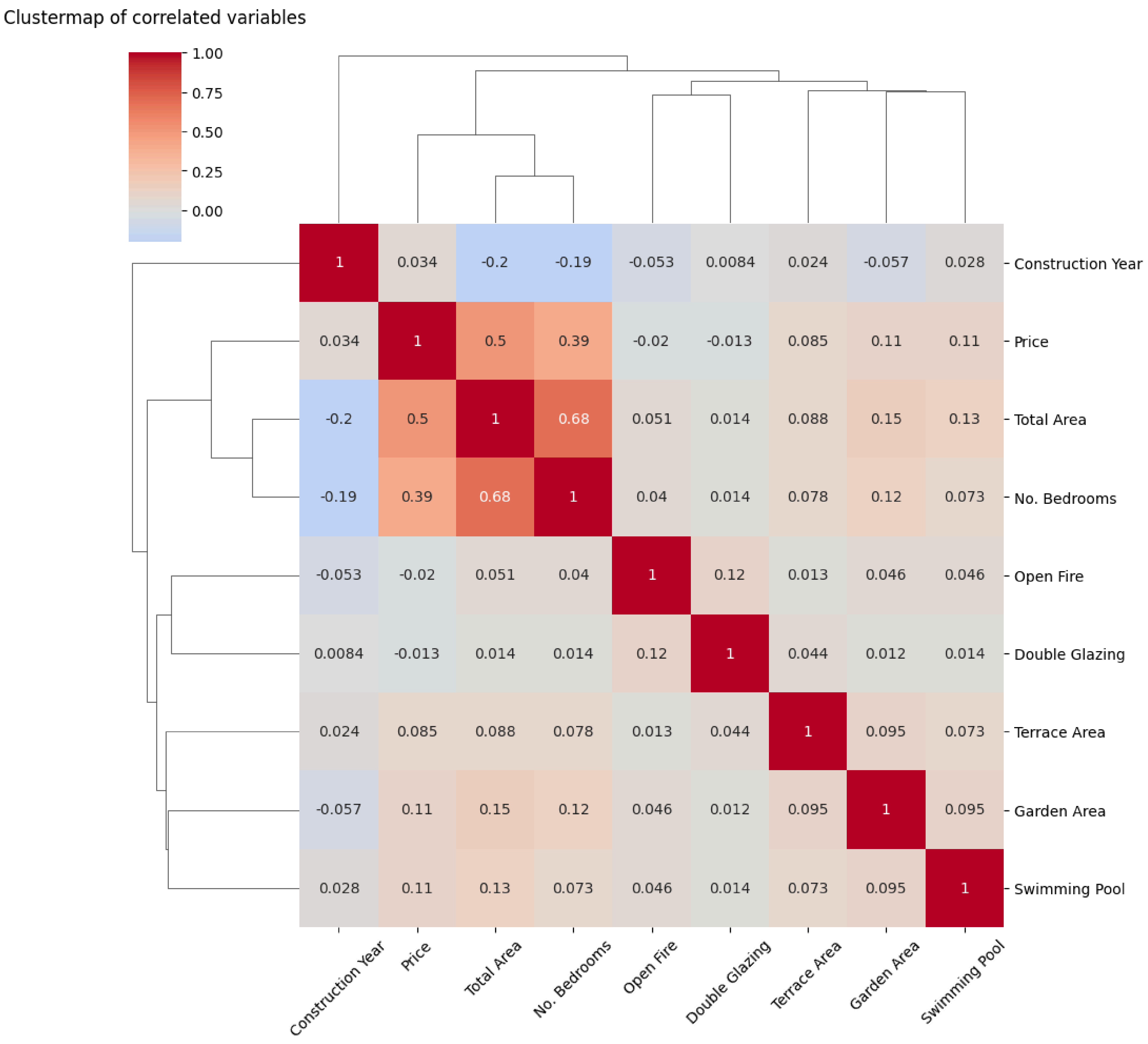


Method:Spearmans

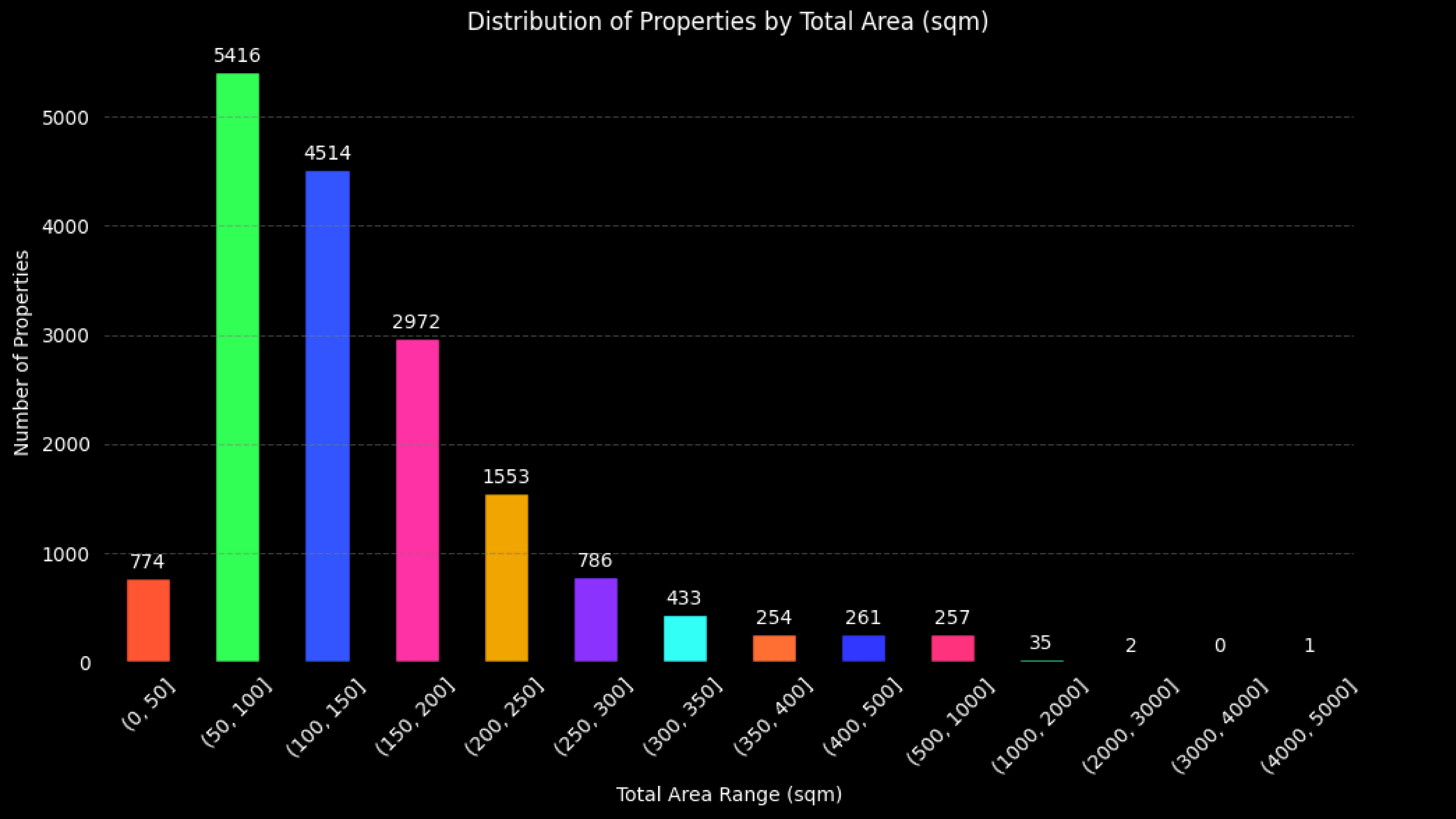


Method: Point biserial


Counts of Variables with Outliers Highlighted
1290
455
917
1997
435
2589
1492
4077
286
0
712
4271
Non-Outliers
Outliers
Counts
16000
14000
12000
10000
8000
6000
4000
2000
0
Price
Construction Year
Total Area
No. Bedrooms
Furnished
Open Fire
Terrace Area
Garden Area
Swimming Pool
Floodzone
Primary Energy Consumption
Double Glazing

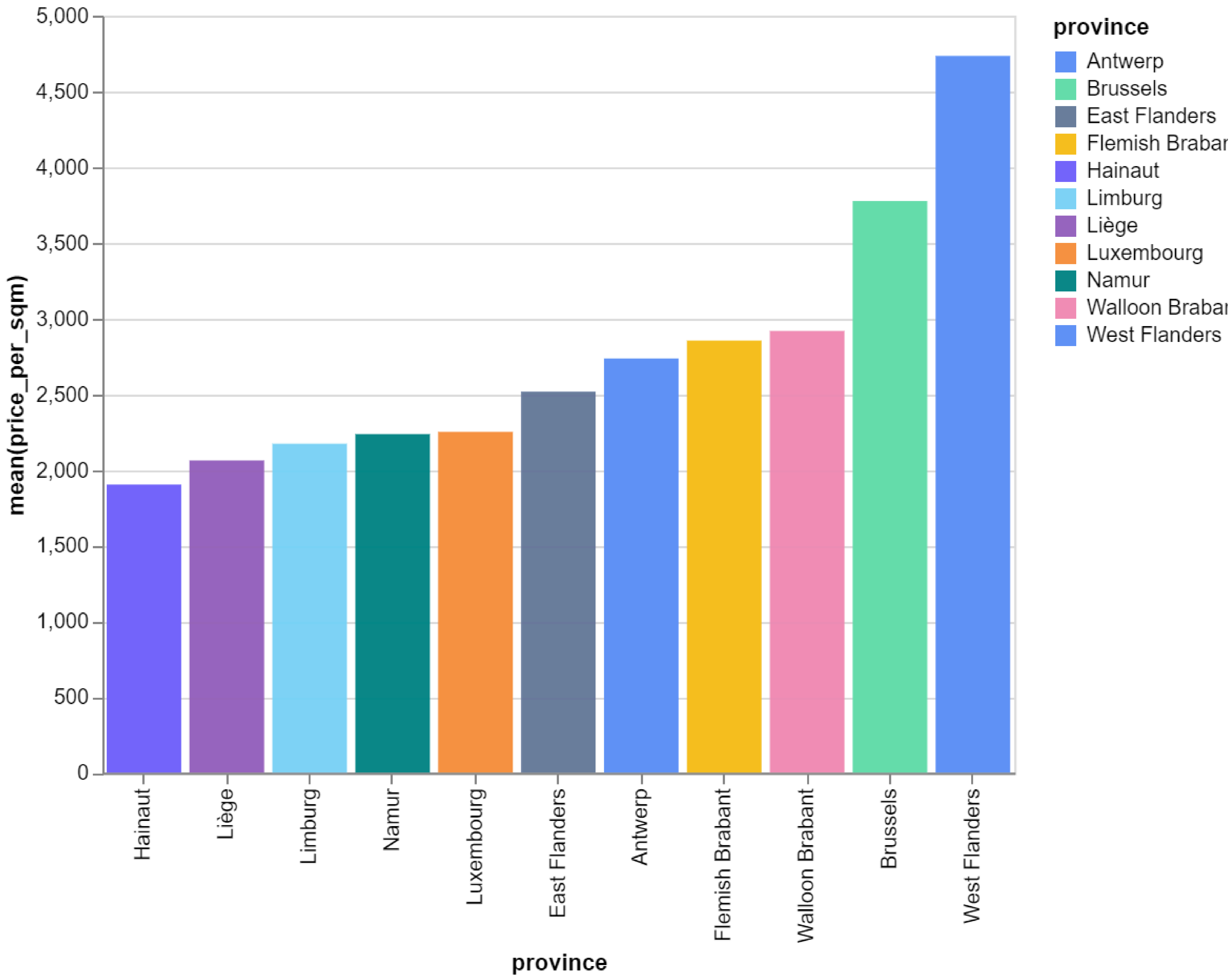

Clustermap of correlated variables
1.00
0.75
0.50
0.25
0.00
1 0.034 -0.2 -0.19 -0.053 0.0084 0.024 -0.057 0.028 Construction Year
0.034 1 0.5 0.39 -0.02 -0.013 0.085 0.11 0.11 Price
-0.2 0.5 1 0.68 0.051 0.014 0.088 0.15 0.13 Total Area
-0.19 0.39 0.68 1 0.04 0.014 0.078 0.12 0.073 No. Bedrooms
-0.053 -0.02 0.051 0.04 1 0.12 0.013 0.046 0.046 Open Fire
0.0084 -0.013 0.014 0.014 0.12 1 0.044 0.012 0.014 Double Glazing
0.024 0.085 0.088 0.078 0.013 0.044 1 0.095 0.073 Terrace Area
-0.057 0.11 0.15 0.12 0.046 0.012 0.095 1 0.095 Garden Area
0.028 0.11 0.13 0.073 0.046 0.014 0.073 0.095 1 Swimming Pool
Construction Year
Price
Total Area
No. Bedrooms
Open Fire
Double Glazing
Terrace Area
Garden Area
Swimming Pool

Top 5 variables

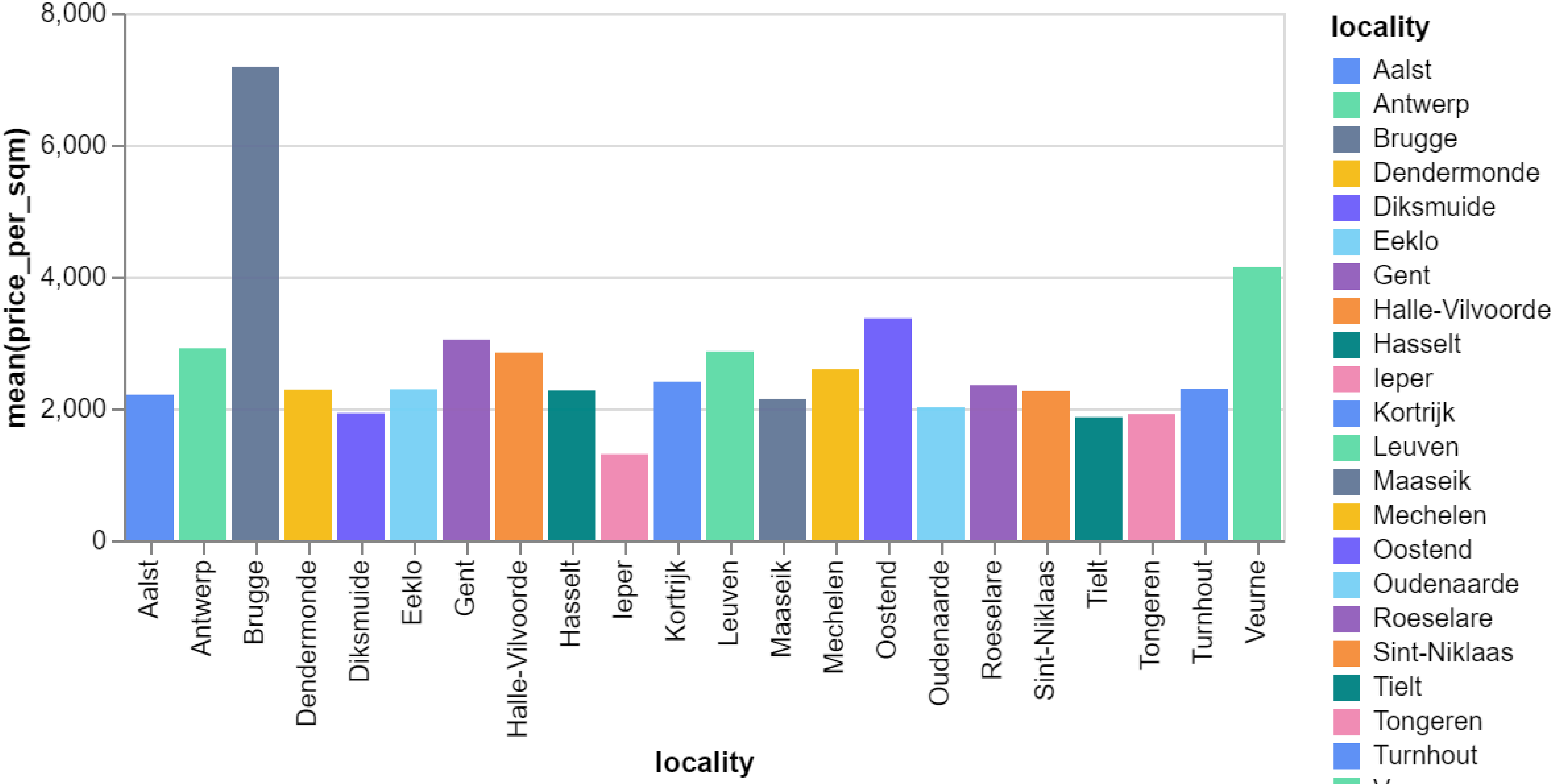


Distribution of Properties by Total Area (sqm)
Number of Properties
0
1000
2000
3000
4000
5000
774
5416
4514
2972
1553
786
433
254
261
257
35
2
0
1
(0, 50]
(50, 100]
(100, 150]
(150, 200]
(200, 250]
(250, 300]
(300, 350]
(350, 400]
(400, 500]
(500, 1000]
(1000, 2000]
(2000, 3000]
(3000, 4000]
(4000, 5000]
Total Area Range (sqm)

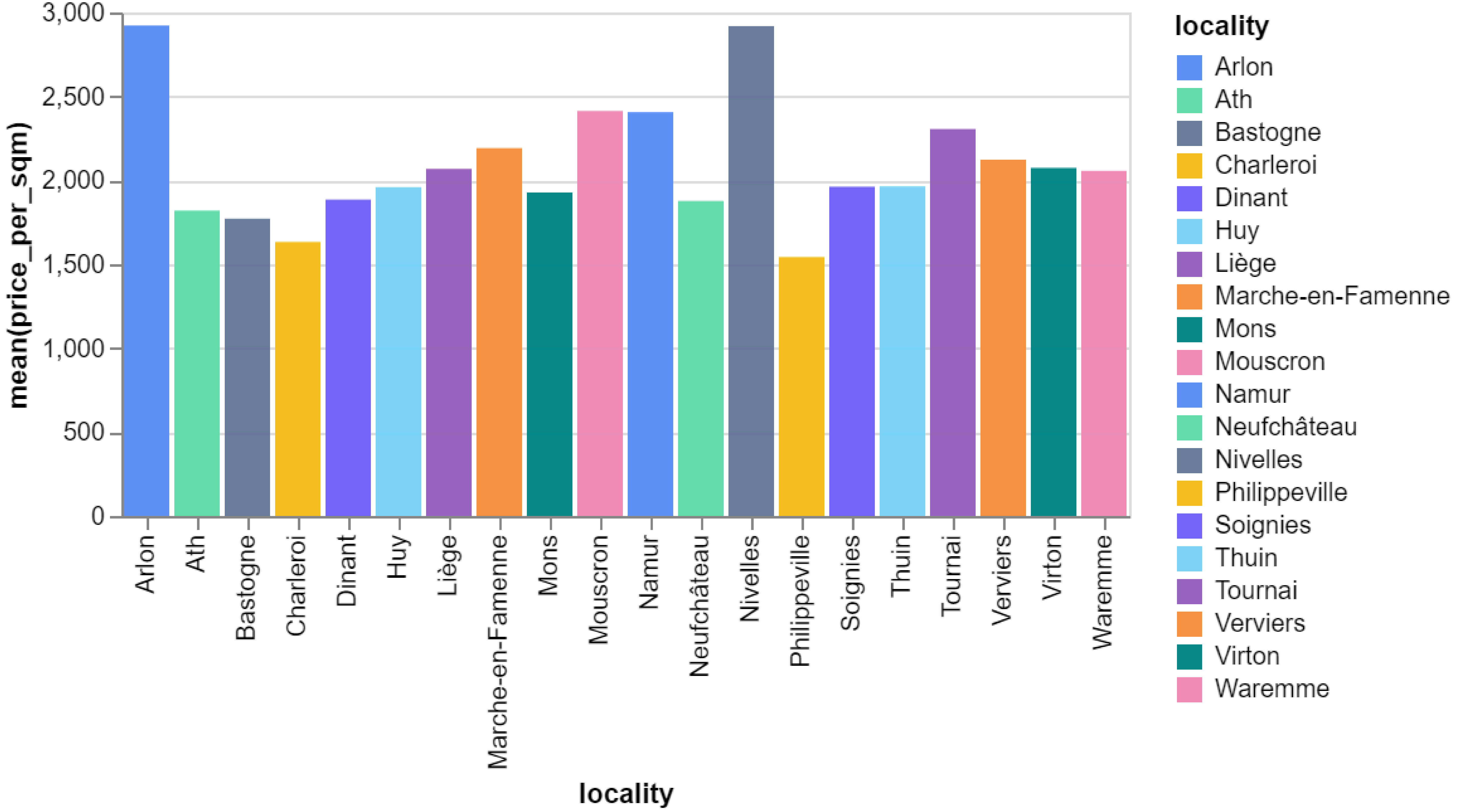
Average Prices per sqm in Belgium



Average Prices per sqm in Flanders



Average Prices per sqm in Wallonia



Conclusion

- **Price** is most correlated with the values of **total area** and **number of bedrooms**.
- Ordinal variables in our dataset have low correlation with price.
- **Garden area** has the most outliers.
- **Brugge** is the most expensive and **Ieper** is the least expensive locality in Flanders.
- **Nivelles** is the most expensive and **Philippeville** is the least expensive locality in Wallonia.
- **West Flanders** is the most expensive and **Hainaut** is the least expensive province in all of Belgium.

THANK YOU