renting_vizualisation

July 14, 2021

1 Exploration of all buildings in the Netherlands

1.1 by Yannick Mariman

1.2 Preliminary Wrangling

This dataset does preliminary explore a dataset of rented buildings in the Netherlands. The data and outcomes of this notebook are confidential should not be shared.

```
Index(['Bron', 'Bouwjaar', 'M2HuurPrijs', 'EnergieLabel',
       'GebruiksOppervlakte', 'AanmeldDatum', 'Postcode', 'TypeWoning',
       'Looptijd', 'TransactieHuurPrijs', 'TransactieDatumOndertekeningAkte',
       'OnderhoudsNiveauBinnen', 'OnderhoudsNiveauBuiten', 'GemeenteNaam',
       'GemeenteCat'],
      dtype='object')
```

1.2.1 What is the structure of your dataset?

```
(534085, 15)
Bron                                        object
Bouwjaar                                   float64
M2HuurPrijs                                float64
EnergieLabel                              category
GebruiksOppervlakte                          int32
AanmeldDatum                        datetime64[ns]
Postcode                                    object
TypeWoning                                  object
Looptijd                                   float64
TransactieHuurPrijs                        float64
TransactieDatumOndertekeningAkte    datetime64[ns]
OnderhoudsNiveauBinnen                    category
OnderhoudsNiveauBuiten                    category
GemeenteNaam                                object
GemeenteCat                               category
dtype: object
```

```
    Bron  Bouwjaar  M2HuurPrijs EnergieLabel  GebruiksOppervlakte  \
0  TIARA    1997.0     8.500000            B                  150
1  TIARA    1937.0     7.307692            C                  130
3  TIARA    1785.0    10.583333            C                   60
```

```
4  TIARA    1937.0     7.727273            D                  110
5  TIARA    2004.0     6.168831            B                  154
7  TIARA    1970.0     6.818182            C                  110
8  TIARA    1970.0     5.576923            D                  130
9  TIARA    1900.0     7.160000            G                  125

  AanmeldDatum Postcode           TypeWoning  Looptijd  TransactieHuurPrijs  \
0   2006-01-10   2851CH  2-onder-1-kapwoning      23.0               1275.0
1   2006-01-02   2025RB         Tussenwoning      25.0                950.0
3   2006-01-10   2312XV          Bovenwoning     163.0                635.0
4   2006-01-10   9722GP          Bovenwoning     133.0                850.0
5   2006-01-12   2152EX           Hoekwoning      53.0                950.0
7   2006-01-05   4538AM  2-onder-1-kapwoning     749.0                750.0
8   2006-01-05   7741GT          Bovenwoning      68.0                725.0
9   2006-01-13   2805AG           Hoekwoning      33.0                895.0

  TransactieDatumOndertekeningAkte OnderhoudsNiveauBinnen  \
0                       2006-02-02                   Goed
1                       2006-01-27                   Goed
3                       2006-06-22                   Goed
4                       2006-05-23                   Goed
5                       2006-03-06             Uitstekend
7                       2008-01-24                   Goed
8                       2006-03-14                   Goed
9                       2006-02-15                   Goed

  OnderhoudsNiveauBuiten    GemeenteNaam GemeenteCat
0                   Goed  Krimpenerwaard      Overig
1                   Goed         Haarlem         G40
3                   Goed          Leiden         G40
4                   Goed       Groningen         G40
5             Uitstekend  Haarlemmermeer         G40
7                   Goed       Terneuzen      Overig
8                   Goed       Coevorden      Overig
9                   Goed           Gouda         G40
```

There are 15 features in the dataset:

Bron is the source from which the data is coming.

Bouwjaar is the year in which the building is build.

M2HuurPrijs is the square metre price for renting.

GebruiksOppervlakte is the usable square metres in a house.

AanmeldDatum is the date at which the building is put online.

Postcode is zipcode/postalcode.

TypeWoning is the type of house people live in.

Looptijd is the time between putting the house online and the signing of the contract.

TransactieHuurPrijs is the total amount of euros paid.

TransactieDatumOndertekeningAkte is the date that the contract is signed.

OnderhoudsNiveaBinnen is the maintenance level inside of the building.

OnderhoudsNiveaBuiten is the maintenance level outside of the building.

GemeenteNaam is the name of the municipality.

GemeenteCat is the category of municipality. The big four, number 5 until 45 and the rest.

OnderhoudsNiveauBinnen, OnderhoudsNiveauBuiten and Energielabels are ordinal features. They are sorted from 'bad' to 'better'. In the case of GemeenteCat (Municipality Category) it is sorted from bigger municipality category to smaller. 'EnergieLabel': ['G','F','E','D','C', 'B', 'A'] 'OnderhoudsNiveauBinnen': ['Slecht', 'Slecht tot matig', 'Matig', 'Matig tot redelijk', 'Redelijk', 'Redelijk tot goed', 'Goed', 'Goed tot uitstekend', 'Uitstekend'] 'OnderhoudsNiveauBuiten': ['Slecht', 'Slecht tot matig', 'Matig', 'Matig tot redelijk', 'Redelijk', 'Redelijk tot goed', 'Goed', 'Goed tot uitstekend', 'Uitstekend'] 'GemeenteCat': ['G4', 'G40', 'Overig']

1.2.2 What is/are the main feature(s) of interest in your dataset?

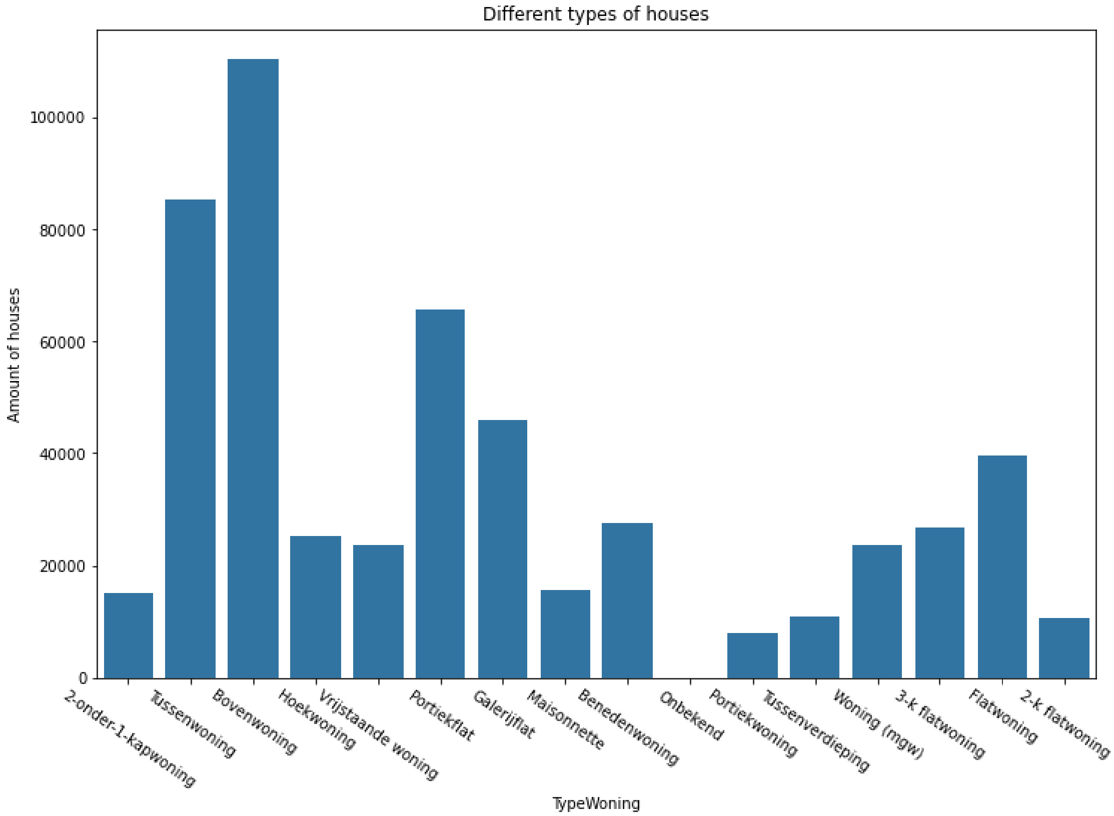
The main feature of interest is 'M2HuurPrijs', else 'TransactieHuurPrijs'. Translated this means rental price which is actually being paid.

1.2.3 What features in the dataset do you think will help support your investigation into your feature(s) of interest?

The assumption is that Gebruiksoppervlakte (square metres), WijkNaam (districtname) and TypeWoning (House type) have 'relatively much' influence. The following features also have some influence but less EnergieLabel (EnergyLabel), SoortObject, OnderhoudsNiveauBinnen (maintenance level inside) and OnderhoudsNiveauBuiten (maintenance level outside).

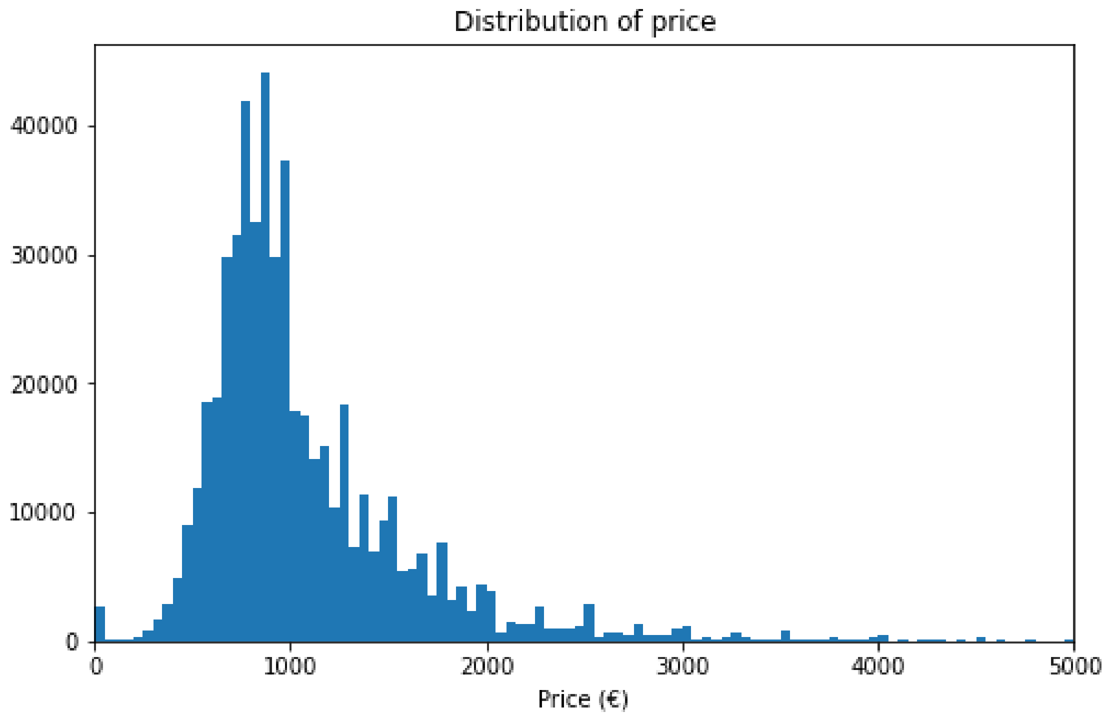
It is hard to find enough datapoints in the WijkNaam column. Therefore, the main focus of the Univariate Exploration will be TypeWoning, EnergieLabel and GebruiksOppervlakte.

1.3 Univariate Exploration

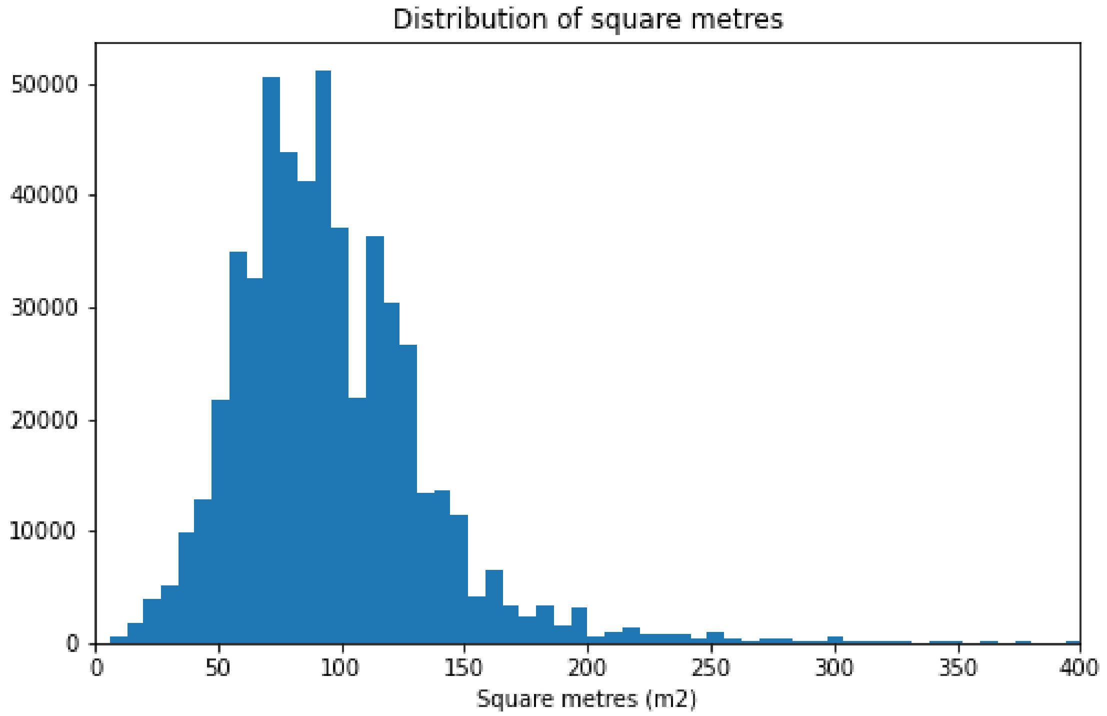


```
C    249838
A    121523
B     58778
D     42996
E     29412
F     17224
G     14314
Name: EnergieLabel, dtype: int64
```

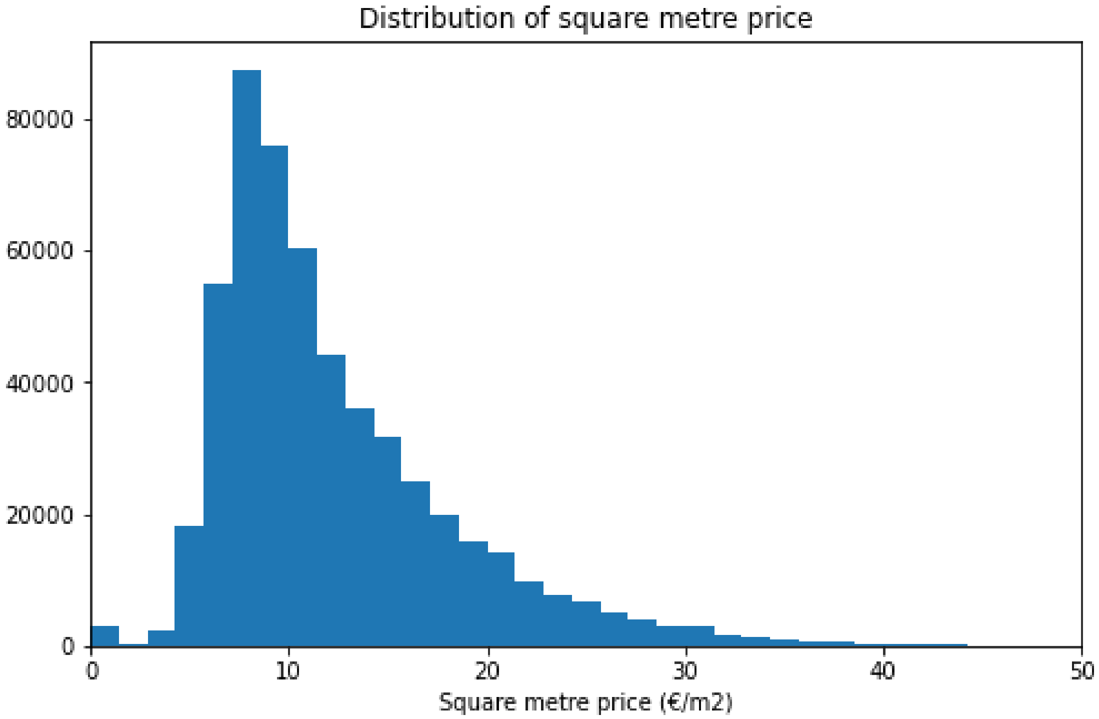
Better energylabels are more common. Recall that A is best and G is worst.



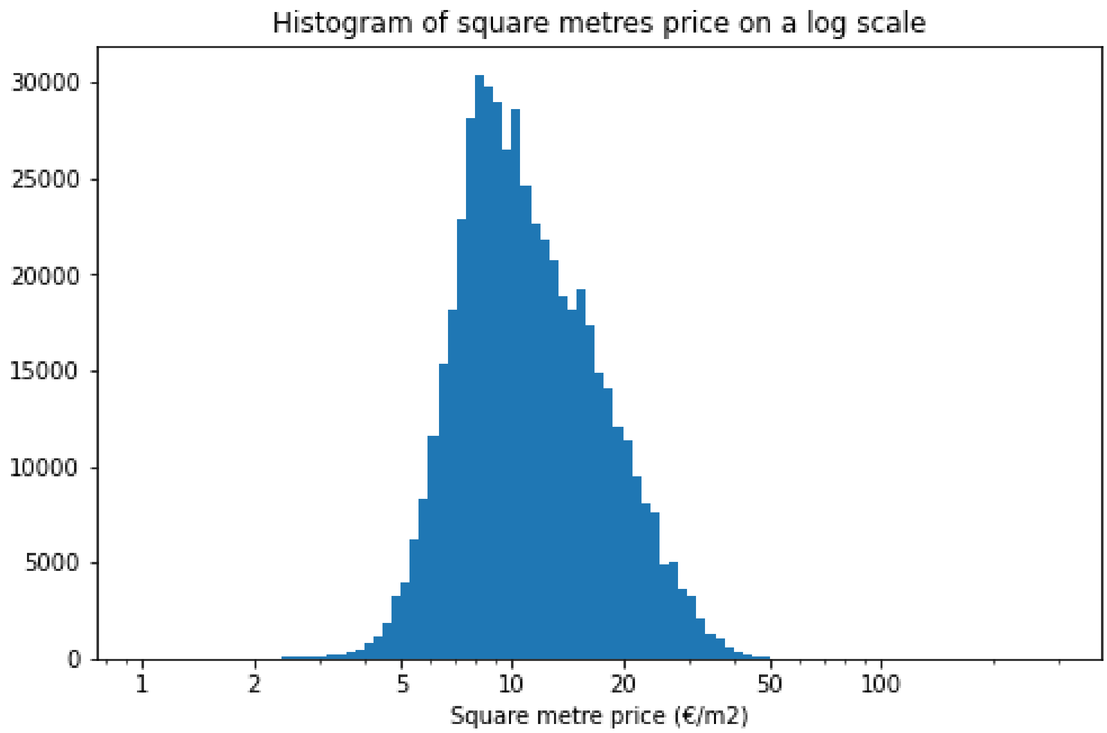
Long tail on the price distribution. Maybe the square metre price can give us more information



Almost all houses are between 30 and 150 square meters. Long tail which are probably some outliers.



Really long tail. Try this with a logscale.



```
Overig    194919
G4        170945
G40       168221
Name: GemeenteCat, dtype: int64
```

This shows that there are roughly an equal amount of rental transactions in the G4, G40 and the rest.

1.3.1 Discuss the distribution(s) of your variable(s) of interest. Were there any unusual points? Did you need to perform any transformations?

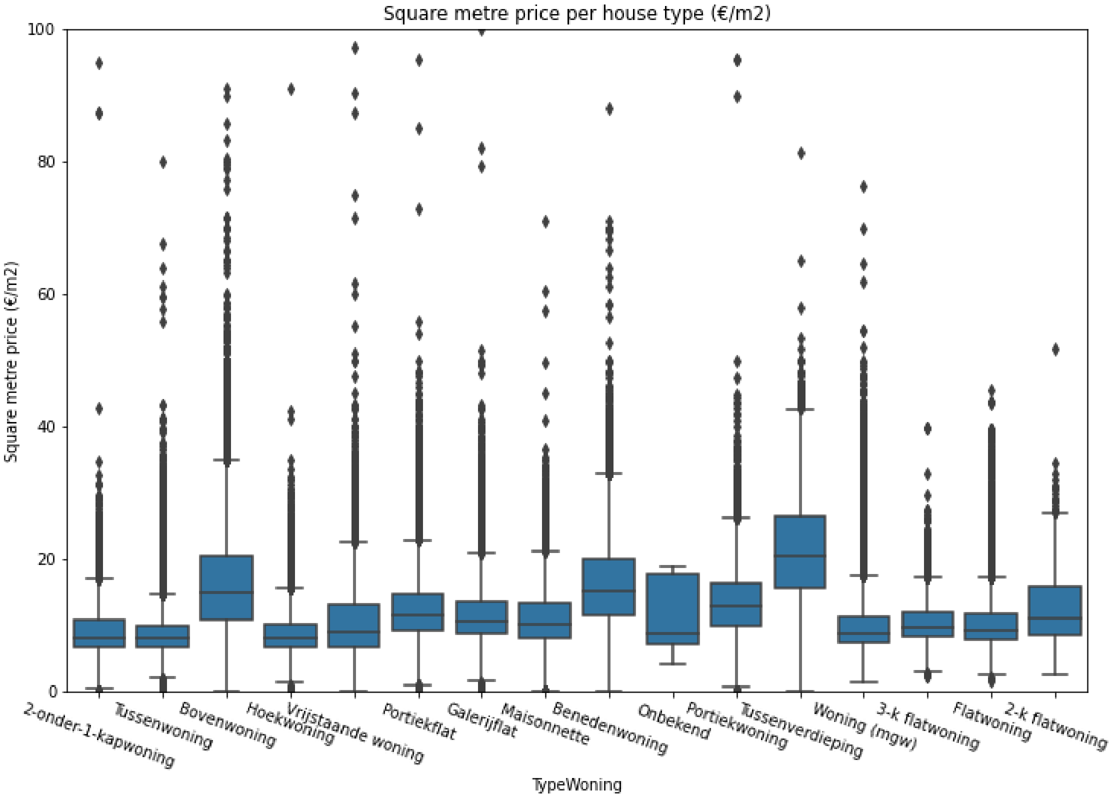
There is a log transformation on the M2HuurPrijs column.

1.3.2 Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

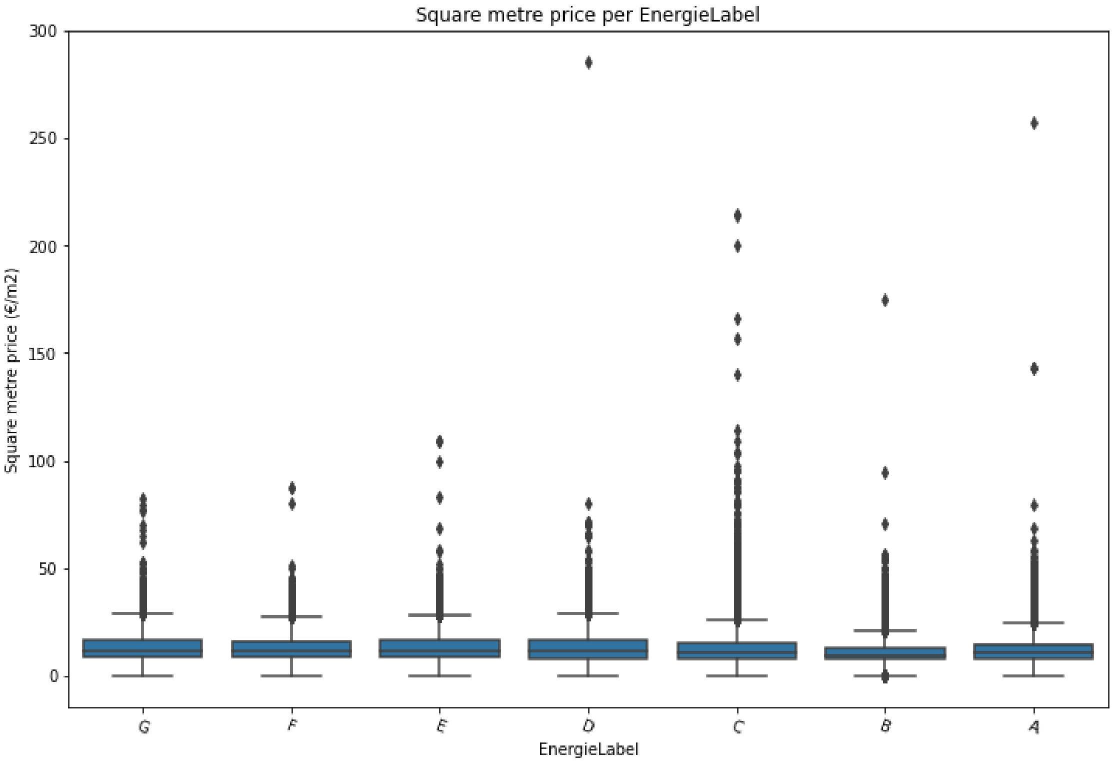
Since it was my own dataset. All the operations are performed in the beginning.

1.4 Bivariate Exploration

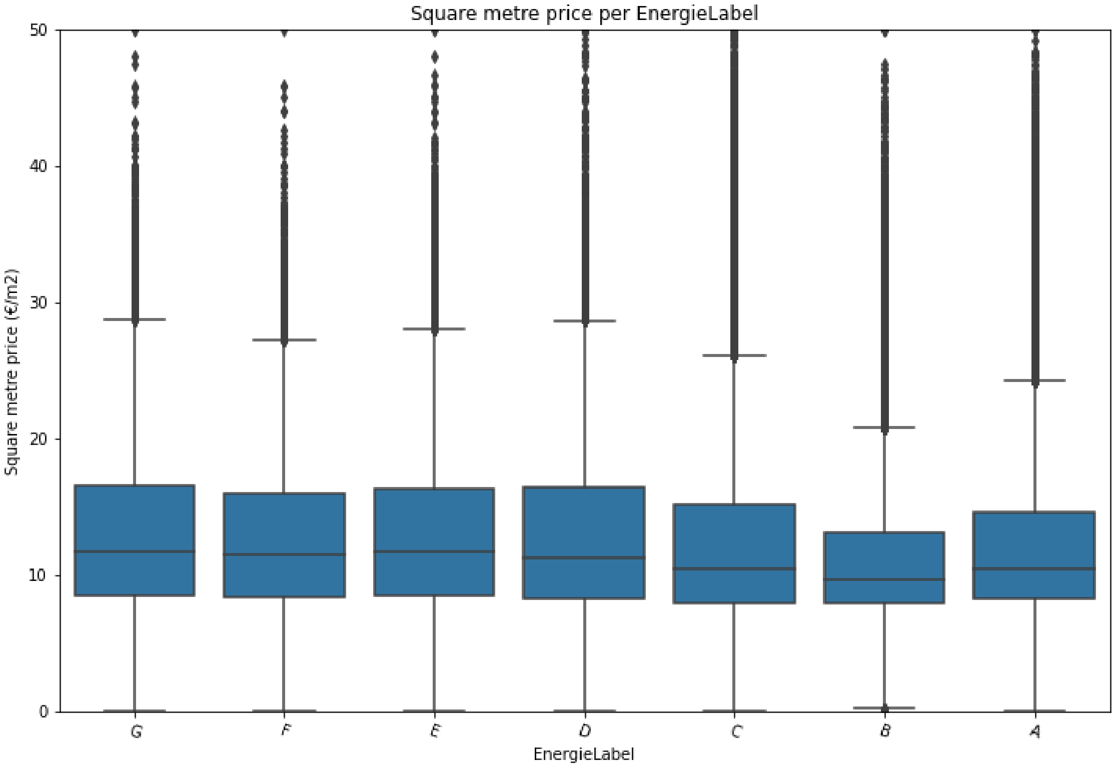
In this section, investigate relationships between pairs of variables in your data. Make sure the variables that you cover here have been introduced in some fashion in the previous section (univariate exploration).



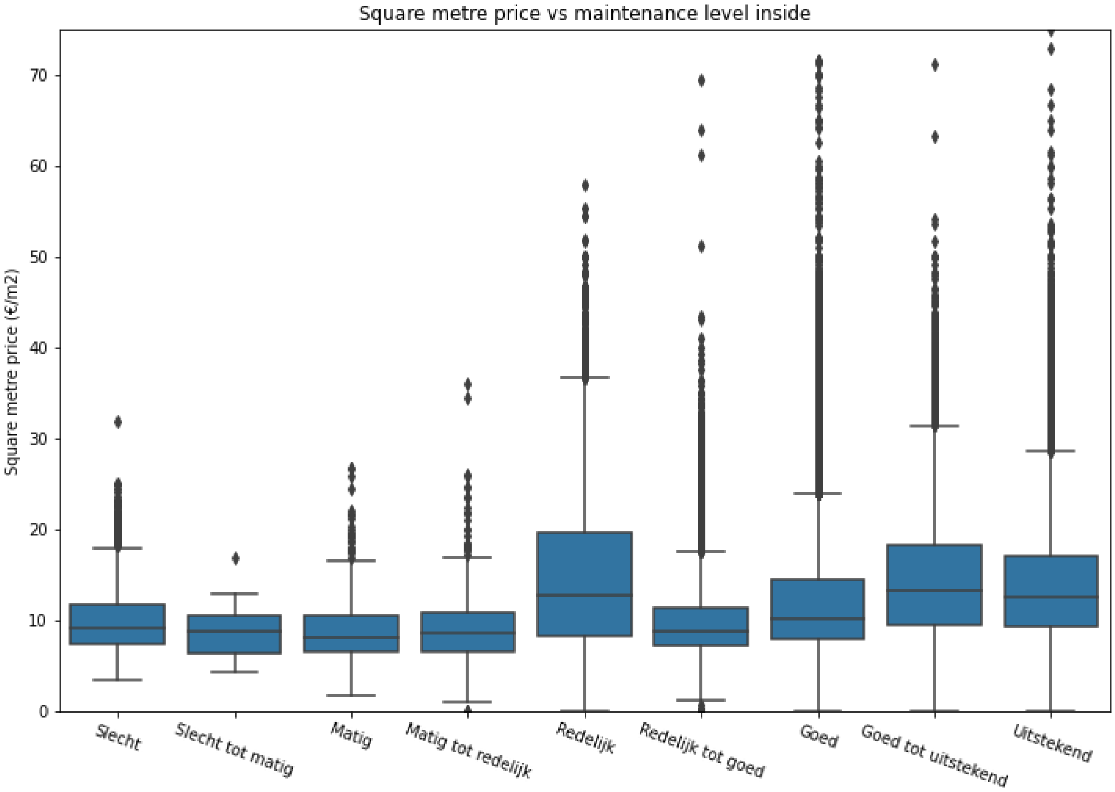
From the data above it is evident that there is no significant difference between the types. There are a lot of outliers in every category, especially ‘Bovenwoning’ has a lot of outliers.

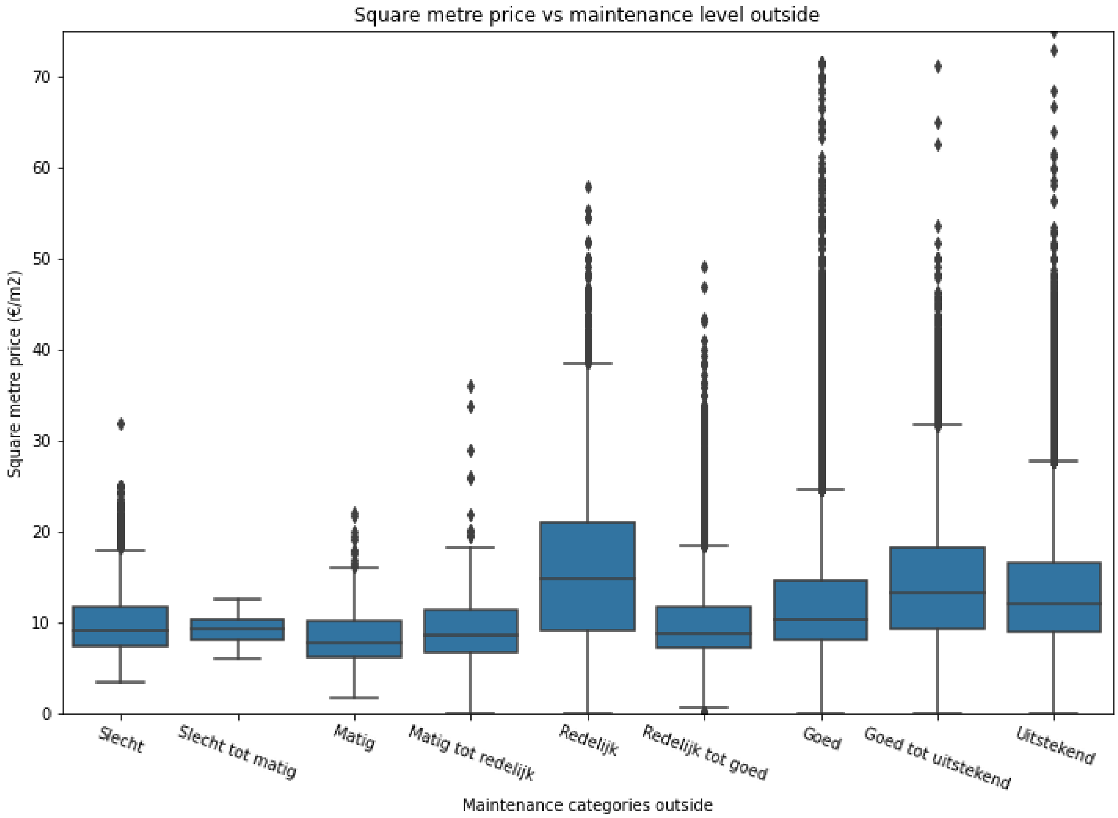


Not a lot of difference is visible. Maybe this is clearer when M2HuurPrijs is capped.

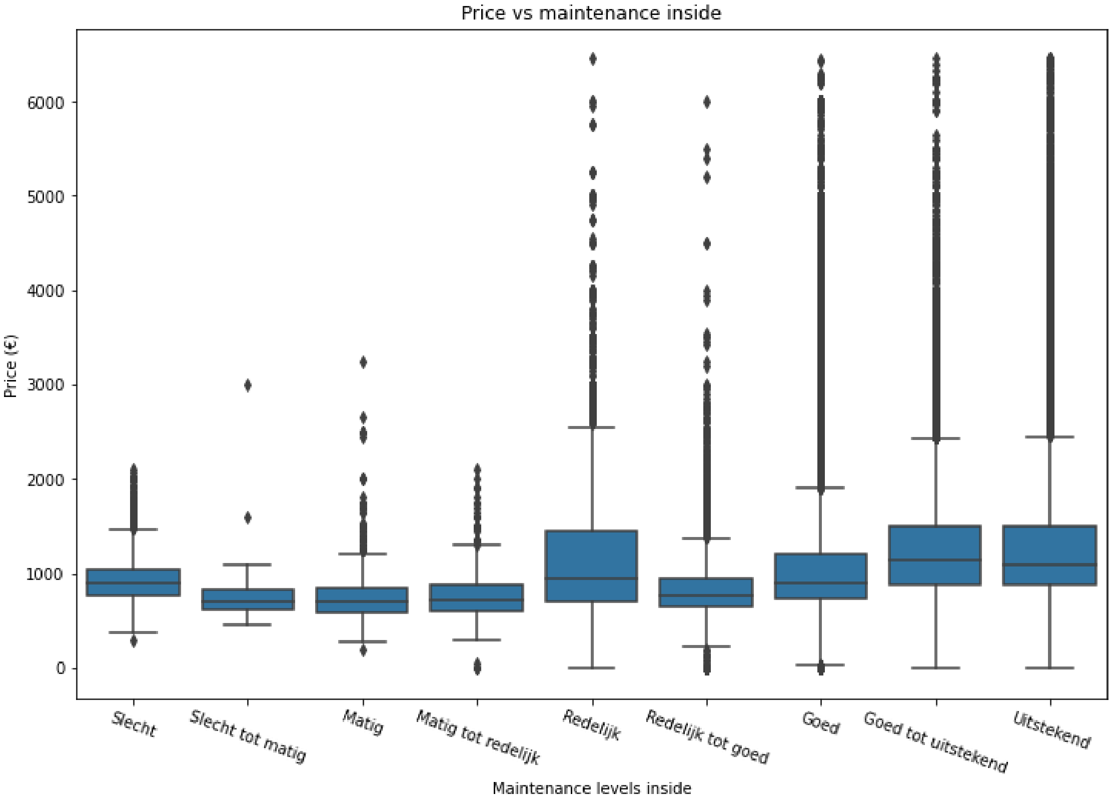


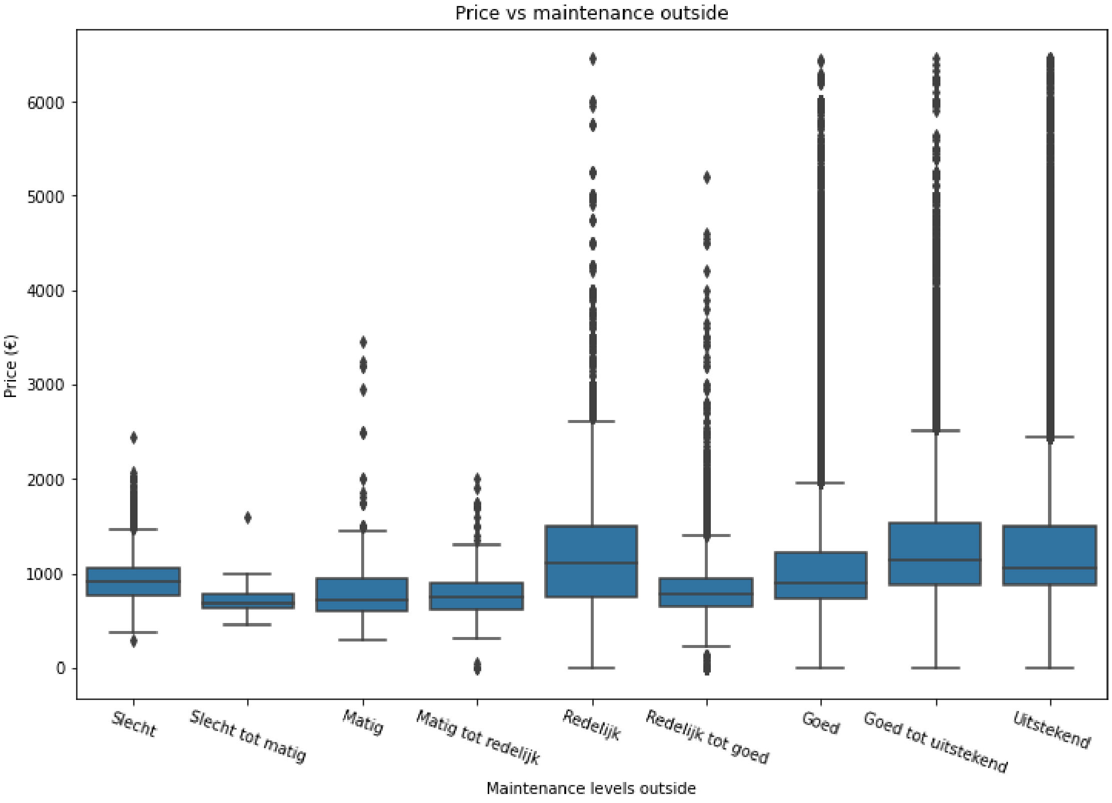
Closer investigation shows that the M2HuurPrijs (recall that this is square metre price) also shows no difference.

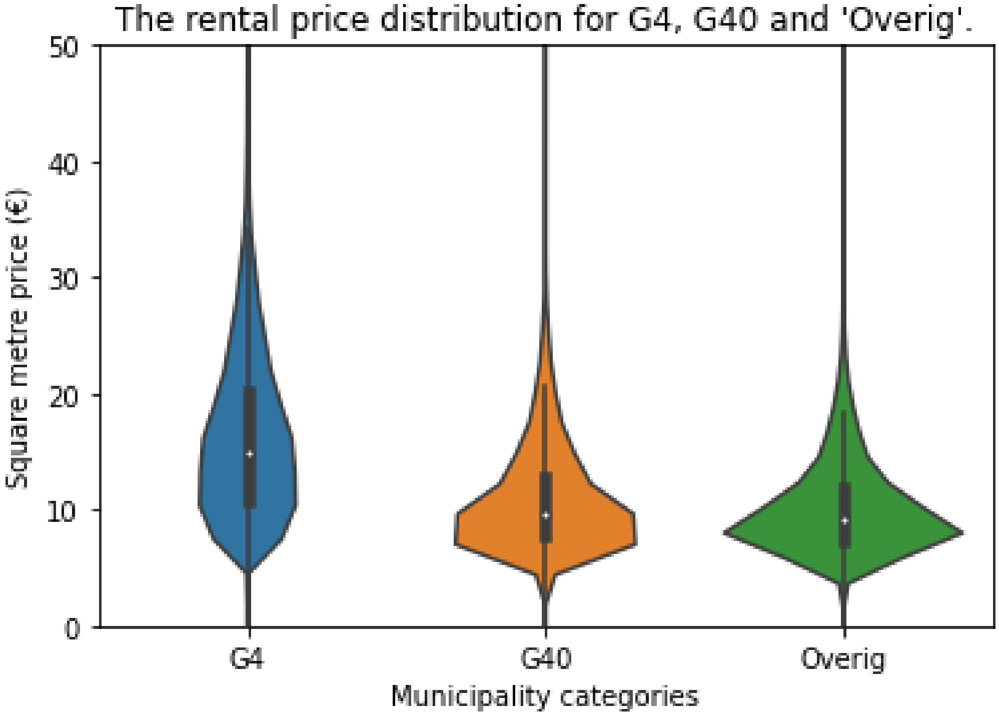




Maintenance levels inside and outside show the same trend. Therefore, I won't make any difference in this in further exploration.


Price vs maintenance inside
Price (€)
6000
5000
4000
3000
2000
1000
0
Slecht
Slecht tot matig
Matig
Matig tot redelijk
Redelijk
Redelijk tot goed
Goed
Goed tot uitstekend
Uitstekend
Maintenance levels inside


Price vs maintenance outside
Price (€)
6000
5000
4000
3000
2000
1000
0
Slecht
Slecht tot matig
Matig
Matig tot redelijk
Redelijk
Redelijk tot goed
Goed
Goed tot uitstekend
Uitstekend
Maintenance levels outside


The rental price distribution for G4, G40 and 'Overig'.
Square metre price (€)
50
40
30
20
10
0
G4
G40
Overig
Municipality categories

The distributions of G40 and Overig are somewhat comparable. However, G4 has a completely different shape. The shape is longer which means that there is more variation of the square metre prices in the G4. Whereas a mean is more obvious in the other two categories.

```
    Bron  Bouwjaar  M2HuurPrijs EnergieLabel  GebruiksOppervlakte  \
0  TIARA    1997.0     8.500000            B                  150
1  TIARA    1937.0     7.307692            C                  130
3  TIARA    1785.0    10.583333            C                   60
4  TIARA    1937.0     7.727273            D                  110
5  TIARA    2004.0     6.168831            B                  154

  AanmeldDatum Postcode           TypeWoning  Looptijd  TransactieHuurPrijs  \
0   2006-01-10   2851CH  2-onder-1-kapwoning      23.0               1275.0
1   2006-01-02   2025RB         Tussenwoning      25.0                950.0
3   2006-01-10   2312XV          Bovenwoning     163.0                635.0
4   2006-01-10   9722GP          Bovenwoning     133.0                850.0
5   2006-01-12   2152EX           Hoekwoning      53.0                950.0

  TransactieDatumOndertekeningAkte OnderhoudsNiveauBinnen  \
0                       2006-02-02                   Goed
1                       2006-01-27                   Goed
3                       2006-06-22                   Goed
4                       2006-05-23                   Goed
5                       2006-03-06             Uitstekend

  OnderhoudsNiveauBuiten    GemeenteNaam GemeenteCat
0                   Goed  Krimpenerwaard      Overig
1                   Goed         Haarlem         G40
3                   Goed          Leiden         G40
4                   Goed       Groningen         G40
5             Uitstekend  Haarlemmermeer         G40
```

1.4.1 Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

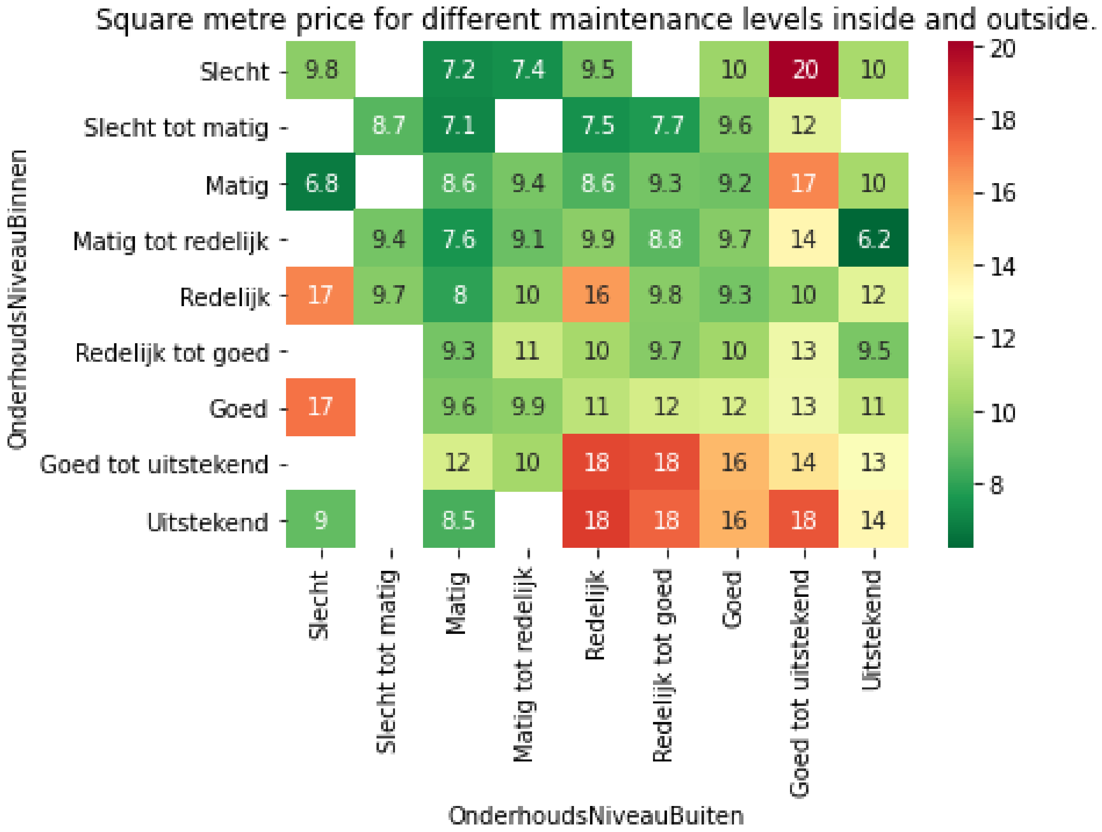
Maintenance inside and outside have no significant different effect. Plotting for maintenance vs price/price per square metre seems to show the same trend.

1.4.2 Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

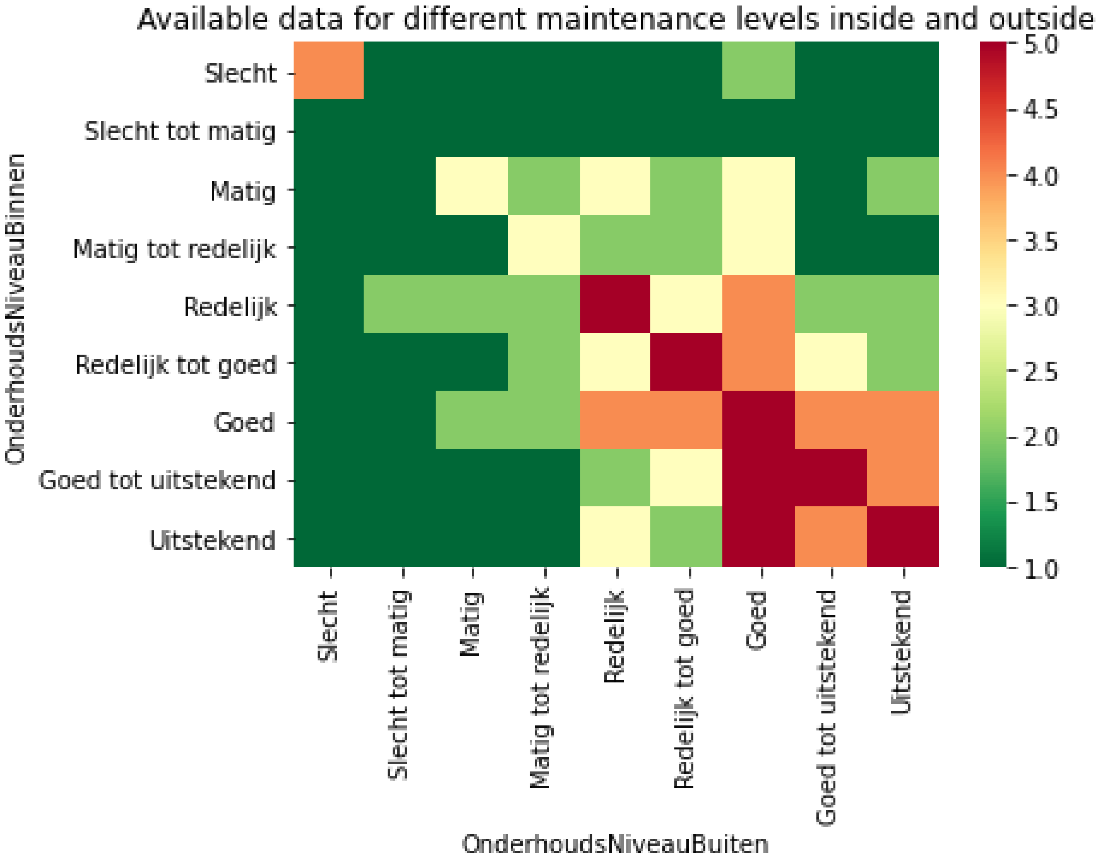
no

1.5 Multivariate Exploration

Create plots of three or more variables to investigate your data even further. Make sure that your investigations are justified, and follow from your work in the previous sections.

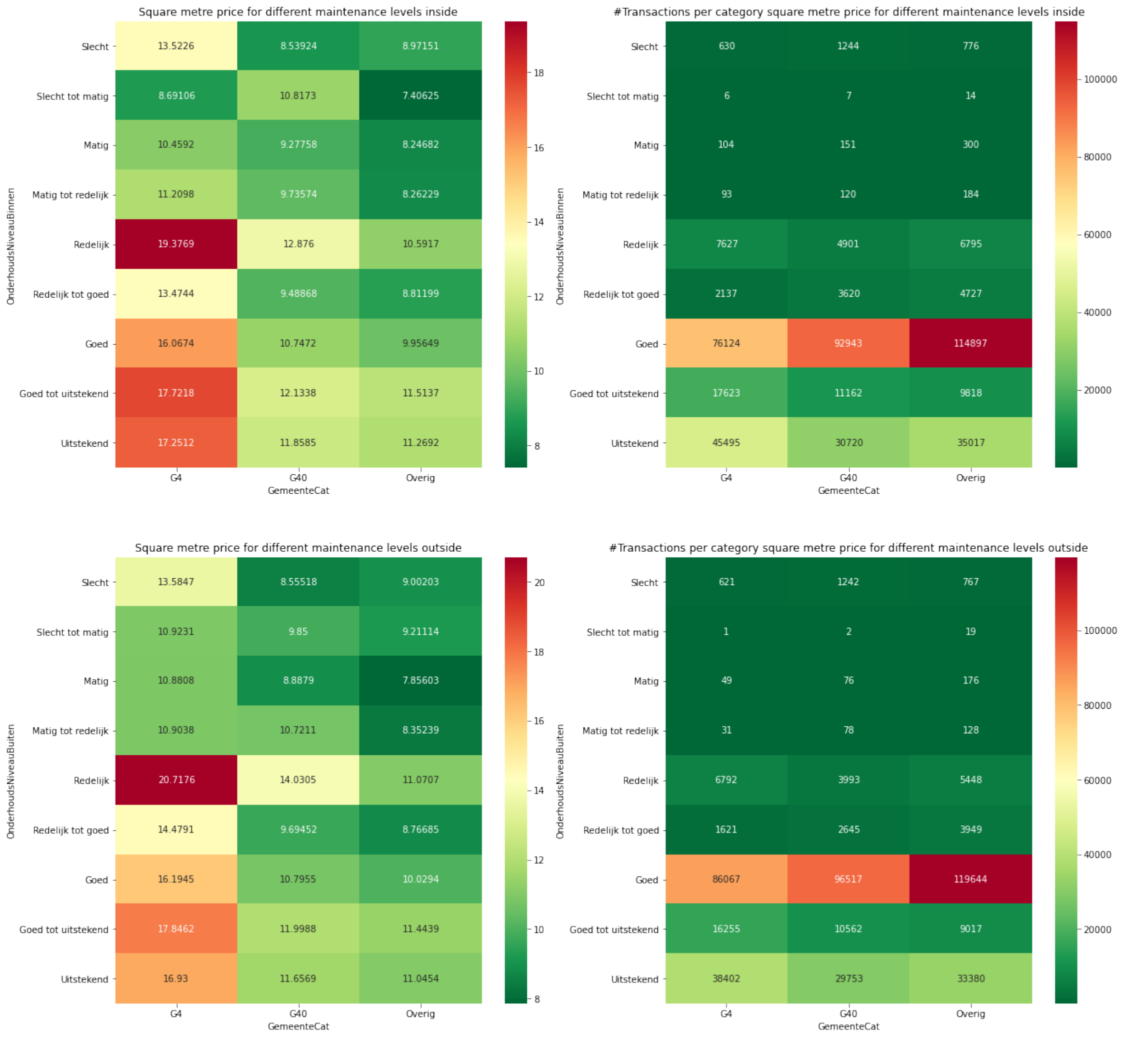


Since the data is more granular. Check if there are still enough data points in every category.



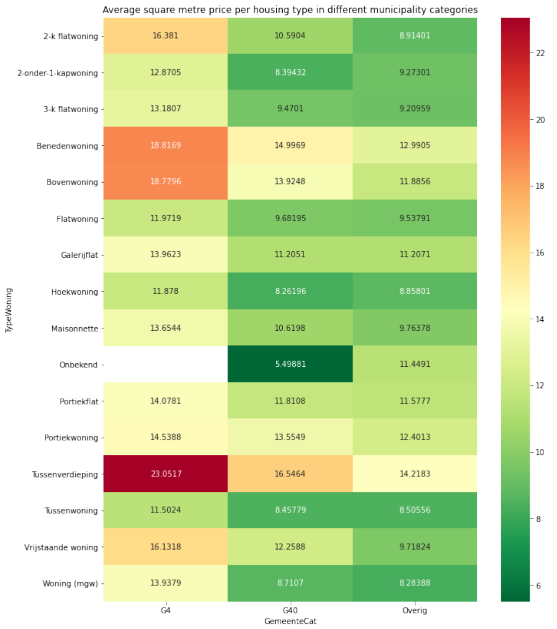
Less than 10 data points is not sufficient (dark green), between 10 and 100 data points is low reliability (light green). More than 100 data points is seen as sufficient to make assumptions. It is evident that maintenance level outside has to be atleast 'Matig tot redelijk' and maintenance level inside has to be atleast 'matig'.

Conclusions can not be drawn if maintenance levels inside as well as outside are 'Slecht' or 'Slecht tot matig'. (The worst two categories).

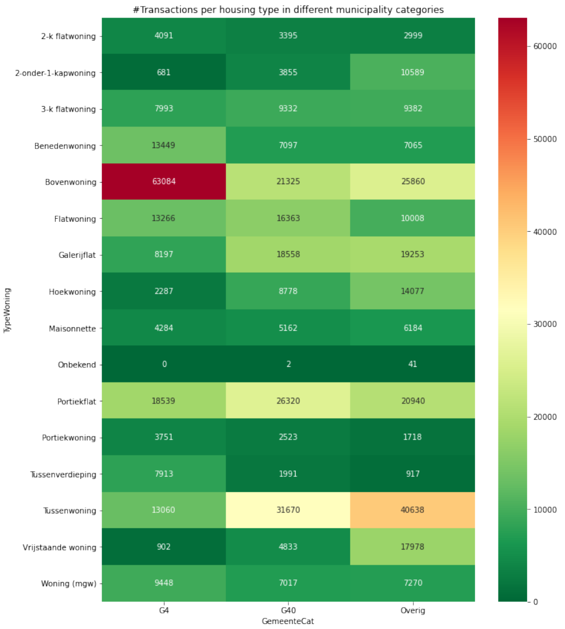


Investigating different municipality categories shows that maintenance levels can not be simply put together. G4 (the biggest 4 municipalities) have as expected the highest renting prices per square metre. ‘Overig’ (municipality 45 until 355 in order of magnitude) have the lowest renting prices.

For some reason ‘Redelijk’ (average) is a square metre price outlier, inside and outside. As is visible in the right heatmaps, this is not due to not sufficient data. The exact problem is most likely due to the ‘standard’ values of the form which is being filled in when a transaction is made. A domain expert is needed to explain this.



Average square metre price per housing type in different municipality categories



#Transactions per housing type in different municipality categories

It is clearly seen that 'Benedenwoning', 'Bovenwoning' and 'Tussenverdieping' are the more expensive housing types. In all municipality categories.

Even though 'Vrijstaande woning' is one of the most luxurious ones. 'Benedenwoning', 'Bovenwoning' and 'Tussenverdieping' are appartment types and are more common in the city centre. Therefore, it could be more expensive. However, this is just an assumption.

1.5.1 Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

Maintenance levels inside as well as outside strengthened each other. A 'better' municipality categorie also strenghtened the square metre price on top off that.

1.5.2 Were there any interesting or surprising interactions between features?

It was surprising that the most expensive house types were 'Benedenwoning', 'Bovenwoning' and 'Tussenverdieping'. Which means lower, middel and upper floor. Not 'vrijstaande woning' which is the most luxurious type of housing.

Maintenance level inside/outside 'mediocre' has a higher square metre price than 'mediocre/good', 'good', 'good/fantastic' and 'fantastic'. This is really contraintuitive and a domain expert indicated that 'mediocre' is the standard value for this field. Thus my assumption is that that could be the cause of the outlier.