Analiza zbioru danych mieszkania

Wielowymiarowa analiza danych

Przemysław Peplinski, Wiktor Galewski, Mikołaj Zalewski

2026-01-18

Spis treści

1 Wstęp

1.1 Autorzy

Autorami są... Wkład w projekt prezentował się następująco...

1.2 Motywacja i cele

Coś tam...

1.3 Opis projektu

Coś tam...

2 Opis danych

Dane pochodzą... /Dane zostały zebrane...

```
## Rows: 106,169
## Columns: 19
## $ Link               <chr> "https://domy.pl/mieszkanie/gdansk-siedlce-malczews~
## $ `Data dodania`     <dttm> 2021-02-02 10:24:00, 2021-02-02 11:08:14, 2021-02-~
## $ `Data modyfikacji` <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ Tytuł              <chr> "Trzypokojowe mieszkanie na sprzedaż:", "Gdańsk, Si~
## $ Opis               <chr> "Młyny Gdańskie Gdańsk, ul. Malczewskiego Młyny Gda~
## $ Cena               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ Powierzchnia       <dbl> 60.02, 89.63, 60.49, 64.89, 104.66, 72.04, 75.21, 6~
## $ `Cena za metr`     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
## $ `Liczba pokoi`     <dbl> 3, 4, 3, 3, 4, 4, 4, 3, 3, 3, 4, 3, 1, 2, 3, 3, 3, ~
## $ Piętro             <dbl> 4, 1, 0, 0, 1, 3, 1, 2, 3, 3, 2, 1, 3, 0, 1, 3, 4, ~
## $ `Liczba pięter`    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ `Rok budowy`       <dbl> 2023, 2023, 2023, 2023, 2023, 2023, 2023, 2023, 202~
## $ Ocena              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Dzielnica          <chr> "Siedlce", "Siedlce", "Siedlce", "Siedlce", "Siedlc~
## $ `Rodzaj budynku`   <chr> "Apartamentowiec", "Pozostałe", "Pozostałe", "Pozos~
## $ Rynek              <chr> "Pierwotny", "Pierwotny", "Pierwotny", "Pierwotny",~
## $ Ogłoszeniodawca    <chr> "Agencja", "Agencja", "Agencja", "Agencja", "Agencj~
## $ Portal             <chr> "Domy", "Morizon", "Morizon", "Morizon", "Morizon",~
## $ `Numer telefonu`   <chr> NA, "585055401", "585055401", "585055401", "5850554~
```

Opisać zmienne...

Wśród zmiennych niepotrzebnych do analizy można wymienić link, tytuł, opis oraz numer telefonu. Zmienne te zostaną zatem usunięte.

3 Czyszczenie danych

3.1 Walidacja

Tak

```
## Rows: 106,169
## Columns: 19
## $ Link               <chr> "https://domy.pl/mieszkanie/gdansk-siedlce-malczews~
## $ `Data dodania`     <dttm> 2021-02-02 10:24:00, 2021-02-02 11:08:14, 2021-02-~
## $ `Data modyfikacji` <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ Tytuł              <chr> "Trzypokojowe mieszkanie na sprzedaż:", "Gdańsk, Si~
## $ Opis               <chr> "Młyny Gdańskie Gdańsk, ul. Malczewskiego Młyny Gda~
## $ Cena               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ Powierzchnia       <dbl> 60.02, 89.63, 60.49, 64.89, 104.66, 72.04, 75.21, 6~
## $ `Cena za metr`     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ `Liczba pokoi`     <dbl> 3, 4, 3, 3, 4, 4, 4, 3, 3, 3, 4, 3, 1, 2, 3, 3, 3, ~
## $ Piętro             <dbl> 4, 1, 0, 0, 1, 3, 1, 2, 3, 3, 2, 1, 3, 0, 1, 3, 4, ~
## $ `Liczba pięter`    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ `Rok budowy`       <dbl> 2023, 2023, 2023, 2023, 2023, 2023, 2023, 2023, 202~
## $ Ocena              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Dzielnica          <chr> "Siedlce", "Siedlce", "Siedlce", "Siedlce", "Siedlc~
## $ `Rodzaj budynku`   <chr> "Apartamentowiec", "Pozostałe", "Pozostałe", "Pozos~
## $ Rynek              <chr> "Pierwotny", "Pierwotny", "Pierwotny", "Pierwotny",~
## $ Ogłoszeniodawca    <chr> "Agencja", "Agencja", "Agencja", "Agencja", "Agencj~
## $ Portal             <chr> "Domy", "Morizon", "Morizon", "Morizon", "Morizon",~
## $ `Numer telefonu`   <chr> NA, "585055401", "585055401", "585055401", "5850554~
```

3.2 Obsługa braków wartości

Kilka zmiennych do usunięcia, dla niektórych trzeba usunąć obserwacje, w innych imputować wg obliczeń.

```
##   missForest iteration 1 in progress...done!
##     estimated error(s): 1.379711e-05 0.03532471
##     difference(s): 9.168541e-11 0.03261746
##     time: 29.01 seconds
##
##   missForest iteration 2 in progress...done!
##     estimated error(s): 1.235244e-05 0.03504698
##     difference(s): 5.99119e-13 0.004266807
##     time: 28.19 seconds
##
##   missForest iteration 3 in progress...done!
##     estimated error(s): 1.242352e-05 0.03537198
##     difference(s): 2.322088e-13 0.004550982
##     time: 28.44 seconds
##
##   missForest iteration 4 in progress...done!
##     estimated error(s): 1.233419e-05 0.0350588
##     difference(s): 2.511408e-13 0.004973066
##     time: 29.28 seconds
```

3.3 Obsługa obserwacji odstających

Jako tako zostawiamy

4 Analiza eksploracyjna

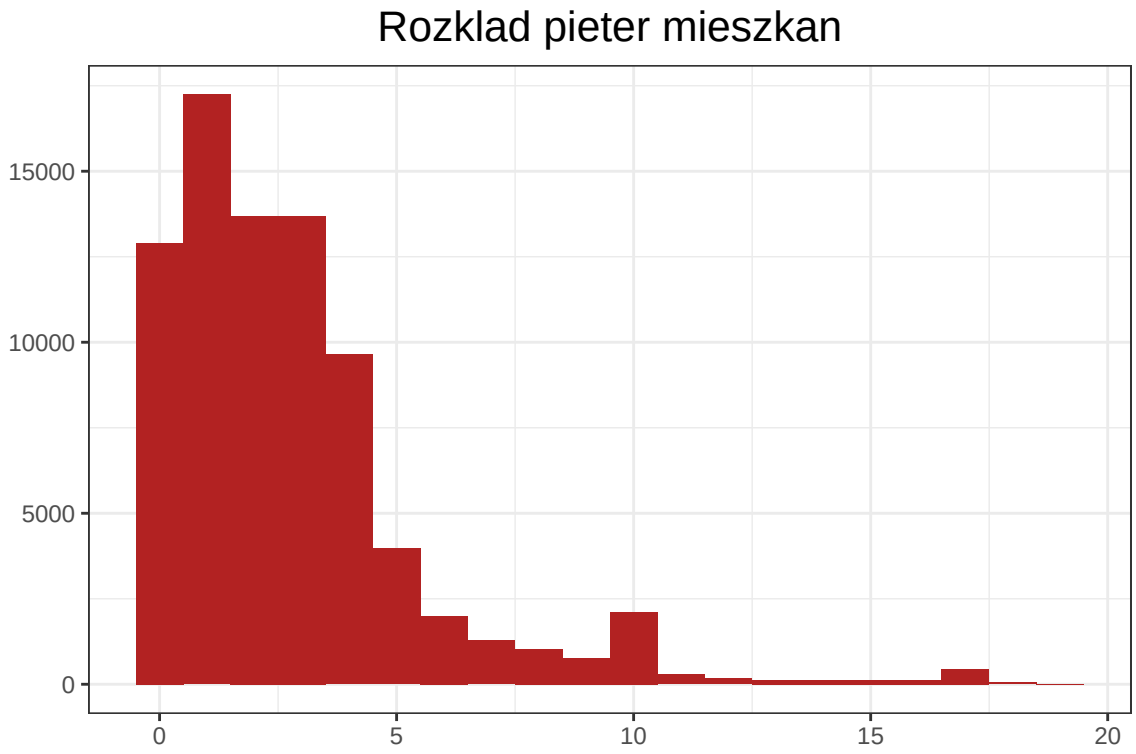
4.1 Statystyki opisowe

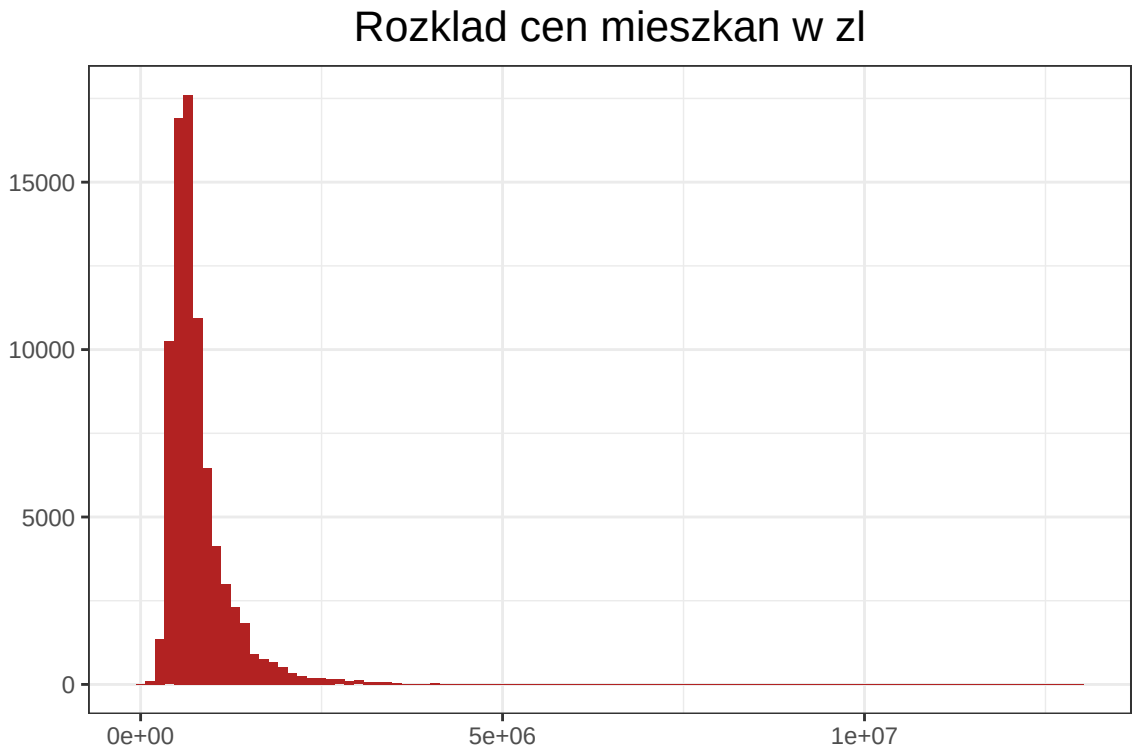
```
##   Data dodania   Data modyfikacji      Cena           Powierzchnia
##  Min.   :18666   Min.   :0.0000   Min.   :   33000   Min.   : 14.00
##  1st Qu.:19733   1st Qu.:0.0000   1st Qu.:  526925   1st Qu.: 40.00
##  Median :20058   Median :0.0000   Median :  669000   Median : 50.60
##  Mean   :19980   Mean   :0.3913   Mean   :  794244   Mean   : 54.38
##  3rd Qu.:20249   3rd Qu.:1.0000   3rd Qu.:  899000   3rd Qu.: 64.60
##  Max.   :20437   Max.   :1.0000   Max.   :13000000   Max.   :553.00
##
##   Cena za metr    Liczba pokoi        Piętro         Rok budowy
```

```
##  Min.   : 1121   Min.   : 1.000   Min.   : 0.000   Min.   :1780
##  1st Qu.:11200   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.:1986
##  Median :13414   Median : 2.000   Median : 2.000   Median :2020
##  Mean   :14664   Mean   : 2.535   Mean   : 2.786   Mean   :2004
##  3rd Qu.:16667   3rd Qu.: 3.000   3rd Qu.: 4.000   3rd Qu.:2024
##  Max.   :62298   Max.   :24.000   Max.   :19.000   Max.   :2028
##
##                Dzielnica             Rodzaj budynku      Rynek
##  Śródmieście        :11064   Apartamentowiec:24873   Min.   :0.0000
##  Ujeścisko-Łostowice:10920   Blok           :47654   1st Qu.:0.0000
##  Jasień             : 9296   Kamienica      : 7236   Median :0.0000
##  Wrzeszcz           : 5576                           Mean   :0.3485
##  Przymorze          : 5052                           3rd Qu.:1.0000
##  Siedlce            : 4915                           Max.   :1.0000
##  (Other)            :32940
##  Ogłoszeniodawca                  Portal
##  Min.   :0.0000   Otodom              :25920
##  1st Qu.:1.0000   Trojmiasto          :16197
##  Median :1.0000   Okolica             : 7993
##  Mean   :0.8734   Nieruchomosci-online: 7119
##  3rd Qu.:1.0000   Allegro             : 6505
##  Max.   :1.0000   Adresowo            : 3257
##                   (Other)             :12772
```

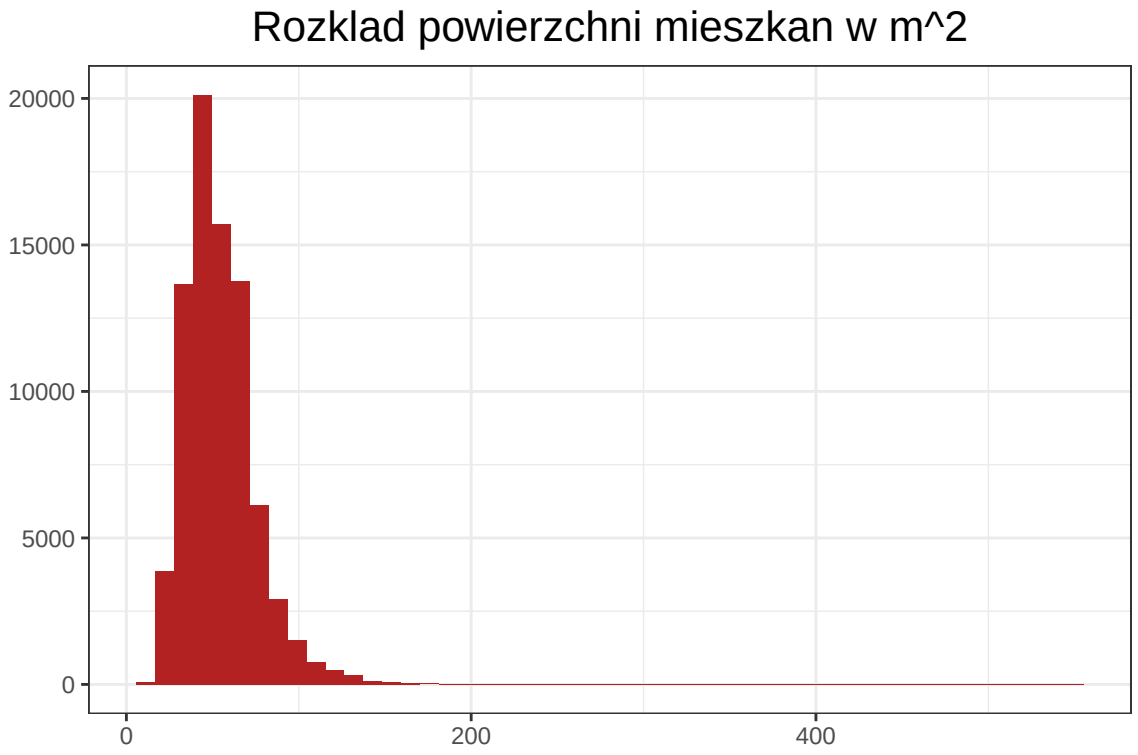
4.2 Analiza rozkładów

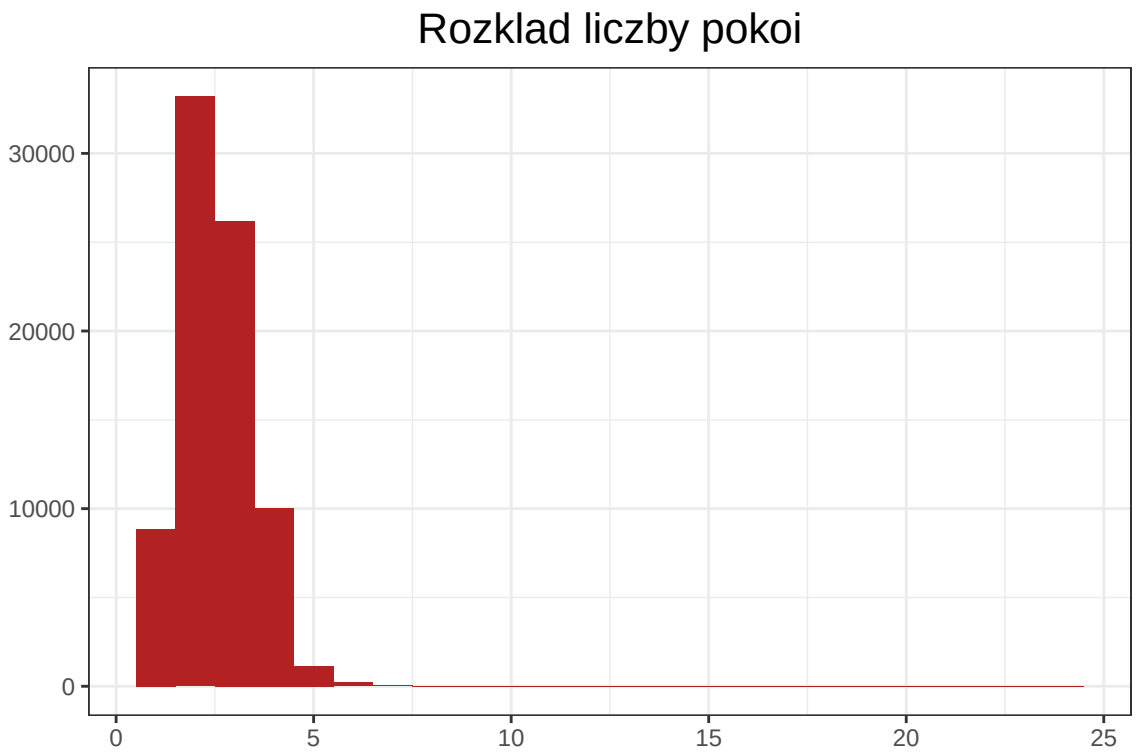
Histogramy zmiennych numerycznych, wykresy kolumnowe zmiennych kategorycznych


Rozklad pieter mieszkan
15000
10000
5000
0
0
5
10
15
20


Rozklad cen mieszkan w zl
15000
10000
5000
0
0e+00
5e+06
1e+07

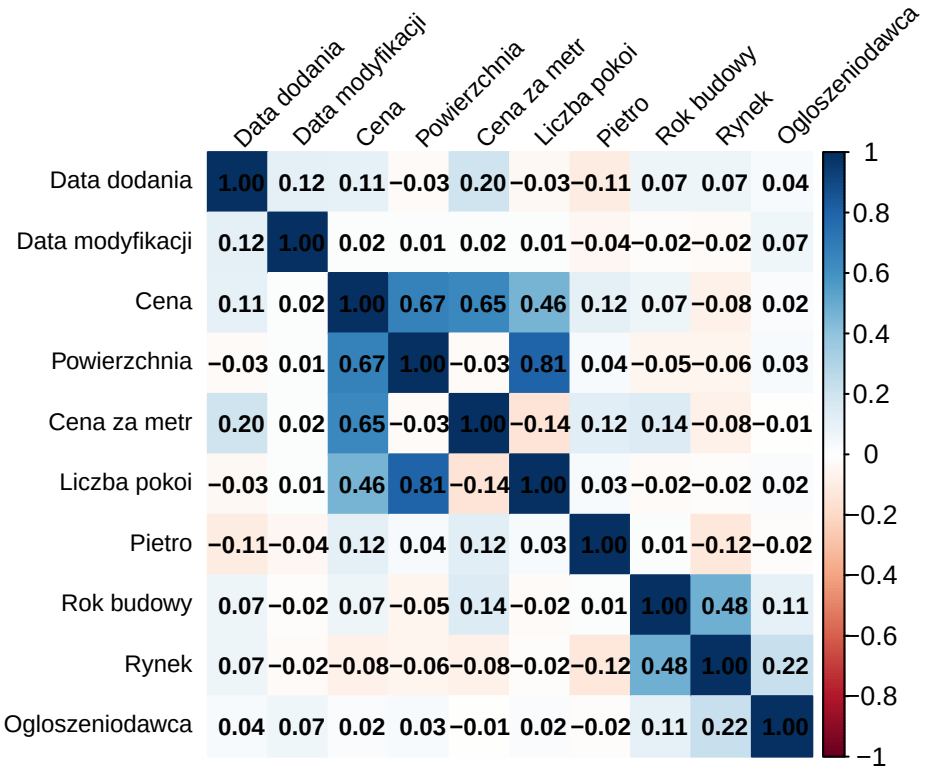

Rozklad ceny za metr kwadratowy mieszkan w zl/m^2
12000
9000
6000
3000
0
0
20000
40000
60000


Rozklad powierzchni mieszkan w m^2
20000
15000
10000
5000
0
0
200
400



4.3 Analiza korelacji

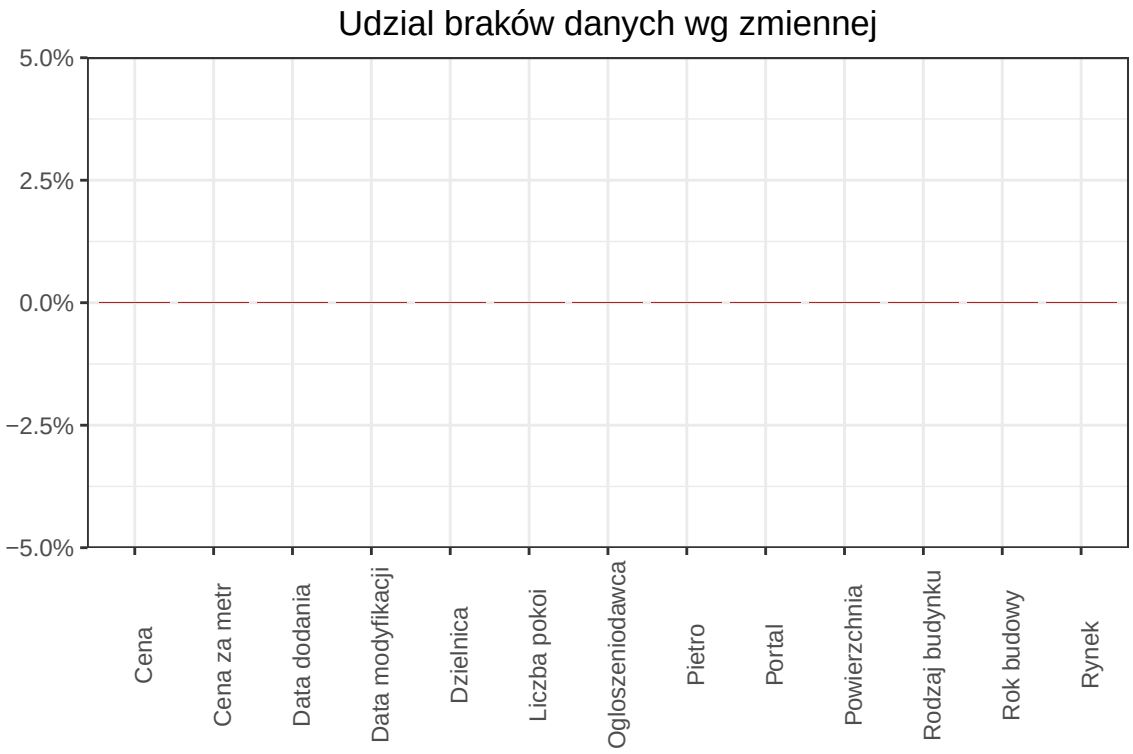
Macierz korelacji



	Data dodania	Data modyfikacji	Cena	Powierzchnia	Cena za metr	Liczba pokoi	Pietro	Rok budowy	Rynek	Ogloszeniodawca
Data dodania	1.00	0.12	0.11	−0.03	0.20	−0.03	−0.11	0.07	0.07	0.04
Data modyfikacji	0.12	1.00	0.02	0.01	0.02	0.01	−0.04	−0.02	−0.02	0.07
Cena	0.11	0.02	1.00	0.67	0.65	0.46	0.12	0.07	−0.08	0.02
Powierzchnia	−0.03	0.01	0.67	1.00	−0.03	0.81	0.04	−0.05	−0.06	0.03
Cena za metr	0.20	0.02	0.65	−0.03	1.00	−0.14	0.12	0.14	−0.08	−0.01
Liczba pokoi	−0.03	0.01	0.46	0.81	−0.14	1.00	0.03	−0.02	−0.02	0.02
Pietro	−0.11	−0.04	0.12	0.04	0.12	0.03	1.00	0.01	−0.12	−0.02
Rok budowy	0.07	−0.02	0.07	−0.05	0.14	−0.02	0.01	1.00	0.48	0.11
Rynek	0.07	−0.02	−0.08	−0.06	−0.08	−0.02	−0.12	0.48	1.00	0.22
Ogloszeniodawca	0.04	0.07	0.02	0.03	−0.01	0.02	−0.02	0.11	0.22	1.00

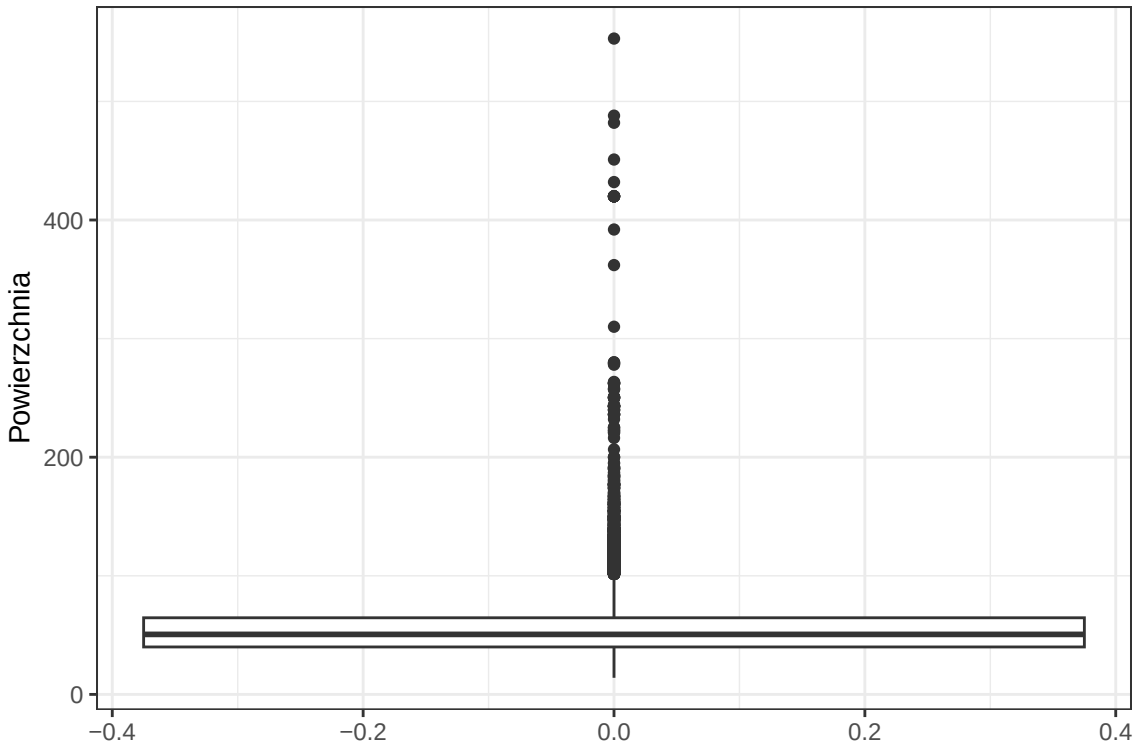
4.4 Analiza braków danych

Krótkie podsumowanie braków danych (wykres słupkowy?)



4.5 Analiza obserwacji odstających

Zrobić boxploty



5 Obróbka zmiennych

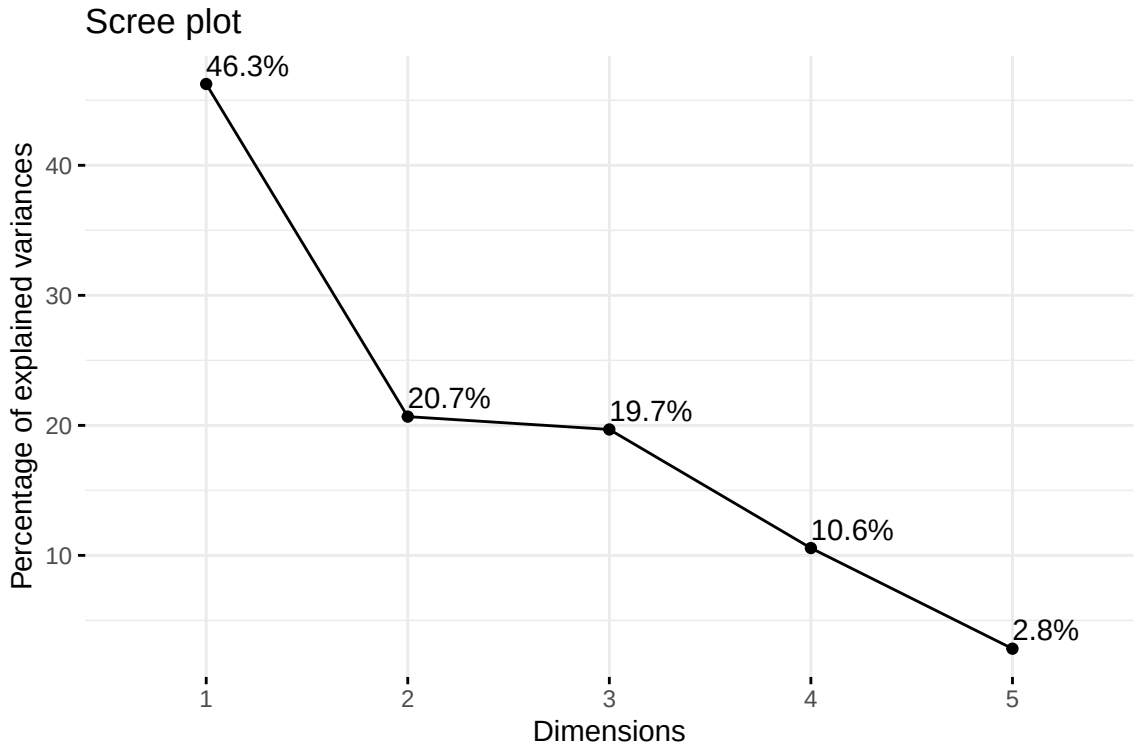
5.1 Kodowanie zmiennych kategorycznych

5.2 Normalizacja zmiennych numerycznych

6 2 metody

6.1 Analiza głównych składowych (PCA)

```
## $chisq
## [1] 167.9269
##
## $p.value
## [1] 7.451025e-31
##
## $df
## [1] 10
```



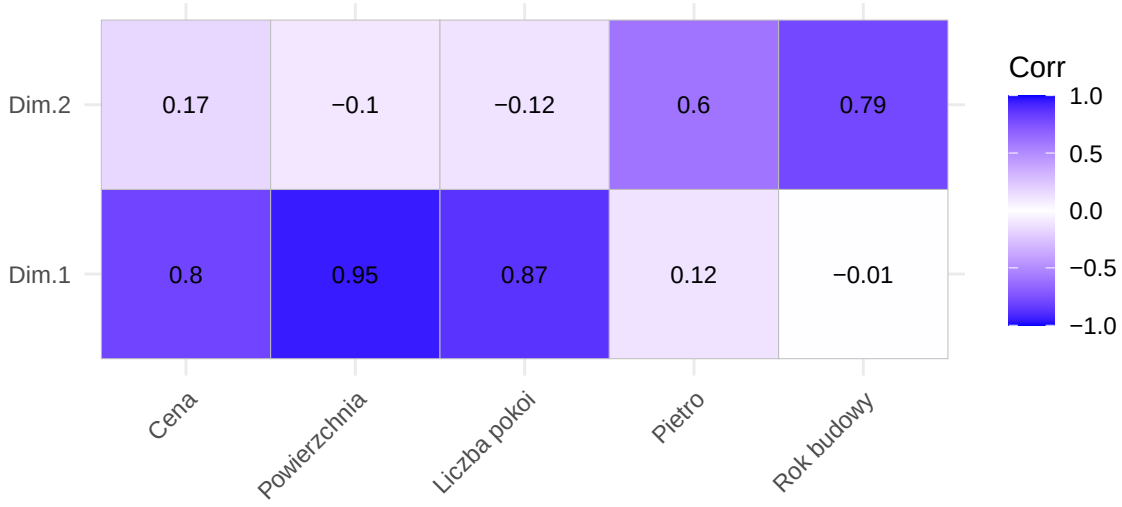

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1  2.3127535        46.255070                    46.25507
## Dim.2  1.0336234        20.672469                    66.92754
## Dim.3  0.9842327        19.684654                    86.61219
## Dim.4  0.5281741        10.563482                    97.17567
## Dim.5  0.1412163         2.824326                   100.00000
```

Korzystając z kryterium Keisera można określić, że mieszkania można określić za pomocą dwóch zmiennych, ponieważ dla nich wartości własne są większe od 1. Taki sam wniosek można wyciągnąć na podstawie oceny wykresu osypiska, gdyż po drugim wymiarze zmiana wariancji jest znacznie mniejsza.

6.2 Analiza korespondencji (CA)

7 Wizualizacja metod

7.1 Wizualizacje PCA



7.2 Wizualizacje CA

8 Wnioski