

Analiza ofert sprzedaży mieszkań w Gdańsku

Wielowymiarowa analiza danych

Przemysław Peplinski, Wiktor Galewski, Mikołaj Zalewski

2026-01-20

Spis treści

1 Wstęp	2
1.1 Autorzy	2
1.2 Motywacja i cele	2
1.3 Opis projektu	2
2 Opis danych	2
3 Czyszczenie danych	3
3.1 Braki danych	3
3.2 Walidacja danych i obsługa braków wartości	4
4 Analiza eksploracyjna	5
4.1 Statystyki opisowe	5
4.2 Analiza rozkładów	5
4.3 Analiza korelacji	11
4.4 Analiza obserwacji odstających	12
5 Obróbka zmiennych	15
5.1 Normalizacja zmiennych numerycznych	15
6 Analiza wielowymiarowa	15
6.1 Analiza głównych składowych (PCA)	15
6.2 Analiza skupień (clustering)	17

7 Wizualizacja metod	18
7.1 Wizualizacje PCA	18
7.2 Wizualizacje analizy skupień	20
8 Wnioski	21

1 Wstęp

1.1 Autorzy

Przedstawiono autorów oraz ich wkłady.

- Przemysław Peplinski: załączek analizy eksploracyjnej, analizy wielowymiarowe oraz ich wizualizacje.
- Wiktor Galewski: zebranie danych, ich opis i czyszczenie
- Mikołaj Zalewski: analiza eksploracyjna

1.2 Motywacja i cele

Rynek mieszkań jest rynkiem istotnym dla młodych ludzi - chociażby dla studentów oraz absolwentów studiów. Przystępujemy więc do analizy rzeczywistych danych na temat rynku mieszkań w Gdańsku w celu lepszego zrozumienia charakteru tego rynku, zarówno w celach praktycznego zastosowania metod wielowymiarowej analizy danych, jak i dla pogłębienia własnej wiedzy życiowej w tym ważnym aspekcie naszego życia.

1.3 Opis projektu

Wśród ofert sprzedaży mieszkań w Gdańsku mogą występować inherentne zależności, których nie sposób zobaczyć przeglądając oferty mieszkań na stronach internetowych. Za pomocą analizy danych chcemy zatem sprawdzić, czy takie zależności występują, a jeśli tak, to jaki mają charakter.

2 Opis danych

Dane zostały pozyskane przy pomocy Skanera Okazji – funkcjonalności platformy Investoro, aplikacji monitorującej rynek nieruchomości w Polsce. Skaner Okazji scrapuje oferty sprzedaży mieszkań z różnych serwisów ogłoszeniowych, między innymi Otodom, OLX, Allegro, Morizon czy Nieruchomosci-online, aby następnie umożliwić wygenerowanie i pobranie pliku programu Excel. Wśród zmiennych znajdziemy chociażby datę dodania oferty, cenę mieszkania, metraż,

liczbę pokoi, lokalizację, piętro, typ budynku, rok budowy i inne. Do analizy przekazano łącznie 13 zmiennych. Obserwacji było początkowo ponad 100 tysięcy, po czyszczeniu danych zostało ich około 80 tysięcy.

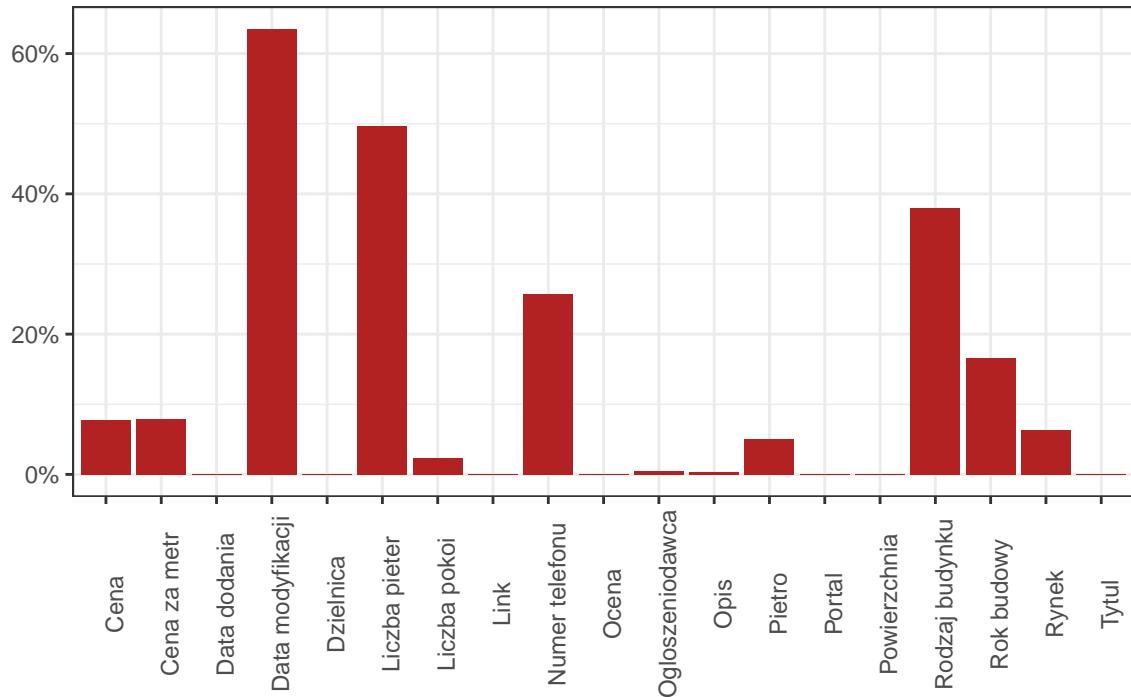
```
## Rows: 106,169
## Columns: 19
## $ Link <chr> "https://domy.pl/mieszkanie/gdansk-siedlce-malczews~
## $ `Data dodania` <dttm> 2021-02-02 10:24:00, 2021-02-02 11:08:14, 2021-02-~<dttm> NA, NA~
## $ `Data modyfikacji` <dttm> NA, NA~<dttm> NA, NA~<chr> "Trzypokojowe mieszkanie na sprzedaż:", "Gdańsk, Si~<chr> "Młyny Gdańskie Gdańsk, ul. Malczewskiego Młyny Gda~<dbl> NA, NA,~<dbl> 60.02, 89.63, 60.49, 64.89, 104.66, 72.04, 75.21, 6~<dbl> NA, NA,~<dbl> 3, 4, 3, 3, 4, 4, 4, 3, 3, 3, 4, 3, 1, 2, 3, 3, 3, ~<dbl> 4, 1, 0, 0, 1, 3, 1, 2, 3, 3, 2, 1, 3, 0, 1, 3, 4, ~<dbl> NA, NA,~<dbl> 2023, 2023, 2023, 2023, 2023, 2023, 2023, 2023, 202~<dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~<chr> "Siedlce", "Siedlce", "Siedlce", "Siedlce", "Siedlc~<chr> "Apartamentowiec", "Pozostałe", "Pozostałe", "Pozos~<chr> "Pierwotny", "Pierwotny", "Pierwotny", "Pierwotny",~<chr> "Agencja", "Agencja", "Agencja", "Agencja", "Agencj~<chr> "Domy", "Morizon", "Morizon", "Morizon", "Morizon",~<chr> NA, "585055401", "585055401", "585055401", "5850554~
```

3 Czyszczenie danych

3.1 Braki danych

Na poniższym wykresie przedstawiono udział braków danych w wartościach każdej zmiennej.

Udział braków danych wg zmiennej



Występuje wiele braków danych, więc potrzebna jest ich kompleksowa obsługa.

3.2 Walidacja danych i obsługa braków wartości

Usunięto zduplikowane obserwacje. Usunięto zmienne uznane za nieprzydatne do analizy (ocena - trudna w interpretacji zmienna syntetyczna dodawana przez platformę dla inwestorów, dużo braków danych (zera) - ocena zaczęła być wyświetlaną dopiero jakiś czas temu; liczba pięter budynku - bardzo dużo braków danych, ciężka do imputacji; reszta usuniętych zmiennych miała znikomą wartość informacyjną). Data modyfikacji ogłoszenia (część ogłoszeń ulegała edycji) zamieniona na zmienną binarną - informuje, czy edycja w ogóle występowała. Cena za metr - usunięto NA (których było bardzo niewiele) i wartości skrajnie odstające uznane za jawne błędy - tym samym poradzono sobie z tożsamymi przypadkami w zmiennej Cena. Podobnie usunięto błędy w innych zmiennych, a także NA, jeśli uznano, że jest ich stosunkowo niewiele, a nie nadają się do imputacji poprzez relacje z resztą zmiennych (w zmiennej Liczba pokoi poprawiono również kilka outlierów ręcznie na podstawie opisu ogłoszenia). Zmodyfikowano typ niektórych zmiennych, na przykład Data jako zmienna numeryczna czy zmienne jakościowe zamienione na factory lub binarne. Na koniec imputowano braki danych w kolumnach Liczba pokoi, Rok budowy i Rodzaj budynku metodą Random Forest.

4 Analiza eksploracyjna

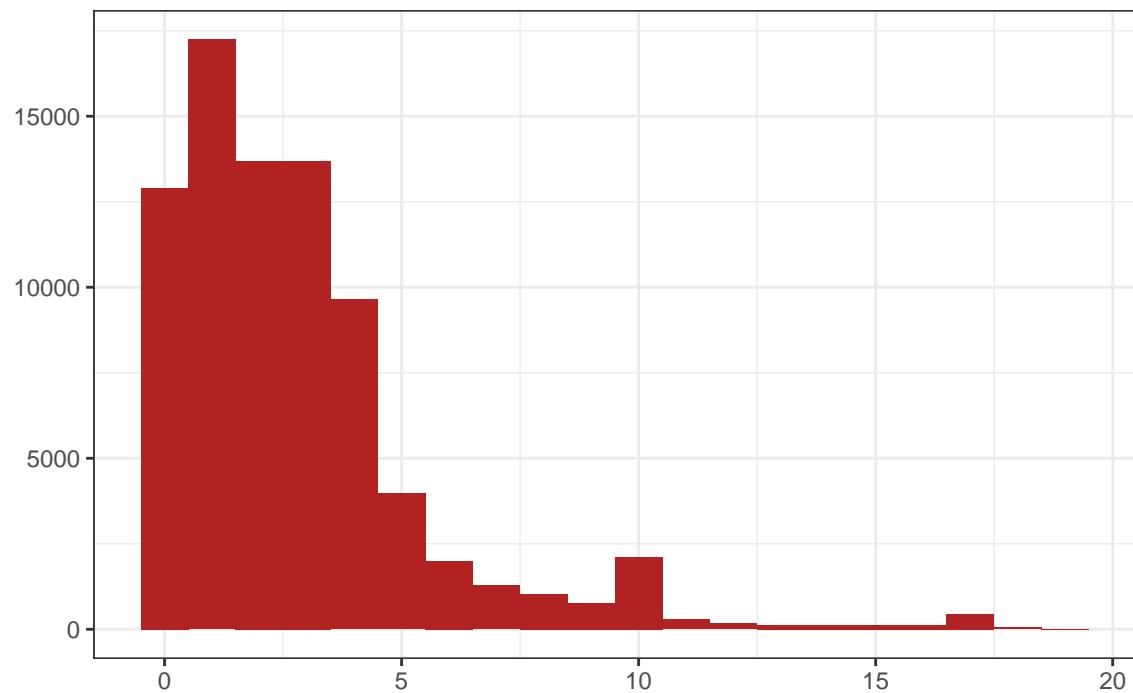
4.1 Statystyki opisowe

```
##   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:24879   Min.    :0.0000
##   Ujeścisko-Łostowice:10920   Blok            :47657   1st Qu.:0.0000
##   Jasień             : 9296   Kamienica        : 7227   Median :0.0000
##   Wrzeszcz          : 5576
##   Przymorze          : 5052
##   Siedlce            : 4915
##   (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
```

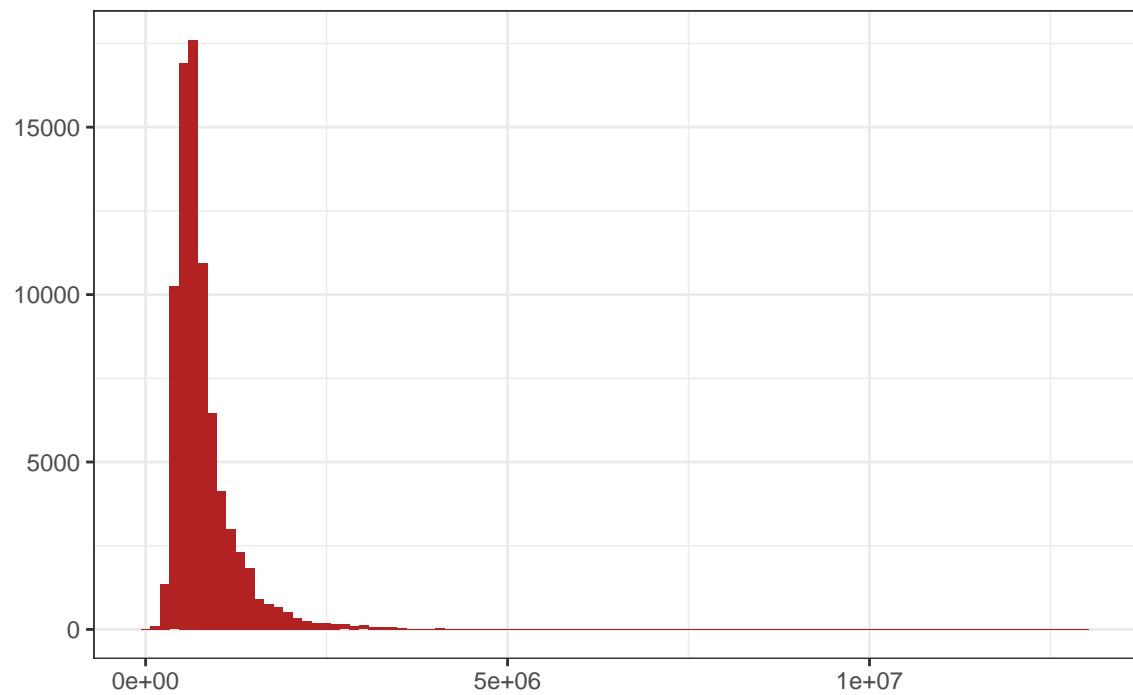
4.2 Analiza rozkładów

Dla zmiennych numerycznych stworzono histogramy, natomiast dla zmiennych kategorycznych wykonano wykresy kolumnowe.

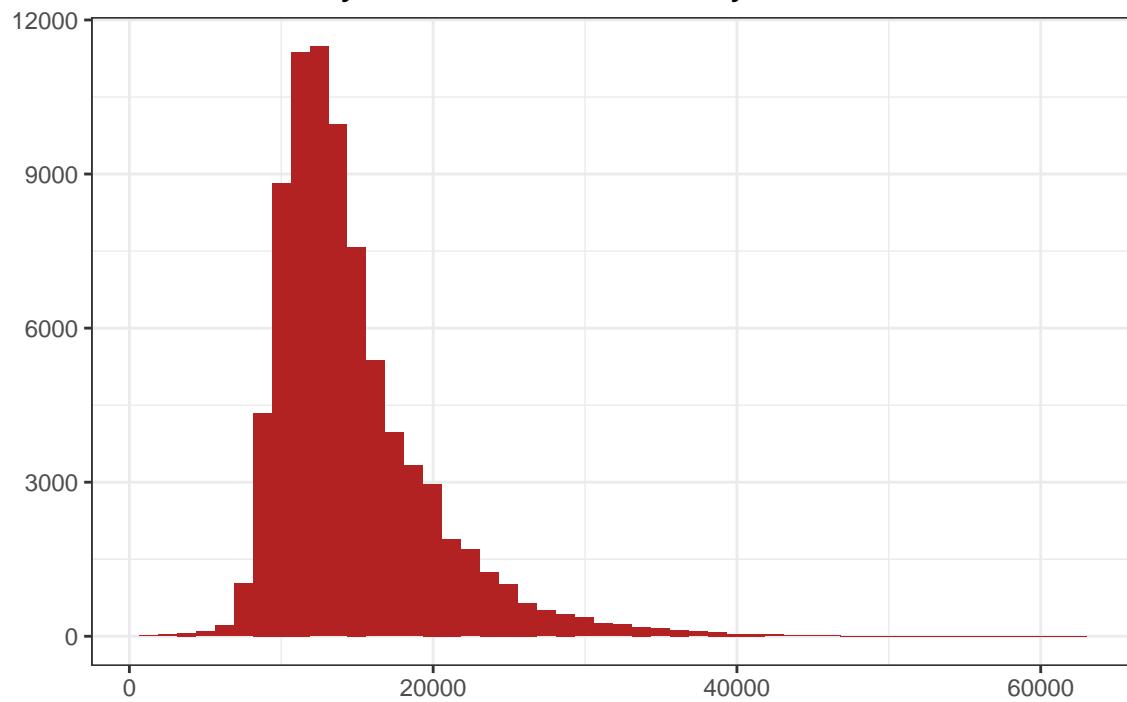
Rozkład pieter mieszkań



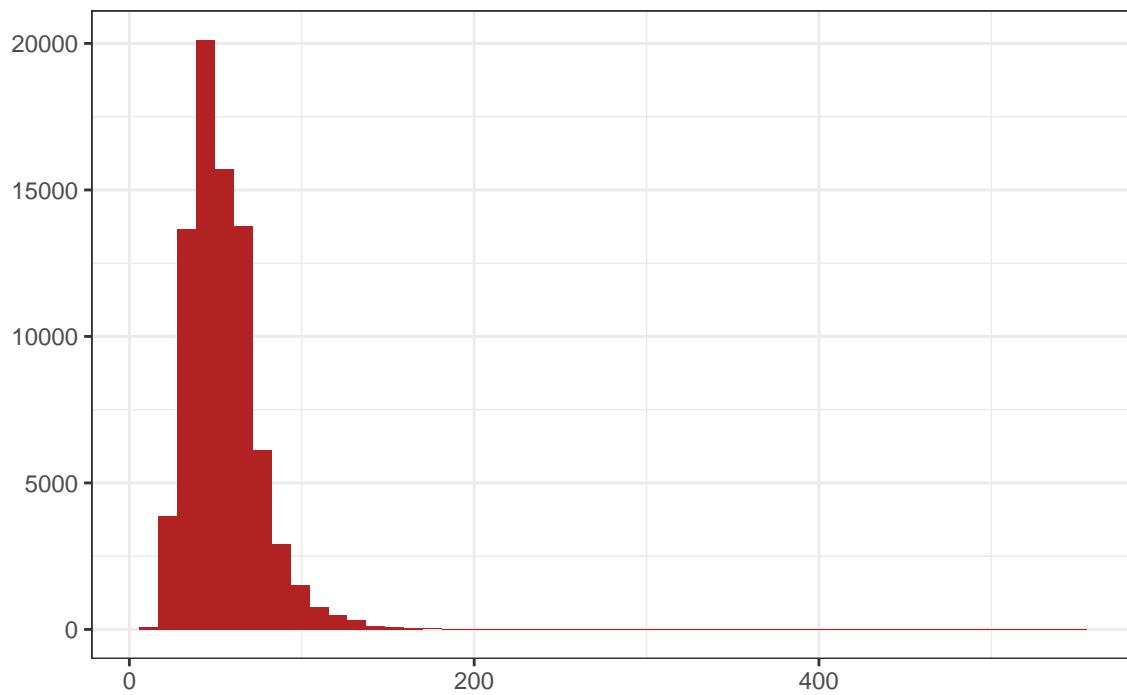
Rozkład cen mieszkań w zł



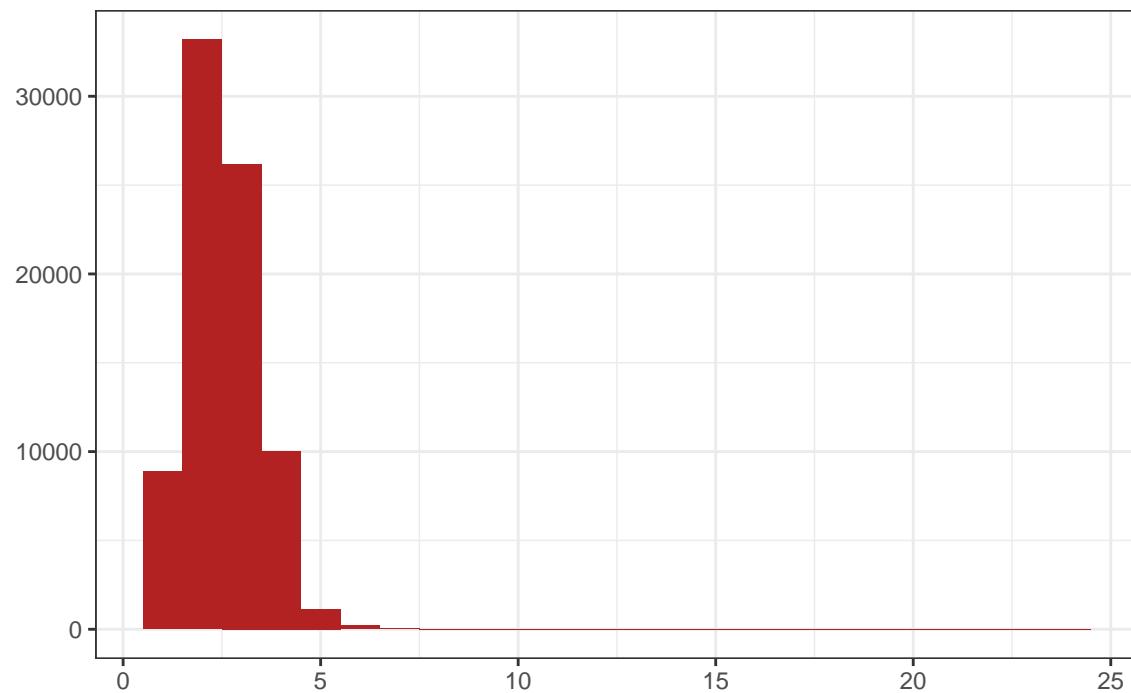
Rozkład ceny za metr kwadratowy mieszkań w zł/m²



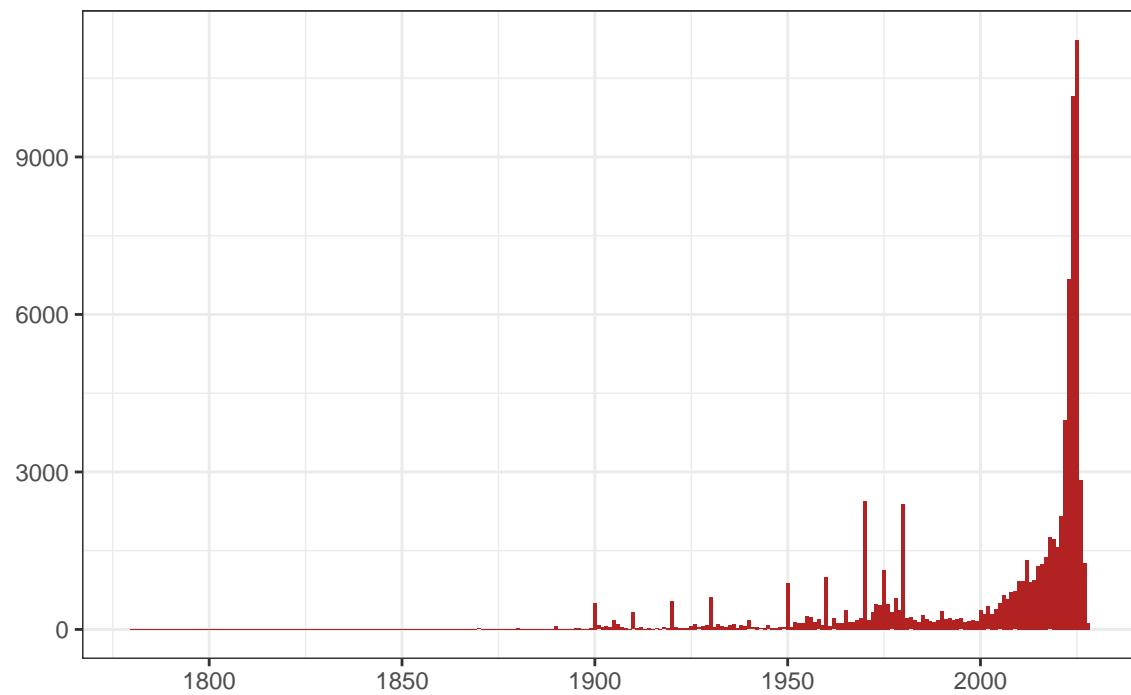
Rozkład powierzchni mieszkań w m²



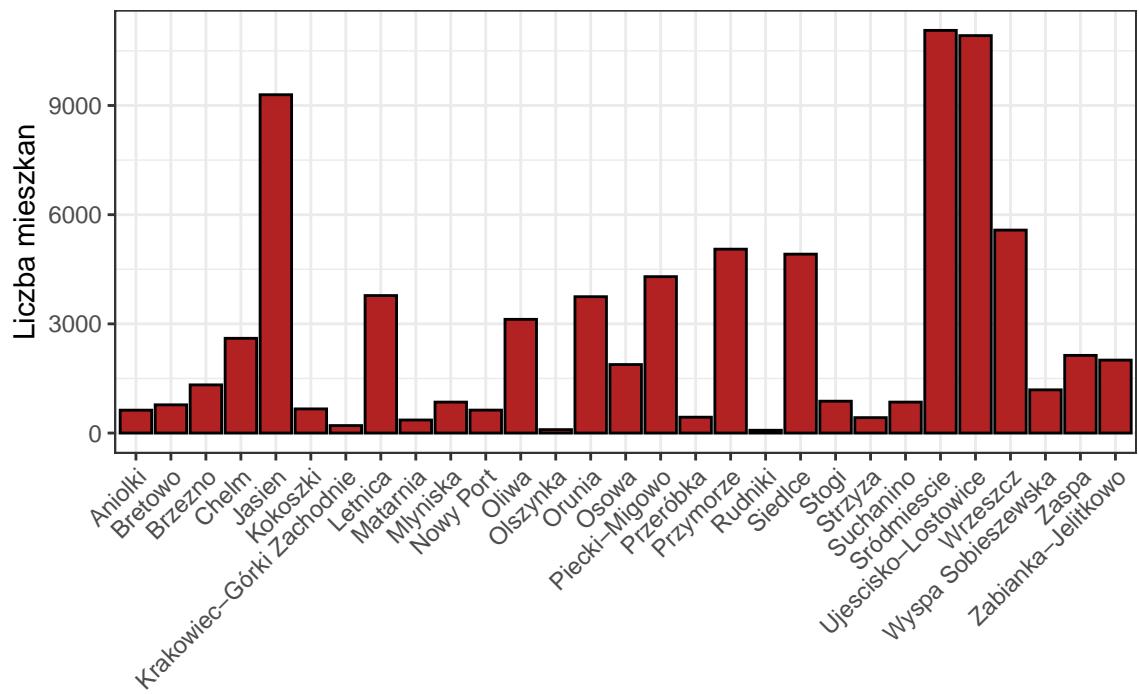
Rozkład liczby pokoi



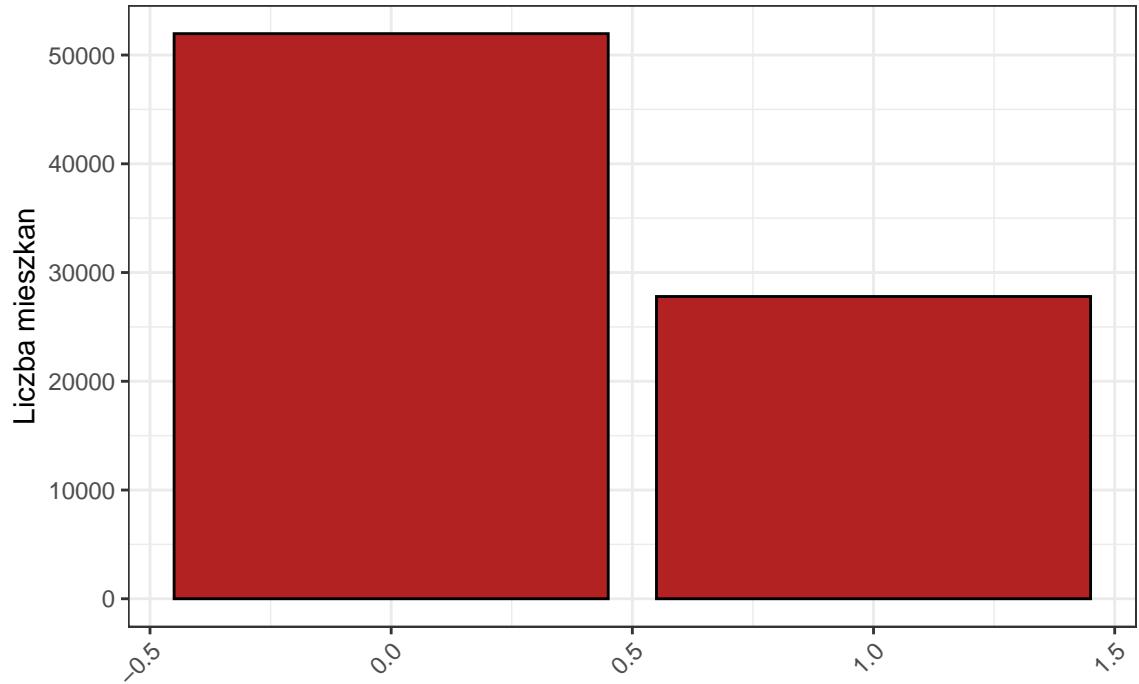
Rozkład mieszkań ze względu na rok budowy



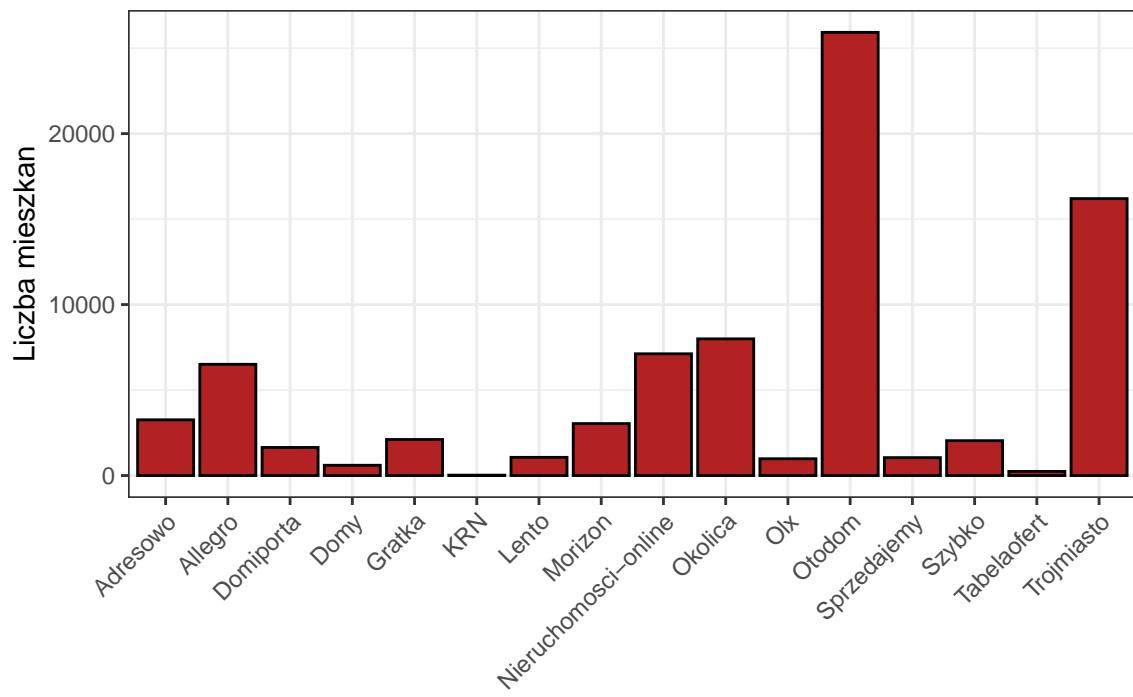
Podział mieszkańców ze względu na dzielnice



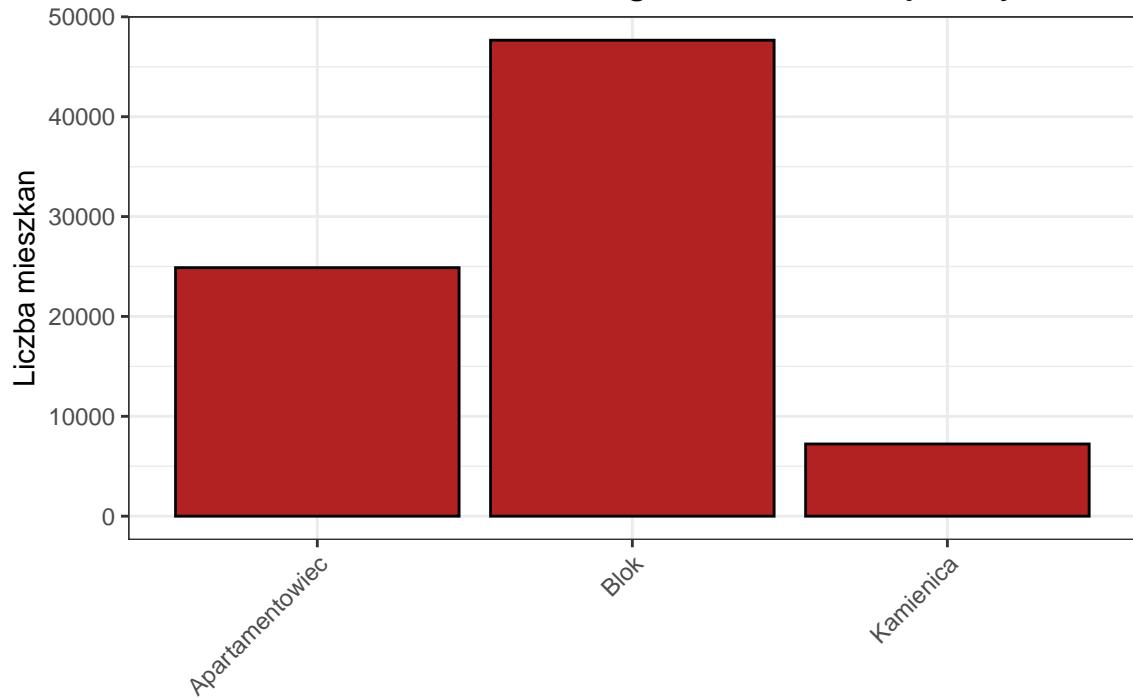
Podział mieszkańców ze względu na rodzaj rynku



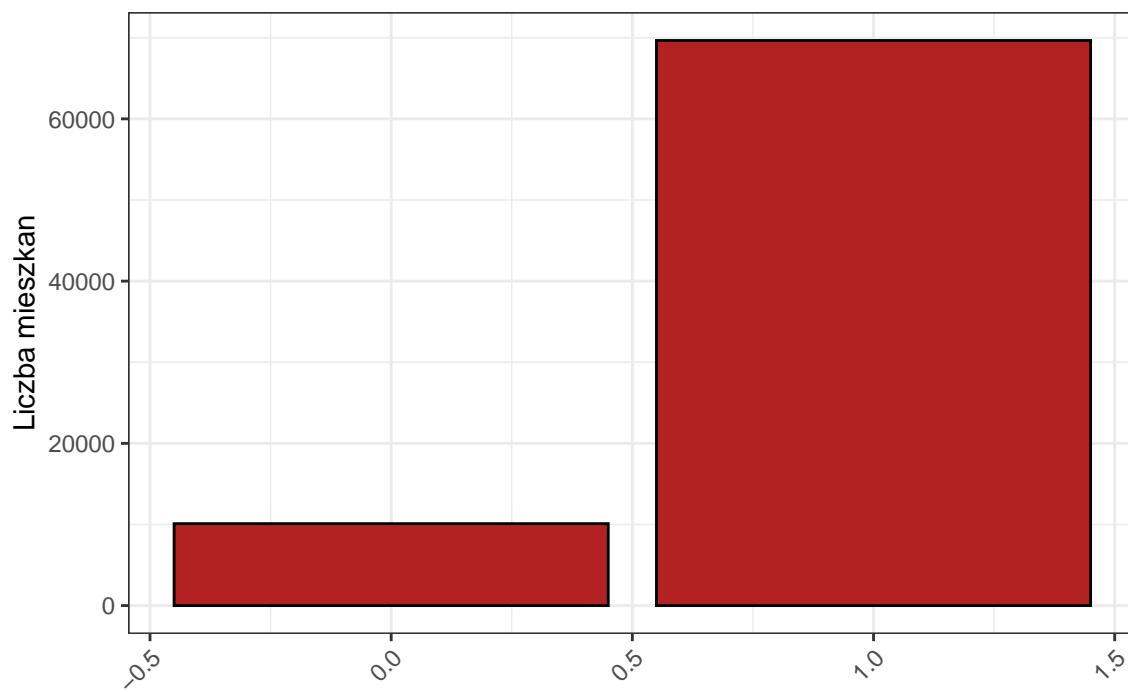
Podzial mieszkani ze wzgledu na portal wystawienia ogloszen



Podzial mieszkani ze wzgledu na rodzaj budynku

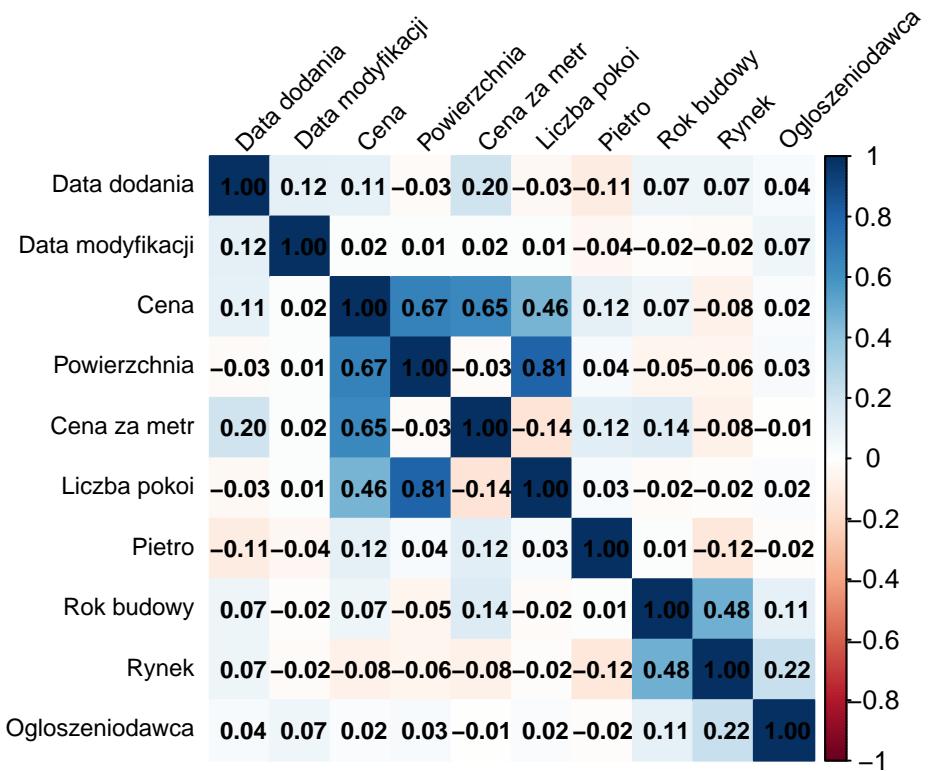


Podzial mieszkani ze wzgledu na ogloszeniodawce



4.3 Analiza korelacji

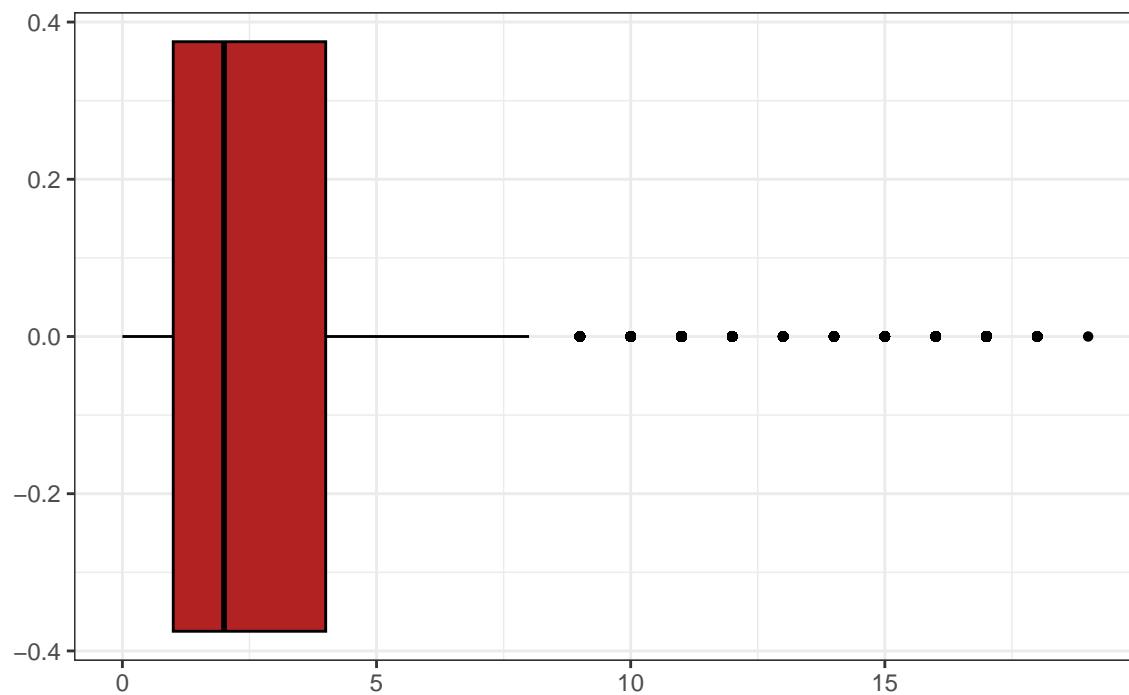
Poniżej przedstawiona została macierz korelacji.



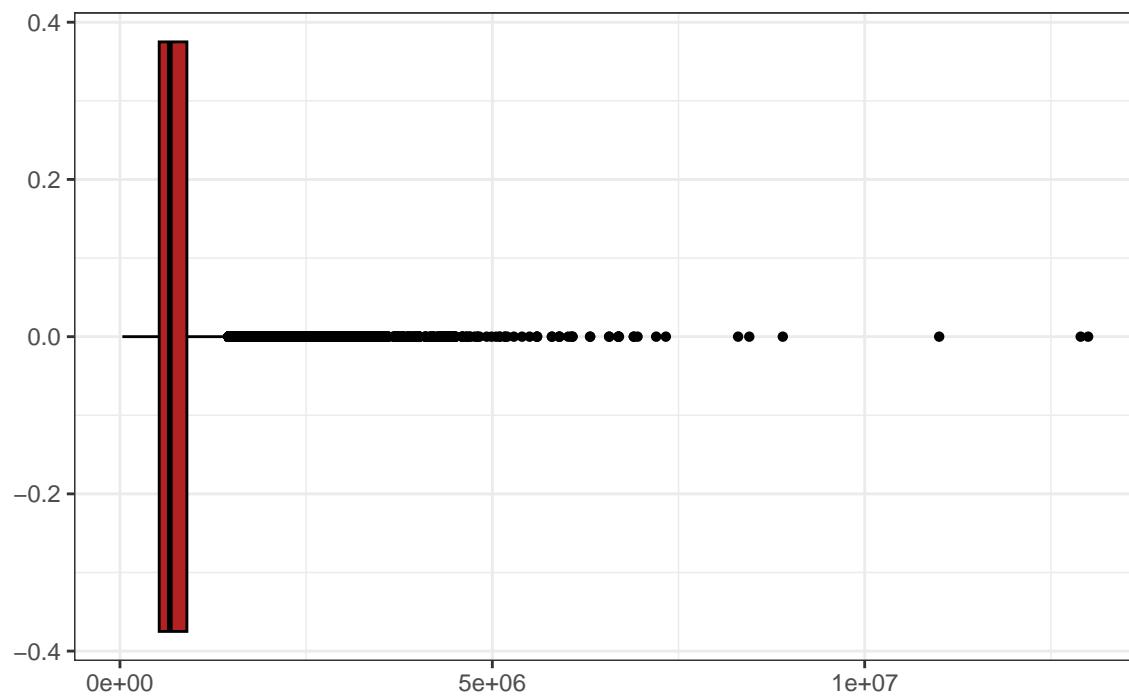
4.4 Analiza obserwacji odstających

Wartości odstające zostały zbadane za pomocą wykresów pudełkowych.

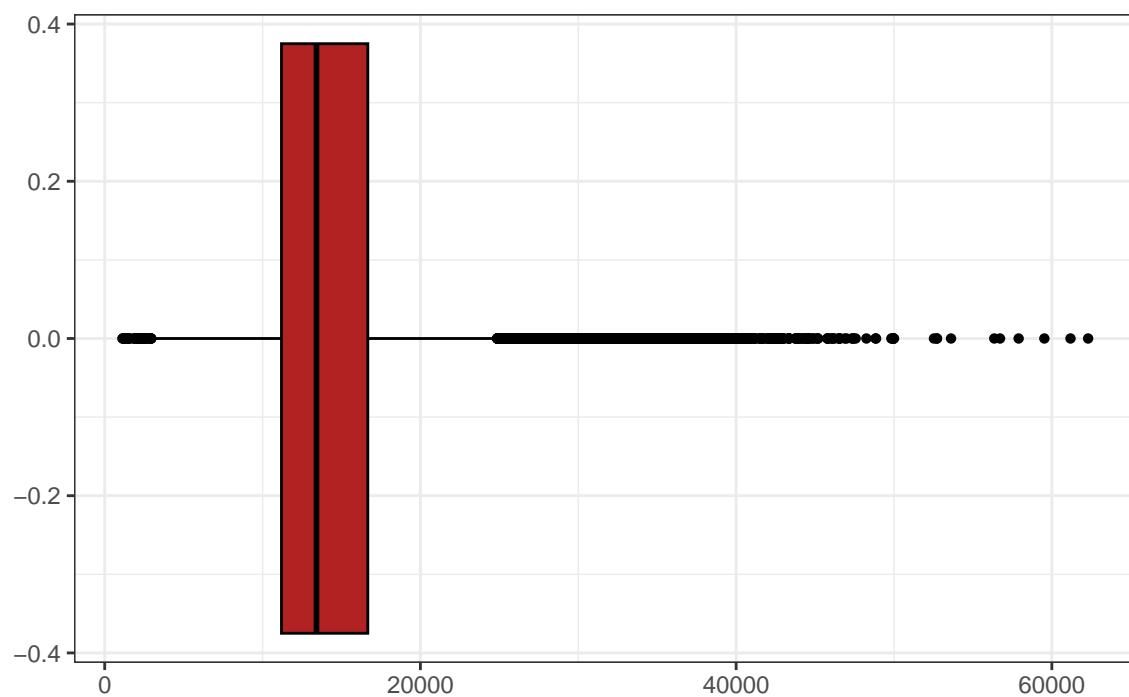
Wykres pudelkowy pieter mieszkń



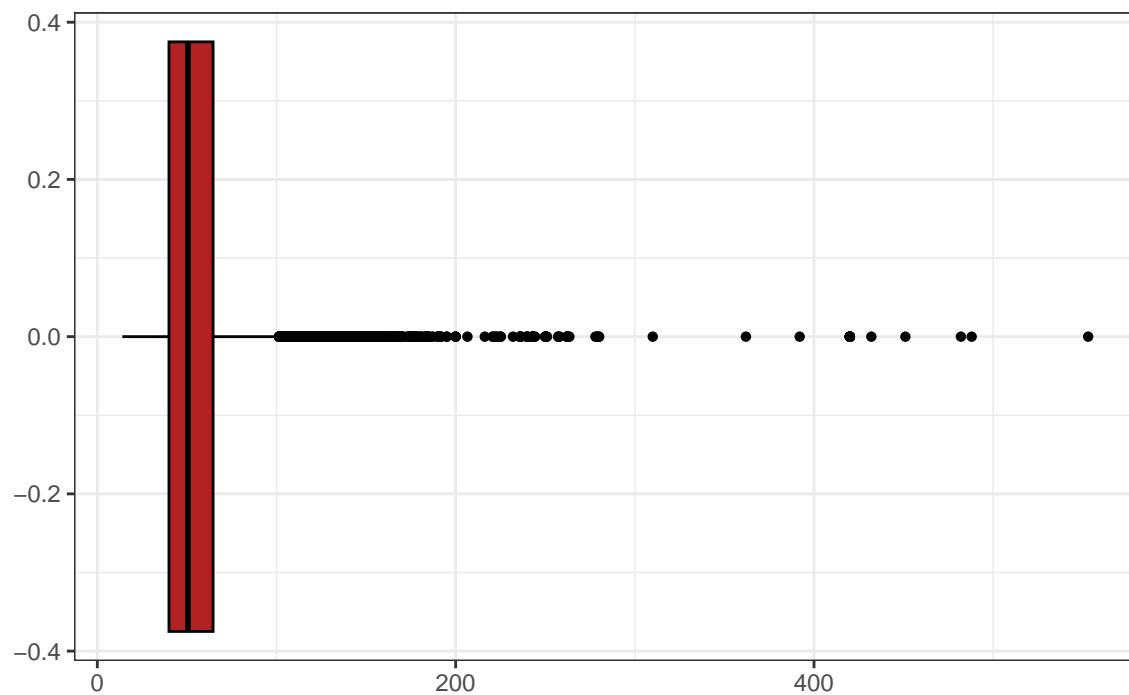
Wykres pudelkowy cen mieszkań w zł



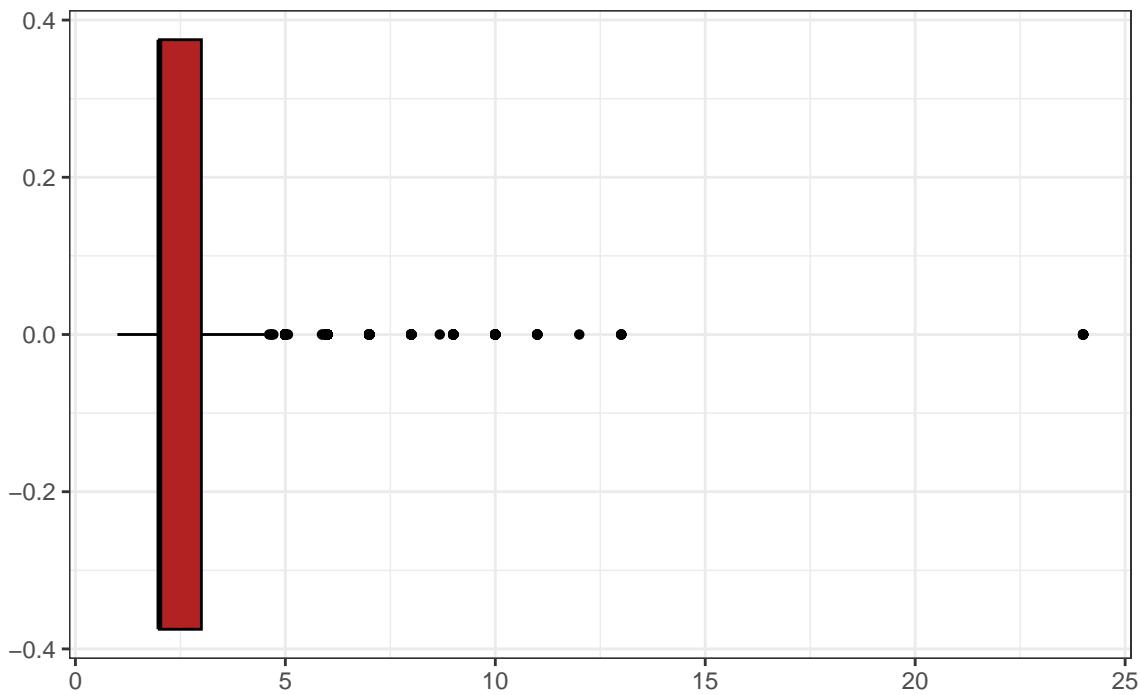
Wykres pudelkowy ceny za metr kwadratowy mieszkań w zł/n



Wykres pudelkowy powierzchni mieszkań w m²



Wykres pudelkowy liczby pokoi



5 Obróbka zmiennych

5.1 Normalizacja zmiennych numerycznych

Przeprowadzamy normalizację zmiennych, które zostaną uwzględnione w analizie. To jedyne przekształcenie, które zostało wykonane.

6 Analiza wielowymiarowa

6.1 Analiza głównych składowych (PCA)

W zbiorze danych tylko pięć ze zmiennych można wykorzystać do przeprowadzenia analizy głównych składowych. Ponieważ PCA wykorzystuje kowariancję, którą oblicza się za pomocą wariancji, wśród zmiennych uwzględnionych w analizie mogą znaleźć się tylko zmienne, dla których istnieje logiczny sens obliczenia różnicę pomiędzy jedną a drugą wartością. W tym zbiorze danych występuje 5 takich zmiennych: cena, powierzchnia, liczba pokoi, piętro oraz rok budowy. Cena za metr kwadratowy zostaje pominięta ze względu na redundancję informacji ze zmiennymi ceny oraz powierzchni. Podane 5 zmiennych zostało wykorzystane do analizy głównych składowych.

```
## $chisq
```

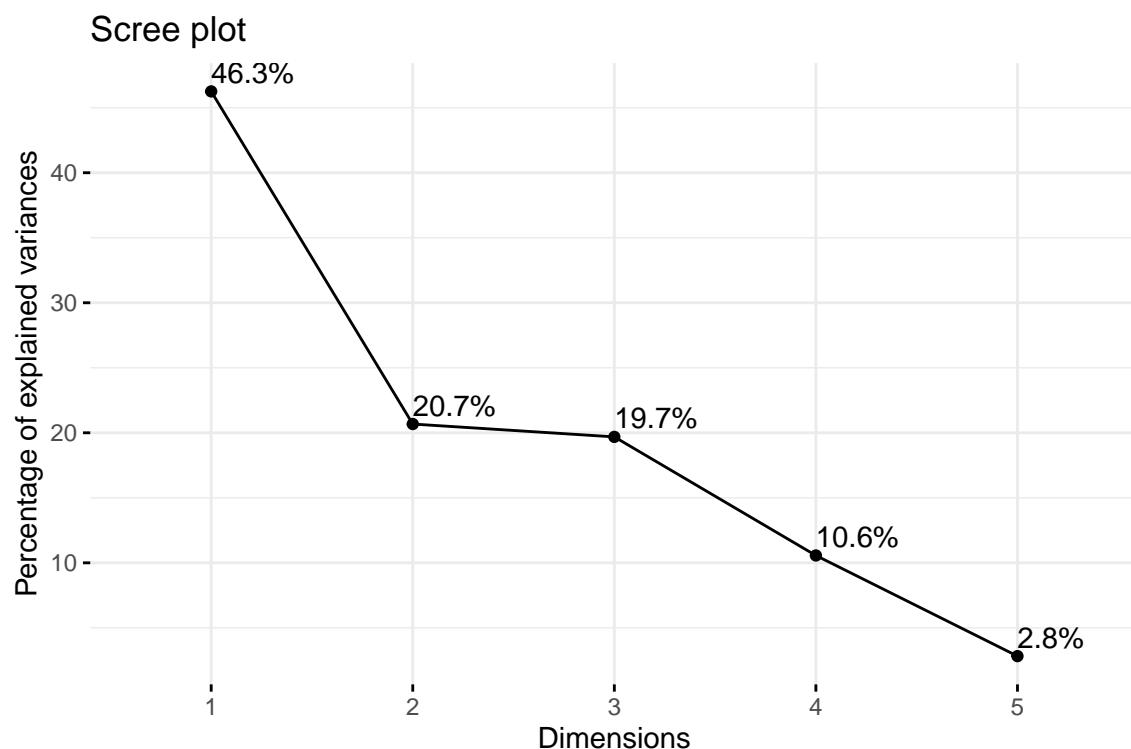
```

## [1] 167.9161
##
## $p.value
## [1] 7.489387e-31
##
## $df
## [1] 10

```

Za pomocą testu Bartletta sprawdzamy czy sensowne jest wykonywanie PCA. Hipoteza zerowa tego testu mówi, że macierz korelacji jest macierzą jednostkową, czyli zmienne nie są skorelowane, a w tym przypadku nie ma sensu wykonywać PCA. Hipoteza alternatywna z kolei mówi, że macierz korelacji nie jest macierzą jednostkową, więc zmienne są skorelowane, a zatem istnieje sens wykonywania PCA.

Na podstawie wykonanego testu Bartletta widać, że wartość p wynosi ok. 7,5e-31, a zatem odrzucamy hipotezę zerową. Z tego względu w kolejnym kroku wykonujemy PCA.



```

##      eigenvalue variance.percent cumulative.variance.percent
## Dim.1  2.3127383        46.25477                46.25477
## Dim.2  1.0336062        20.67212                66.92689
## Dim.3  0.9842534        19.68507                86.61196
## Dim.4  0.5281681        10.56336                97.17532
## Dim.5  0.1412340        2.82468                100.00000

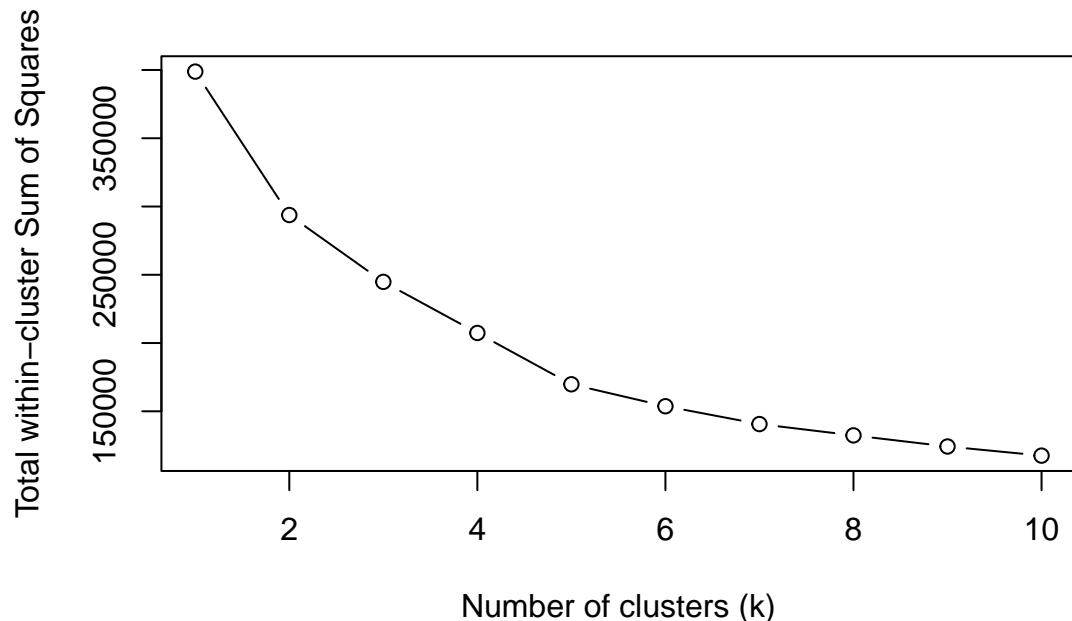
```

Korzystając z kryterium Keisera można określić, że mieszkania można określić za pomocą pierwszych dwóch głównych składowych, 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 skupień (clustering)

Drugą wykonaną analizą jest analiza skupień. Została tutaj wykorzystana metoda k-średnich. Analiza została przeprowadzona dla tych samych zmiennych, co PCA.

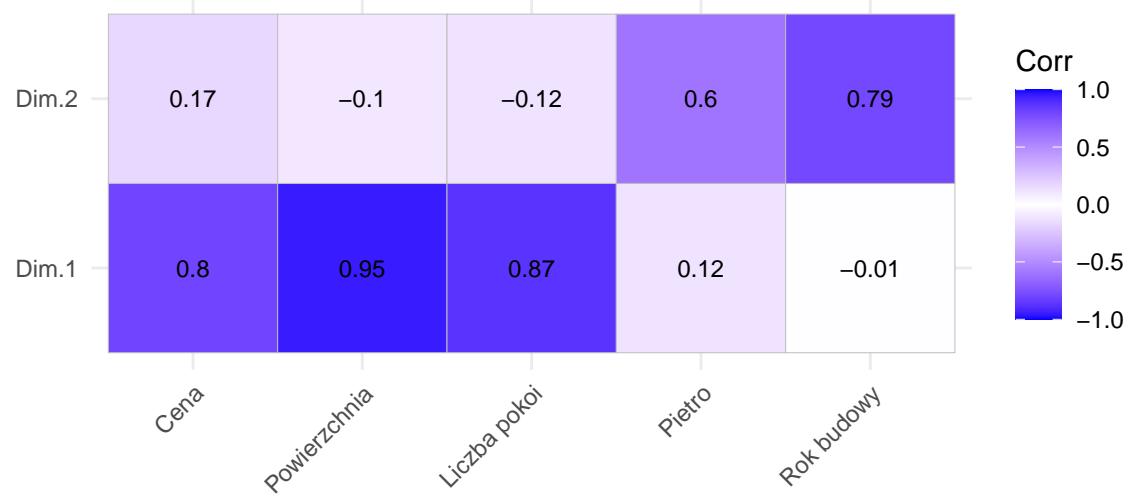
Ze względu na rozmiar zbioru danych, niemożliwe jest stworzenie wykresów za pomocą funkcji fviz_nbclust, nawet dla metody lokciowej WSS, ze względu na ich złożoność obliczeniową związaną z wielokrotnym klastrowaniem i zapisywaniem tych wyników. Z tego względu, jedyną alternatywą jest manualne zastosowanie metody k-średnich, przy zapisywaniu tylko aktualnego obiektu kmeans oraz sumy kwadratów z każdej liczby klastrów, a następnie przedstawienie sum kwadratów wewnętrz klastrów na wykresie, co wykonano poniżej.



Wykres nie daje bardzo konkluzywnej odpowiedzi na pytanie ile klastrów użyć, natomiast na podstawie oceny wykresu można powiedzieć, że 3 klastry powinny być dobrą opcją. Analiza została zatem przeprowadzona dla trzech klastrów.

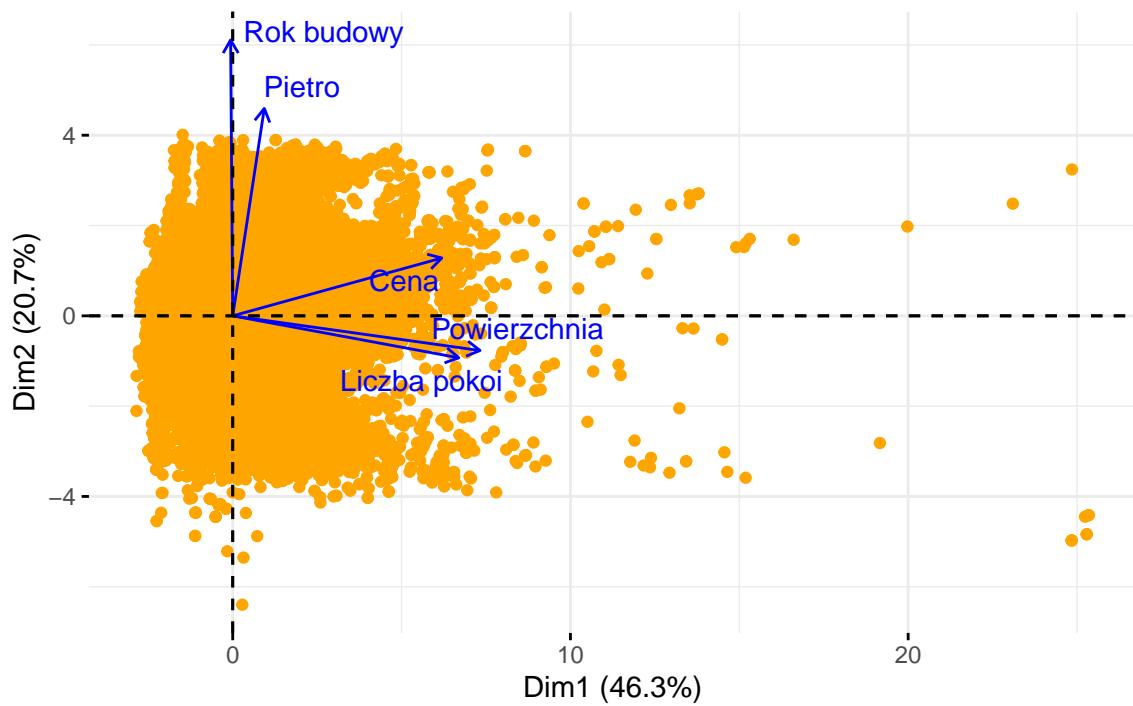
7 Wizualizacja metod

7.1 Wizualizacje PCA



Mieszkania można zatem opisać za pomocą dwóch składowych głównych. Pierwsza z nich dotyczy parametrów samego mieszkania, które dotyczą rozmiaru oraz wartości rynkowej nieruchomości. Obserwacje z niskimi wartościami tej składowej to małe, tańsze mieszkania, np. kawalerki, natomiast wyższe wartości mogą oznaczać duże, wielopokojowe nieruchomości. Druga składowa główna dotyczy właściwości samego budynku, w którym znajduje się to mieszkanie. Niskie wartości tej składowej oznaczają niskie, stare budynki, a wyższe wartości wskazują na nowsze budynki w wysokiej zabudowie.

PCA – Biplot

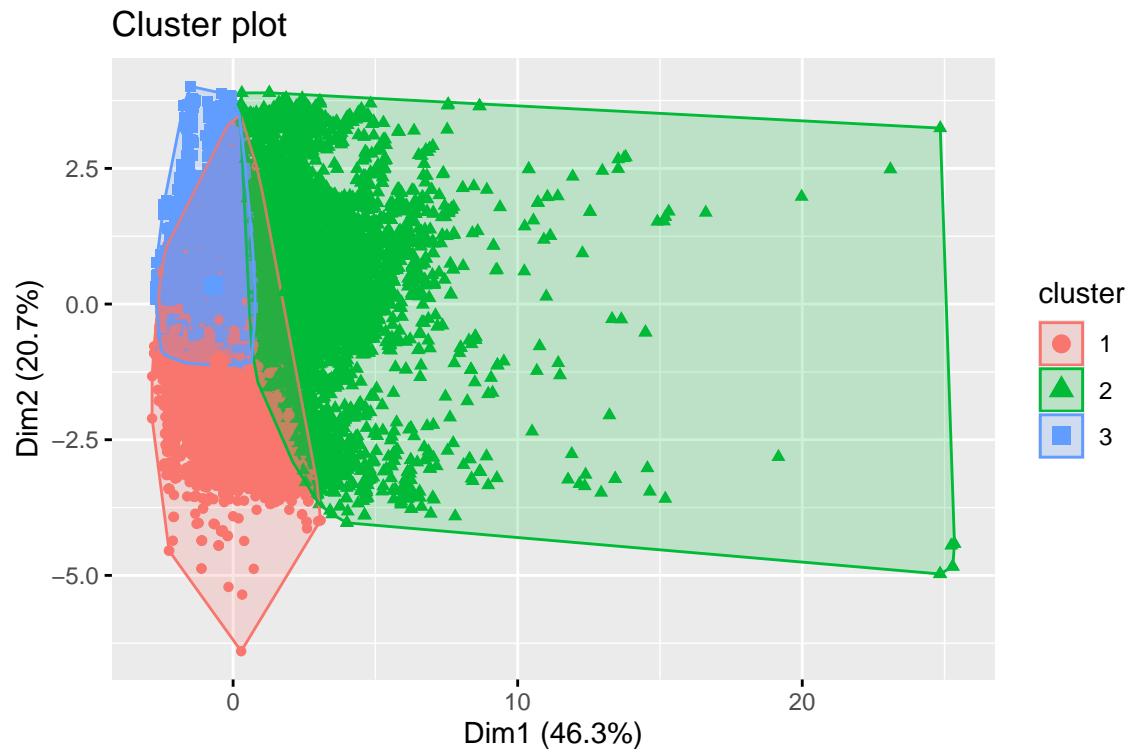


Wykres biplot pozwala na wyciągnięcie dwóch istotnych wniosków.

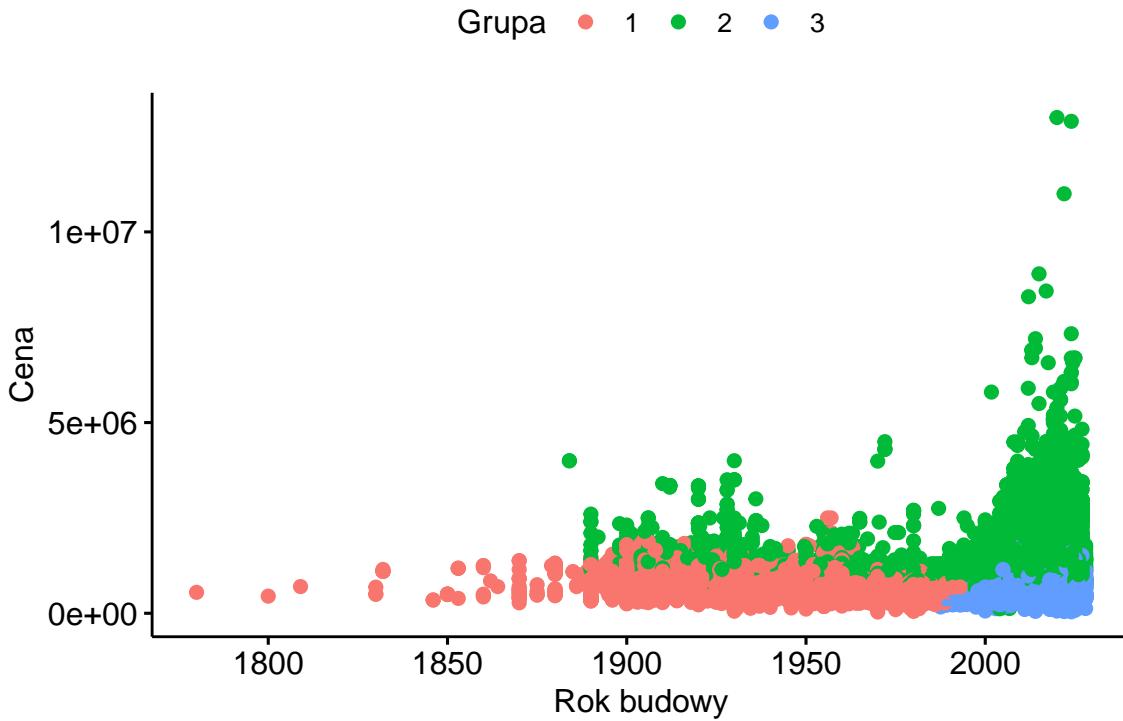
Po pierwsze, potwierdza się wniosek wyciągnięty z macierzy korelacji o sile i kierunku wpływu zmiennych na poszczególne główne składowe.

Po drugie, przedstawienie obserwacji za pomocą punktów wskazuje, że występuje mała ilość silnie odstających obserwacji w pierwszym z wymiarów. Mogą to być wielopokojowe apartamenty przeznaczone na wynajem pojedynczych pokoi, a być może i hostele lub hotele. W drugim wymiarze nie ma tak silnie odstających obserwacji, lecz pojawiają się stare nieruchomości o niskiej zabudowie.

7.2 Wizualizacje analizy skupień



Ze względu na powiązanie metody k-średnich z metodą PCA, wyniki analizy skupienia można nałożyć na rozrzut według głównych składowych. Na podstawie wykresu można ocenić, że powstałe grupy to tańsze, starsze mieszkania, tańsze, nowsze mieszkania oraz droższe mieszkania ogółem.



Na powyższym rysunku przedstawiono wykres punktowy roku budowy oraz ceny, z zaznaczeniem grup dla każdej obserwacji. Tutaj również widać wniosek, że tańsze mieszkania zostały podzielone na te nowsze i starsze, a droższe mieszkania stanowią swoją odrębną grupę.

8 Wnioski

Na podstawie wykonanych analiz można powiedzieć, że mieszkania w Gdańsku można opisać za pomocą ich własnych parametrów, a także cech budynku, w którym się znajdują. Ponadto zauważalny jest wyraźny podział pomiędzy tańszymi i droższymi mieszkańami, przy czym w tych tańszych mieszkańach występuje dodatkowy podział pomiędzy starszymi i droższymi mieszkańami. Oznacza to, że dla tańszych mieszkań rok budowy jest istotnym czynnikiem różnicowania mieszkań. Zależność ta jednak zanika dla droższych mieszkań.

Niestety nie udała się analiza wielu zmiennych ze zbioru danych, ze względu na brak znajomości odpowiednich metod pozwalających na jednoczesną analizę zmiennych numerycznych i kategorycznych. Ze względu na brak wystarczającej liczby kategorii, analiza korespondencji nie mogła zostać zastosowana. W przypadku kontynuacji prac nad projektem analiza została by rozszerzona o wielowymiarowe metody dla zmiennych kategorycznych oraz metody dla mieszanych zbiorów danych.