AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Statystyczna analiza danych

Redukcja wymiarowości i statystyczne testowanie hipotez

Przedmiot: Statystyczna analiza danych

Kierunek studiów: Informatyka i ekonometria

Wojciech Liberacki

Opis danych

Dane, wykorzystane w projekcie pochodzą ze strony kaggle.com, odnoszą się do Krakowskich mieszkań wystawionych na sprzedaż w 2024 roku.

Zmienne:

- price – cena mieszkania [PLN]
- squareMeters – metry kwadratowe powierzchni mieszkania
- poiCount – ilość użytecznych miejsc (szkoły, sklepy, apteki itp.) w odległości do 500m od mieszkania
- rooms – ilość pokojów w mieszkaniu
- centreDistance – odległość od centrum miasta [km]

Statystyki opisowe

Tabela 1. Statystyki opisowe

	squareMeters	centreDistance	poiCount	price	rooms
Średnia	55.4	4.3	20.6	861330	2.59
Mediana	49.9	4.39	14	755830	2
Odchylenie standardowe	22.5	2.13	26.4	357922	0.988
Minimum	25.1	0.54	0	379000	1
Maksimum	146	9.67	186	2693600	6
Skośność	1.7	0.183	3.3	2.16	1.08
Kurtoza	3.41	-0.676	12.8	6.64	1.73

Metraż mieszkań waha się od 25.1 m² do 146 m², ze średnią wynoszącą 55.4 m² i medianą 49.9 m². Metraże są zróżnicowane, choć większość mieszkań skupia się wokół średniej, co może sugerować różnorodność rozmiarów mieszkań.

Odległość od centrum miasta, wyrażona w kilometrach, obejmuje zakres od 0.54 km do 9.67 km. Skośność bliska zeru sugeruje, że większość mieszkań jest umieszczona w okolicach centrum, jednak wartości maksymalne wskazują na obecność mieszkań oddalonych od centrum.

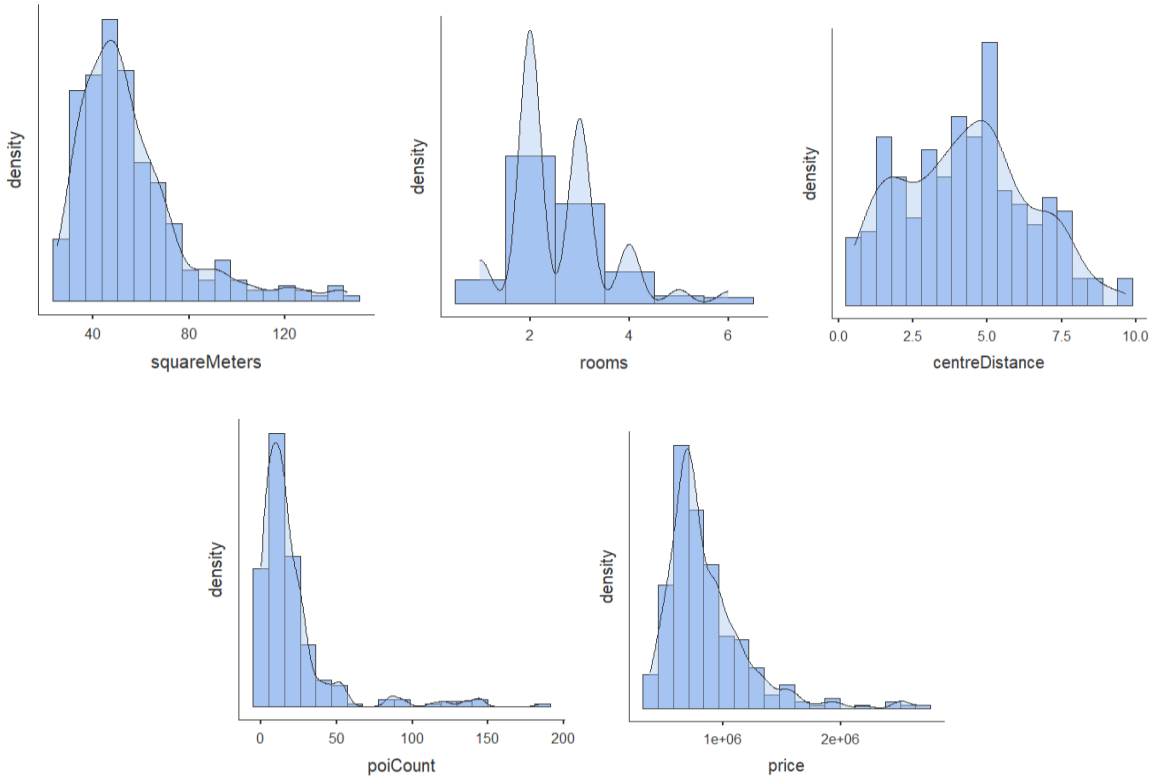
Liczba punktów interesujących (POI – point of intrest) w okolicy mieszkań oscyluje między 0 a 186, ze średnią wynoszącą 20.6 i medianą 14. Wysoka skośność i kurtoza sugerują znaczną zmienność liczby POI, co może oznaczać zróżnicowane otoczenie.

Ceny mieszkań w przedstawionym zbiorze danych mieszczą się w zakresie od 379000 PLN do 2693600 PLN, z średnią wartością wynoszącą 861330 PLN i medianą 755830 PLN. Skośność wynosząca 2.16 wskazuje na obecność cen mieszkań silnie przesuniętych w prawo, co może oznaczać dominację droższych nieruchomości.

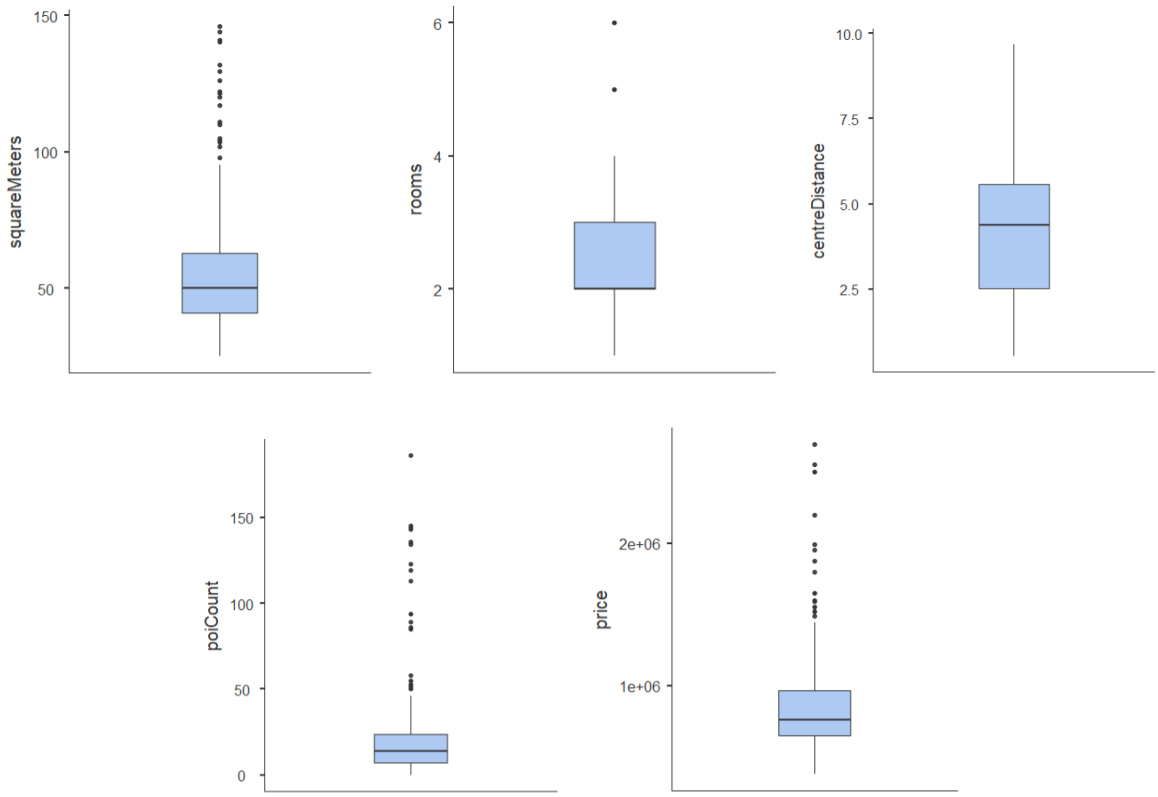
Liczba pokoi w mieszkaniach oscyluje od 1 do 6, ze średnią wynoszącą 2.59 i medianą 2. Wartość kurtozy sugeruje, że rozkład liczby pokoi ma grubsze ogony niż rozkład normalny, co może oznaczać obecność mieszkań z nietypową liczbą pokoi.

Histogramy i wykresy pudełkowe

Rysunek 1. Histogramy zmiennych



Rysunek 2. Wykresy pudełkowe zmiennych



Dzięki analizie histogramów i wykresów pudełkowych, można zidentyfikować wartości odstające i się ich pozbyć, żeby nie sprawiały problemów w dalszych analizach (zredukowano dane z 298 do 249 obserwacji).

Macierz korelacji

	squareMeters	rooms	centreDistance	poiCount	price
squareMeters	—				
rooms	0.762	—			
centreDistance	0.087	0.081	—		
poiCount	-0.1	-0.124	-0.47	—	
price	0.732	0.549	-0.269	0.081	—

Widzimy, że każda ze zmiennych jest dobrze skorelowana ze zmienną price (docelowo - zmienna objaśniana). Między zmiennymi objaśniającymi nie istnieje współliniowość.

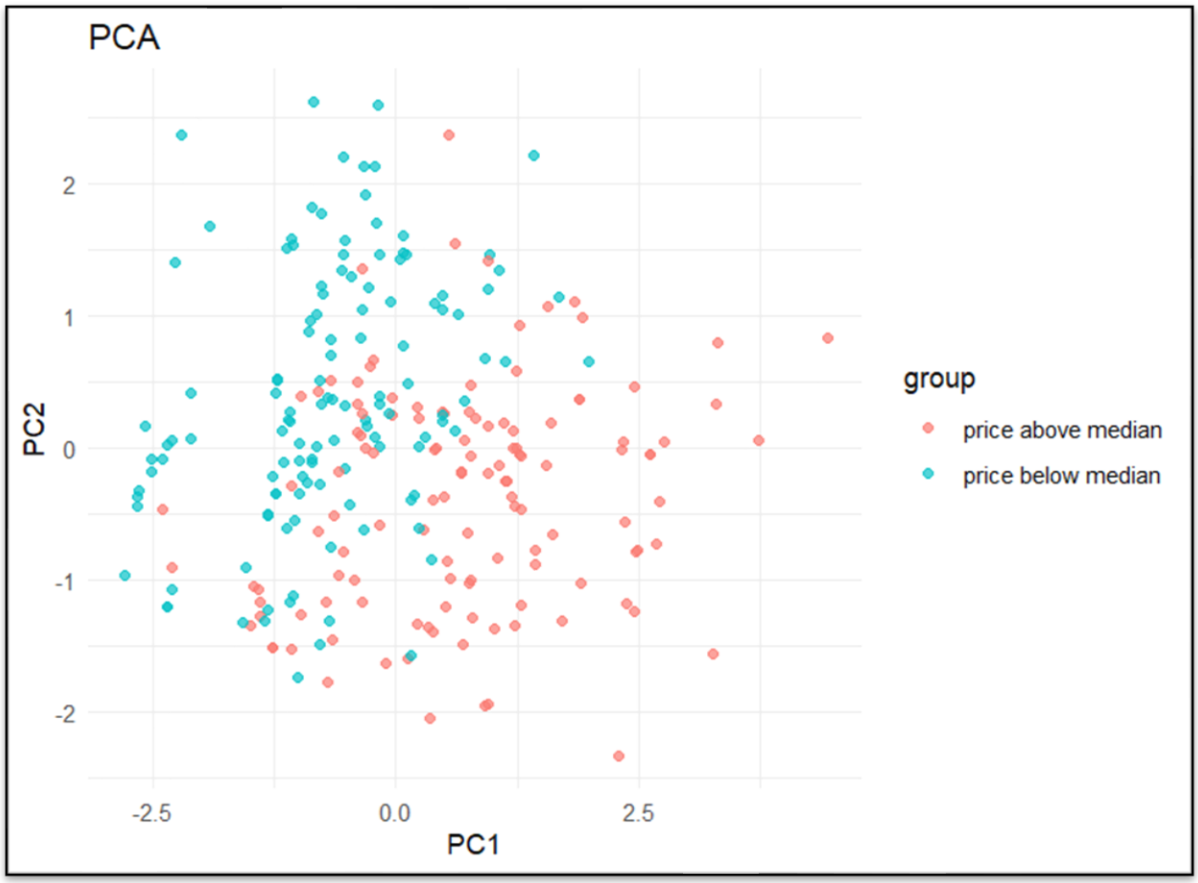
Redukcja wielowymiarowości

Tabela 2. Procent wariancji

explained_variance	0.46522367	0.34300155	0.13256841	0.05920637
cumulative_explained_variance	47%	81%	94%	100%

Z racji na to, że trzy pierwsze składowe wyjaśniają 94% wariancji a ostatnia tylko 6%, to je wykorzystam do PCA(analiza składowych głównych). Punkty zostały podzielone na te reprezentujące mieszkania o cenie wyższej lub niższej niż ogólna mediana cen.

Rysunek 3. Analiza składowych głównych



Dane zredukowane do dwóch wymiarów, tworzą wykres, na którym można dostrzec pewną oś podziału mieszkań – te wycenione poniżej mediany i te powyżej. Oczywiście podział nie jest idealnie dokładny, bo nieruchomości na aukcje wystawiają różne jednostki, a wycena nieruchomości to bardzo złożone zagadnienie.

Regresja liniowa

Celem analizy jest określenie, które zmienne najlepiej estymują cenę mieszkania. Z racji na specyfikę danych do estymacji modelu zastosuję model regresji liniowej, a do analizy dwóch modeli - **ANOVĘ**.

Tabela 3. 1 Model regresji liniowej

Współczynnik	Estymacja	Błąd Std.	Wartość t	Pr(>|t|)
(Intercept)	348421.8	41676.32	8.36	4.81e-15 ***
squareMeters	11842.31	904.5	13.093	<2e-16 ***
rooms	-2647.76	15769.23	-0.168	0.867
centreDistance	-34787.3	4458.46	-7.803	1.77e-13 ***
poiCount	-20.93	817.4	-0.026	0.98

Stosuję ANOVĘ, aby sprawdzić słuszność podejrzenia, że model byłby efektywniejszy bez zmiennej poiCount.

Tabela 4. 1 ANOVA

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	245	3.70E+12				
2	244	3.70E+12	1	9937696	7.00E-04	0.9796

Suma kwadratów reszt (RSS) dla obu modeli jest praktycznie identyczna, co wskazuje, że dodanie poiCount nie zmniejszyło błędu modelu.

Bardzo niska wartość statystyki F i wysokie p-value (0.9796) wskazuje, że dodanie zmiennej poiCount do modelu nie jest statystycznie istotne.

Zatem usuwam zmienną poiCount z modelu.

Współczynnik	Estymacja	Błąd Std.	Wartość t	Pr(>|t|)
(Intercept)	347813.2	34161.1	10.182	<2e-16 ***
squareMeters	11842	902.6	13.12	<2e-16 ***
rooms	-2618.4	15695.2	-0.167	0.868
centreDistance	-34734.2	3939	-8.818	<2e-16 ***

Tabela 5. 2 Model regresji liniowej

Następnie przy pomocy ANOVY, testuję kolejną zmienną wobec, której mam podejrzenia, że nie jest optymalna dla modelu - rooms.

Tabela 6. 2 ANOVA

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
3	245	3.70E+12				
4	246	3.70E+12	-1	-4.2E+08	0.0278	0.8676

ANOVA pokazuje, że nie ma istotnej różnicy między modelem z rooms a modelem bez tej zmiennej (p-value = 0.8676), zmienna rooms nie wnosi znaczącej wartości do modelu.

Finalny model i podsumowanie

Tabela 7. 3 Model regresji liniowej

Współczynnik	Estymacja	Błąd Std.	Wartość t	Pr(>|t|)
(Intercept)	347284.2	33946.4	10.23	<2e-16 ***
squareMeters	11727.5	584.8	20.054	<2e-16 ***
centreDistance	-34748.7	3930.2	-8.841	<2e-16 ***

Obie zmienne wyjaśniające w modelu są istotne statystycznie. Model wyjaśnia około 64.8% zmienności ceny nieruchomości, co jest dość dobrym wynikiem, wskazującym na umiarkowanie silną zależność między zmiennymi a ceną nieruchomości. Model ten skutecznie pokazuje, że zarówno metraż, jak i odległość od centrum mają istotny wpływ na cenę nieruchomości.