Raport Ceny Mieszkań

Natan Warelich 318159

Spis treści

1 Wprowadzenie

Do raportu cen domów. Wykorzystano model regresji liniowej.

2 Eksploracyjna analiza danych

2.1 Wczytanie danych

Podobnie jak w pierwszym raporcie, dane zostały wczytane przy użyciu biblioteki `pandas`. Ze względu na problemy z interpretacją separatorów dziesiętnych, zastosowano zamianę przecinków na kropki w kolumnach zawierających liczby zmiennoprzecinkowe. Przykładowo:

```
df['Lot_Size'] = df['Lot_Size'].str.replace(',', '.').astype(float)
df['House_Price'] = df['House_Price'].str.replace(',', '.').astype(float)
```

2.2 Weryfikacja zgodności danych z opisem

Wszystkie dane zostały zweryfikowane pod kątem kompletności i zgodności z założeniami. Wyniki weryfikacji przedstawiają się następująco:

- brak brakujących danych,
- brak duplikatów,
- brak obserwacji odstających,
- brak wartości niezgodnych z założeniami, tj.:
  - brak wartości ujemnych tam, gdzie są one niedozwolone,
  - zmienne zawierają wartości w zdefiniowanych zakresach.

2.3 Rozkład zmiennych kategorialnych

Zmiennym kategorialnym przypisano następujące zakresy:

- liczba sypialni: od 1 do 5,
- liczba łazienek: od 1 do 5,
- liczba miejsc parkingowych (garaż): od 0 do 2,
- jakość osiedla: od 1 do 10 (zgodnie z opisem danych).

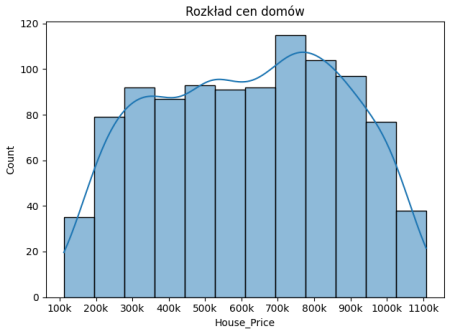
Rozkład częstości występowania poszczególnych wartości zmiennej `Num_Bedrooms` przedstawia się następująco:

```
Num_Bedrooms
1:    201
2:    215
3:    182
4:    197
5:    205
```

Rozkład ten sugeruje względnie równomierne rozłożenie liczby sypialni w analizowanej próbie.

2.4 Rozkład cen domów

Rozkład cen domów przedstawiono na rysunku 1.

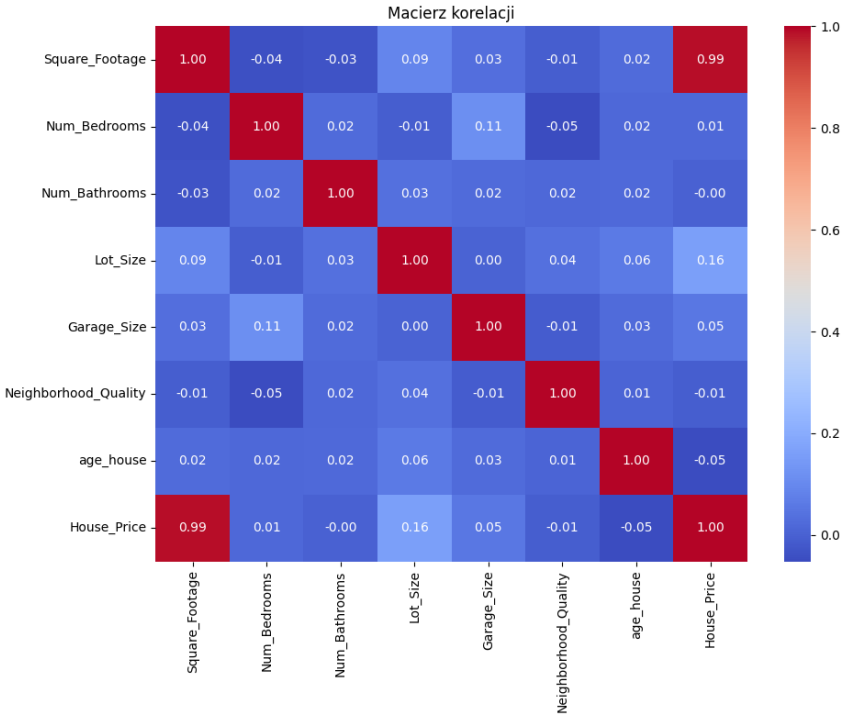


Rysunek 1: Rozkład cen domów

Rozkład przypomina rozkład normalny — domy o najniższych i najwyższych cenach występują rzadziej, a największe zagęszczenie obserwacji przypada na przedział około 700–750 tysięcy. Należy jednak zaznaczyć, że występują nieregularności w przedziale 350–650 tysięcy, co wskazuje na odchylenie od klasycznego rozkładu normalnego.

2.5 Macierz korelacji i współczynników determinacji

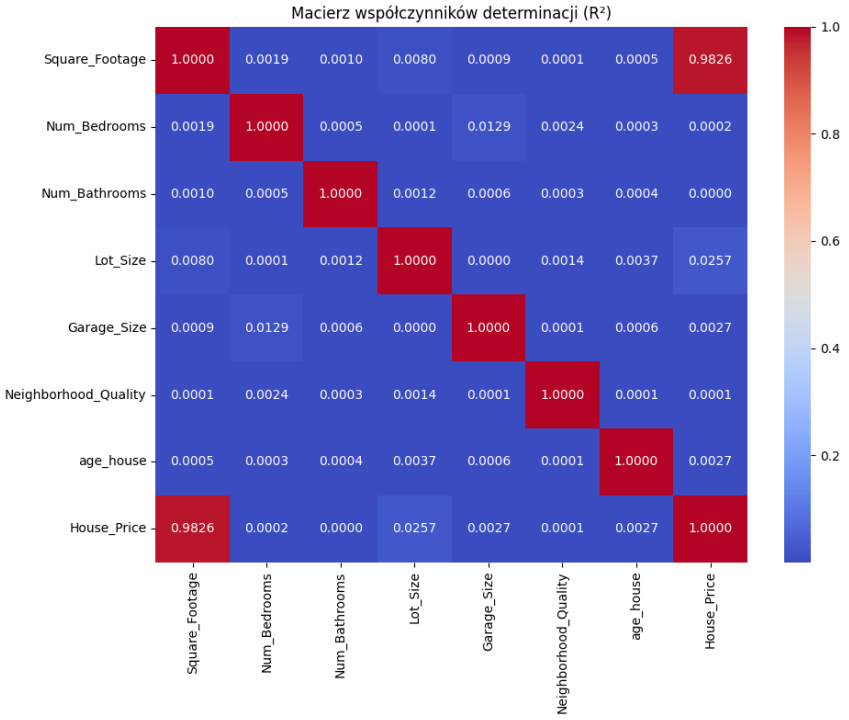
Macierz korelacji liniowej Pearsona została przedstawiona na rysunku 2.



Rysunek 2: Macierz korelacji Pearsona

Zaobserwowano silną dodatnią korelację pomiędzy ceną domu a powierzchnią użytkową (`Square_Footage`) oraz wielkością działki (`Lot_Size`). Warto zauważyć, że niektóre zmienne wykazują korelację ujemną, co oznacza, że ich wzrost może wiązać się ze spadkiem ceny domu. Dotyczy to m.in. wieku budynku (`age_house`), którego współczynnik korelacji wynosi $r = -0{,}05$. Wynik ten jest zgodny z intuicją – starsze domy są zazwyczaj mniej atrakcyjne ze względu na większe zużycie techniczne.

Z kolei macierz współczynników determinacji (R^2), czyli kwadratów korelacji Pearsona, została przedstawiona na rysunku 3.



Rysunek 3: Macierz współczynników determinacji (R^2)

Wartości współczynników determinacji potwierdzają wcześniejsze obserwacje — zmienne silnie skorelowane z ceną mają również wysokie wartości R^2.

2.5.1 Zależność ceny od powierzchni domu

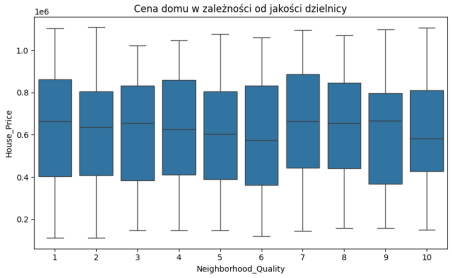
Jedną z najbardziej oczywistych zależności jest ta między ceną domu a jego powierzchnią. Zależność ta została przedstawiona na rysunku 4.

Rysunek 4: Zależność między ceną a powierzchnią domu

Widoczna jest wyraźna liniowa zależność – większość obserwacji układa się wzdłuż jednej prostej, co sugeruje zasadność zastosowania liniowego modelu regresji w dalszej analizie.

2.5.2 Zależność ceny od jakości osiedla

Zależność między ceną domu a jakością osiedla została przedstawiona na rysunku 5.



Rysunek 5: Zależność między ceną a jakością osiedla

Na wykresie zauważalna jest pozornie paradoksalna zależność — osiedla o najwyższej jakości (10) mają niższą medianę oraz górny kwartyl cen niż te o najniższej jakości (1). Po dokładniejszej analizie okazuje się, że średnia powierzchnia domów w osiedlach o najniższej jakości jest wyższa, co może tłumaczyć wyższe ceny w tej grupie. Poniżej przedstawiono fragment kodu i wybrane wyniki:

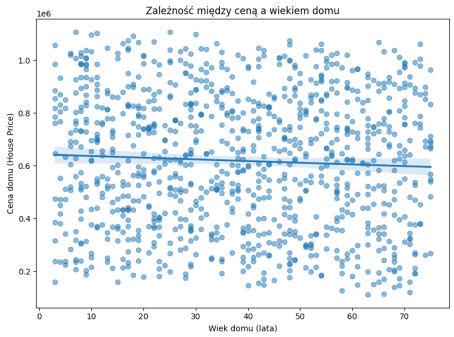
```
print(df[df['Neighborhood_Quality'] == 10]['Square_Footage'].mean())
print(df[df['Neighborhood_Quality'] == 1]['Square_Footage'].mean())
```

```
Neighborhood Quality = 10: mean = 2760.57
Neighborhood Quality = 1: mean = 2876.97
```

Jak widać, domy w osiedlach o niższej jakości są przeciętnie większe, co może wpływać na ich wyższą cenę pomimo gorszej lokalizacji.

2.5.3 Zależność ceny od wieku budynku

Zależność ceny domu od jego wieku została przedstawiona na rysunku 6.



Rysunek 6: Zależność ceny domu od wieku budynku

Dane są rozproszone i nie wykazują jednoznacznej zależności liniowej. Brak wyraźnego trendu sugeruje, że wiek budynku ma ograniczoną wartość predykcyjną przy tworzeniu prostego modelu regresji. Niemniej jednak, ze względu na możliwe nieliniowe zależności lub interakcje z innymi zmiennymi, zmienna ta może okazać się użyteczna w bardziej zaawansowanych modelach.

3 Prosta regresja liniowa

Pierwszym zastosowanym modelem predykcyjnym była prosta regresja liniowa. Wybór ten został podyktowany silną zależnością pomiędzy ceną nieruchomości a jej powierzchnią użytkową, co intuicyjnie wskazuje na zasadność wykorzystania liniowego modelu regresyjnego.

3.1 Parametry modelu

Model został uruchomiony z wykorzystaniem następujących parametrów:

```
random_state = 318153
test_size = 0.25
```

3.2 Wstępna ocena jakości modelu

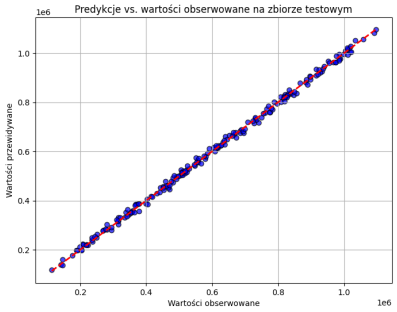
Na podstawie wyników uzyskanych dla danych treningowych i testowych, oszacowano następujące metryki jakości modelu:

Metryka	Zbiór treningowy	Zbiór testowy
RMSE	9626.07	10190.66
MAE	7599.89	8271.13
MAPE	1.53%	1.73%

Tabela 1: Porównanie metryk regresji dla zbioru treningowego i testowego

Wyniki te należy uznać za bardzo zadowalające – odchylenie względne rzędu 1.73% w zbiorze testowym wskazuje na wysoką trafność predykcji.

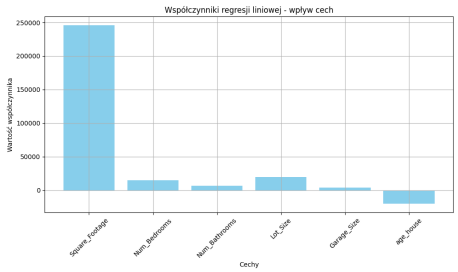
Na rysunku 7 przedstawiono porównanie wartości predykowanych oraz rzeczywistych w zbiorze testowym:



Rysunek 7: Porównanie wartości predykowanych i rzeczywistych (zbiór testowy)

Większość obserwacji skupiona jest w bezpośrednim sąsiedztwie linii regresji, co dodatkowo potwierdza trafność modelu.

Na kolejnej wizualizacji (rysunek 8) przedstawiono wpływ poszczególnych cech na wartości predykowane:



Rysunek 8: Wpływ zmiennych objaśniających na wynik regresji

Zgodnie z przewidywaniami, zmienna dotycząca jakości osiedla wykazuje zerowy wpływ na wynik modelu. W związku z powyższym, została ona wykluczona z dalszych analiz.

3.3 Identyfikacja obserwacji nietypowych

Pomimo zadowalających wyników, istotne jest przeprowadzenie analizy pod kątem obecności obserwacji nietypowych, które mogą zafałszować ocenę modelu. Do głównych zagrożeń dla regresji liniowej zalicza się:

- obserwacje o wysokiej dźwigni (ang. *high leverage points*),
- obserwacje wpływowe (ang. *influential points*),
- obserwacje odstające (ang. *outliers*).

Do identyfikacji powyższych przypadków zastosowano analizę z wykorzystaniem odległości Cooka. Poniżej zaprezentowano najważniejsze wyniki tej analizy, przeprowadzonej za pomocą skryptu `prosta_regresja_liniowa_poszukiwania_pkt_nietypowych.py`:

```
Top 5 punktów wysokiej dźwigni:
[531 736 486 670 697]

Obserwacje wpływowe (Cook's distance > 0.0053):
[  7  19  20  36  54  72  84  92 156 172 179 180 229 233 255 277 321 340
 370 377 391 460 462 500 502 570 585 616 617 632 634 648 651 660 710 720
 726]

Liczba punktów odstających powyżej +3: 3
Index: 391, Reszta standaryzowana: 0.04
Index: 460, Reszta standaryzowana: 0.12
Index: 585, Reszta standaryzowana: 1.07

Liczba punktów odstających poniżej -3: 2
```

```
Index: 180, Reszta standaryzowana: -0.34
Index: 377, Reszta standaryzowana: 0.11
```

Pomimo że liczba obserwacji nietypowych nie jest znaczna, zdecydowano się na ich usunięcie i ponowne przeszkolenie modelu. Uzyskane rezultaty przedstawiono w tabeli 2:

Metryka	**Zbiór treningowy**	**Zbiór testowy**
RMSE	9703.30	10186.16
MAE	7661.62	8270.61
MAPE	1.55%	1.73%

Tabela 2: Porównanie metryk regresji po usunięciu obserwacji nietypowych

Uzyskane zmiany są marginalne. Wartość MAPE w zbiorze treningowym wzrosła, a w zbiorze testowym pozostała praktycznie niezmieniona. Wskazuje to na minimalny wpływ usuniętych punktów na ogólną jakość modelu.

3.4 Walidacja modelu

W celu zwiększenia wiarygodności oceny modelu, przeprowadzono walidację z podziałem danych w proporcji 60% treningowe, 20% walidacyjne, 20% testowe. Otrzymane wyniki przedstawiono w tabeli 3:

Tabela 3: Metryki błędów regresji dla różnych zbiorów danych

Zbiór danych	**RMSE**	**MAE**	**MAPE**
Treningowy	9549.33	7572.03	1.51%
Walidacyjny	9845.29	7762.65	1.64%
Testowy	10369.97	8422.99	1.75%

Zastosowanie walidacji nie przyczyniło się do poprawy wyników. Wręcz przeciwnie – wartości błędów, w tym MAPE dla zbioru testowego, uległy pogorszeniu, co sugeruje, że model w swojej pierwotnej formie był już optymalny względem dostępnych danych.

3.5 Podsumowanie

Podsumowując, prosta regresja liniowa okazała się modelem wysoce skutecznym w kontekście analizowanego zbioru danych. Wykluczenie zmiennej opisującej jakość osiedla — ze względu na jej zerowy wpływ — okazało się zasadne. Brak przeuczenia oraz niska wartość MAPE (1.73% dla danych testowych) wskazują, że model ten stanowi solidną bazę dla dalszych analiz predykcyjnych.