

Maciej Kopczyński

Nr albumu: 402810

Zastosowanie algorytmów regresji w analizie i predykcji  
cen nieruchomości

Praca inżynierska  
na kierunku Analiza Danych

Praca wykonywana pod kierunkiem  
dr. Moniki Bartkiewicz  
Katedra Równań Różniczkowych  
i Informatyki

Łódź, 2025

**Słowa kluczowe:** Uczenie maszynowe, uczenie nadzorowane, uczenie nienadzorowane, modele liniowe, modele predykcyjne, modele automatycznej wyceny, regresja, regresja liniowa, regresja wieloraka, regresja lasso, regresja grzbietowa, estymacja OLS, błędy oszacowania, parametryzacja modelu, analiza reszt, heteroskedastyczność, autokorelacja, multikolinearność

**Title in English:** Application of Regression Algorithms in the Analysis and Prediction of Real Estate Prices

**Keywords:** Machine learning, supervised learning, unsupervised learning, linear models, predictive models, automated valuation models (AVMs), regression, linear regression, multiple regression, lasso regression, ridge regression, OLS (ordinary least squares) estimation, estimation errors, model parameterization, residual analysis, heteroscedasticity, autocorrelation, multicollinearity

Spis treści

[1. Wprowadzenie 5](#_Toc196676143)

[1.1. Zakres pracy 5](#_Toc196676144)

[1.2. Struktura pracy 6](#_Toc196676145)

[2. Kontekst teoretyczny i technologiczny 8](#_Toc196676146)

[3. Definicje i przykłady 9](#_Toc196676147)

[3.1. Uczenie maszynowe 10](#_Toc196676148)

[3.2. Uczenie nienadzorowane 12](#_Toc196676149)

[3.3. Uczenie nadzorowane 13](#_Toc196676150)

[3.4. Regresja 15](#_Toc196676151)

[3.5. Model regresji liniowej 18](#_Toc196676152)

[3.6. Prosta regresja liniowa 20](#_Toc196676153)

[3.7. Normalność błędów losowych 24](#_Toc196676154)

[3.8. Homoskedastyczność 25](#_Toc196676155)

[3.9. Autokorelacja 26](#_Toc196676156)

[3.10. Regresja wieloraka 27](#_Toc196676157)

[3.11. Multikolinearność 32](#_Toc196676158)

[3.12. Regresja grzbietowa (Ridge) 33](#_Toc196676159)

[3.13. Regresja Lasso 34](#_Toc196676160)

[3.14. Standaryzacja zmiennych - normalizacja 36](#_Toc196676161)

[3.15. Kodowanie one-hot encoding 36](#_Toc196676162)

[3.16. Wartości odstające 37](#_Toc196676163)

[3.17. Ocena jakości modelu - wzory 38](#_Toc196676164)

[3.18. Ocena jakości modelu - interpretacja 40](#_Toc196676165)

[4. Opis zbioru danych i środowiska pracy 43](#_Toc196676166)

[4.1. Źródło danych i sposób pozyskania 43](#_Toc196676167)

[4.2. Zmienna zależna SalePrice – definicja i znaczenie 45](#_Toc196676168)

[4.3. Charakterystyka i klasyfikacja zmiennych niezależnych 49](#_Toc196676169)

[4.4. Środowisko pracy i konfiguracja narzędzi 52](#_Toc196676170)

[4.5. Biblioteki i pakiety analityczne 53](#_Toc196676171)

[5. Analiza danych i implementacja modeli predykcyjnych 54](#_Toc196676172)

[5.1. Czyszczenie i przygotowanie danych 54](#_Toc196676173)

[5.2. Eksploracyjna analiza danych 61](#_Toc196676174)

[5.3. Transformacje i inżynieria danych 67](#_Toc196676175)

[5.4. Konstruowanie modeli regresji 73](#_Toc196676176)

[6. Podsumowanie wyników i znaczenie praktyczne 82](#_Toc196676177)

[7. Potencjalne ulepszenia i perspektywy dalszego rozwoju 86](#_Toc196676178)

[8. Bibliografia 87](#_Toc196676179)

1. Wprowadzenie

Rynek nieruchomości jest bardzo istotny dla gospodarki i ma kluczowe znaczenie dla sytuacji finansowej wielu gospodarstw domowych. W Polsce od kilku lat można zauważyć duże zmiany cen mieszkań – przede wszystkim ich wyraźny wzrost. Według danych Głównego Urzędu Statystycznego (GUS), w 2023 roku ceny mieszkań wzrosły średnio o 8,8% w porównaniu z poprzednim rokiem, a od roku 2015 ten wzrost wyniósł już prawie 80% [1]. Tak szybki wzrost cen sprawia, że analiza rynku mieszkaniowego oraz przewidywanie cen stają się szczególnie ważne i bardzo aktualne.

W 2022 roku znacznie zaostrzono warunki przyznawania kredytów hipotecznych, co poskutkowało gwałtownym spadkiem liczby transakcji na rynku. Według analiz NBP, w tym okresie sprzedało się około 35 tysięcy mieszkań – niemal dwukrotnie mniej niż rok wcześniej [2]. Mimo to ceny mieszkań nie spadły znacząco, co pokazuje, że ciągle istnieje silne zapotrzebowanie na nieruchomości [2]. Dlatego właśnie coraz większe znaczenie ma monitorowanie rynku oraz skuteczne przewidywanie cen. Jest to ważne zarówno dla inwestorów, deweloperów czy banków, jak i dla badaczy zajmujących się tym tematem. Żeby lepiej przewidywać ceny mieszkań, coraz częściej stosuje się zaawansowane metody analizy danych. Wśród nich szczególnie popularne stały się modele regresyjne, które są jednymi z podstawowych metod uczenia maszynowego. Modele regresji pozwalają analizować, jak różne cechy mieszkania – takie jak lokalizacja, wielkość, standard czy otoczenie – wpływają na jego cenę. Dzięki temu możliwe jest przewidywanie, ile będzie warte mieszkanie na podstawie wybranych parametrów wsadowych.

1.1. Zakres pracy

Głównym celem niniejszej pracy inżynierskiej jest zastosowanie algorytmów regresji w analizie i prognozowaniu cen mieszkań. Celem pracy jest stworzenie uniwersalnego modelu predykcji regresyjnej, który będzie mógł być szeroko zastosowany na danych rzeczywistych.  
Zakres praktyczny pracy obejmuje:

1. Przygotowanie zbioru danych do modelowania regresyjnego z wykorzystaniem technik transformacji danych. Studium przypadku zostało wykonane w oparciu o podstawowy zbiór danych nieruchomości Ames Housing rozszerzony o uzupełniające informacje przestrzenne.
2. Przygotowanie bazowych modeli regresji opartych o metody OLS oraz regularyzację Ridge i Lasso.
3. Ocenę jakości podstawowych modeli bazowych i zastosowanie hiperparametryzacji w celu znalezienia optymalnych wartości parametrów.
4. Porównanie wyników modeli i selekcję modelu cechującego się najmniejszymi błędami predykcji.
5. Podsumowanie rozważań praktycznych i wskazanie najważniejszych predyktorów ceny nieruchomości dla najoptymalniejszego modelu regresji.

Praca obejmuje cały proces od przygotowania i eksploracji danych, przez budowanie modeli regresyjnych, aż po ich ocenę pod kątem dokładności prognoz. Dodatkowo przeanalizowano także znaczenie różnych cech mieszkań, dzięki czemu udało się wyciągnąć praktyczne wnioski dotyczące tego, co najbardziej wpływa na ich cenę. Tak zdefiniowany cel oraz zakres pracy umożliwiają ocenę przydatności algorytmów regresyjnych jako narzędzi wspomagających prognozowanie cen mieszkań.

1.2. Struktura pracy

Praca ustrukturyzowana jest w następujący sposób (z wyłączeniem **1.** Wprowadzenia):

1. Kontekst teoretyczny i technologiczny – rozdział zawiera przegląd dotychczasowego stanu wiedzy i technologii związanych z tematem pracy. Wyjaśniony został szerszy kontekst powstania pracy.
2. Definicje i przykłady – rozdział zawiera definicje wykorzystywanych w pracy teoretycznych pojęć oraz praktyczne przykłady użycia wraz z niezbędnymi obliczeniami.
3. Opis zbioru danych i środowiska pracy – część pracy, w której znajduje się dokładny opis zbiorów danych wykorzystywanych do prowadzonych analiz i predykcji wraz ze źródłami ich pochodzenia. W sposób szczegółowy opisano strukturę, wymiarowość oraz ogólną jakość danych. W rozdziale omówiono również wykorzystywane narzędzia i środowisko pracy.
4. Analiza danych i implementacja modeli predykcyjnych – sekcja pracy, gdzie szczegółowo omówiono proces badawczo-projektowy począwszy od etapu przygotowania danych, przez budowanie modeli predykcyjnych, aż po końcową ocenę ich skuteczności w oparciu o wybrane metody.
5. Podsumowanie wyników i znaczenie praktyczne – rozdział podsumowujący rozważania i przedstawiający wyniki pracy wraz z osadzeniem rezultatów w świecie rzeczywistym.
6. Potencjalne ulepszenia i perspektywy dalszego rozwoju – ostatni rozdział merytoryczny, w którym wskazane zostały możliwe dalsze kierunki rozwoju modeli.
7. Bibliografia

2. Kontekst teoretyczny i technologiczny

Rynek nieruchomości zmienia się bardzo dynamicznie i staje się coraz trudniejszy do analizy. Żeby skutecznie przewidywać ceny mieszkań, potrzebne są odpowiednie narzędzia analityczne. Ma to szczególne znaczenie wtedy, gdy sytuacja ekonomiczna jest niepewna – na przykład zmieniają się stopy procentowe, inflacja albo popyt na mieszkania. W takich warunkach coraz częściej wykorzystuje się Automatyczne Modele Wyceny (AVM – Automated Valuation Models). Modele AVM opierają się przede wszystkim na metodach statystycznych, analizie dużych zbiorów danych oraz algorytmach uczenia maszynowego, dzięki czemu pozwalają szybko i skutecznie oszacować wartość nieruchomości [3].

Wśród tych metod największą popularnością cieszą się algorytmy uczenia maszynowego, np. drzewa decyzyjne, sieci neuronowe czy właśnie regresja [4;5]. Szczególnie ważną rolę odgrywa tutaj regresja wieloraka. Popularność tej metody bierze się przede wszystkim stąd, że łatwo jest zrozumieć jej wyniki – od razu widać, jak poszczególne cechy mieszkania (np. lokalizacja, powierzchnia, standard) wpływają na jego cenę [6]. Poza tym regresja ma jasno określone zasady działania oraz szeroko dostępne narzędzia diagnostyczne, takie jak analiza reszt, współczynnik determinacji R² czy testy na współliniowość, które pozwalają ocenić, czy model jest dokładny i wiarygodny [7].

Dodatkową zaletą regresji wielorakiej jest to, że jest stosunkowo łatwa do zaimplementowania i nie wymaga dużych mocy obliczeniowych, w przeciwieństwie do bardziej złożonych metod, np. głębokich sieci neuronowych. Dzięki temu można szybko przygotować skuteczny model i regularnie go aktualizować. Dodatkowo regresja stanowi naturalną podstawę podejścia hedonicznego, które traktuje cenę nieruchomości jako sumę udziałów jej poszczególnych atrybutów, co jest powszechnie stosowaną praktyką wyceny [8].

Regresja wieloraka jest skutecznym i szeroko wykorzystywanym narzędziem analitycznym na rynku nieruchomości. Niniejszej praca koncentruje się na wykorzystaniu właśnie tej metody, analizując jej przydatność i efektywność w prognozowaniu cen mieszkań.

3. Definicje i przykłady

W niniejszym rozdziale zostały przedstawione podstawowe terminy i definicje, które stanowią fundament dla dalszych części pracy. Zrozumienie tych pojęć jest kluczowe, aby właściwie interpretować metody, modele oraz uzyskane wyniki omówione w kolejnych rozdziałach. Opisane zostały zarówno zagadnienia teoretyczne, jak i przykłady praktyczne, które pomogą w lepszym zobrazowaniu omawianej problematyki. W ten sposób uzyskana została wspólna płaszczyzna pojęciowa, niezbędna do dalszej analizy i zrozumienia omawianych wniosków.

Wszystkie poniższe obliczenia zostały oparte o przykładowy zestaw danych dotyczących nieruchomości ukazany na rysunku 3.1. poniżej. W rzeczywistości tabela składa się ze 100 obserwacji. Każdą obserwację opisuje 5 cech:

1. Powierzchnia – powierzchnia wewnętrzna nieruchomości wyrażona w m2.
2. LiczbaPokoi – liczba wszystkich pomieszczeń w nieruchomości.
3. Lokalizacja – kategorialna zmienna informująca o lokalizacji położenia nieruchomości.
4. Wiek – wiek nieruchomości wyrażany w latach od momentu jej wybudowania.
5. OdlOdCentrum – zmienna liczbowa podająca dystans do centrum wyrażona w km.

Dodatkowo każdy rekord ma przypisaną cenę nieruchomości wyrażoną w PLN. Zmienną tą należy traktować jako zmienną celu, której zależność od zmiennych objaśniających zostanie przeanalizowana w następujących przykładach.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **Powierzchnia** | **LiczbaPokoi** | **Lokalizacja** | **Wiek** | **OdlOdCentrum** | **Cena** |
| 1 | 54 | 3 | Centrum | 17 | 0,91 | 571 200 |
| 2 | 85 | 4 | Peryferia | 63 | 8,36 | 402 500 |
| 3 | 74 | 4 | Peryferia | 53 | 14,11 | 330 900 |
| 4 | 47 | 3 | Centrum | 15 | 1,16 | 538 300 |
| 5 | 61 | 3 | Centrum | 20 | 0,66 | 736 800 |
| 6 | 95 | 5 | Peryferia | 55 | 9,66 | 468 300 |
| 7 | 34 | 2 | Śródmieście | 17 | 2,72 | 254 500 |
| 8 | 64 | 3 | Śródmieście | 22 | 2,46 | 516 100 |
| 9 | 89 | 5 | Centrum | 10 | 0,5 | 1 029 700 |
| 10 | 57 | 2 | Śródmieście | 31 | 2,96 | 508 800 |
| 11 | 60 | 3 | Peryferia | 73 | 13,22 | 217 700 |
| 12 | 69 | 4 | Centrum | 10 | 0,76 | 827 500 |
| … | … | … | … | … | … | … |

Wycinek zbioru danych używanych do objaśnienia definicji na praktycznych przykładach.

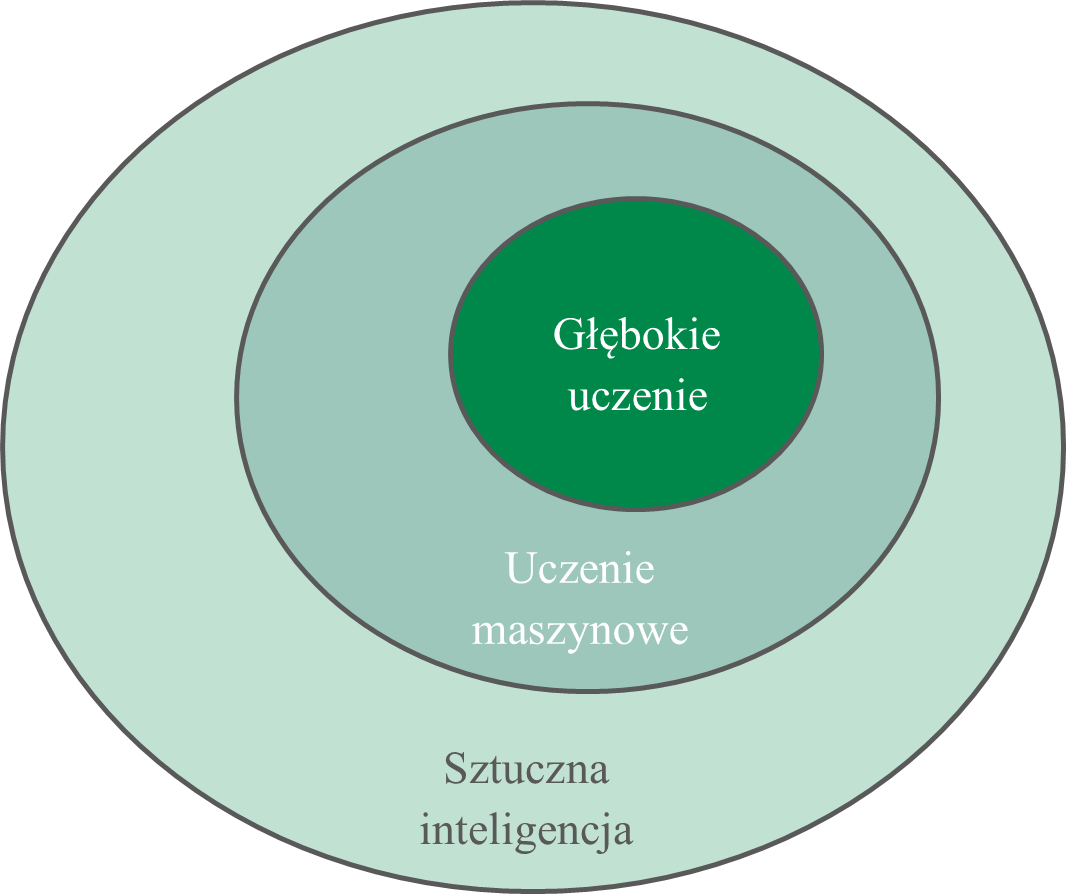
Rysunek 3.1.

3.1. Uczenie maszynowe

Uczenie maszynowe(ang. Machine Learning – ML) jest gałęzią sztucznej inteligencji, która skupia się na tworzeniu algorytmów umożliwiających komputerom odkrywanie nowej wiedzy poprzez dokonywanie generalizacji na bazie przykładów. Poglądowy schemat relacji między uczeniem maszynowym a sztuczną inteligencją (ang. AI – Artificial Intelligence) i uczeniem głębokim (ang. DL – Deep Learningiem) ukazuje ilustracja 3.1.

Wykorzystuje ono narzędzia statystyki, analizy danych oraz teorii prawdopodobieństwa, aby odnajdywać wzorce w zbiorach obserwacji. Metody uczenia maszynowego dzieli się najczęściej na uczenie nadzorowane, nienadzorowane oraz uczenie przez wzmacnianie, co odzwierciedla różne rodzaje dostępnych danych i odmienny charakter celów analitycznych.

„Uczenie maszynowe to nauka o tym, jak sprawić, by komputery uczyły się bez wyraźnego zaprogramowania” Arthur Samuel[9]



Schemat relacji między sztuczną inteligencją, uczeniem maszynowym i uczeniem głębokim.

Rysunek 3.2.

3.1. Przykład

Firma deweloperska planuje dużą inwestycję na przedmieściach dużego miasta. Przed podjęciem decyzji o zainwestowaniu pieniędzy i rozpoczęciem prac zostaje przeprowadzona analiza lokalnego rynku nieruchomości. Deweloperzy od lat specjalizują się w realizacji projektów jednorodzinnych domów wolnostojących z ogrodami. W wybranej lokalizacji tego typu nieruchomości jest już kilka. Z tego względu firma postanawia podzlecić wykonanie modelu predykcyjnego, który na bazie rzeczywistych dostępnych danych z regionu będzie przewidywał wysokość cen planowanych przez deweloperów nieruchomości. Zadaniem analityków jest także zidentyfikowanie kluczowych predyktorów mających wpływ na cenę. Tego typu analiza pozwoli oszacować potencjalne zyski, przy założeniu, że inwestycja dojdzie do skutku oraz zaistnieje zainteresowanie ze strony kupujących.

Obraz zawierający zrzut ekranu, design

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

W jaki sposób mogą zmieniać się ceny nieruchomości w zależności od ich parametrów?

Rysunek 3.3.

Tworzenie modeli uczenia maszynowego, w oparciu o które analitycy zamierzają dokonywać predykcji, wymaga w pierwszej kolejności jednoznacznego określenia charakteru dostępnych danych (w tym ich formatu, reprezentacji oraz jakości), a także zdefiniowania docelowego sposobu wykorzystania modelu – zarówno w kontekście wyboru metody uczenia (nadzorowanego lub nienadzorowanego), jak i zasięgu zadania predykcyjnego. Odpowiedni dobór technik i procedur treningu ma kluczowe znaczenie dla precyzji i wiarygodności prognoz, co z kolei przekłada się na efektywność całego projektu. Kluczowy jest zatem wybór stosowanych metod:

1. **Uczenie nienadzorowane**
2. **Uczenie nadzorowane**

3.2. Uczenie nienadzorowane

Uczenie nienadzorowane (ang. unsupervised learning) to metoda uczenia maszynowego, w której model analizuje dane bez uprzednio zdefiniowanych etykiet czy wyników, odkrywając ukryte wzorce i struktury. Dzięki temu techniki te umożliwiają segmentację danych, wykrywanie klastrów oraz identyfikację anomalii, co jest szczególnie przydatne przy analizie złożonych zbiorów danych. W kontekście wyceny nieruchomości, uczenie nienadzorowane może pomóc w identyfikacji grup nieruchomości o podobnych cechach oraz w wykrywaniu niestandardowych zachowań cenowych. Metody takie jak klasteryzacja czy analiza głównych składowych (ang. PCA – Principal Component Analysis) pozwalają na redukcję wymiarowości danych, co ułatwia ich interpretację. Zastosowanie uczenia nienadzorowanego może stanowić wartościowe uzupełnienie tradycyjnych modeli predykcyjnych.

3.2. Przykład

Firma deweloperska planuje rozszerzyć swoją działalność na nowe segmenty rynku nieruchomości. Przed podjęciem decyzji o kierunku rozwoju, zespół analityków przeprowadza szczegółową analizę lokalnego rynku, aby zidentyfikować naturalne grupy obiektów o podobnych cechach, niezależnie od wcześniej ustalonych kategorii. W związku z tym firma decyduje się na wykorzystanie metod uczenia nienadzorowanego, takich jak klasteryzacja k-średnich, aby automatycznie wyodrębnić segmenty nieruchomości na podstawie ich cech, takich jak metraż, lokalizacja, standard wykończenia czy cena.

Obraz zawierający linia, diagram, tekst, Wykres

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Rysunek 3.4.

Wykres punktowy przedstawiający klasteryzacją danych na 3 klastry wraz z wykresem słupkowym pokazującym liczbę nieruchomości w każdym z nich.

Zastosowano algorytm k-średnich (ang. k-means) w celu sprawdzenia, czy da się wyodrębnić grupy nieruchomości na podstawie ich ceny oraz powierzchni mieszkalnej. Powyższe wykresy  
(Rysunek 3.4.) pokazują podział nieruchomości na klastry (n = 3) oraz liczbę nieruchomości w każdym z nich. Uzyskane klastry pozwolą określić, które segmenty charakteryzują się największym potencjałem wzrostu oraz umożliwią lepsze dopasowanie oferty do potrzeb różnych grup klientów. Taka analiza pozwoli deweloperom na efektywniejsze planowanie inwestycji, optymalizację strategii marketingowej oraz precyzyjne targetowanie oferty. Model uczenia nienadzorowanego wykorzystując klasteryzację (np. algorytm K-średnich, ang. K-means) mógłby pomóc w grupowaniu podobnych mieszkań w klastry. Przykładowo:

1. **Klaster 0**: Mieszkania średniej wielkości o umiarkowanej cenie – obie cechy przyjmują wartości umiarkowane w stosunku do pozostałych klastrów.
2. **Klaster 1**: Najdroższe nieruchomości, które cechują się jednocześnie większą powierzchnią mieszkaniową. Dane z wyrazistym trendem liniowym z kilkoma wartościami odstającymi.
3. **Klaster 2**: Nieruchomości o najniższej cenie, jednocześnie o mniejszych powierzchniach, chociaż z wyraźną dużą wariancją i odchyleniem standardowym (ok. 17 m2 w porównaniu z 14 m2 w klastrze 0 oraz 12,6 m2 w klastrze 1.

Powyższe grupowanie może stanowić dobre podłoże pod późniejszą predykcję cen nieruchomości. Dzięki takim segmentom można również rozważyć trenowanie modelu predykcyjnego dla każdego klastra oddzielnie. Może to zapewnić lepszą precyzję prognozy, ponieważ model jest lepiej dopasowany do specyfiki mieszkań w danej grupie.

3.3. Uczenie nadzorowane

Uczeniem nadzorowanym (ang. supervised learning) nazywane są metody uczenia maszynowego, w których model jest trenowany na danych wejściowych (cechach) oraz odpowiadających im etykietach (wartościach docelowych). Celem jest jak najlepsze nauczenie modelu przewidywania wyniku na podstawie dostarczonych danych, tak by mógł on generalizować rezultat dla nowych nieznanych przykładów.Podejście uczenia nadzorowanego zakłada naukę na bazie parametrów wejściowych (ang. input parameters) aby móc określić wartość zmiennej wyjściowej (ang. output variable).

Dzieli się ono na dwa główne typy problemów:

1. **Klasyfikacja** – której celem jest przypisanie danych wejściowych do jednej z predefiniowanych (dyskretnych) klas.
2. **Regresja** – której celem jest przewidywanie ciągłej wartości na podstawie danych wejściowych. Niniejsza praca kładzie nacisk właśnie na ten typ problemu.

Po wytrenowaniu modelu na danych treningowych, można stosować go przy dokonywaniu prognoz dla rzeczywistych danych produkcyjnych.

3.3. Przykład

Przykładowe zbiór danych z rysunku 3.1. stanowi wartość zarówno dla osób kupujących, jak i handlujących nieruchomościami. Odpowiednio duża liczba informacji może pozwolić na znalezienie odpowiedzi na pytanie: „Ile może / powinna kosztować nieruchomość o danych parametrach?”.

Zbiór opisuje podstawowe parametry nieruchomości, takie jak powierzchnia (m2), liczba pokoi, lokalizacja , wiek, czy odległość od centrum.

Opisane parametry zostały zaznaczone w tabeli na rysunku 3.5. Każda z opisywanych właściwości nazywana jest cechą (ang. feature).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **Powierzchnia** | **LiczbaPokoi** | **Lokalizacja** | **Wiek** | **OdlOdCentrum** | **Cena** |
| 1 | 54 | 3 | Centrum | 17 | 0,91 | 571 200 |
| 2 | 85 | 4 | Peryferia | 63 | 8,36 | 402 500 |
| 3 | 74 | 4 | Peryferia | 53 | 14,11 | 330 900 |
| 4 | 47 | 3 | Centrum | 15 | 1,16 | 538 300 |
| 5 | 61 | 3 | Centrum | 20 | 0,66 | 736 800 |
| 6 | 95 | 5 | Peryferia | 55 | 9,66 | 468 300 |
| 7 | 34 | 2 | Śródmieście | 17 | 2,72 | 254 500 |
| 8 | 64 | 3 | Śródmieście | 22 | 2,46 | 516 100 |
| 9 | 89 | 5 | Centrum | 10 | 0,5 | 1 029 700 |
| 10 | 57 | 2 | Śródmieście | 31 | 2,96 | 508 800 |
| 11 | 60 | 3 | Peryferia | 73 | 13,22 | 217 700 |
| 12 | 69 | 4 | Centrum | 10 | 0,76 | 827 500 |
| … | … | … | … | … | … | … |

Przykładowy zbiór właściwości (cech) wykorzystywanych do predykcji ceny nieruchomości.

Rysunek 3.5.

Cechy służą modelowi do predykcji etykiet (ang. labels). W modelach opartych na technikach regresji częściej usłyszeć można określenie wartość docelowa lub zmienna celu. Zmienną celu w przykładzie jest Cena zaznaczona w tabeli na rysunku 3.6.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **Powierzchnia** | **LiczbaPokoi** | **Lokalizacja** | **Wiek** | **OdlOdCentrum** | **Cena** |
| 1 | 54 | 3 | Centrum | 17 | 0,91 | 571 200 |
| 2 | 85 | 4 | Peryferia | 63 | 8,36 | 402 500 |
| 3 | 74 | 4 | Peryferia | 53 | 14,11 | 330 900 |
| 4 | 47 | 3 | Centrum | 15 | 1,16 | 538 300 |
| 5 | 61 | 3 | Centrum | 20 | 0,66 | 736 800 |
| 6 | 95 | 5 | Peryferia | 55 | 9,66 | 468 300 |
| 7 | 34 | 2 | Śródmieście | 17 | 2,72 | 254 500 |
| 8 | 64 | 3 | Śródmieście | 22 | 2,46 | 516 100 |
| 9 | 89 | 5 | Centrum | 10 | 0,5 | 1 029 700 |
| 10 | 57 | 2 | Śródmieście | 31 | 2,96 | 508 800 |
| 11 | 60 | 3 | Peryferia | 73 | 13,22 | 217 700 |
| 12 | 69 | 4 | Centrum | 10 | 0,76 | 827 500 |
| … | … | … | … | … | … | … |

Etykieta (zmienna celu), której prawidłowa estymacja jest celem tworzenia modelu.

Rysunek 3.6.

W przypadku rzeczywistych analiz należy zadbać o to, aby wolumen danych był znacząco większy. Zbyt mała lub niereprezentatywna próbka danych, na których model jest uczony może doprowadzić do niepożądanego efektu przeuczenia (ang. overfittingu) w skutek czego model może nie działać prawidłowo na danych rzeczywistych. W oparciu o tak przygotowane dane można stworzyć model regresji, którego zadaniem będzie wyjaśnienie zmiennej celu na postawie posiadanych wartości predyktorów.

3.4. Regresja

Analiza regresji jest jednym z najczęściej wykorzystywanych narzędzi w analizie danych i statystyce. W swojej podstawowej formie, analiza regresji pozwala na wykrywanie relacji między pojedynczą zmienną niezależną a pojedynczą zmienną zależną. Zmienna zależna jest najczęściej wynikiem, na którym zależy analitykowi, natomiast zmienna niezależna pełni funkcję pozwalającą określić wartość zmiennej zależnej. Kluczowe korzyści wynikające z zastosowania analizy regresji polegają na tym, że pozwala ona:

1. Określić, czy zmienne niezależne mają istotny związek ze zmienną zależną.
2. Ocenić względną siłę wpływu poszczególnych zmiennych niezależnych na zmienną zależną.
3. Dokonywać prognoz i predykcji.

Wyróżnia się wiele modeli regresji. Wybór odpowiedniego jej rodzaju zależy od typu problemu, który chce się opisać przy jej pomocy oraz od struktury danych, które są przedmiotem analizy. Należy przy tym jednocześnie pamiętać, że nie każda struktura danych pozwala na sensowne wykorzystanie metod regresji. Za podstawowy podział uznaje się:

1. **Regresja liniowa –** modeluje zależność między zmienną zależną a jedną lub wieloma zmiennymi niezależnymi , przy założeniu, że relacja jest liniowa. Dokładniejsze omówienie założeń modeli regresji liniowej znajduje się w kolejnym podrozdziale.
2. **Regresja nieliniowa –** modeluje sytuacje, w których zależność między zmiennymi nie może być wyrażona funkcją liniową względem parametrów. Model może przyjmować postać potęgową, wykładniczą, logarytmiczną lub inną funkcję nieliniową, np.:

Estymacja parametrów wymaga numerycznych metod optymalizacji (np. algorytmu Gaussa-Newtona), ponieważ nie ma analitycznego rozwiązania jak w regresji liniowej. Modele nieliniowe są bardziej elastyczne, ale wymagają większej ostrożności przy dopasowywaniu i interpretacji. Regresja nieliniowa ma zastosowanie w zjawiskach, gdzie występuje np. zmienność tempa wzrostu. Wyróżnia się m.in.:

* Regresję wykładniczą
* Regresję potęgową
* Regresję sigmoidalną

Na rysunkach 3.7. oraz 3.8. pokazane zostały cztery różne przypadki danych ocenianych pod kątem przydatności do budowy modelu regresyjnego (liniowego oraz wielomianowego).

1. Pierwszy wykres ilustruje przypadek, w którym dane tworzą kilka oddzielnych skupisk, nie wykazując spójnego trendu – skutkuje to bardzo niskim współczynnikiem korelacji r (Rysunek 3.7.) oraz w efekcie również niskim R2, co wskazuje, że taka struktura danych nie nadaje się do modelowania regresyjnego.
2. Drugi wykres pokazuje słabą zależność z dużym szumem, gdzie mimo istnienia pewnego trendu, wysoki poziom losowych fluktuacji utrudnia precyzyjne dopasowanie modelu.
3. Trzeci wykres prezentuje umiarkowaną zależność z zakłóceniami, co prowadzi do wyraźniejszego trendu i umiarkowanej wartości współczynnika korelacji Pearsona (dla modeli prostej regresji) oraz R2.
4. Czwarty wykres obrazuje idealny przypadek – silną, liniową / wielomianową zależność z minimalnym szumem, co skutkuje wysokim współczynnikiem determinacji i doskonałą przydatnością tych danych do budowy modelu regresyjnych.

Znaczenie współczynnika korelacji Pearsona oraz współczynnika determinacji R2 zostało dokładniej omówione w dalszej części rozdziału.

Obraz zawierający linia, Wykres, diagram, tekst

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Rysunek 3.7.

Wpływ struktury danych na ogólny poziom przydatności analizy regresji - prosta regresja liniowa.

Obraz zawierający linia, Wykres, diagram, zrzut ekranu

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Rysunek 3.8.

Wpływ struktury danych na ogólny poziom przydatności analizy regresji – regresja wielomianowa.

Rysunek 3.9. pokazuje przykłady zależności między zmiennymi dla aproksymacji innych funkcji elementarnych (kolejno wykładniczej, logarytmicznej, sinusoidalnej, potęgowej).

Obraz zawierający diagram, linia, Wykres, tekst

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Rysunek 3.9.

Przykłady mniej standardowych zależności między predyktorami a zmienną celu – funkcje elementarne.

3.5. Model regresji liniowej

Regresja liniowa modeluje zależność między zmienną zależną a jedną lub wieloma zmiennymi niezależnymi , przy założeniu, że relacja jest liniowa. Model opisuje tę zależność równaniem:

gdzie:

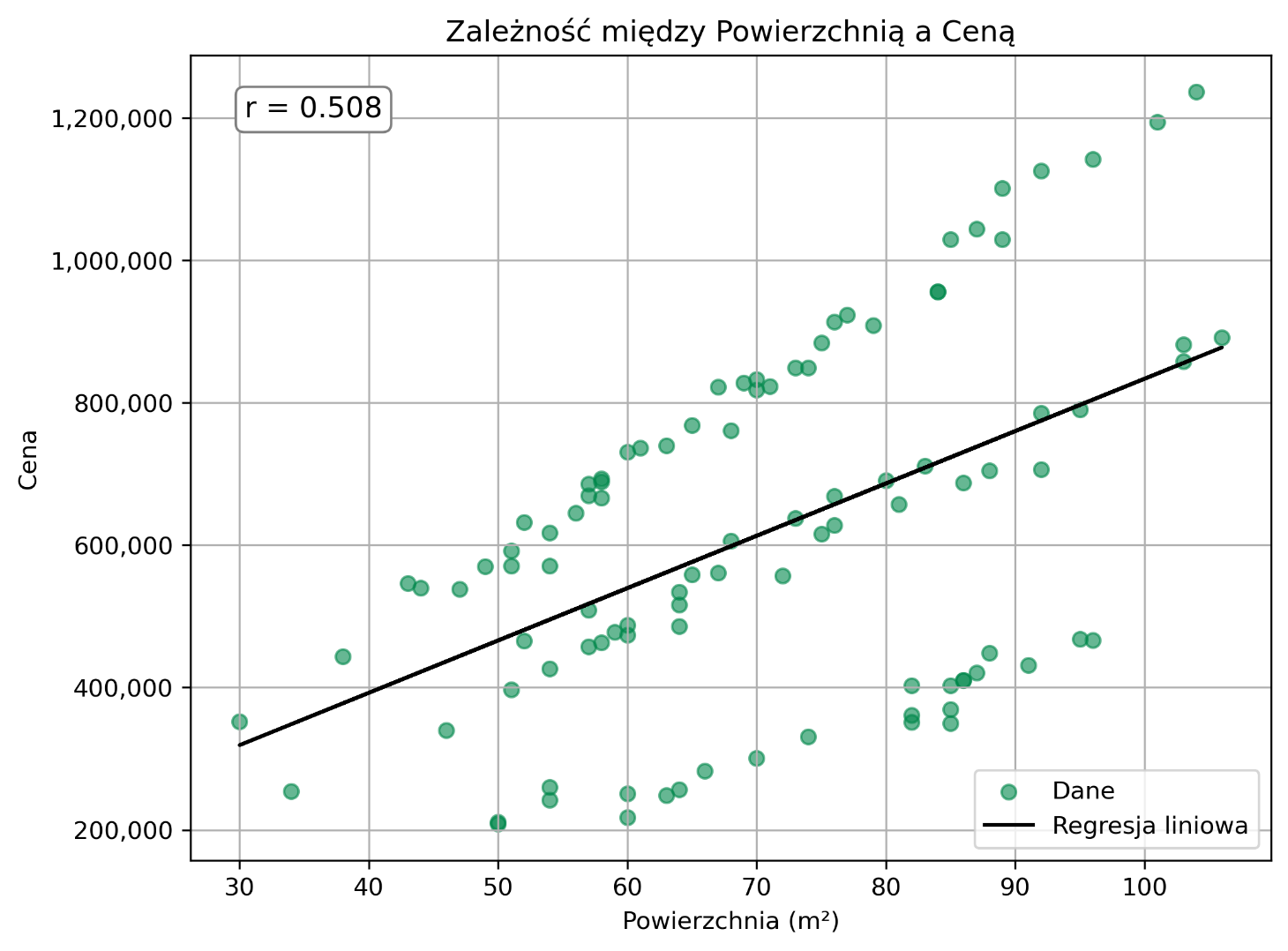
* - są współczynnikami kolejnych zmiennych niezależnych
* - jest składnikiem losowym odpowiedzialnym za błąd oszacowania w modelu.

Najważniejszymi założeniami w modelu regresji liniowej są:

1. Związek między a jest liniowy.
2. Wartości zmiennych objaśniających są ustalone (nie są losowe). Losowość wartości pochodzi wyłącznie z błędu losowego.
3. Błędy losowe mają rozkład normalny o średniej 0 i stałej wariancji . Stała wariancja błędów nosi nazwę homoskedastyczność (ang. homoscedascity).
4. Błędy losowe są od siebie niezależne. Nie występuje autokorelacja (ang. residual autocorrelation).

1. W przypadku regresji wielorakiej należy zadbać o brak multikolinearności (ang. multicollinearity).

W podstawowych modelach prostej regresji liniowej (z jedną zmienną) wykorzystuje się współczynnik korelacji Pearsona do sprawdzenia kierunku oraz siły liniowej zależności. W modelu regresji liniowej pomaga ocenić, czy warto budować model predykcyjny.



Badanie zależności liniowej dla zmiennej objaśniającej Powierzchnia i zmiennej celu Cena.

Rysunek 3.10.

Obliczony współczynnik korelacji Pearsona między Powierzchnią a Ceną wynosi r = 0,508. Oznacza to umiarkowaną dodatnią liniową zależność między powierzchnią a ceną nieruchomości. W praktyce oznacza to, że wraz ze wzrostem powierzchni rośnie cena, ale zależność nie jest idealnie liniowa. Dla p-value < 0,0001 korelacja jest statystycznie istotna, czyli nie jest wynikiem przypadku. Im większa wartość , tym lepsze dopasowanie modelu i potencjalnie wyższy współczynnik determinacji R2 (dla jednej zmiennej r2 = R2). Współczynnik korelacji Pearsona dla populacji wyraża się wzorem:

Gdy:

* r > 0 – mówimy o zależności dodatniej
* r < 0 – mówimy o zależności ujemnej
* r = 0 – mówimy o braku zależności liniowej

Dla modelu o większej liczbie zmiennych objaśniających w celu określenia zasadności tworzenia modelu regresji wykorzystuje się głównie właśnie współczynnik determinacji R2. Współczynnik determinacji został dokładniej omówiony w dalszej części rozdziału. W tym przypadku widać, że na wykresie formują się tak na prawdę 3 potencjalne skupiska danych, dla których można zbudować modele regresji. Potencjalnie naruszona zostaje zatem zasada liniowej zależności, homogeniczność modelu.

3.6. Prosta regresja liniowa

Modele regresji liniowej, nazywane również modelami zależności liniowej, wykorzystują zależności liniowe pomiędzy danymi w zbiorze. Rysunek 3.11. ukazuje zależnosci między ceną a powierzchnią nieruchomości oraz ceną a odległością od centrum.

Obraz zawierający tekst, linia, zrzut ekranu, diagram

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Przykład rosnącej (po lewo) i malejącej (po prawo) zależności pomiędzy ceną a innymi zmiennymi. Wykresy uzyskano na bazie wycinku danych ze zbioru przykładowego.

Rysunek 3.11.

W obu przypadkach zbadano w jaki sposób na cenę wpływają zmienne opisujące. Widocznym jest, że punkty skupiają się wzdłuż pewnej prostej – na lewym wykresie prosta ta jest rosnąca, a na prawym malejąca. Bez względu na jej kierunek nasuwa się pytanie, w jaki sposób dopasować taką prostą, aby optymalnie opisywała zależność w danych?

Niech szukana prosta będzie opisana wzorem:

gdzie:

* - wyraz wolny
* - współczynnik kierunkowy prostej
* - sumaryczny błąd losowy modelu
* - zmienna niezależna
* - przewidywana zmienna zależna

Powyższą prostą można wykreślić dobierając nieznane parametry i .

Prawidłowy dobór tych parametrów jest niezbędny w celu jak najlepszego przybliżenia relacji pomiędzy zmiennymi dla ogółu populacji na podstawie dostępnej próby.

Celem nadrzędnym jest w takim razie ustalenie wartości nieznanych parametrów i tak, by prosta jak najlepiej opisywała dane. Istnieje kilka metod, które na to pozwalają. Podstawową z nich jest metoda najmniejszych kwadratów.

Metoda najmniejszych kwadratów (ang. OLS – Ordinary Least Squares) polega na znalezieniu takich wartości i ​, które minimalizują sumę kwadratów błędów. Niech równanie:

opisuje zależność między jedną zmienną zależną a jedną zmienną niezależną w modelu prostej regresji liniowej. Aby znaleźć optymalne i należy zminimalizować funkcję straty daną wzorem:

Minimalizacja przebiega poprzez obliczenie pochodnych funkcji S względem oraz i przyrównaniu ich do zera:

Rozwiązując ten układ równań (tak zwane równania normalne) wyliczane są wartości liniowych estymatorów i :

Otrzymane liniowe estymatory minimalizują sumę kwadratów błędów i mają własności nieobciążoności oraz minimalnej wariancji wśród wszystkich liniowych estymatorów. [10]

3.6. Przykład

Analityk założył prostą zależność liniową między powierzchnią nieruchomości a jego ceną. Na tej podstawie zbudowano model prostej regresji liniowej. Znalezienie najlepiej dopasowanej do danych prostej regresji dokonano przy pomocy metody OLS.

Dostępne są poniższe dane z rynku nieruchomości:

(77; 923500), (68; 760900), (51; 591900), (57; 685600), (76; 913700), (57; 670000), (79; 908800), (96; 1142400), (69; 827500), (89; 1029700), (60; 730700), (52; 466000)

W tym miejscy należy założyć liniową zależność zmiennej zależnej od niezależnej.

Wykorzystując wzory na współczynniki i wyliczono wzór prostej regresji:

Na rysunku 3.12. pokazano zamodelowany scenariusz, a w tabeli na rysunku 3.13. minimalizowane kwadraty różnic.

Obraz zawierający tekst, zrzut ekranu, linia, diagram

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Zależność ceny nieruchomości od powierzchni w modelu regresji liniowej.

Rysunek 3.12.

Po wielokrotnych iteracjach algorytm OLS pozwolił na znalezienie optymalnej (minimalnej) sumy odchyleń wartości przewidywanych (ang. SSE – Error Sum of Squares) przez model od rzeczywistych punktów danych:

Wartość SSE, choć dla wybranych parametrów minimalna, i tak jest wartością bardzo dużą. W kontekście cen nieruchomości, wyliczana zmienna celu faktycznie przyjmuje duże wartości, dlatego suma kwadratów odchyleń również będzie dużą wartością. SSE samo w sobie nie jest jednak wystarczające do oceny jakości modelu. W dalszej części rozdziału omówione zostały pozostałe miary jakości dopasowania modelu. Model stworzony w przykładzie w rzeczywistości przyjąłby postać:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **xi (Powierzchnia)** | **yi (Cena)** | **ŷi (Cena przewidywana)** | **ei² (Błąd kwadratowy)** |
| 1 | 77 | 923 500 | 903 315 | 407 432 293 |
| 2 | 68 | 760 900 | 788 243 | 747 625 114 |
| 3 | 51 | 591 900 | 570 884 | 441 675 638 |
| 4 | 57 | 685 600 | 647 599 | 1 444 091 558 |
| 5 | 76 | 913 700 | 890 529 | 536 884 340 |
| 6 | 57 | 670 000 | 647 599 | 501 813 972 |
| 7 | 79 | 908 800 | 928 887 | 403 474 437 |
| 8 | 96 | 1 142 400 | 1 146 245 | 14 787 776 |
| 9 | 69 | 827 500 | 801 029 | 700 737 832 |
| 10 | 89 | 1 029 700 | 1 056 745 | 731 421 174 |
| 11 | 60 | 730 700 | 685 956 | 2 002 004 668 |
| 12 | 52 | 466 000 | 583 670 | 13 846 165 869 |

Proces wyliczania średniego błędu kwadratowego (SSE) w modelu prostej regresji liniowej.

Rysunek 3.13.

3.7. Normalność błędów losowych

W klasycznym modelu regresji liniowej (prostej i wielorakiej) zakłada się, że błędy losowe mają rozkład normalny ze średnią 0 oraz stałą wariancją - .

Normalność błędów nie wpływa na estymację współczynników, ale jest niezbędna do prawidłowego przeprowadzania testów istotności i tworzenia przedziałów ufności.

Aby sprawdzić spełnienie tego warunku stosuje się testy statystyczne:

* Shapiro-Wilk
* Kolmogorov-Smirnov
* Anderson-Darling

3.7. Przykład

Analizę rozkładu reszt można przeprowadzić przy pomocy wykresu Q-Q (kwantyl – kwantyl). Wykres pozwala zobaczyć jak bardzo różnią się rzeczywiste obserwowane reszty od teoretycznych wartości przewidywanych przez rozkład normalny. Na rysunku 3.12. pokazano wykres Q-Q dla modelu:

Obraz zawierający linia, Wykres, diagram, tekst

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Rysunek 3.14.

Wykres typu kwantyl – kwantyl błędów modelu prostej regresji liniowej zależności ceny od powierzchni.

Widać, że punkty układają się wzdłuż prostej rozkładu normalnego. Zaobserwowano jeden punkt odstający, gdzie cena rzeczywista jest zdecydowanie niższa niż model przewiduje.Nie stwierdzono istotnego naruszenia założenia normalności rozkładu reszt. Formalne potwierdzenie zapewnia test Shapiro-Wilka, który dla p-value na poziomie 0,6081 zwraca statystykę równą 0,9480. Nie ma zatem podstaw to odrzucenia hipotezy o normalności reszt.

3.8. Homoskedastyczność

W modelu regresji liniowej zakłada się, że wariancja błędów losowych (reszt) jest stała dla wszystkich wartości zmiennych niezależnych. Oznacza to, że:

Jeśli wariancja błędów zmienia się w zależności od , dochodzi do naruszenia tego założenia, tj. występuje heteroskedastyczność i w związku z tym istnieje niebezpieczeństwo niepoprawnego oszacowania współczynników modelu.

3.8. Przykład

Aby sprawdzić występowanie homoskedastyczności przeprowadzony został test Breuscha-Pagana:

* H0​: homoskedastyczność (wariancja stała)
* H1​: heteroskedastyczność (wariancja zmienna)

Dla poziomu istotności 0,05 nie stwierdzono podstaw do odrzucenia hipotezy H0. P-value = 0,9941, więc zachowana jest stabilność błędów estymacji.

Obraz zawierający tekst, zrzut ekranu, linia, numer

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Wykres reszt względem wartości przewidywanych. Zapewnia możliwość graficznej oceny zjawiska homoskedastyczności.

Rysunek 3.15.

3.9. Autokorelacja

Autokorelacja to zjawisko polegające na występowaniu zależności między błędami losowymi (resztami) w modelu regresji liniowej. Formalnie oznacza to, że kowariancja między resztami dla różnych obserwacji nie jest równa zeru:

Autokorelacja jest szczególnie istotna w modelach opartych na szeregach czasowych lub danych przestrzennych, gdzie kolejność obserwacji ma znaczenie. Jej obecność narusza założenie niezależności błędów, co prowadzi do zaniżenia błędów standardowych współczynników regresji i może skutkować fałszywie pozytywnymi wynikami testów istotności. Autokorelację bada się przy pomocy testów statystycznych:

* Durbin-Watson
* Breusch-Godfrey

3.9. Przykład

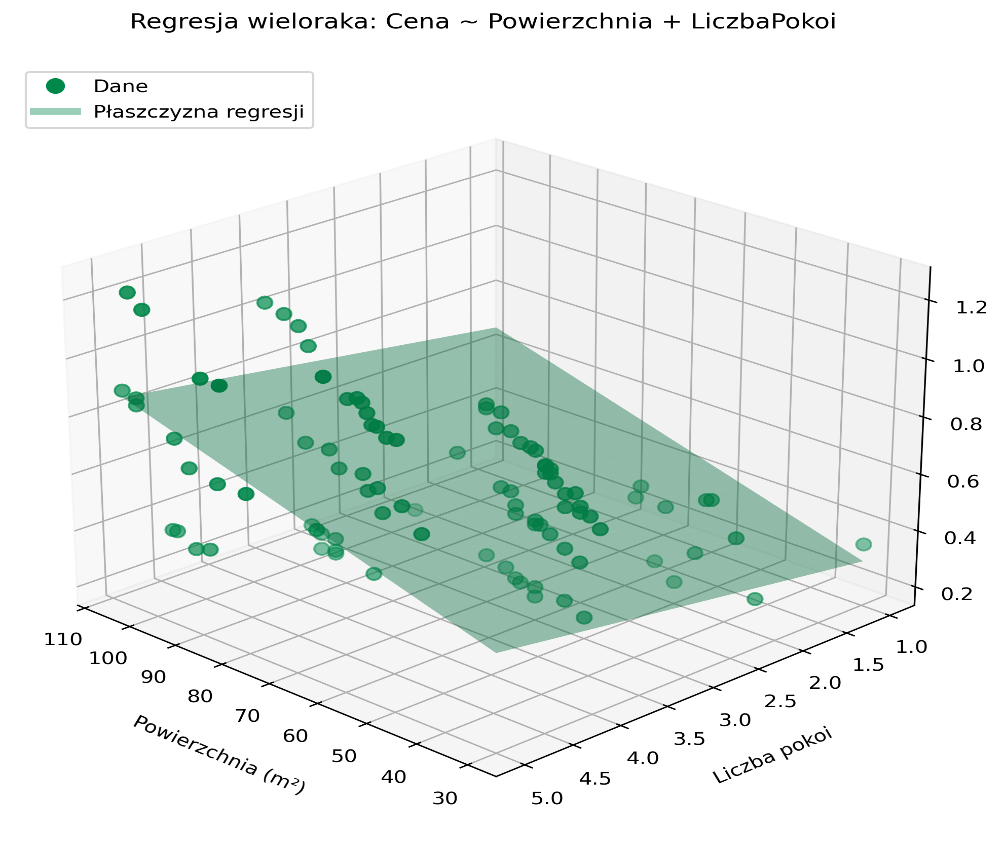
W przypadku przykładowych danych nie zachodzi uzasadniona potrzeba przeprowadzania analizy autokorelacji, gdyż dane nie są w żaden logiczny sposób uporządkowane chronologicznie.

3.10. Regresja wieloraka

Modele regresji oparte o jeden predyktor są najprostszą formą modeli regresji liniowej. W praktyce znacznie częściej zmienność zmiennej wynikowej jest uzależniona od wielu aspektów. W analizowanym wcześniej przykładzie przyjęto założenie, że cena nieruchomości będzie kalkulowana na podstawie jedynie informacji o powierzchni nieruchomości. Należy jednak sprawdzić, czy oraz w jakim stopniu inne zmienne ze zbioru danych mogą wpłynąć na cenę nieruchomości.

Tego typu modele oparte o wiele zmiennych niezależnych są nazywane modelami regresji wielorakiej (ang. multiple regression). Wzór opisujący matematycznie prostą regresji zaprezentowany w podrozdziale 3.5. można uogólnić dla wariantu z n predyktorami i sprowadzić równanie do postaci:

W powyższym wzorze oznacza pierwszy predyktor, natomiast oznacza n-ty predyktor. Analogicznie oznacza współczynnik pierwszego predyktora, jest współczynnikiem n-tego predyktora, a niezmiennie pozostaje wyrazem wolnym. Podobnie jak w przypadku prostej regresji liniowej, parametry , , …, wyliczane są przy pomocy np. metody OLS w wariancie macierzowym dostosowanym dla wielu zmiennych.



Trójwymiarowa przestrzeń zależności między powierzchnią (m²), liczbą pokoi a ceną nieruchomości (w mln).

Rysunek 3.16.

Podobnie jak w przypadku zaprezentowanej wcześniej zwykłej metody najmniejszych kwadratów, wielokrotna regresja liniowa również dąży do oszacowania wartości współczynników , , …, , które minimalizują resztową sumę kwadratów (SSE).

Aby pokazać, w jaki sposób stosuje się metodę OLS dla modelu regresji wielorakiej powyższy wzór przekształcony został do notacji macierzowej.

Niech:

* oznacza macierz projektową (ang. design matrix) o wymiarach

, przedstawiająca niezależnych zmiennych dla obserwacji. Założono, że model zawiera pewną stałą (wyraz wolny). Z tego względu jedna kolumna macierzy składa się z samych 1. Kolumna ta jest traktowana, jak każda inna kolumna macierzy .

* oznacza wektor o wymiarach opisujący zmienną zależną modelu.
* oznacza wektor o wymiarach współczynników modelu.
* oznacza wektor o wymiarach reprezentujący błędy modelu.

Powyższe założenia pozwalają na określenie macierzowego wzoru dla modelu regresji wielorakiej:

lub prościej:

Powyższy wzór pozwala na dokładne opisanie rzeczywistości. Model uwzględnia składnik systematyczny (ang. systematic component) oraz składnik stochastyczny (ang. stochastic component). Składnik systematyczny odpowiada za przewidywalną część modelu, natomiast składnik stochastyczny obejmuje losowe, nieprzewidywalne czynniki oraz błędy.

Oba te czynniki tworzą pełny obraz badanego zjawiska. Celem – tak samo jak w metodzie OLS dla pojedynczego predyktora – jest uzyskanie oszacowań wartości dla wektora , gdzie SSE osiąga minimalną wartość.

Niech funkcja straty będzie opisana jako:

gdzie:

* oznacza wartość funkcji straty dla wektora .
* oznacza wektor reszt modelu

Znalezienie minimum funkcji straty wymaga zastosowania metody wykorzystującej gradient funkcji.

Gradient to pojęcie wywodzące się z analizy matematycznej, które opisuje kierunek i tempo wzrostu funkcji wielu zmiennych. Można go przedstawić jako wektor pochodnych cząstkowych względem każdej zmiennej.

Dla funkcji gradient jest wektorem postaci:

Gradient wskazuje kierunek, w którym funkcja rośnie najszybciej. Norma gradientu (długość) wskazuje na szybkość wzrostu.

W metodach optymalizacji, takich jak metoda gradientu prostego, gradient funkcji celu jest używany do określania kierunku, w którym należy zmieniać zmienne, aby zminimalizować lub zmaksymalizować wartości funkcji.

Punkty, w których gradient wynosi 0, czyli punkty gdzie każda składowa wektora gradientu wynosi 0 – wektor zerowy – sygnalizują potencjalne ekstremum lokalne funkcji : minimum, maksimum albo punkt siodłowy.

Zatem zapis:

Oznacza, że:

Zdefiniowanie pojęcia gradientu jest niezbędne w celu wyznaczenia minimum funkcji straty . Aby je znaleźć obliczono gradient funkcji względem oraz przyrównano wynik do 0:

gdzie:

* oznacza gradient funkcji względem wektora
* oznacza transpozycję
* Wynikowy gradient wskazuje kierunek wzrostu funkcji

Przyrównując otrzymany gradient do 0:

otrzymano równanie postaci:

nazywane równaniem normalnym.

Równania normalne są zatem wynikiem ustawienia gradientu funkcji celu równego zeru, co daje warunek stacjonarności - punkt, w którym funkcja osiąga minimum (przy ważnym założeniu, że funkcja ta jest funkcją wypukłą).

Jeśli macierz jest macierzą odwracalną - co zachodzi zawsze kiedy macierz ma pełną kolumnową rangę, a zatem kiedy wszystkie kolumny są od siebie liniowo niezależne – równania normalne można rozwiązać poprzez wymnożenie obu stron przez macierz odwrotną :

Otrzymane równanie jest ostatecznym wzorem na estymator parametrów regresji metodą najmniejszych kwadratów (OLS) w regresji wielorakiej. należy zatem rozumieć, jako wektor współczynników (wraz z wyrazem wolnym) kolejnych predyktorów modelu.

3.10. Przykład

Stworzony wcześniej prosty model regresji liniowej okazał się niewystarczający. Na cenę nieruchomości ma wpływ znacznie więcej parametrów niż tylko powierzchnia. Postanowiono stworzyć model w oparciu o 4 zmienne objaśniające: powierzchnię, liczbę pokoi, wiek oraz odległość od centrum. Zależność tę można przedstawić w ten sposób:

Korzystając z metody obliczania najmniejszej resztowej sumy kwadratów dla modelu 4 predyktorów i wyrazu wolnego, po podstawieniu danych z wykresu na rysunku 3.19. otrzymujemy następujące rezultaty dla współczynników zmiennych:

|  |  |
| --- | --- |
|  | **Wartość współczynnika** |
|  | 191 445 |
|  | 10 762 |
|  | -21 131 |
|  | -4 392 |
|  | -32 332 |

Rezultaty działania algorytmu gradientowego do wyliczenia wartości współczynników.

Rysunek 3.17.

W rzeczywistości model przyjmuje postać:

Interpretacja powyższych estymacji jest następująca:

1. **Powierzchnia**: każdemu dodatkowego m² odpowiada wzrost ceny o 10 762, przy założeniu stałych pozostałych zmiennych.
2. **Liczba pokoi**: każdy dodatkowy pokój obniża cenę o 21 131 – jest to bez wątpienia anomalia, która może wynika z kolinearności.
3. **Wiek**: każdy dodatkowy rok wieku budynku obniża cenę o 4 392.
4. **OdlOdCentrum**: każdy dodatkowy km od centrum obniża cenę o 32 332.

3.11. Multikolinearność

Multikolinearność to termin opisujący sytuację, gdy co najmniej jedna zmienna niezależna w modelu jest liniową kombinacją pozostałych, np.

Zjawisko to zaburza estymację współczynników. Sprawia, że są niestabilne i nieintuicyjne. Powoduje również wzrost wartości błędów standardowych. Model może przez to, pomimo wysokiej wartości R2, być trudno interpretowalny. Wykrywanie multikolinearności przebiega przy pomocy VIF – Variance Inflation Factor.

3.11. Przykład

Korzystając z modelu opracowanego w poprzednim podrozdziale obliczony został wskaźnik VIF dla każdej zmiennej objaśniającej .

|  |  |  |
| --- | --- | --- |
| **X** | **VIF** | **Interpretacja** |
| Powierzchnia | 83,33 | Bardzo silna multikolinearność |
| LiczbaPokoi | 77,51 | Bardzo silna multikolinearność |
| Wiek | 6,98 | Umiarkowana multikolinearność |
| OdlOdCentrum | 3,72 | Niska lub dopuszczalna kolinearność |

Wyniki analizy wskaźnika VIF (Variance Inflation Factor) i ich interpretacja.  
  
.

Rysunek 3.18.

Klasycznym przykładem kolinearności prowadzącej do multikolinearności jest para zmiennych Powierzchnia i LiczbaPokoi. VIF > 10 to poważny sygnał ostrzegawczy. W takiej sytuacji należy rozważyć usunięcie jednej ze zmiennych, np. LiczbaPokoi. Po usunięciu źródeł powstania multikolinearności poziom interpretowalności modelu wzrośnie. Alternatywnym podejściem do usuwania zmiennych jest korzystanie z metod regresji, które są odporne na to zjawisko:

1. Regresja grzbietowa (Ridge)
2. Regresja Lasso

3.12. Regresja grzbietowa (Ridge)

Regresja grzbietowa (ang. RidgeRegression) stanowi odmianę klasycznej regresji liniowej, w której do funkcji kosztu (zazwyczaj minimalizowanej OLS) wprowadza się dodatkowy wyraz kary (ang. penalty term). Ma to na celu ograniczenie wpływu współliniowości zmiennych objaśniających oraz przeciwdziałanie zjawisku przeuczenia (overfitting). W literaturze technicznej i statystycznej metoda ta bywa określana także mianem regresji z regularyzacją L2 [11].

W metodzie OLS estymacja odbywa się przez minimalizację sumy kwadratów reszt. Macierzowo:

W regresji grzbietowej do funkcji kosztu dodajemy wyraz kary (regularizacji) w postaci normy L2 wektora . Całkowita minimalizowana funkcja przyjmuje postać, którą macierzowo można zapisać jako:

gdzie jest hiperparametrem kontrolującym siłę regularyzacji. Wyraz wolny zazwyczaj nie jest regularyzowany. Im większa wartość tym większa kara za duże wartości współczynników wektora . Jednym z głównych powodów stosowania regresji grzbietowej jest poprawa stabilności oszacowań. Regresję grzbietową stosuje się w następujących przypadkach:

1. Liczba zmiennych objaśniających jest porównywalna lub większa od liczby obserwacji .
2. Występuje silna współliniowość między zmiennymi objaśniającymi w macierzy .
3. Chęć ograniczenia złożoności modelu i osiągnięcia stabilności estymacji współczynników – zmniejszenie wariancji dla wektora .

Na rysunku 3.19. przedstawiono, jak wartości współczynników regresji zmieniają się wraz ze wzrostem parametru regularyzacji . Wykres ten ukazuje istotę regresji grzbietowej – zmniejszanie i stabilizację wartości współczynników. Pozwala to uzyskać kompromis między dopasowaniem modelu a redukcją wariancji estymatorów.

Obraz zawierający tekst, diagram, linia, Wykres

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Rysunek 3.19.

Wpływ zwiększania wartości parametru regularyzacji λ na wartości współczynników β. Skala logarytmiczna.

3.13. Regresja Lasso

Regresja Lasso (ang. Least Absolute Shrinkage and Selection Operator) to metoda regularyzacji w regresji liniowej, która wprowadza karę w postaci normy L1 do funkcji kosztu. Dzięki temu podejściu możliwe jest zarówno ograniczenie wielkości współczynników regresji, jak i ich selekcja (tj. część współczynników może być dokładnie równa zero).

W regresji Lasso do funkcji kosztu dodajemy karę w postaci normy L1, czyli sumy wartości bezwzględnych współczynników ​. Problem estymacji przy użyciu Lasso można zapisać jako:

gdzie założono, że:

1. ponownie jest parametrem kontrolującym siłę regularyzacji L1.
2. Funkcja nie jest różniczkowalna w punkcie .
3. Dzięki karze L1, Lasso przyjmuje rzadkie rozwiązania, co umożliwia selekcję zmiennych, pozwala na eliminację nieistotnych predyktorów. Jest to szczególnie widoczne przy wysokowymiarowych danych treningowych.
4. Wprowadzenie kary zwiększa bias estymatorów, ale w zamian za to redukuje wariancję i często prowadzi do lepszej zdolności generalizacji modelu przy odpowiedniej wartości hiperparametru .

Ze względu na brak analitycznego rozwiązania problemu Lasso, standardowo stosuje się algorytmy optymalizacyjne, np. algorytm coordinate descent. Rysunek 3.20. pokazuje wykres Lasso Trace Plot, który ilustruje zmiany wartości współczynników w zależności od parametru regularizacji . W miarę wzrostu zaobserwowano, że niektóre współczynniki maleją do zera, co odzwierciedla selekcję zmiennych i redukcję złożoności modelu.

Obraz zawierający tekst, linia, Wykres, diagram

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Działanie efektu shrinkage na wartości poszczególnych współczynników β modelu.

Rysunek 3.20.

3.14. Standaryzacja zmiennych - normalizacja

Standaryzacja zmiennych polega na przekształceniu danych do postaci o średniej 0 i odchyleniu standardowym 1, według wzoru:

gdzie:

* – średnia normalizowanej zmiennej
* – odchylenie standardowe zmiennej

Celem standaryzacji jest sprowadzenie zmiennych do porównywalnej skali, szczególnie gdy mają różne jednostki miary (np. m² vs PLN). Jest to wymóg konieczny w modelach z regularyzacją (Ridge, Lasso), aby zapewnić równomierny wpływ zmiennych na estymację parametrów.

3.14. Przykład

Zmienną Powierzchnia cechują średnia i odchylenie standardowe . Wybrano 3 punkty z przykładowego zbioru danych:

Wystandaryzowane wartości tych punktów wynoszą odpowiednio:

3.15. Kodowanie one-hot encoding

W analizie regresji i uczeniu maszynowym zmienne kategoryczne (np. „Lokalizacja”) muszą zostać zakodowane numerycznie. Popularnym sposobem kodowania jest metoda one-hot encoding, w której dla zmiennej kategorialnej o kategoriach tworzone jest binarnych zmiennych pomocniczych. Każda przyjmuje wartość dla danej kategorii i dla pozostałych. Jedna kategoria musi zostać pominięta – oznacza się ją jako kategorię referencyjną.

3.15. Przykład

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Lokalizacja** |  |  | **Lokalizacja** | **Lokalizacja\_Śródmieście** | **Lokalizacja\_Peryferia** |
| Śródmieście |  |  | Śródmieście | 1 | 0 |
| Peryferia |  | | Peryferia | 0 | 1 |
| Centrum |  |  | Centrum | 0 | 0 |

Przekształcenie zmiennej kategorialnej Lokalizacja metodą one-hot encoding. Kategoria Centrum jest kategorią referencyjną (brak zmiennej – obie kolumny mają 0).

Rysunek 3.21.

3.16. Wartości odstające

Wartość odstająca (ang. outlier) to obserwacja, która znacząco odbiega od pozostałych danych – jej wartość jest nietypowa w kontekście rozkładu zmiennej. Outliery mogą wynikać z:

* Błędów pomiaru lub wprowadzania danych
* Naturalnej zmienności danych (wielkość próbki, rozkład)
* Specyfiki obiektu (np. wyjątkowo duża, świetnie wyposażona nieruchomość)

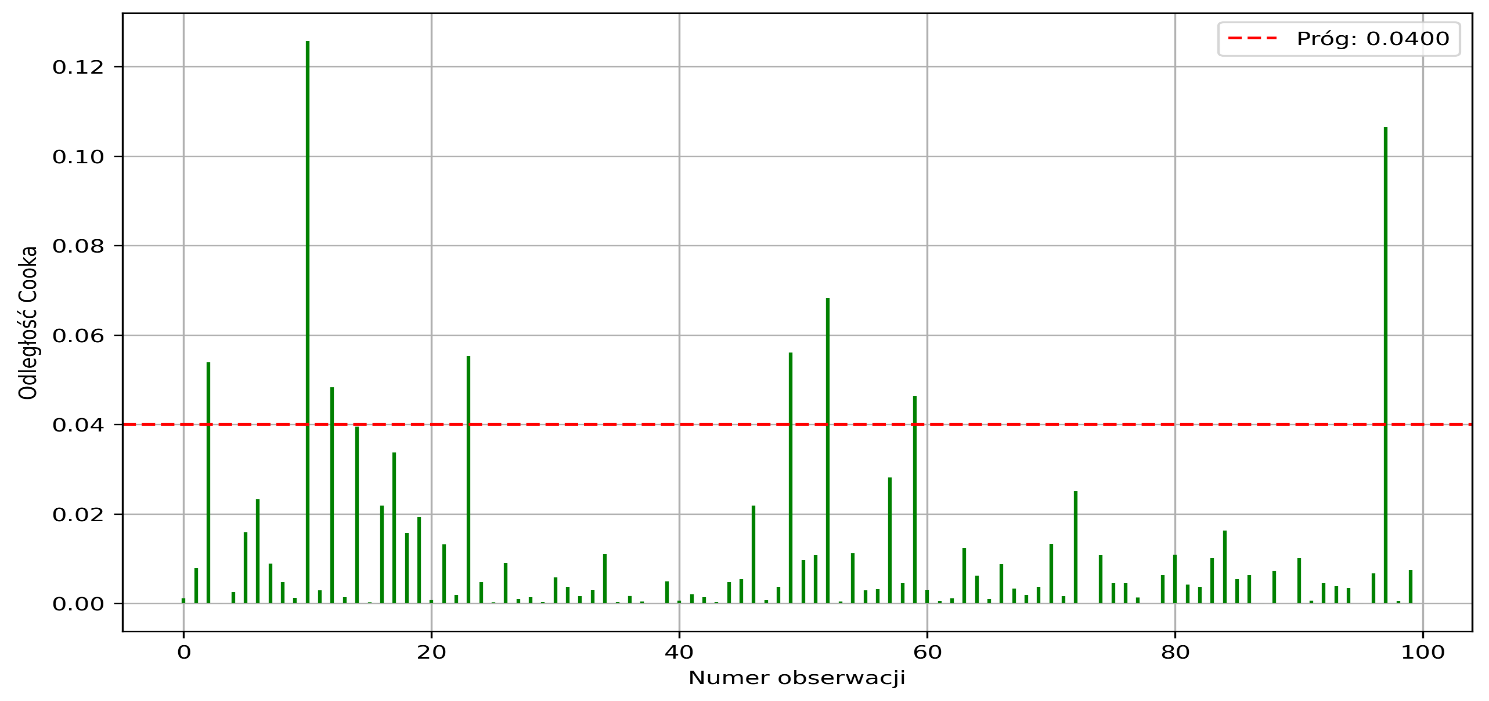
Wartości odstające wpływają negatywnie na model regresji. Zakłócają one estymację współczynników regresji – pojedynczy outlier może silnie przesunąć prostą regresji. Takie punkty danych należy zidentyfikować i podjąć decyzję o dalszych krokach. Identyfikację outlierów ułatwiają metody:

* Reszt studentyzowanych
* Odległości Cooka

Po zidentyfikowaniu potencjalnych wartości odstających należy przeprowadzić merytoryczną analizę tego, czy tak skrajna wartość ma uzasadnienie w rzeczywistości (np. luksusowa posiadłość). Jeśli punkt ten wpływa w znacznym stopniu na wyniki całego modelu można podjąć decyzję o usunięciu go lub skorzystaniu z modelowania odpornego na outliery, np. regresji ważonej.

3.16. Przykład

Przeprowadzono analizę wartości odstających wykorzystując odległość Cook’a. Przy progu istotności 0,04 (dla 100 obserwacji) uzyskano potencjalnie 8 obserwacji odstających:



Metoda odległości Cook’a pozwoliła na identyfikację 8 potencjalnych wartości odstających.

Rysunek 3.22.

3.17. Ocena jakości modelu - wzory

|  |  |
| --- | --- |
| **Nazwa wskaźnika** | **Wzór** |
| Współczynnik korelacji r |  |
| Współczynnik determinacji R2 |  |
| Skorygowany współczynnik determinacji |  |
| SST |  |
| SSR |  |
| SSE |  |
| Błąd standardowy rezyduum  RSE |  |
| Przedziały ufności dla |  |
| MAE |  |
| MAPE |  |
| RMSE |  |
| Kryterium Akaike (AIC) |  |
| Kryterium Schwarza (BIC) |  |
| Kryterium Mallowsa Cp |  |

Tabela zawierająca kompleksowy zbiór wskaźników wykorzystywanych do oceny jakości modelu regresji liniowej wraz z ich matematycznym wzorem (\* Objaśnienia oznaczeń znajdują się na końcu podrozdziału).

Rysunek 3.23.

3.18. Ocena jakości modelu - interpretacja

|  |  |
| --- | --- |
| **Nazwa wskaźnika** | **Opis i interpretacja** |
| Współczynnik korelacji r | Miara siły liniowej zależności pomiędzy dwiema zmiennymi. Wartość r ∈ [−1,1].   * r > 0 – korelacja dodatnia, * r < 0 – korelacja ujemna, * r = 0 – brak liniowej zależności.   Im bliżej wartości 1 lub -1, tym silniejszy związek. |
| Współczynnik determinacji R2 | Wskazuje, jaka część całkowitej zmienności zmiennej zależnej jest wyjaśniona przez model regresji.   * R2 ∈ [0,1]. * Im wyższy R2, tym lepsze dopasowanie modelu.   Przykład: R2 = 0,85 oznacza, że 85% zmienności zmiennej zależnej jest wyjaśnione przez zmienne objaśniające. |
| Skorygowany współczynnik determinacji | Koryguje klasyczne R2 o liczbę zmiennych w modelu. Zapobiega sztucznemu zawyżaniu R2 przy dodawaniu nieistotnych zmiennych.  Przy porównywaniu modeli o różnej liczbie predyktorów preferowane jest . |
| SST | Mierzy całkowitą zmienność zmiennej zależnej względem jej średniej. Stanowi punkt odniesienia – dzieli się na SSR i SSE. |
| SSR | Miara zmienności zmiennej zależnej wyjaśniona przez model regresji.   * Wraz z SSE składa się na SST. * Im większy SSR, tym większa efektywność modelu. |
| SSE | Suma kwadratów odchyleń wartości rzeczywistych od przewidywanych. Mierzy całkowity błąd modelu.  Im mniejsze SSE, tym lepsze dopasowanie. Wyrażone w jednostkach zmiennej zależnej do kwadratu. |
| Błąd standardowy rezyduum  RSE | Średni błąd estymacji, skorygowany o liczbę parametrów w modelu.   * Pokazuje przeciętną odległość między wartością przewidywaną a rzeczywistą. * Im mniejszy RSE, tym większa precyzja prognoz. |
| Przedziały ufności dla | Zakres wartości, w którym z określonym prawdopodobieństwem znajduje się prawdziwa wartość (nieznana) parametru regresji . Typowy poziom ufności: 95%.   * Wąski przedział → precyzyjna estymacja, * Szeroki przedział → duża niepewność. |
| MAE | Średnia bezwzględna różnica między wartością rzeczywistą a przewidywaną.   * Mniej wrażliwy na obserwacje odstające niż RMSE. * Intuicyjnie pokazuje „średni błąd” w PLN, m² itd. |
| MAPE | Średni procentowy błąd predykcji. Pokazuje, o ile procent przeciętnie model myli się w prognozie.   * Umożliwia porównywanie modeli niezależnie od skali danych. * Problem: wrażliwość przy . |
| RMSE | Średni błąd kwadratowy predykcji, wyrażony w jednostkach zmiennej zależnej.   * Im niższy RMSE, tym lepsza dokładność modelu. * Czuły na obserwacje odstające (bo błędy są kwadratowane). |
| Kryterium Akaike (AIC) | Miara dopasowania z karą za złożoność modelu (liczbę zmiennych).   * Niższe AIC = lepszy kompromis dopasowania i prostoty modelu. * Używane do porównywania modeli, nie do interpretacji bezwzględnej. |
| Kryterium Schwarza (BIC) | Podobne do AIC, ale bardziej penalizuje złożoność – faworyzuje prostsze modele przy dużych próbach.   * Niższe BIC = lepszy model. * BIC rośnie szybciej niż AIC przy zwiększaniu liczby parametrów. |
| Kryterium Mallowsa Cp | Służy do wyboru najlepszego modelu spośród wielu.   * Dobre dopasowanie: (liczba zmiennych) * Wartość : – model przeparametryzowany * – niedostateczne dopasowanie |

Tabela zawierająca kompleksowy zbiór wskaźników wykorzystywanych do oceny jakości modelu regresji liniowej wraz z ich opis wraz z interpretacją.

Rysunek 3.24.

**\* Objaśnienia oznaczeń wzorów do tabeli z rysunku 3.23.**

* **–** obserwacja rzeczywista
* – wartość przewidywana przez model
* – średnia z wartości rzeczywistych
* – liczba obserwacji
* – liczba zmiennych objaśniających
* – funkcja wiarygodności modelu
* – błąd standardowy estymatora
* – suma kwadratów reszt z modelu z p zmiennymi
* – średni błąd kwadratowy modelu pełnego

# 4. Opis zbioru danych i środowiska pracy

Rozdział ten stanowi wprowadzenie do części praktycznej niniejszej pracy i koncentruje się na przedstawieniu danych oraz środowiska analitycznego, w którym realizowano analizę. W pierwszej kolejności zaprezentowane zostanie źródło danych oraz proces ich pozyskania, a następnie scharakteryzowane zostaną zmienne, które posłużą do budowy modeli regresyjnych. Szczególną uwagę poświęcono zmiennej zależnej – cenie nieruchomości (SalePrice) – która stanowi podstawę do oceny skuteczności modeli predykcyjnych. Kolejno opisano środowisko obliczeniowe oraz narzędzia programistyczne, które umożliwiły efektywne przetwarzanie danych i generowanie wyników. Informacje zawarte w tym rozdziale stanowią punkt wyjścia do dalszej analizy eksploracyjnej danych, której celem będzie weryfikacja założeń oraz identyfikacja cech mających istotne znaczenie dla predykcji cen nieruchomości.

## 4.1. Źródło danych i sposób pozyskania

Dane wykorzystane w niniejszej analizie pochodzą ze zbioru Ames Housing Dataset, który stanowi rozszerzoną i bardziej szczegółową alternatywę dla klasycznego zbioru Boston Housing. Zbiór ten został opracowany przez profesora Truman State University (Stany Zjednoczone) Dean De Cock’a, wykładającego na wydziale matematyki. Oficjalnie zbiór ukazał się w publikacji Journal of Statistics Education [12] w roku 2011. Zawiera wyjątkowo szczegółowe informacje na temat sprzedaży nieruchomości przeprowadzonych w mieście Ames w stanie Iowa (Stany Zjednoczone) w latach 2006–2010. Podstawowy zbiór danych obejmuje łącznie 2930 obserwacji i ponad 80 zmiennych opisujących cechy nieruchomości, w tym ich lokalizację, parametry techniczne, sposób wykończenia oraz otoczenie. Bogata struktura i wielowymiarowość Ames Housing Dataset powoduje, że zbiór jest szeroko wykorzystywany w literaturze naukowej do celów modelowania regresyjnego oraz oceny jakości budowanych w oparciu o niego modeli predykcyjnych. Dane zostały pierwotnie udostępnione publicznie na stronie internetowej American Statistical Association [13] w formie pliku tekstowego [14] oraz arkusza kalkulacyjnego [15]. Niestety obecnie dostęp nie jest dłużej wspierany. Istnieją jednak inne rzetelne źródła, z których można pozyskać zbiór Ames Housing Dataset. Na potrzeby niniejszej pracy zbiór został pobrany z oficjalnego repozytorium OpenML [16] – platformy często wykorzystywanej w uczeniu maszynowym, która zapewnia pełną integrację z Pythonem. Zbiór został pobrany i dodatkowo zapisany w formie pliku csv.

W swojej podstawowej formie zbiór danych nie uwzględnia m.in. współrzędnych geograficznych. Uznano, że położenie geograficzne może być wartościową dodatkową informacją, która rozszerzy możliwości analizy a także potencjalnie wzbogaci modele regresji o dodatkowe predyktory. Sam autor informuje w swojej publikacji [12] o możliwości wykorzystania zmiennej PID (ang. Parcel Identification Number) czyli numeru identyfikacyjnego działki do pozyskania dokładnych współrzędnych geograficznych i innych danych przestrzennych. Numer PID pozwala na identyfikację działki w porozumieniu z biurem rzeczoznawcy majątkowego [17] lub wykorzystując systemy informacji przestrzennej (ang. GIS – Geographic Information System) dostawcy Beacon Inc. [18]. W celu pozyskania danych przestrzennych na podstawie numeru PID skonstruowany został web-scrapper [19], który potrafił pobierać dane dla pojedynczych rekordów. Ograniczenia w ilości możliwych zapytań do serwera spowodowały jednak, że lepszym rozwiązaniem okazało się wyeksportowanie plików pozwalających na zbudowanie własnej lokalnej bazy danych na silniku SZBD MySQL [20], który umożliwia przechowywanie danych przestrzennych w formie skompresowanych obiektów. Numery PID ze zbioru danych Ames Housing Dataset zostały wykorzystane do dopasowania odpowiedniego adresu (zmienna address) oraz współrzędnych szerokości geograficznej (zmienna latitude) i długości geograficznej (zmienna longitude) nieruchomości. Uzyskany zbiór, wzbogacony o dane przestrzenne, stanowił podstawę do dalszej analizy eksploracyjnej oraz budowy modeli predykcyjnych.

Obraz zawierający tekst, mapa

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna. Obraz zawierający mapa, tekst

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Fragment ortofotomapy obszaru Ames (Iowa) z naniesionymi granicami działek ewidencyjnych, przygotowany w systemie informacji przestrzennej (GIS). Zdjęcia zostały wyeksportowane z serwisu Beacon Inc.

Rysunek 4.1.

## 4.2. Zmienna zależna SalePrice – definicja i znaczenie

Podstawowym celem niniejszej analizy jest stworzenie modelu umożliwiającego predykcję ceny sprzedaży nieruchomości. Z tego względu jako zmienną zależną wybrano zmienną SalePrice, która reprezentuje rzeczywistą cenę transakcyjną uzyskaną przy sprzedaży nieruchomości w mieście Ames (stan Iowa, USA) w latach 2006–2010. Wartości zmiennej wyrażone są w dolarach amerykańskich (USD) i odnoszą się do pełnych kwot transakcyjnych odnotowanych w rejestrach publicznych, co czyni je szczególnie wiarygodnymi.

Zmienna SalePrice jest zmienną ilościową ciągłą i charakteryzuje się istotnym rozrzutem wartości – ceny w zbiorze mieszczą się w zakresie od około 12 000 USD do ponad 750 000 USD. Średnia cena sprzedaży nieruchomości wynosi około 180 000 USD przy odchyleniu standardowym na poziomie 79 000 USD. Mediana wynosi 160 000 USD, co wskazuje na obecność wartości odstających i prawoskośność rozkładu. SalePrice pełni rolę zmiennej objaśnianej w modelach regresyjnych i stanowi punkt odniesienia do oceny efektywności i trafności prognozowania.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **N** | **Min** | **Q1** | **Mediana** | **Q3** | **Max** |
| 2 930 | 12 789 | 129 500 | 160 000 | 213 500 | 755 000 |

Podstawowe statystyki opisowe dla rozkładu zmiennej objaśnianej SalePrice w zbiorze.

Rysunek 4.2.

Obraz zawierający tekst, diagram, Wykres, linia

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Histogram zmiennej SalePrice. Szerokość pojedynczego przedziału na histogramie wynosi 24 740 USD.

Rysunek 4.2.

Rozstęp międzykwartylowy (ang. IQR – Interquartile Range) wynosi 84 000 USD. Zgodnie ze wzorami:

wyznaczono odpowiednio:

1. Upper Bound = 339 500 USD, co przekłada się na potencjalnych 137 (4,7%) outlierów.
2. Lower Bound = 3 500 USD, w wyniku czego nie uzyskano żadnych outlierów.

Obraz zawierający tekst, zrzut ekranu, linia, Wykres

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Rysunek 4.3.

Wykres boxplot wraz z zaznaczonymi wartościami Lower i Upper Bound. Wykres pokazuje potencjalnych 137 wartości odstających zmiennej SalePrice.

Zmienna SalePrice cechuje się asymetrycznym prawostronnym rozkładem – oznacza to, że większość obserwacji znajduje się w dolnym zakresie cen, natomiast niewielka liczba transakcji dotyczy nieruchomości o wysokiej wartości, co pokazuje wykres gęstości na rysunku 4.4.

Obraz zawierający linia, Wykres, diagram

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Wykres gęstości rozkładu zmiennej SalePrice w postaci violin chart. Widoczna prawoskośność.

Rysunek 4.4.

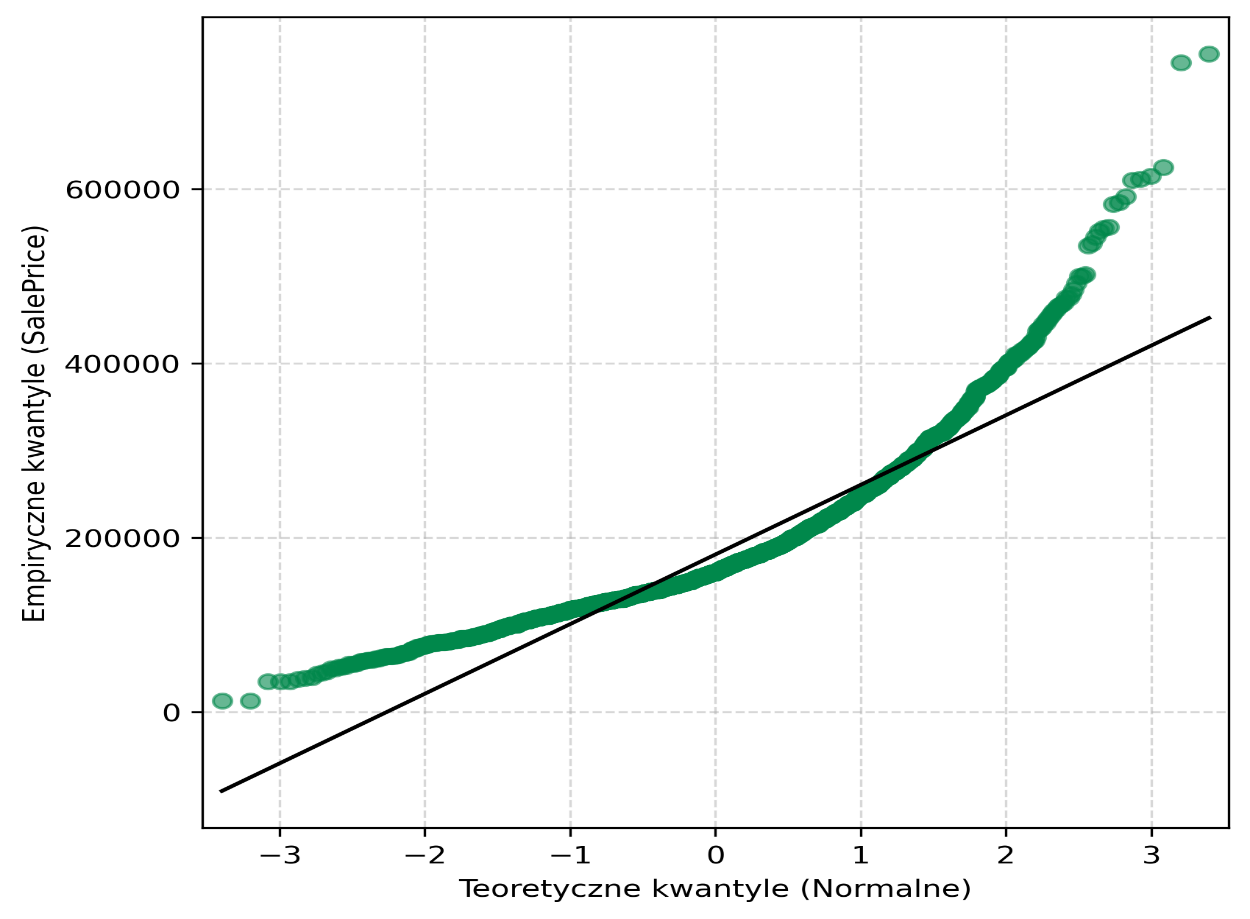
Estymacja gęstości rozkładu (Kernel Density Estimation – KDE) wskazuje, że największa koncentracja cen sprzedaży mieści się w przedziale 120 000–180 000 USD. Widoczny jest również długi ogon rozkładu w prawą stronę, co oznacza obecność transakcji o znacznie wyższej wartości (powyżej 300 000 USD), klasyfikowanych jako outliery.. Współczynnik zmienności na poziomie 0,42 świadczy o wysokiej zmienności. Prawoskośność potwierdza wartość współczynnika skośności na poziomie 1,74. Dodatkowo wysoka wartość kurtozy (5,1) wskazuje na potencjalnie gorsze dopasowanie modelu regresji liniowej do danych, ponieważ rozkład nie przypomina precyzyjnie rozkładu normalnego.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Średnia** | **Odchylenie standardowe** | **Współczynnik zmienności** | **Skośność** | **Kurtoza** |
| 180 796 | 79 873 | 0,442 | 1,744 | 5,119 |

Pozostałe statystyki opisowe dla zmiennej SalePrice wraz ze skośnością i wartością kurtozy.

Rysunek 4.5.

W celu ustalenia, czy rozkład istotnie różni się od rozkładu normalnego przeprowadzono test Shapiro-Wilka. Jako hipotezę H0 przyjęto normalność rozkładu SalePrice. Statystyka testowa (W) na poziomie 0,87 oraz p-value = 0 dla poziomu istotności równego 0,05 potwierdzają, że rozkład istotnie różni się od rozkładu normalnego. Hipoteza H0 zostaje więc odrzucona. Wykres kwantyl-kwantyl na rysunku 4.6. potwierdza brak rozkładu normalnego.



Wykres kwantyl-kwantyl zmiennej SalePrice pokazujący wyraźne różnice w stosunku od rozkładu normalnego.

Rysunek 4.6.

Wykazana prawoskośność uzasadnia zastosowanie przekształcenia logarytmicznego. Transformacja przy pomocy funkcji logarytmu naturalnego zmniejsza wpływ wartości ekstremalnych i przybliża rozkład do rozkładu normalnego. Zastosowano przekształcenie postaci:

Rozkład przekształconej zmiennej SalePrice zdecydowanie bardziej przypomina rozkład normalny. Potwierdzają to wykresy porównawcze kwantyl-kwantyl na rysunku 4.7. Punkty dla przekształconej zmiennej układają się zdecydowanie bliżej prostej będącej punktem odniesienia.

Obraz zawierający diagram, linia, Wykres, tekst

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Wykres porównawczy kwantyl-kwantyl dla bazowej zmiennej SalePrice (po lewo) oraz przekształconej funkcją logarytmu naturalnego zmiennej LogSalePrice (prawo).

Rysunek 4.7.

Ponownie przeprowadzono test Shapiro-Wilka na przekształconej zmiennej. Test również tym razem wskazał na statystycznie istotne odstępstwo od rozkładu normalnego (W = 0,9858; p < 0,000001), co oznacza formalne odrzucenie hipotezy o normalności. Niemniej jednak, należy podkreślić, że testy normalności przy dużych próbach (n > 1000) są wyjątkowo czułe na nawet niewielkie odchylenia od normalności, które nie mają znaczenia praktycznego. Wizualna analiza rozkładu przy pomocy Q–Q plot wskazują na istotną poprawę rozkładu po transformacji logarytmicznej, w szczególności w zakresie redukcji skośności i kurtozy. Z tego względu, pomimo formalnego odrzucenia normalności, rozkład LogSalePrice jest wystarczająco zbliżony do normalnego, aby możliwe było dalsze stosowanie modeli regresji liniowej, które są odporne na niewielkie naruszenia założeń normalności błędów losowych, zwłaszcza przy dużych próbach. W dalszej części pracy przedstawiono porównanie wyników modeli opartych o zmienną SalePrice i LogSalePrice.

Dalsza analiza zmiennej zależnej została przeprowadzona w podrozdziale 5.1, w którym, w oparciu o metody eksploracyjnej analizy danych, przeprowadzono selekcję obserwacji do dalszego modelowania regresyjnego. Obserwacje uznane za odstające (outliery) zostały wykluczone, natomiast pozostałe posłużyły do budowy modeli predykcyjnych w kolejnych etapach pracy.

## 4.3. Charakterystyka i klasyfikacja zmiennych niezależnych

Szczegółowy opis zmiennych niezależnych zbioru Ames Housing Dataset został przygotowany przez jego autora Dean De Cock’a i opublikowany na w zasobach American Statistical Association [12]. Dla celów analitycznych zmienne zostały sklasyfikowane na podstawie ich typu danych i sposobu pomiaru. Klasyfikacja ta ma istotne znaczenie przy dalszym przetwarzaniu danych, w tym podczas kodowania zmiennych kategorycznych oraz standaryzacji zmiennych liczbowych. Dla celów przejrzystości w niniejszym rozdziale przedstawiono wybrane zmienne objaśniające. Pełne zestawienie znajduje się w repozytorium [19]. Zbiór danych wykorzystywany w tej pracy składa się z 81 zmiennych pochodzących ze zbioru podstawowego oraz trzech zmiennych pozyskanych na podstawie PID nieruchomości. Poniżej przedstawiono podzbiór dziesięciu z nich wraz z opisem.

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Nazwa** | **Typ** | **Opis** |
| 1 | Gr\_Liv\_Area | Ciągła (stopy²) | Silnie dodatnia korelacja z ceną, kluczowy miernik wielkości użytkowej |
| 2 | Overall\_Qual | Porządkowa (1–10) | Ocena jakości wykończenia – najczęściej koreluje z ceną, subiektywna ale istotna. |
| 3 | Garage\_Area | Ciągła (stopy²) | Powierzchnia garażu – wpływa na cenę, często analizowana zmienna |
| 4 | Total\_Bsmt\_SF | Ciągła (stopy²) | Powierzchnia piwnicy – zwiększa użyteczność i wartość nieruchomości |
| 5 | Year\_Built | Dyskretna (rok) | Rok budowy – silny predyktor (młodsze nieruchomości = wyższa cena) |
| 6 | Full\_Bath | Dyskretna (liczba) | Liczba pełnych łazienek – wpływa na komfort i wartość |
| 7 | TotRms\_AbvGrd | Dyskretna (liczba) | Liczba pomieszczeń – miernik funkcjonalności i wielkości |
| 8 | Garage\_Cars | Dyskretna (liczba) | Liczba miejsc garażowych – koreluje z GarageArea, użyteczne do oceny |
| 9 | First\_Flr\_SF | Ciągła (stopy²) | Powierzchnia parteru – często skorelowana z ceną i GrLivArea |
| 10 | Mas\_Vnr\_Area | Ciągła (stopy²) | Powierzchnia okładziny murarskiej – estetyka zewnętrzna, wpływ na wartość |

Wycinek 10 przykładowych zmiennych objaśniających ze zbioru Ames Housing Dataset.

Rysunek 4.8.

Dodatkowo zbiór został poszerzony o zmienne:

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Nazwa** | **Typ** | **Opis** |
| 1 | Address | Kategoryczna | Adres nieruchomości, uzupełniająca informacja identyfikacyjna, niewykorzystywana w modelowaniu |
| 2 | Latitude | Ciągła (ilościowa zmiennoprzecinkowa) | Szerokość geograficzna nieruchomości, użyteczna do określenia lokalizacji |
| 3 | Longitude | Ciągła (ilościowa zmiennoprzecinkowa) | Długość geograficzna nieruchomości, wraz z lattitude umożliwia analizę przestrzenną |

Dodatkowe zmienne pozyskane dzięki PID rozszerzające zbiór Ames Housing Dataset.

Rysunek 4.9.

Zmienne Latitude oraz Longitude posłużą w dalszej części pracy do wyznaczenia zmiennej Dist\_from\_Center\_ft, czyli odległości od centrum miasta metodą odległości Haversine'a. Adres może być wykorzystany przez wiele bibliotek pozwalających na wizualizację przestrzenna np. w postaci mapy / planu miasta. W ramach klasyfikacji cech przeprowadzono wstępną analizę współzależności pomiędzy zmiennymi liczbowymi a zmienną celu SalePrice. Celem było określenie siły i kierunku zależności liniowych, co umożliwia identyfikację potencjalnych predyktorów ceny nieruchomości.

Obraz zawierający tekst, diagram, zrzut ekranu, Plan

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Mapa cieplna dla wybranych zmiennych pokazujących siłę i kierunek zależności liniowych względem zmiennej celu SalePrice.

Rysunek 4.10.

Analiza współczynnika korelacji Pearsona pozwoliła określić siłę oraz kierunek liniowych zależności pomiędzy wybranymi zmiennymi objaśniającymi a zmienną zależną SalePrice. Poniżej przedstawiono szczegółowe wnioski z przeprowadzonej analizy:

1. **Silna dodatnia korelacja (r > 0,6)**

* Overall\_Qual (r ≈ 0,81) – ogólna jakość wykończenia domu to zmienna o najsilniejszym wpływie liniowym na cenę. Wzrost jakości o jeden punkt wiąże się z wyraźnym wzrostem wartości nieruchomości. Wskazuje to na znaczenie subiektywnej oceny wykończenia dla rynku nieruchomości.
* Gr\_Liv\_Area (r ≈ 0,71) – powierzchnia mieszkalna nad poziomem gruntu. Zmienna ta wykazuje silną dodatnią korelację. Większe domy osiągają wyższe ceny.
* Garage\_Area (r ≈ 0,65) oraz Garage\_Cars (r ≈ 0,64) – powierzchnia garażu i liczba miejsc w garażu są również silnie skorelowane z ceną.
* Total\_Bsmt\_SF (r ≈ 0,61) – Wielkość piwnicy wpływa na cenę.
* First\_Flr\_SF (r ≈ 0,61) – cena zależy w dużym stopniu od wielkości parteru. Łączy się to ze wcześniejszym wnioskiem dotyczącym łącznej powierzchni mieszkania.

1. **Umiarkowana dodatnia korelacja (0,4 < r ≤ 0,6)**

* Full\_Bath (r ≈ 0,56) – Im więcej łazienek, tym wyższa cena nieruchomości.
* TotRms\_AbvGrd (r ≈ 0,53) – Zmienna ta wykazuje umiarkowaną korelację.
* Year\_Built (r ≈ 0,52) – Nowsze domy osiągają wyższe ceny.
* Mas\_Vnr\_Area (r ≈ 0,47) – Umiarkowany stopień korelacji.

1. **Słaba korelacja i brak związku liniowego**

* longitude (r ≈ -0,02) – współrzędna geograficzna długości nie wykazuje istotnego związku z ceną. Samodzielnie nie niesie wartości predykcyjnej.
* lattitude (r ≈ -0,50) – współrzędna szerokości geograficznej wykazuje istotną ujemną korelację, co sugeruje przestrzenny trend cenowy (np. nieruchomości położone bardziej na północ miasta są tańsze). Zmienna ta powinna zostać rozważona przy dalszej analizie przestrzennej i może pełnić rolę dodatkowego predyktora.

Zmienne Garage\_Area i Garage\_Cars są silnie skorelowane ze sobą, co może wskazywać na potencjalną multikolinearność w modelach regresji wielorakiej. Przeprowadzona analiza pozwoliła wskazać zestaw zmiennych, które charakteryzują się najsilniejszą korelacją z ceną nieruchomości i mogą zostać uznane za potencjalnie wartościowe predyktory w dalszych analizach. Uzyskane wyniki stanowią podstawę do eksploracyjnej analizy danych i konstrukcji modeli predykcyjnych w kolejnych etapach pracy.

Poza zmiennymi liczbowymi w zbiorze danych znajdują się zmienne kategorialne. Te z nich, które zostały wykorzystane w modelowaniu będą w dalszym etapie pracy transformowane metodami one-hot encoding lub label-encoding, / mappingu liczbowego. Np. zmienna MS\_SubClass oraz Sale\_Condition. Dodatkowo w podrozdziale 5.1. poświęconym eksploracyjnej analizie danych rozważono konieczność transformacji funkcyjnych niektórych zmiennych, aby poprawić normalność ich rozkładu.

## 4.4. Środowisko pracy i konfiguracja narzędzi

Analizy przedstawione w niniejszej pracy zostały przeprowadzone w środowisku lokalnym opartym na systemie operacyjnym Windows 11 (64-bit). Proces pozyskiwania, przetwarzania i analizy danych obejmował kilka etapów oraz wykorzystanie szeregu specjalistycznych narzędzi programistycznych i analitycznych.

Podstawowy zbiór danych dotyczący nieruchomości został pobrany z publicznego repozytorium OpenML (Ames Housing Dataset). Niezależnie od tego, w celu rozszerzenia danych o informacje przestrzenne, pozyskano zbiór danych geograficznych w postaci plików konfiguracyjnych bazy danych, zawierających współrzędne geograficzne oraz inne informacje lokalizacyjne dotyczące poszczególnych działek. Na podstawie tych plików utworzono lokalną bazę danych w systemie MySQL, umożliwiającą przechowywanie danych przestrzennych w formie obiektów.

W kolejnym etapie, przy użyciu języka Python 3.12.7 (dystrybucja Anaconda) oraz środowiska Jupyter Notebook, dokonano integracji obu źródeł danych w jeden spójny zbiór analityczny. Dane zostały oczyszczone i przekształcone z wykorzystaniem bibliotek pandas i numpy, a następnie zapisane w formacie CSV w celu dalszej analizy. Prace wykonywane były w specjalnym wirtualnym środowisku systemu Windows, przy pomocy dystrybucji Anaconda.

Do analizy i wizualizacji danych zastosowano zarówno arkusze kalkulacyjne Excel (wraz z wbudowanym narzędziem do pobierania i przekształcania danych Power Query), które posłużyły do wykonania podstawowych analiz statystycznych i eksploracyjnych (EDA), jak i narzędzia programistyczne w języku Python – szczególnie w przypadku bardziej złożonych analiz i modelowania regresyjnego. Modele oraz analizy statystyczne, w tym testy normalności, diagnostyka reszt oraz ocena jakości modeli, wykonano przy pomocy bibliotek statsmodels, scipy.stats oraz scikit-learn.

Do analizy i wizualizacji danych przestrzennych, a także do weryfikacji położenia działek i ich identyfikatorów, wykorzystano oprogramowanie QGIS. Narzędzie to umożliwiło przegląd i analizę danych przestrzennych bezpośrednio z utworzonej bazy danych MySQL.

Kod źródłowy, w tym skrypty do pozyskiwania, łączenia danych oraz generowania wykresów i modeli regresyjnych, została udostępniona w publicznym repozytorium GitHub. Repozytorium to umożliwia w dużym stopniu replikację przeprowadzonych analiz oraz stanowi dokumentację kroków obliczeniowych.

## 4.5. Biblioteki i pakiety analityczne

Analiza danych i modelowanie zostały przeprowadzone z wykorzystaniem szeregu bibliotek analitycznych w języku Python. Poniżej przedstawiono kluczowe pakiety. Umożliwiły one przeprowadzenie analizy eksploracyjnej danych, budowę modeli regresyjnych oraz ocenę ich jakości.

* + 1. **Operacje pomocnicze**
* os
* warnings
* itertools
* datetime

**2. Pobieranie i integracja danych**

* mysql-connector-python
* sqlalchemy

**3. Prztwarzanie i EDA**

* pandas
* numpy

1. **Analiza statystyczna i modelowanie**

* scipy.stats
* statsmodels
* sklearn

**5. Wizualizacja danych**

* matplotlib
* matplotlib.ticker
* mpl\_toolkits.mplot3d
* seaborn

**6.**  **Obsługa danych przestrzennych i generowanie map**

* geopandas
* shapely
* fiona
* folium

# 5. Analiza danych i implementacja modeli predykcyjnych

Surowe dane dotyczące cen nieruchomości w Ames wymagały wstępnego oczyszczenia i przygotowania przed analizą. Przeprowadzono szereg kroków, obejmujących identyfikację brakujących informacji oraz ich uzupełnienie, usunięcie nieprawidłowych rekordów, przekształcenia wybranych zmiennych, standaryzację formatów danych tekstowych, korektę błędów w danych (np. lat budowy) i weryfikację poprawności. Poniżej opisano kolejne etapy procesu czyszczenia oraz uzasadnienie podjętych działań.

## 5.1. Czyszczenie i przygotowanie danych

Pierwszym krokiem była identyfikacja zakresu brakujących danych w zbiorze. Zbadano każdy atrybut pod kątem występowania wartości pustych (NULL/NaN). oraz potencjalnych symboli zastępczych (np. pusty ciąg znaków lub specjalne kody). Stwierdzono, że na pierwszy rzut oka wiele pól zawiera braki, jednak zgodnie z dokumentacją zbioru dane te nie są brakami losowymi, lecz oznaczają brak danej cechy w nieruchomości​. Przykładowo, brak wartości w kolumnie Pool\_QC (ogólna ocena basenu) oznacza, że dom nie posiada basenu, a brak wpisu dla zmiennej. Alley oznacza brak dostępu do alejki dojazdowej​. Podobnie, pola dotyczące piwnicy (Bsmt\_Qual, Bsmt\_Cond itp.), garażu (Garage\_Type, Garage\_Finish, Garage\_Yr\_Blt itp.), kominka (Fireplace\_Qu) czy ogrodzenia (Fence) mają wartość NA właśnie wtedy, gdy dany dom nie ma takiej infrastruktury. Po uwzględnieniu tej zamierzonej pustki okazało się, że rzeczywiste braki danych dotyczą tylko kilku zmiennych: m.in. Lot\_Frontage (długość działki od ulicy), Mas\_Vnr\_Type (rodzaj okładziny kamiennej) oraz sporadycznie pojedynczych wpisów w innych kolumnach (np. Electrical)​.

* + 1. **Ujednolicenie oznaczeń braków**

Przed przystąpieniem do imputacji upewniono się, że wszystkie braki zostały konsekwentnie oznaczone jako NaN. W tym celu zamieniono ewentualne puste ciągi znaków lub nietypowe kody na wartość NaN. Dzięki temu dalsze procedury uzupełniania mogły traktować brakujące dane jednolicie.

* + 1. **Imputacja wartości brakujących**

Dla każdej zmiennej przyjęto strategię uzupełniania odpowiednią do charakteru danych i przyczyny braku

* 1. **Brak cechy - wartości NA, ang. Not Applicable**

W kategoriach tekstowych, gdzie brak oznacza brak danej cechy w domu, brakujące pola zastąpiono jednoznacznym identyfikatorem None. Takie podejście zastosowano m.in. dla kolumn: Pool\_QC (brak basenu), Misc\_Feature (brak dodatków np. szopy), Alley (brak alejki), Fence (brak ogrodzenia)​ oraz wszystkich jakościowych informacji o piwnicy i garażu. W ten sposób zamierzone braki zostały zastąpione kategorią None, co informuje algorytmy, że dana cecha nie występuje, zamiast traktować te wiersze jako przypadkowo nieuzupełnione. Dla przykładu, dokumentacja wskazuje wprost, że w kolumnie Alley wartość NA oznacza brak dostępu do alejki​. Zgodnie z tym wszystkie NaN w Alley zamieniono na kategorię None. Podobnie, brakujące wpisy w MasVnrType zinterpretowano jako brak okładziny i oznaczono jako None.

* 1. **Brak cechy – pola numeryczne**

W przypadku zmiennych numerycznych, gdzie NaN oznacza brak danej struktury, brakujące wartości uzupełniono zerem, które w tym kontekście pełni rolę wartości oznaczającej nieistnienie powierzchni lub liczby sztuk. Dotyczy to przede wszystkim parametrów piwnic i garaży. Na przykład: brak powierzchni piwnicy (Total\_Bsmt\_SF) oznacza, że dom nie ma piwnicy – w takim przypadku zastąpienie pustej wartości zerem jest logiczne (0 sf² piwnicy). Analogicznie postąpiono dla innych pól liczbowych powiązanych z nieistniejącą piwnicą (Bsmt\_Fin\_SF1, Bsmt\_Fin\_SF2, Bsmt\_Unf\_SF, Bsmt\_FullBath, Bsmt\_HalfBath) oraz garażem (Garage\_Cars – liczba miejsc = 0, Garage\_Area = 0). Zastąpienie braków zerem zastosowano także dla Mas\_Vnr\_Area w przypadkach, gdy brak okładziny wynikał z Mas\_Vnr\_Type="None"​. Powyższe działania spowodowały, że wszystkie techniczne braki oznaczające brak cechy zostały uzupełnione odpowiednio, nie zniekształcając znaczenia danych.

* 1. **Brak rzeczywisty - prawdziwe braki danych**

Dla pól, w których brak wartości oznacza rzeczywisty brak informacji (pominięcie w rejestracji danych), zastosowano metodę imputacji opartą na analizie danych:

* + Zmienna Electrical (typ instalacji elektrycznej) miała dokładnie jeden brakujący wpis. Uzupełniono go w najprostszy, uzasadniony sposób – poprzez podstawienie wartości modalnej (najczęstszej) dla tej zmiennej w zbiorze. W tym przypadku brak został zastąpiony dominującą kategorią "SBrkr" (Standard Circuit Breakers), co jest rozsądne, gdyż zdecydowana większość domów miała ten typ instalacji.
  + Zmienna Lot\_Frontage (długość frontu działki w stopach) posiadała znaczną liczbę brakujących wartości (w danych ok. 17%, a w całym zbiorze około 486 braków​). Ponieważ brak ten nie wynika z braku cechy (każdy dom ma dostęp do jakiejś ulicy), uznano, że wartości te zostały opuszczone przypadkowo i należy je oszacować. Zastosowano strategię imputacji z wykorzystaniem informacji o lokalizacji – długość frontu działki uzupełniono medianą obliczoną w obrębie danej dzielnicy (Neighborhood). Intuicyjnie, domy położone w tej samej okolicy mają podobne wielkości działek, zatem użycie mediany z danej dzielnicy pozwala uchwycić lokalne zależności lepiej niż jedna globalna średnia dla całego miasta. Implementacyjnie, pogrupowano dane po zmiennej Neighborhood i dla każdego podzbioru obliczono medianę Lot\_Frontage, którą następnie przypisano do brakujących obserwacji z tej grupy. Takie grupowe uzupełnienie braków zapewnia, że np. w prestiżowej dzielnicy North Ames brakujące wartości frontu otrzymają medianę obliczoną z działek w North Ames, a w tańszej okolicy Edwards – medianę dla Edwards, co odzwierciedla rzeczywistość lepiej niż przypisanie jednej wartości wszystkim. (Alternatywnie rozważano uwzględnienie dodatkowo kształtu działki przy imputacji – np. mediana w grupie (Neighborhood, LotShape). Zrezygnowano jednak z tego pomysłu z uwagi na małą liczbę braków.
  + Zmienne Longitude i Latitude uzupełniono podobną metoda, co zmienną Lot\_Frontage. Również wykorzystano medianę dla Neighborhood.
  + Zmienną Address uzupełniono ręcznie dla odpowiadających wartości zmiennych Longitude i Latitude.

Po wykonaniu powyższych czynności większość zmiennych nie posiadała już wartości pustych. Wszystkie brakujące dane zostały uzupełnione albo odpowiednią wartością zastępczą (0, „None”) informującą o braku cechy, albo oszacowane na podstawie innych danych (mediana lub moda). W efekcie, zgodnie z oczekiwaniami, po imputacji tylko nieliczne brakujące wpisy mogłyby pozostać. Końcowa kontrola (opisano w dalszej części pracy) potwierdziła, że w zbiorze nie pozostały żadne nieuzupełnione braki. Istotne było, aby wprowadzając powyższe uzupełnienia nie zaburzyć logiki danych. Dlatego każdy przypadek traktowano indywidualnie. Szczególną uwagę zwrócono na zmienną Lot\_Frontage. Czasem praktykowanym, lecz niewłaściwym podejściem jest uzupełnianie brakujących długości działki wartością 0. Taka operacja wprowadzałaby do danych domy o zerowej szerokości działki przy ulicy, co jest fizycznie niemożliwe i oznaczałoby poważny błąd. Zamiast tego zastosowano metodę opartą na medianie, dzięki czemu uniknięto sprzeczności z logiką domeny. Podobnie, upewniono się, że zero jako wartość zastępcza pojawia się tylko tam, gdzie rzeczywiście oznacza nie dotyczy (np. 0 garaży, 0 piwnic), a nie w przypadku prawdziwych braków wymagających estymacji.

**3. Usuwanie odstających i niepełnych rekordów**

Kolejnym etapem było usunięcie z danych nietypowych obserwacji (outlierów) lub rekordów zawierających sprzeczne informacje, o ile takie występowały. Celem było poprawienie jakości danych do modelowania poprzez eliminację przypadków, które mogłyby nadmiernie zaburzać rozkłady lub relacje między zmiennymi. Analiza danych ujawniła istnienie kilku domów o skrajnych wartościach cech, mogących negatywnie wpłynąć na trening modelu regresyjnego. W szczególności zwrócono uwagę na dwa domy o wyjątkowo dużej powierzchni użytkowej (Gr\_Liv\_Area ponad 4000 sf2) połączonej z relatywnie niską ceną sprzedaży. Takie punkty odstają na wykresie zależności ceny od powierzchni – podczas gdy dla większości domów większa powierzchnia wiąże się z wyższą ceną, te trzy obiekty miały zaskakująco niskie ceny jak na swoje rozmiary. Istniało podejrzenie, że mogły to być np. budynki w bardzo złym stanie lub sprzedaże pod przymusem, niemniej statystycznie działały one jak obserwacje odstające o dużej dźwigni. W związku z powyższym podjęto decyzję o usunięciu tych rekordów ze zbioru, aby model uczony na danych nie był zaburzony przez te nieliczne ekstremalne przypadki​. Usunięcie ich jest uzasadnione, gdyż stanowią one < 0,5% danych, a mogą nieproporcjonalnie wpływać na dopasowanie modelu (np. współczynniki regresji). Po ich usunięciu zależność ceny od powierzchni mieszkalnej staje się bardziej regularna. Decyzja o usunięciu opierała się na analizie tzw. z-score oraz rozkładów – usuwano te rekordy, które ekstremalnie odbiegały od reszty i miały marginalny wpływ na statystyczną reprezentatywność zbioru. Dodatkowo sprawdzono również, czy w danych nie występują duplikaty rekordów, np. powtarzające się order\_id lub PID. Zbiór nie zawierał żadnych duplikatów. Nie stwierdzono również brakujących wartości w zmiennej celu SalePrice. Dodatkowo, na etapie czyszczenia danych postanowiono usunąć kolumnę identyfikatora order\_id. Była to kolumna zawierająca unikalny numer każdego domu, użyteczna do administracyjnego oznaczenia wierszy, lecz nie niosła informacji przydatnej w przewidywaniu ceny.

Obraz zawierający tekst, diagram, Wykres

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Wykres rozrzutu pokazujący relację między Gr\_Liv\_Area oraz SalePrice. Na wykresie widać 3 wartości odstające.

Rysunek 5.1.

**4. Transformacje i kodowanie zmiennych**

W ramach czyszczenia danych dokonano także wybranych transformacji zmiennych, aby przygotować je do modelowania i zapewnić spójność typów danych. Kilka cech w zbiorze Ames wymaga interpretacji lub zmiany typu, zamiast pozostawienia ich w formie surowej:

* 1. **Konwersja zmiennych numerycznych na kategorie**
  + Zmienna MS\_SubClass (klasa zabudowy) pierwotnie jest zapisana jako liczba całkowita, jednak liczby te odpowiadają kodom kategorii (np. 20 = jednorodzinny bungalow, 30 = jednorodzinny wiekowy). W takiej sytuacji bardziej odpowiednie jest potraktowanie jej jako zmiennej kategorycznej, dlatego dokonano konwersji MS\_SubClass na typ kategoryczny. Dzięki temu model nie będzie interpretował różnicy między 20 a 30 jako 10. Podobnie postąpiono z kilkoma innymi atrybutami kodowanymi liczbowo, które tak naprawdę reprezentują kategorie bez naturalnego porządku: np. MoSold (miesiąc sprzedaży, 1–12) potraktowano jako kategorię nominalną a nie zmienną ciągłą, ponieważ miesiące są cykliczne i nie rosną liniowo.

**4.2. Kodowanie zmiennych porządkowych**

* Wiele cech jakościowych domu (materiały, wykończenie, stan) jest ocenianych w skali tekstowej, np. Exter\_Qual, Bsmt\_Qual, Heating\_QC mają możliwe wartości: Ex (Excellent), Gd (Good), TA (Typical/Average), Fa (Fair), Po (Poor) albo Functional: Typ (Typcial), Min1, Min2, Maj1. Te kategorie mają naturalny porządek jakościowy, który należy zachować. Dlatego zamieniono je na skale numeryczne odpowiadające rankingowi. Przykładowo, dla kolumn oceniających jakość zastosowano mapowanie: Ex = 5, Gd = 4, TA = 3, Fa = 2, Po = 1 (oraz „None” = 0, jeśli cecha nie występuje). Dzięki temu modele uczące się na danych liczbowych będą mogły wykorzystać informację, że np. Excellent > Good > Average w sensie jakości, zamiast traktować te etykiety jako odrębne, nieuporządkowane klasy. Wprowadzenie takiego kodowania jest uzasadnione, ponieważ różnice w jakości mają nieliniowy, ale uporządkowany wpływ na cenę. Zachowanie tej informacji w postaci stopni to lepsze rozwiązanie niż metoda one-hot encoding, która gubi relacje porządkowe.

**4.3. Transformacja zmiennej docelowej**

* Analizie poddano również rozkład zmiennej SalePrice. Okazało się, że ceny domów są silnie prawoskośne (zostało to odnotowane już na wcześniejszym etapie w podrozdziale 4.2.). Aby przybliżyć rozkład do normalnego i zmniejszyć wpływ skrajnie drogich nieruchomości na model, zastosowano transformację logarytmiczną ceny sprzedaży. Zastąpiono SalePrice jej logarytmem naturalnym ln(SalePrice). Dzięki temu rozkład stał się bardziej symetryczny, spełniając lepiej założenia wielu modeli regresyjnych. Transformacja ta nie wpływa na relacje porządkowe (rosnąca funkcja logarytmiczna zachowuje kolejność cen), a stabilizuje wariancję i redukuje heteroskedastyczność. W związku z tym wszelkie oceny modeli opartych o ln(SalePrice) liczone są później w skali logarytmicznej lub wyniki są transformowane z powrotem na skalę oryginalną w celu interpretacji. W dalszej części pracy porównano model z SalePrice przed i po transformacji.

Podsumowując, w ramach tego kroku doprowadzono do zgodności typy danych z ich znaczeniem, co pozwoliło uniknąć błędnych interpretacji przez algorytmy uczące się. Transformacje te są ważnym elementem preprocessingu.

1. **Oczyszczanie i standaryzacja pól tekstowych**

Zadbano o oczyszczenie danych tekstowych w zbiorze, głównie wartości kategorii zapisanych jako ciągi znaków. Celem było usunięcie ewentualnych literówek, niespójności formatów oraz zapewnienie jednolitości oznaczeń. Zmienne kategorialne zostały poddane operacji trimowania i usuwania białych znaków na początku i na końcu. M.in. w kolumnach Exterior\_1st i Exterior\_2nd kategoria "Brk Cmn" pojawiała się zamiennie z "BrkComm" – obie oznaczają Brick Common. Ujednolicono zapis, zastępując wszystkie wystąpienia "Brk Cmn" na "BrkComm"​. Dzięki temu zredukowano ryzyko traktowania przez model dwóch oznaczeń jako odrębnych kategorii, mimo że oznaczają to samo.

1. **Korekta niespójnych danych i walidacja końcowa**

W ramach czyszczenia danych zwrócono także uwagę na spójność logiczną informacji w obrębie rekordu oraz poprawność danych liczbowych. Wykryto i naprawiono następujące kwestie:

* Zidentyfikowano jeden rekord, w którym rok wybudowania garażu (Garage\_Yr\_Blt) był podany jako 2207, co wyraźnie wykracza poza realistyczny zakres (sklasyfikowano to jako błąd przy wprowadzaniu danch). Poprawiono na 2007.
* Sprawdzenie chronologii budowy i remontu: Zweryfikowano, czy dla żadnego domu rok ostatniego remontu Year\_Remod\_Add nie jest wcześniejszy niż rok budowy Year\_Built. Nie stwierdzono takiej sytuacji.
* Upewniono się, że dokonane wcześniej uzupełnienie braków nie wprowadziło sprzeczności między powiązanymi kolumnami – szczególnie w kontekście zmiennych związanych z garażem i piwnicą.

Po wykonaniu wszystkich powyższych operacji przeprowadzono końcową walidację danych. Sprawdzono ponownie kompletność danych: wyliczono sumy braków dla każdej kolumny, upewniając się, że żadna zmienna nie zawiera już wartości NaN. Wszystkie kolumny mają pełne dane – brakujące pola zostały uzupełnione bądź oznaczone jako None/0. Po zakończeniu procesu czyszczenia uzyskano zbiór danych z 82 zmiennymi oraz 2927 obserwacjami.

## 5.2. Eksploracyjna analiza danych

Wej części pracy przeprowadzono eksploracyjną analizę danych (EDA) zbioru Ames Housing. Analiza obejmuje ponowne zbadanie rozkładu zmiennej SalePrice (po dokonanym oczyszczaniu danych) i jej transformacji logarytmicznej, obliczenie podstawowych statystyk opisowych wybranych cech, oraz zbadanie zależności pomiędzy ceną a wybranymi cechami numerycznymi i kategorycznymi. Szczegółową analizę SalePrice wykonano już w rozdziale 4.2. dlatego w tym miejscu skupiono się na porównaniu SalePrice z jej odpowiednikiem po zastosowaniu logarytmu.

**Obraz zawierający diagram, tekst, Wykres, zrzut ekranu

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna. Obraz zawierający diagram, Wykres, tekst

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.**

Histogramy połączone z KDE porównujące rozkłady zmiennej SalePrice i Log\_SalePrice.

Rysunek 5.2.

Po transformacji logarytmicznej rozkład ceny staje się zbliżony do normalnego. Zniknął długi ogon widoczny wcześniej – wartości Log\_SalePrice układają się bardziej symetrycznie wokół średniej (~ 11,95, odpowiadającej ok. 155 tys. USD po odwróceniu logarytmu). Wykres pokazuje lekko spłaszczony, prawie symetryczny kształt, co potwierdza redukcję skośności (zmniejsza się z ~ 1,76 do ~ 0,12, zgodnie ze statystykami zbioru). Warto zauważyć, że wartości odstające w oryginalnej skali (bardzo drogie lub bardzo tanie domy) w skali logarytmicznej są mniej odległe od reszty danych – np. najdroższe domy (~ 13,5 w skali log, tj. ~ 730 tys. USD) mieszczą się teraz w granicach ok. 3 odchyleń standardowych od średniej, zamiast ok. 7. Znormalizowanie zmiennej celu w ten sposób jest korzystne przy dalszym modelowaniu (ograniczy wpływ skrajnych obserwacji i ułatwi spełnienie założeń modeli regresyjnych).

Kolejnym etapem jest ogólna charakterystyka danych – obliczono podstawowe statystyki opisowe dla kluczowych zmiennych numerycznych oraz rozkłady częstości dla wybranych cech kategorycznych.

* Zbiór danych składa się z 2927 obserwacji (połączony zestaw treningowy i testowy po czyszczeniu) i 81 cech (38 numerycznych i 43 kategorycznych, po zakodowaniu cech jakościowych liczbowo).
* SalePrice: średnia ~ 180 801, mediana 160 000, odchylenie standardowe ~ 79 927. Minima i maksima od 12 789 do 755 000 (USD) potwierdzają wcześniejsze spostrzeżenia o szerokim zakresie cen.
* Gr\_Liv\_Area (pow. mieszkalna nadziemna w stopach kw.): średnio ~ 1515 sqft, mediana 1442 sqft. Najmniejszy dom ma zaledwie 334 sqft powierzchni, największy – 4476 sqft. Większość domów ma powierzchnię w przedziale od ~ 800 do 2500 sqft.
* Overall\_Qual (ogólna jakość wykończenia, skala ocen 1–10): średnia ~ 6,1; mediana 6. Zatem typowy dom ma dobrą jakość (5–7). Tylko nieliczne domy uzyskały oceny skrajnie niskie (1–2) lub najwyższe 10.
* Year\_Built: mediana 1973 rok, średni rok budowy to 1971. Najstarsze domy pochodzą z lat 1870-tych, a najnowsze z 2010 roku (ostatni rok występujący w zbiorze).

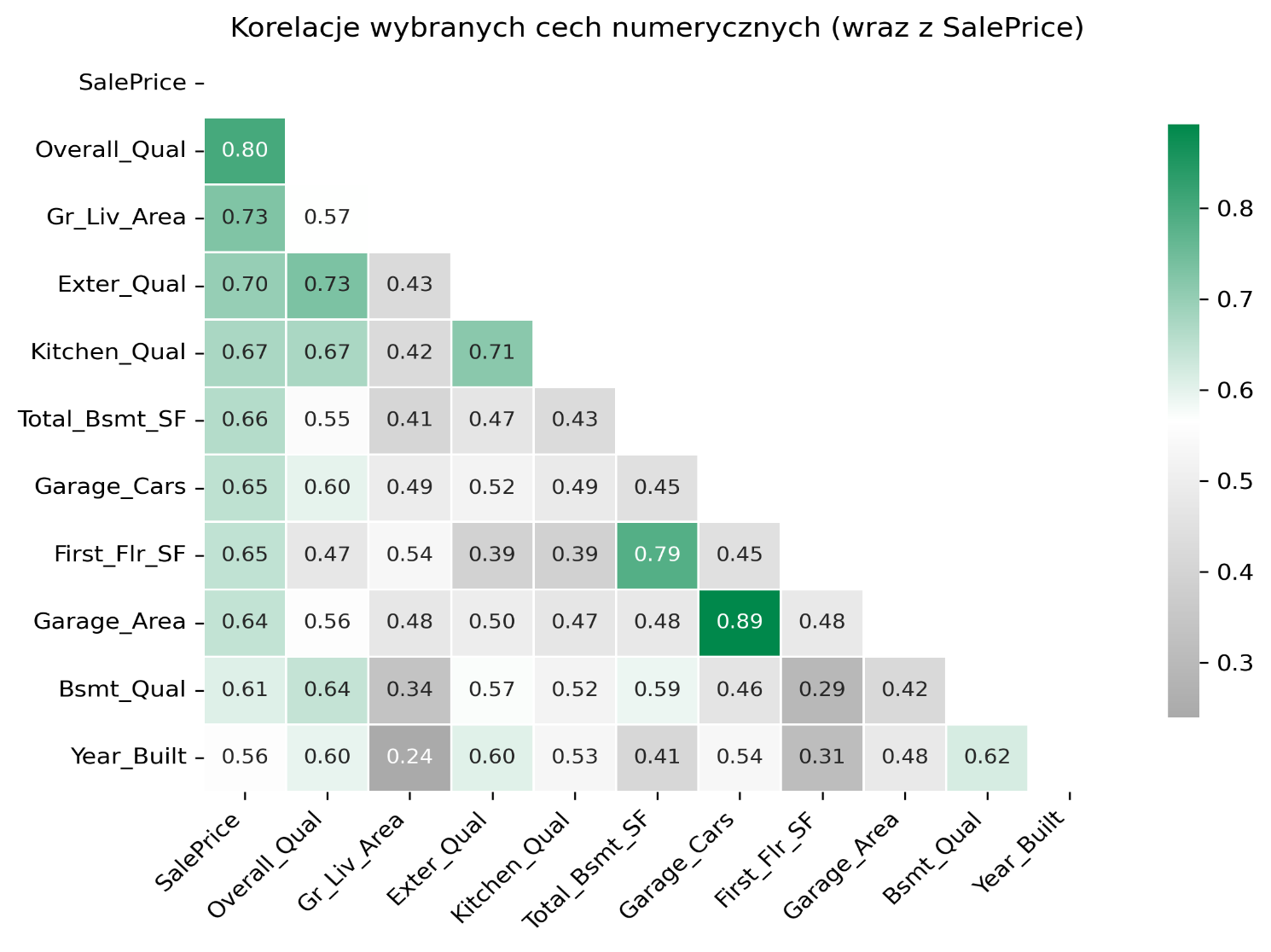
Dla cech kategorialnych stwierdzono m.in.

* Neighborhood (dzielnica): w danych występuje 28 unikalnych dzielnic Ames. Rozkład liczby domów jest nierównomierny – najwięcej nieruchomości znajduje się w dzielnicy NAmes (North Ames, 443 domy, ok. 15% danych). Kolejne pod względem liczby to CollgCr (College Creek, 267 domów) i OldTown (239). Z kolei najmniej reprezentowane są dzielnice ekskluzywne jak Greens czy GrnHill (po kilka domów).
* House Style (styl domu): dominują domy jednopiętrowe (1Story: 1480 domów, ~ 50.6%) oraz dwupiętrowe (2Story: 871 domów, ~ 29,8%). Mniej popularne są domy z półpiętrem (1,5Fin/Unf, łącznie ~ 333 domy) oraz wielopoziomowe typu split-level (SLvl, 128) czy split foyer (SFoyer, 83). Styl budynku może wpływać na cenę poprzez związek z powierzchnią.

Powyższe statystyki dostarczają ogólnego poglądu na strukturę danych. W kolejnych krokach przeanalizowano zależności pomiędzy ceną a poszczególnymi cechami – zarówno numerycznymi, jak i kategorycznymi, żeby zidentyfikować kluczowe atrybuty wpływające na wartość nieruchomości.

* + 1. **Korelacje cech numerycznych z ceną**

Najpierw zbadano, które cechy liczbowe wykazują najsilniejszy związek z ceną sprzedaży. W tym celu obliczono korelacje Pearsona pomiędzy SalePrice a wszystkimi cechami numerycznymi. Do cech najbardziej skorelowanych (|r| wysokie) należą m.in. Overall\_Qual, Gr\_Liv\_Area, Exter\_Qual i Kitchen\_Qual, Total\_Bsmt\_SF, Garage\_Cars oraz Garage\_Area Year\_Built.



Macierz korelacji (współczynniki Pearsona) między SalePrice a wybranymi cechami numerycznymi.

Rysunek 5.3.

Pierwsza kolumna (oraz wiersz) macierzy przedstawia korelacje poszczególnych cech z ceną. Potwierdza się, że Overall\_Qual jest najsilniej skorelowaną cechą (r = 0,80). Kolejne pod względem siły są: Gr\_Liv\_Area (pow. mieszkalna nad ziemią, r = 0,73), Exter\_Qual (zewnętrzna jakość, r = 0,70) oraz Kitchen\_Qual (jakość kuchni, r = 0,67). Stosunkowo wysokie korelacje z ceną mają też: powierzchnia piwnicy (Total\_Bsmt\_SF, r = 0,66), wielkość garażu (Garage\_Cars r = 0,65, Garage\_Area r = 0,64) czy powierzchnia pierwszego piętra (First\_Flr\_SF, r = 0,65). Cechy te są intuicyjnie powiązane z wartością domu – większy metraż i lepsza jakość wykończenia przekładają się na wyższe ceny.

Warto zwrócić uwagę na współzależności między samymi cechami niezależnymi. Niektóre atrybuty są ze sobą silnie skorelowane: np. liczba aut w garażu vs powierzchnia garażu (Garage\_Cars – Garage\_Area, r = 0,89) czy powierzchnia piwnicy vs powierzchnia pierwszego piętra (Total\_Bsmt\_SF – First\_Flr\_SF, r = 0,79). Dla analizy modelu oznacza to możliwość współliniowości – np. Garage\_Cars i Garage\_Area niosą podobną informację. Mimo to, ich korelacje z ceną są również wysokie, więc mogą być użyteczne po odpowiednim przetworzeniu. Warto odnotować umiarkowaną korelację Year\_Built z ceną (r = 0,56) oraz dość wysoką korelację cech jakościowych Exter\_Qual i Kitchen\_Qual z Overall\_Qual (odpowiednio 0,73 i 0,67), co wskazuje, że ogólna ocena jakości jest powiązana z ocenami szczegółowymi.

* + 1. **Wpływ cech kategorycznych na cenę**

Oprócz cech liczbowych istotną rolę w kształtowaniu ceny odgrywają cechy kategoryczne, takie jak lokalizacja czy typ budynku. Zbadano zatem, jak rozkładają się ceny w różnych kategoriach tych zmiennych. Z zestawienia na rysunku 5.4. wynika, że lokalizacja ma bardzo duży wpływ na cenę nieruchomości. Najdroższe domy znajdują się w dzielnicach NoRidge, StoneBr i NridgHt – średnie ceny przekraczają tam 300 tys. USD. NoRidge (Northridge) osiąga najwyższą średnią ~ 315 tys. USD. Odnotowano również wysokie średnie w StoneBr (~ 310 tys.) i NridgHt (~ 290 tys.). Dla porównania, w tańszych dzielnicach (nieujętych na wykresie) średnie ceny są kilkukrotnie niższe – np. najniższa średnia jest w MeadowV (~ 95,8 tys. USD), IDOTRR (~ 103,8 tys.) i BrDale (~ 105,6 tys.). Warto zauważyć, że najliczniej reprezentowana w danych dzielnica NAmes ma średnią cenę ~ 142 tys. USD – poniżej średniej globalnej. Oznacza to, że wiele domów z NAmes jest tańszych (NAmes dostarcza dużo danych, ale o stosunkowo niskich cenach).

Obraz zawierający tekst, zrzut ekranu, linia, diagram

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Średnia wartość SalePrice w piętnastu najdroższych dzielnicach miasta Ames (USD).

Rysunek 5.4.

* + 1. **Rozkłady cen dla różnych typów zabudowy**

Zbadano również rozkłady cen dla różnych typów zabudowy (liczba kondygnacji). Wykorzystano wykres pudełkowy, aby uwzględnić rozrzut cen w ramach każdej kategorii House\_Style

Obraz zawierający diagram, tekst, linia, Wykres

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Rozkłady SalePrice dla różnych typów budynków House\_Style.

Rysunek 5.5.

Z wykresu wynika, że domy jednopiętrowe (1Story) i dwupiętrowe (2Story) znacząco różnią się cenami. Domy 2Story osiągają wyższe mediany cen (ok. 185 tys. USD) niż 1Story (~ 140 tys. USD), co można tłumaczyć tym, że zazwyczaj przy tej samej powierzchni zabudowy dom piętrowy oferuje większy metraż użytkowy. Również rozstęp cen dla 2Story jest większy. Najdroższe domy piętrowe przekraczają 500 tys. USD (widoczne jako punkty odstające powyżej wąsów), podczas gdy dla domów parterowych ceny rzadko dochodzą do 400 tys. USD. Domy z półpiętrem (1.5Fin) oraz wielopoziomowe SLvl i SFoyer mają wyraźnie niższe mediany (100 – 150 tys. USD) i węższe rozkłady. Są to zazwyczaj mniejsze budynki. Kategorie 2.5Unf/Fin są nieliczne, stąd ich pudełka są mniej miarodajne – np. 2.5Fin (tylko 8 domów) pokazuje medianę ~ 230 tys. USD, ale występuje duży rozrzut. Generalnie jednak trend jest zgodny z oczekiwaniami: większe, bardziej złożone konstrukcyjnie domy osiągają wyższe ceny. Styl budynku jest powiązany z innymi cechami (np. domy 2-piętrowe mają też większą powierzchnię, często lepszy standard), co sugeruje, że zmienna House\_Style może pośrednio wskazywać na inne czynniki wpływające na cenę.

* + 1. **Rozkłady cen dla różnych wartości ogólnej oceny jakości domu**



Rysunek 5.6.

Rozkład SalePrice dla różnych wartości oceny jakości domu Overall\_Qual.

Rozkład SalePrice dla różnych wartości oceny jakości domu (Overall\_Qual). Zależność ceny od ogólnej jakości wykończenia jest wyraźnie rosnąca: domy ocenione niżej (np. Overall\_Qual = 2–4) cechują się znacznie niższymi cenami (mediany rzędu 50 –100 tys. USD), podczas gdy domy o najwyższej jakości (Overall\_Qual = 10) osiągają medianę ceny powyżej 350 tys. USD. Dla każdej kolejnej kategorii jakości obserwuje się wzrost zarówno mediany, jak i rozstępu międzykwartylowego cen. Przykładowo, mediana SalePrice rośnie z ok. 80 tys. USD (jakość 3) do ~ 210 tys. USD (jakość 7) i ~ 440 tys. USD (jakość 10). Rozstępy międzykwartylowe również się poszerzają dla wyższych jakości – świadczy to o większym zróżnicowaniu cen wśród domów o najwyższej klasy standardzie (luksusowe domy mogą mieć bardzo różne ceny w zależności od dodatkowych czynników). Brak nakładających się pudełek wskazuje, że Overall\_Qual jest silnym predyktorem ceny – nawet najlepszy dom z jakością 5 nie osiąga poziomów cen typowych dla jakości 8 czy 9.

* + 1. **Usuwanie dodatkowych outlierów wykrytych podczas EDA**

Po transformacji zmiennej celu i usunięciu skrajnych anomalii, dane nie wykazują poważnych odstających wartości mogących zaburzyć dopasowanie modelu. Wykorzystanie logarytmu dodatkowo redukuje wpływ ewentualnych obserwacji o ekstremalnych cenach (wysokich i niskich).  
Pomimo tego, że odległość Cooka wskazywała potencjalnie punkty znaczące, postanowiono nie usuwać żadnego z nich, ponieważ na rynku nieruchomości nietypowe obserwacje się zdarzają i pomimo tego, że usunięcie ich zwiększyłoby dopasowanie modelu, to mogłoby zakłamywać rzeczywistość.

## 5.3. Transformacje i inżynieria danych

Po pierwotnych transformacjach w podrozdziale 5.1. w tym rozdziale dane zostaną ostatecznie przygotowane do modelowania w kolejnym podrozdziale. Konieczne transformacje opisane poniżej są wykonywane równolegle na zbiorze z podstawową zmienną SalePrice, jak i na zbiorze z przekształconą zmienna Log\_SalePrice. Na początku ze zbioru usunięto zmienne uznane za nieprzydatne w dalszym modelowaniu. Obejmuje to kolumny będące unikatowymi opisami poszczególnych obserwacji, cechy o znikomej zmienności (niemal stałe dla wszystkich domów), a także atrybuty zdublowane informacyjnie przez inne kolumny. Poniżej przedstawiono listę usuniętych zmiennych wraz z uzasadnieniem.

* + 1. **Usunięcie nieprzydatnych zmiennych**
* Address - gdyż nie wnosi ona informacji przydatnych przy predykcji regresyjnej.
* Utilities – prawie wszystkie domy posiadają pełne podłączenie do mediów (AllPub), więc zmienna Utilities jest praktycznie stała (99% wartości identycznych). Taka cecha nie różnicuje obserwacji, zatem nie będzie użyteczna w modelu.
* Street – zdecydowana większość (> 99%) domów znajduje się przy utwardzonej drodze (Pave), podczas gdy tylko 12 na 2927 obserwacji ma drogę nieutwardzoną (Grvl). Zmienna Street również została uznana za nieważną z uwagi na skrajnie nierówny rozkład.
* Condition\_2 – w niemal wszystkich przypadkach ma wartość Norm. Inne kategorie pojawiają się łącznie tylko w kilkudziesięciu domach. Informacja ta częściowo dubluje zmienną Condition\_1, dlatego Condition\_2 pominięto w modelowaniu.
* Roof\_Matl – materiał pokrycia dachu jest w ~ 98% domów standardowy (CompShg). Nietypowe materiały dachów występują incydentalnie, więc zmienna ta nie dostarcza istotnych informacji. Z tego powodu Roof\_Matl została usunięta.
* Misc\_Feature oraz Misc\_Val – dodatkowe udogodnienia poza głównymi kategoriami (np. szopa na narzędzia, drugi garaż) występują bardzo rzadko (ponad 96% obserwacji ma wartość None w Misc\_Feature, a Misc\_Val = 0). Ze względu na tak sporadyczne występowanie cech dodatkowych zrezygnowano z wykorzystania tych zmiennych. Pozwala to uprościć model bez istotnej utraty informacji (większość domów i tak nie posiada Misc\_Feature).
* Szczegółowe powierzchnie piwnicy – kolumny BsmtFin\_SF\_1, BsmtFin\_SF\_2 i Bsmt\_Unf\_SF zawierają metraże wykończonej (pierwszej i drugiej części) oraz niewykończonej piwnicy. Ich suma dla każdego domu dokładnie równa się całkowitej powierzchni piwnicy Total\_Bsmt\_SF. Oznacza to, że zmienna Total\_Bsmt\_SF w pełni dubluje informację zawartą w tych trzech. Dla ograniczenia nadmiarowości danych usunięto szczegółowe składowe powierzchni piwnicy, pozostawiając tylko łączną powierzchnię Total\_Bsmt\_SF. Zostawienie wszystkich 3 cech prowadziłoby do powstania multikolinearności.
* Szczegółowe powierzchnie nadziemne – taka sama sytuacja dotyczy powierzchni parteru i piętra: zmienne First\_Flr\_SF, Second\_Flr\_SF oraz Low\_Qual\_Fin\_SF sumują się do całkowitej powierzchni mieszkalnej Gr\_Liv\_Area. W celu uniknięcia współliniowości, kolumny składowe zostały pominięte, zachowując jedynie sumaryczną Gr\_Liv\_Area.
* Garage\_Cars – liczba miejsc garażowych silnie koreluje z powierzchnią garażu Garage\_Area (większy garaż mieści więcej aut). Ta informacja jest więc zdublowana. W modelowaniu wykorzystano ciągłą zmienną Garage\_Area, a usunięto cechę Garage\_Cars.
* Garage\_Yr\_Blt – rok budowy garażu został uznany za zbędny, ponieważ informacja ta jest częściowo zawarta w roku budowy domu oraz stanie garażu (cechy Garage\_Qual, Garage\_Cond, Garage\_Finish). Ponadto wiele domów nie ma garażu (Garage\_Yr\_Blt = 0 po uzupełnieniu braków), co utrudnia bezpośrednie wykorzystanie tej zmiennej. Z tych względów kolumna Garage\_Yr\_Blt również została pominięta.

Po wykonaniu tych operacji liczba kolumn zmniejszyła się. Zbiór danych zawiera teraz wyłącznie potencjalnie przydatne cechy (w tym oryginalną zmienną docelową SalePrice oraz Log\_SalePrice w osobnych zbiorach). W wyniku usunięcia 15 niepotrzebnych kolumn, liczba cech spadła z pierwotnych 82 do 67 cech opisowych ( w tym zmienna celu w obu wariantach w osobnych zbiorach).

* + 1. **Transformacja zmiennych kategorycznych**

Kolejnym krokiem było przygotowanie zmiennych kategorycznych do modelowania. Zgodnie z ustaleniami z wcześniejszych podrozdziałów, cechy porządkowe zostały zakodowane liczbowo, zachowując swoją hierarchię jakości. Jedna z kategorycznych zmiennych porządkowych pozostała do zakodowania na tym etapie – Land\_Slope. Ma ona trzy kategorie jakościowe: Gentle (łagodne nachylenie), Mod (umiarkowane) i Sev (strome nachylenie). Uznano, że mniejszy stopień nachylenia terenu jest korzystniejszy. Dlatego przypisano tym kategoriom wartości numeryczne rosnąco: Sev = 0, Mod = 1, Gtl = 2. Po zakodowaniu wszystkich cech porządkowych, zmienne kategoryczne nominalne przygotowano do modelowania poprzez zakodowanie one-hot. Do zakodowania one-hot wybrano następujące cechy nominalne:

* MS\_SubClass
* MS\_Zoning
* Alley
* Land\_Contour
* Lot\_Config
* Neighborhood
* Condition\_1
* Bldg\_Type
* House\_Style
* Roof\_Style
* Mas\_Vnr\_Type
* Foundation
* Heating
* Electrical
* Garage\_Type
* Sale\_Type
* Sale\_Condition

Do funkcji pd.get\_dummies() przekazano listę powyższych kolumn celem utworzenia zmiennych dummy. Zastosowano opcję drop\_first = True, aby uniknąć pułapki zmiennych dummy. Usunięto jedną kolumnę dummy w każdej grupie kategorii, traktując ją jako kategorię bazową. Dzięki temu nie wprowadzana jest idealna współliniowość. Na przykład dla zmiennej MS\_Zoning (7 kategorii) utworzono 6 kolumn binarnych: MS\_Zoning\_FV, MS\_Zoning\_RH, MS\_Zoning\_RL, MS\_Zoning\_RM, MS\_Zoning\_C (all) oraz MS\_Zoning\_I (all). Brak osobnej kolumny dla strefy bazowej oznacza, że obserwacje z tą kategorią mają wszystkie powyższe kolumny ustawione na 0.

Po wykonaniu przekształceń zbiór danych znacznie się rozszerzył. Liczba kolumn zwiększyła się z 67 cech (po usunięciu zmiennych nieprzydatnych) do ponad 190 cech numerycznych. Dla każdej z 19 cech nominalnych powstało n - 1 nowych kolumn (n to liczba unikalnych kategorii danej cechy). Na przykład zmienna Neighborhood posiada 28 dzielnic, co po kodowaniu dało 27 kolumn dummy. Podobnie Exterior\_1st (16 kategorii) zakodowano w 15 kolumn, Exterior\_2nd w 16 kolumn, Sale\_Type (10 kategorii) w 9 kolumn. Łącznie dodano 135 kolumn binarnych, a dotychczasowe kolumny kategoryczne zostały usunięte.

W efekcie wszystkie zmienne wejściowe w zbiorze danych mają postać liczbową, co umożliwia zastosowanie algorytmów regresji. Zmienne binarne pozostały w tej formie. Nie zostały w żaden sposób przekształcone ani znormalizowane.

* + 1. **Inżynieria nowych cech**

Po uporządkowaniu danych, przeprowadzono wzbogacenie zbioru o nowe cechy stworzone na podstawie istniejących. Celem inżynierii cech jest uchwycenie istotnych informacji, które mogą poprawić skuteczność modelu. Poniżej opisano te zmienne wraz z uzasadnieniem:

* Total\_SF – łączna powierzchnia domu (sq ft). Jest to suma całkowitej powierzchni mieszkalnej nad ziemią (Gr\_Liv\_Area) oraz powierzchni piwnicy (Total\_Bsmt\_SF). Cecha ta reprezentuje ogólną wielkość domu, uwzględniając zarówno powierzchnię nadziemną, jak i podziemną. Wprowadzenie Total\_SF pozwala modelowi uchwycić wpływ łącznej wielkości domu jako całości. Warto zauważyć, że wcześniejsze szczegółowe zmienne składające się na ten metraż zostały już zredukowane.
* House\_Age – wiek domu w latach w momencie sprzedaży. Obliczona jako różnica Yr\_Sold - Year\_Built. Im starszy dom, tym potencjalnie niższa cena. Ta zmienna reprezentuje upływ czasu od wybudowania domu do momentu sprzedaży.
* Remod\_Age – wiek od ostatniego remontu (w latach) w momencie sprzedaży, wyznaczony jako Yr\_Sold - Year\_Remod\_Add. Dla domu, który nigdy nie był remontowany, Year\_Remod\_Add jest równy Year\_Built, więc Remod\_Age będzie równy House\_Age. Natomiast dla domu po remontach Remod\_Age będzie mniejszy niż całkowity wiek domu. Ta cecha odzwierciedla, jak dawno dom przechodził modernizację.
* Remodeled – wskaźnik remontu (zmienna binarna: 1 – jeśli dom był kiedykolwiek remontowany, 0 – jeśli nigdy nie był remontowany.
* Overall\_Qual\_Cond – łączna ocena jakości i stanu domu. Jest to suma dwóch istniejących zmiennych: Overall\_Qual oraz Overall\_Cond. Ponieważ jakość i stan domu razem wpływają na jego wartość, połączenie ich może stworzyć silniejszy predyktor. Przykładowo, dom o jakości 8 i stanie 7 uzyska Overall\_Qual\_Cond = 15.
* Living\_Lot\_Ratio – stosunek powierzchni mieszkalnej do powierzchni działki. Obliczono ją jako Gr\_Liv\_Area / Lot\_Area. Zmienna ta informuje, jak duża część działki jest zajęta przez powierzchnię mieszkalną domu.
* Total\_Bathrooms – łączna liczba łazienek. Powstała przez zsumowanie wszystkich łazienek z uwzględnieniem połówek: Total\_Bathrooms = Full\_Bath + 0.5 \* Half\_Bath + Bsmt\_Full\_Bath + 0.5 \* Bsmt\_Half\_Bath. Ten sposób modelowania upraszcza analizę.
* Total\_Porch\_SF – łączna powierzchnia ganku. Wyliczona jako suma: Wood\_Deck\_SF + Open\_Porch\_SF + Enclosed\_Porch + Three\_Ssn\_Porch + Screen\_Porch.

Po dodaniu tych zmiennych dokonano jeszcze usunięcia cech, które zostały przez nie zastąpione lub zdublowane:

* Year\_Built oraz Year\_Remod\_Add
* Overall\_Cond

Pozostałe cechy wykorzystane do stworzenia nowych pozostawiono, ponieważ mogą nieść dodatkową informację. Po tych operacjach zbiór danych zawiera 200 cech (w tym zmienne celu).

* + 1. **Standaryzacja cech numerycznych**

W zbiorze danych Ames zmienne liczbowe występują w różnych skalach i jednostkach – np. powierzchnie podawane są w stopach kwadratowych, lata są w dziesiątkach lub setkach, wskaźniki jakości w skali 1–5, odległość od centrum miasta Dist\_from\_Center\_ft w setkach stóp. Takie zróżnicowanie skali może utrudniać trenowanie modeli, ponieważ cechy o większych liczbowo wartościach mogłyby zdominować te o mniejszych. Dlatego zdecydowano się ujednolicić skale zmiennych numerycznych poprzez standaryzację. Przeprowadzając standaryzację, wykluczono zmienne binarne. Cechy zakodowane one-hot oraz flagi Remodeled czy Central\_Air już są na jednolitej skali 0–1 i nie ma potrzeby ich przekształcania. Ponadto zmienne binarne po standaryzacji straciłyby interpretację. Standaryzacji nie podlegała także zmienna docelowa. W praktyce zastosowano klasę StandardScaler z biblioteki scikit-learn. Najpierw wybrano listę kolumn do standaryzacji – były to wszystkie cechy numeryczne nie będące zmiennymi binarnymi. Następnie dopasowano skalę do danych i przetransformowano kolumny.

W rezultacie przeprowadzonych transformacji uzyskano dwa finalne zbiory danych gotowe do trenowania modeli regresyjnych:

* Ames\_Prepared.csv – zawierający wszystkie przygotowane cechy oraz oryginalną zmienną docelową SalePrice. Zbiór ten ma 2927 wierszy i 200 kolumn (199 cech objaśniających + 1 kolumna z ceną SalePrice).
* Ames\_Prepared\_Log.csv – zawierający identyczny zestaw cech objaśniających, jednak z przetransformowaną logarytmicznie zmienną docelową Log\_SalePrice. Ten zbiór również ma w sumie 200 kolumn (199 cech + Log\_SalePrice)

Należy podkreślić, że na tym etapie nie dokonano jeszcze podziału na zbiór treningowy i testowy. Podział ten zostanie wykonany w kolejnym podrozdziale. Dzięki temu wszystkie powyższe transformacje (szczególnie obliczenie średnich i odchyleń do standaryzacji) zostały przeprowadzone z wykorzystaniem informacji ze wszystkich dostępnych danych.

## 5.4. Konstruowanie modeli regresji

Celem tego rozdziału jest zbudowanie i przetestowanie różnych modeli regresyjnych do predykcji cen nieruchomości. Wykorzystane zostaną dwa warianty danych: z oryginalnymi cenami sprzedaży (SalePrice) oraz z cenami przetransformowanymi logarytmem naturalnym (Log\_SalePrice). Przygotowany zostanie klasyczny model regresji w oparciu o SalePrice, model na podstawie logarytmicznego przekształcenia Log\_SalePrice oraz modele regresji Ridge i Lasso (z dostrojeniem hiperparametrów).

* + 1. **Metodyka trenowania modeli**

Aby móc ocenić uogólnioną skuteczność modeli, podzielono każdy ze zbiorów na podzbiór treningowy i testowy w proporcji 80/20. Użyto funkcji train\_test\_split z biblioteki scikit-learn, z parametreem random\_state. Po podziale, zbiór treningowy zawiera 2341 obserwacji, a zbiór testowy 586 obserwacji (dla każdego z dwóch przypadków: z oryginalną ceną i z ceną logarytmiczną). Modele zostały wytrenowane wyłącznie na zbiorach treningowych. Zbudowanych zostało 6 modeli:

* Klasyczny model liniowy OLS w wariancie SalePrice i Log\_SalePrice
* Regresja liniowa Ridge w wariancie SalePrice i Log\_SalePrice
* Regresja liniowa Lasso w wariancie SalePrice i LogSalePrice.

W modelach Ridge i Lasso zastosowano regularyzację L2 i L1 poprzez dobranie parametru alfa. Wartość parametru alfa została dostrojona przy pomocy kroswalidacji z wykorzystaniem klasy GridSearchCV z 5-krotną walidacją krzyżową. Przeszukiwana siatka wartości alfa to np.logspace(-4, 2, 20). Kryterium oceny jakości został negatywny błąd średniokwadratowy – minimalne MSE, a najlepsza wartość alfa została wykorzystana w do trenowania modeli Ridge i Lasso. Do ogólnej oceny dopasowania wykorzystano współczynnik determinacji R2 oraz RMSE.

* + 1. **Regresja liniowa OLS (SalePrice)**

Regresja OLS została wytrenowana na zbiorze treningowym z oryginalną ceną sprzedaży jako zmienną objaśnianą. Wykorzystano implementację LinearRegression z biblioteki scikit-learn nie stosując żadnej regularyzacji. Model OLS, dysponując 199 wystandaryzowanymi cechami na 2341 próbkach treningowych, dopasował się niemal perfekcyjnie do danych treningowych. Współczynnik determinacji R2 na zbiorze treningowym wyniósł ~ 0,925. Oznacza to, że model wyjaśnił około 92,5% wariancji ceny sprzedaży w danych uczących. Jednak brak regularyzacji w warunkach potencjalnie współliniowych cech spowodował znaczne przeuczenie modelu. Na zbiorze testowym uzyskano skrajnie niski wynik R2 ~ -6,35E16 , czyli wartość silnie ujemną. Zgodnie z tym RMSE na danych testowych osiągnęło ogromne – nieakceptowalne wartości. Błąd OLS jest miliony razy większy od typowych wartości ceny. Po głębszej analizie wartości reszt okazało się, że za nierealistyczne wartości modelu odpowiadają trzy rekordy danych:

**1.**

Indeks: 2337

Rzeczywista: 319 000,00

Przewidywana: -78 856 014 622,41

Reszta: 78 856 333 622,41

**2.**

Indeks: 2727

Rzeczywista: 148 400,00

Przewidywana: -548 052 867 120 193,06

Reszta: 548 052 867 268 593,06

**3.**

Indeks: 727

Rzeczywista: 44 000,00

Przewidywana: -12 103,19

Reszta: 56 103,19

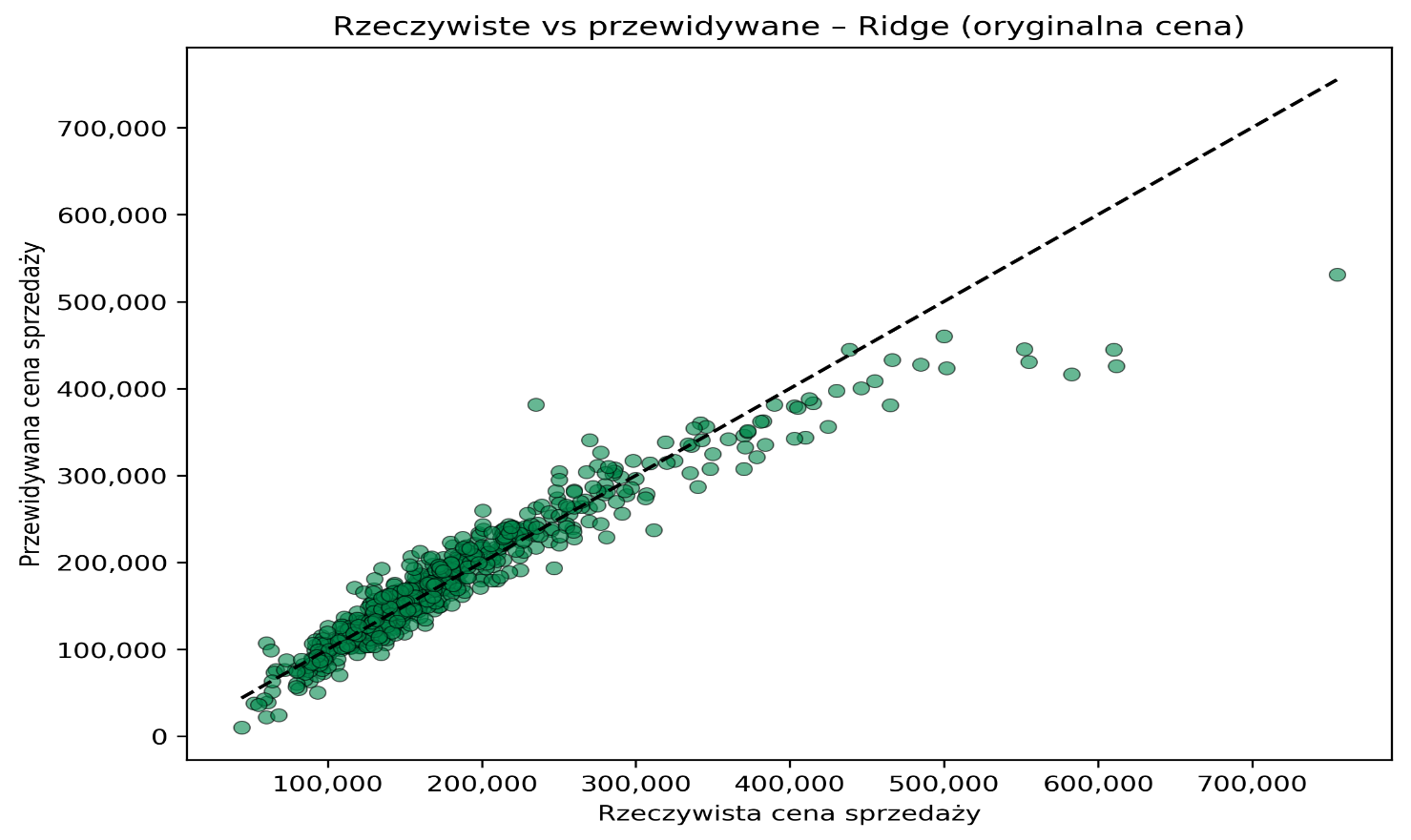
Dwa z trzech ukazanych punktów miara odległości Cook’a wskazała jako punkty wpływowe. Po wykluczeniu tych punktów z analizy model zachowywał się lepiej, natomiast uznano, że model potrafił generować zupełnie nierealistyczne predykcje. Jest to w dużej mierze spowodowane brakiem regularyzacji przy tak wielu zmiennych objaśniających.

* + 1. **Regresja liniowa OLS (Log\_SalePrice)**

Drugi model to również regresja liniowa OLS, tym razem trenowana na danych z logarytmowaną zmienną celu Log\_SalePrice. Procedura trenowania była analogiczna. Zastosowanie logarytmicznej transformacji zmiennej celu często pomaga ustabilizować wariancję i uczynić zależność bardziej liniową. Jednak w tym przypadku nie zapobiegło to przeuczeniu modelu. Na zbiorze treningowym model osiągnął R2 ~ 0,921, co znów wskazywałoby na dobre dopasowanie. Podobnie jak w poprzednim modelu jednak wyniki na zbiorze testowym są fatalne głównie za sprawą tych samych punktów. Ponownie osiągana jest ogromna ujemna wartość R2 oraz zupełnie nieprawdopodobne wartość błędu RMSE na poziomie reszt wskazujących na predykcje logarytmu ceny odbiegające dziesiątkami milionów razy od faktycznych wartości. Wniosek jest podobny do wniosku z modelu OLS dla SalePrice – brak regularyzacji uniemożliwia stosowanie modelu. Logarytmizacja w tym przypadku nie wpłynęła pozytywnie na predykcję i nie wyeliminowała błędów braku regularyzacji.

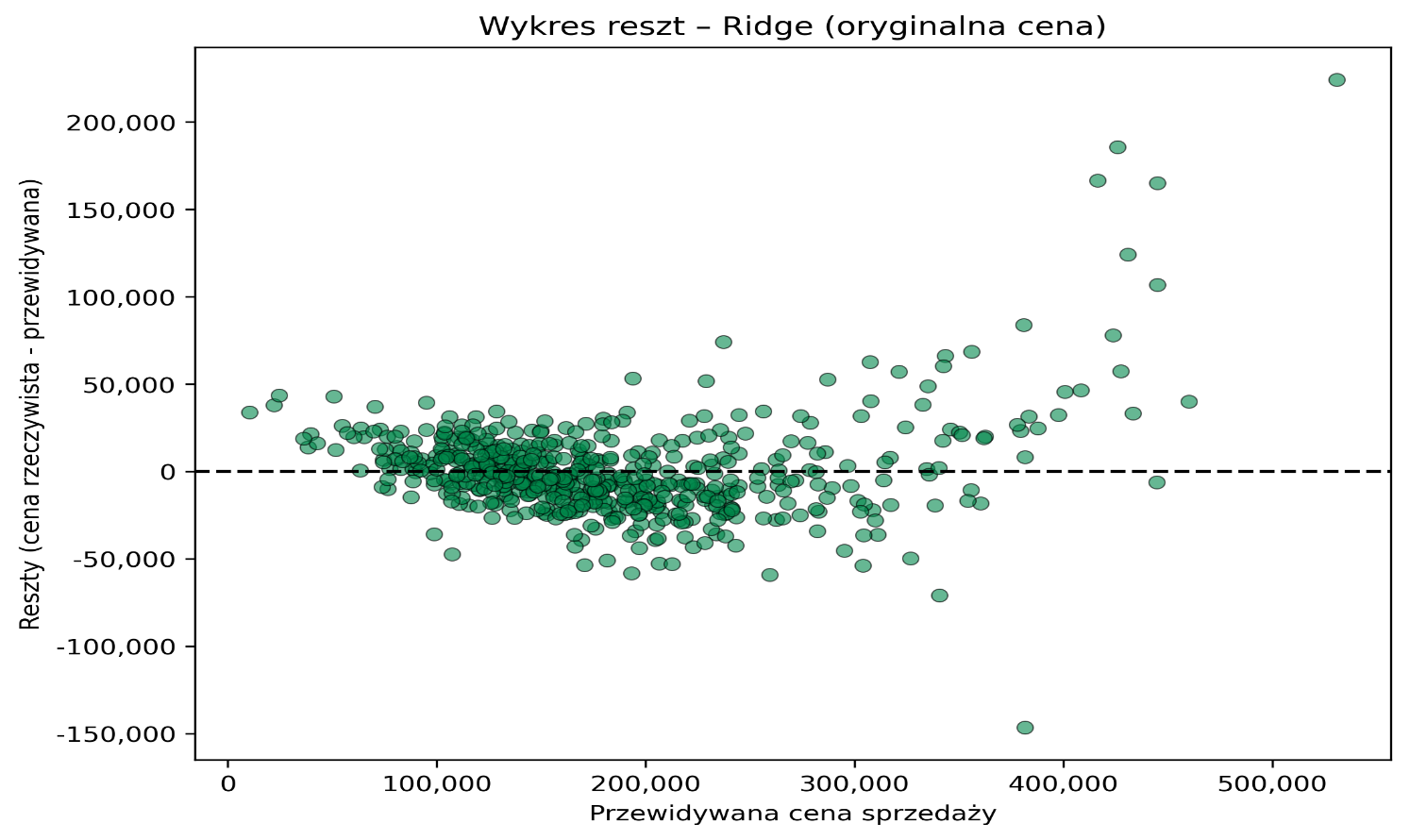
* + 1. **Regresja Ridge (SalePrice)**

Kolejny model to regresja Ridge z karą L2, trenowana na oryginalnych cenach sprzedaży. Regularyzacja grzbietowa dodaje do funkcji kosztu karę za duże wartości wag, co przeciwdziała współliniowości i przeuczeniu kosztem niewielkiego obciążenia. W przypadku wykorzystywanego zbioru wpłynęło to bardzo pozytywnie na wyniki modeli. Proces walidacji wskazał jako optymalny parametr regularyzacji alfa = 11,28. Oznacza to, że do funkcji kosztu dodanie dość silnej kary za wielkość współczynników (11,28 \* suma kwadratów wag) przyniosło najlepszy efekt. Model Ridge z alfa = 11,28; wytrenowany na pełnym zbiorze treningowym, osiągnął na zbiorze testowym bardzo wysoki współczynnik determinacji R2 = 0,906. R2 na treningu wyniósł 0,921 - model dobrze generalizuje. Błąd RMSE na zbiorze testowym wyniósł ok. 27 500 $, co stanowi znacznie bardziej realistyczny poziom błędu predykcji domu. Zastosowanie regularyzacji L2 wyraźnie ustabilizowało model. Współliniowe cechy zostały obarczone karą przez algorytm, redukując wariancję i zapobiegając niekontrolowanemu wzrostowi wag. Chociaż model Ridge dalej wykorzystuje wszystkie cechy, to wpływ mniej istotnych zmiennych jest stłumiony.

****

Wartości przewidywane modelu Ridge względem rzeczywistych wartości SalePrice.

Rysunek 5.6.

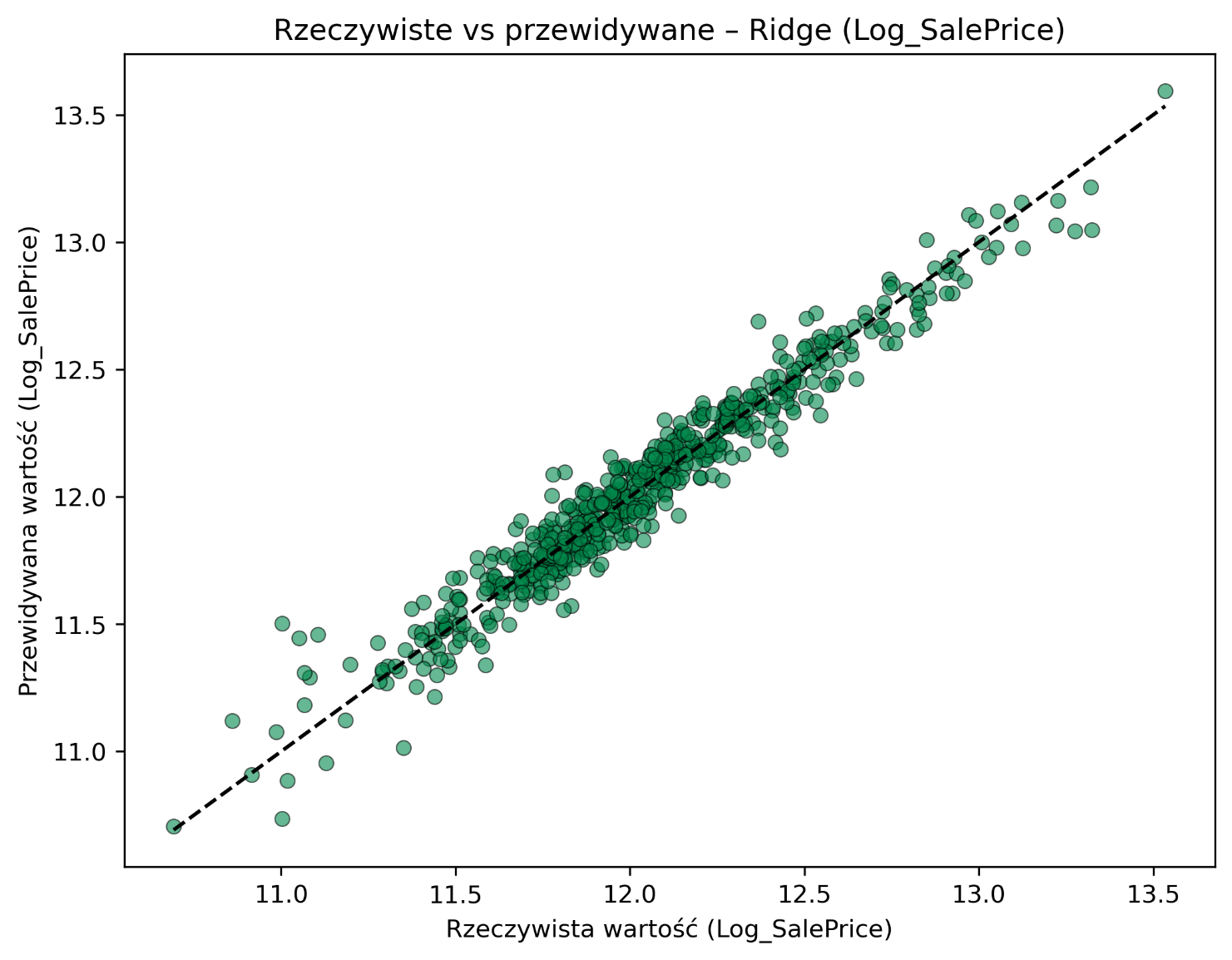
****

Wykres reszt dla modelu Ridge budowanego w oparciu o zmienną celu SalePrice.

Rysunek 5.7.

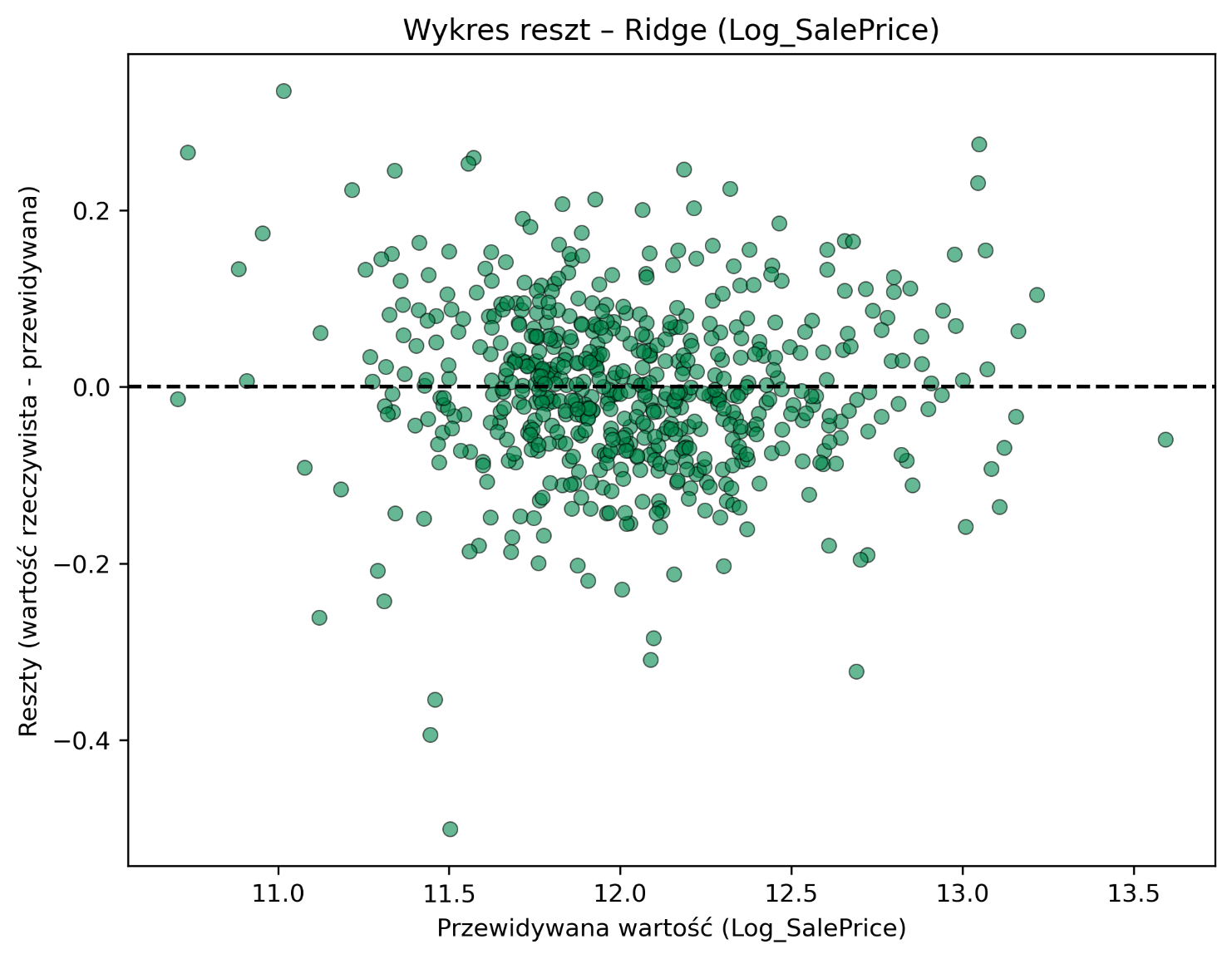
* + 1. **Regresja Ridge (Log\_SalePrice)**

Następnie wytrenowano model Ridge na zbiorze z logarytmowaną ceną sprzedaży. Trenowanie i hiperparametryzacja przebiegają tak samo jak w modelu poprzednim. GridSearch wskazał alfa = 11,28. Model Ridge z tą wartością regularyzacji, dopasowany do danych z logarytmowaną ceną, osiągnął najwyższą jakość predykcji spośród wszystkich rozważanych modeli. Współczynnik determinacji na zbiorze testowym wyniósł R2 = 0,948 wyjaśniając 94,8% wariancji Log\_SalePrice. Błąd RMSE wyniósł około 0,097 w skali logarytmicznej. Jeśli , to błąd 0,097 w tej skali przekłada się na pomyłkę rzędu ~ 1,10 razy w wartości ceny, czyli średnio model myli się o ~10% w przewidywaniu ceny domu. Jest to bardzo dobry wynik. Dla porównania, R2 na zbiorze treningowym wyniosło 0,927. Tak wysoki współczynnik determinacji i niski błąd RMSE oznacza, że logarytmiczna transformacja zmiennej celu w połączeniu z regularyzacją L2 dobrze oddaje strukturę danych.

****

Rysunek 5.8.

Wartości przewidywane modelu Ridge względem rzeczywistych wartości Log\_SalePrice.

****

Wykres reszt dla modelu Ridge budowanego w oparciu o zmienną celu Log\_SalePrice.

Rysunek 5.9.

Wykres wartości przewidywanej a rzeczywistej Log\_SalePrice na rysunku 5.8. pokazuje punkty niemal idealnie skupione wokół linii prostej. Wykres reszt na rysunku 5.9. również potwierdza dopasowanie. Reszty są równomiernie rozłożone i dodatkowo osiągają niskie wartości.

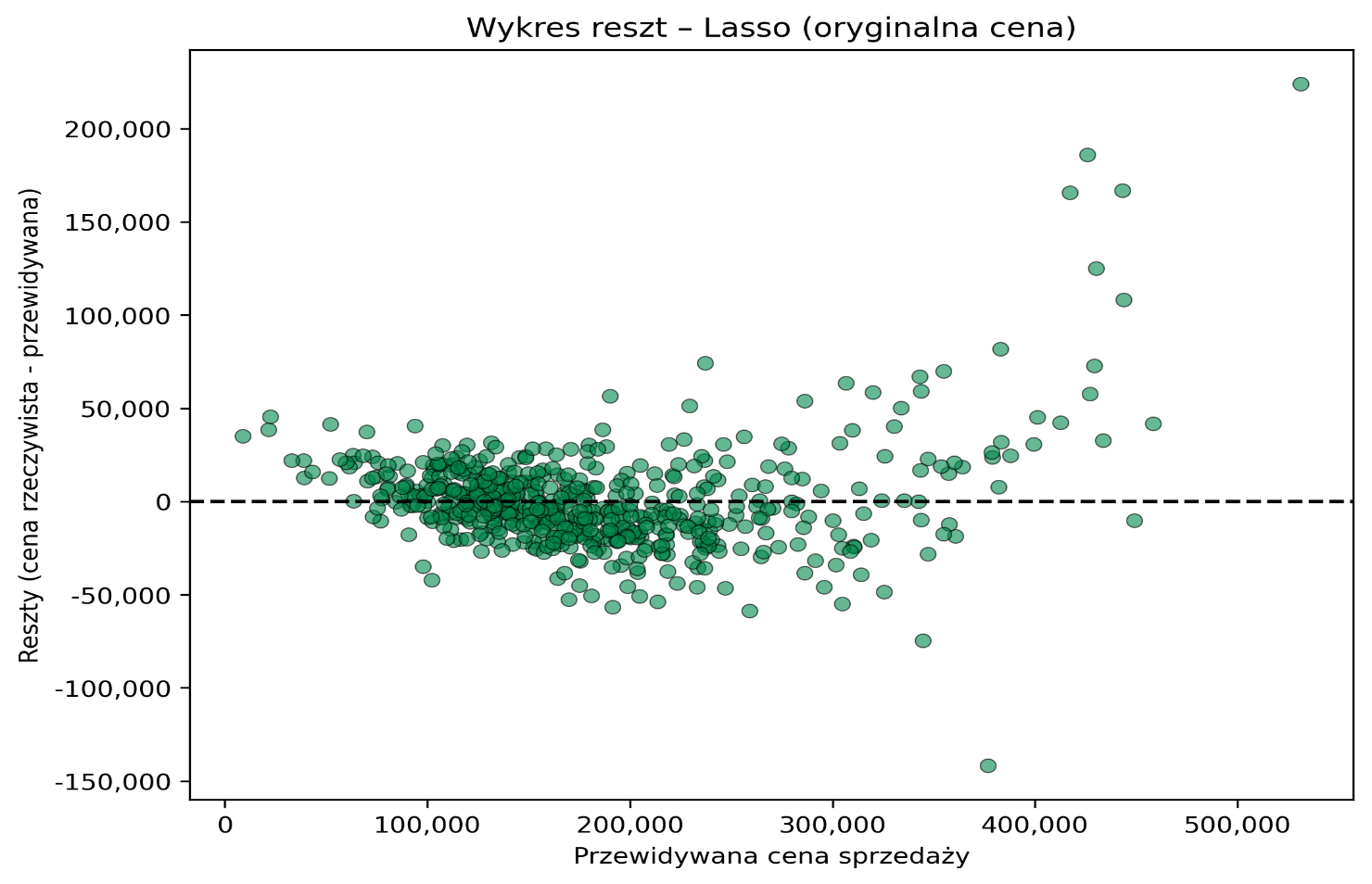
* + 1. **Regresja Lasso (SalePrice)**

Kolejnym modelem jest regresja Lasso wykorzystująca regularyzację L1 na oryginalnym zbirze SalePrice. Kara Lasso prowadzi do wyzerowania współczynników cech o mniejszej ważności - wbudowana selekcja zmiennych. Hiperparametryzacja alfa przebiega znów z wykorzystaniem algorytmu grid search i wykazała optymalną wartość alfa = 23,36. Model Lasso wyzerował 59 cech spośród 199, którymi dysponował. Otrzymany model jest zatem oparty o pozostające 140 cech znaczących. Model osiągnął współczynnik determinacji na poziomie R2 ~ 0,906 i RMSE = 27514 $ na zbiorze testowym. Model zachowuje się bardzo podobnie do wersji Ridge dla klasycznej zmiennej SalePrice. Wskazuje to, że oba rodzaje regularyzacji doprowadziły do zbliżonego poziomu dopasowania. Lasso ułatwia jednak interpretację modeli – wyzerował ~ 30% cech.

****

Wartości przewidywane modelu Lasso względem rzeczywistych wartości SalePrice.

Rysunek 5.10.

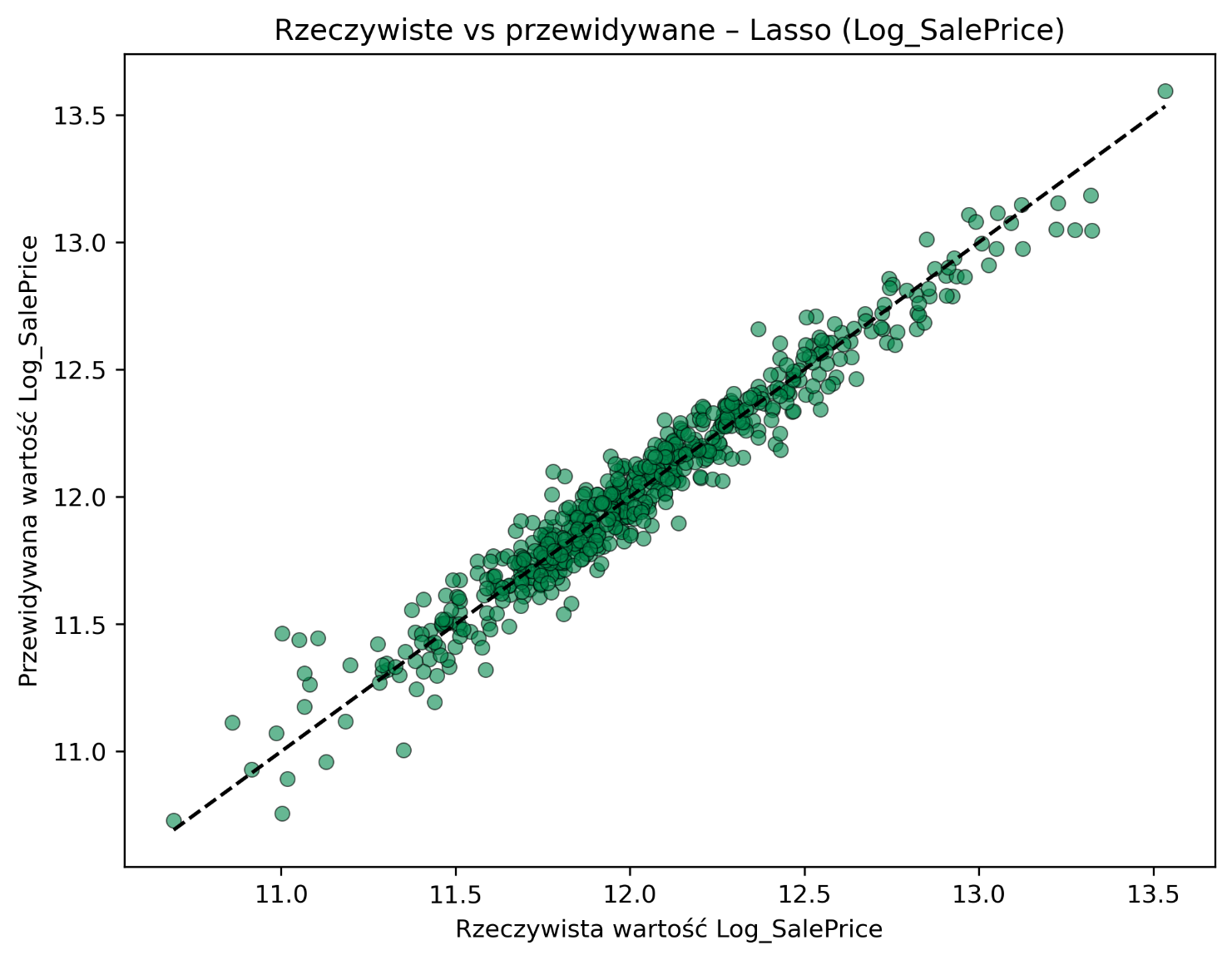
****

Wykres reszt dla modelu Lasso budowanego w oparciu o zmienną celu SalePrice.

Rysunek 5.11.

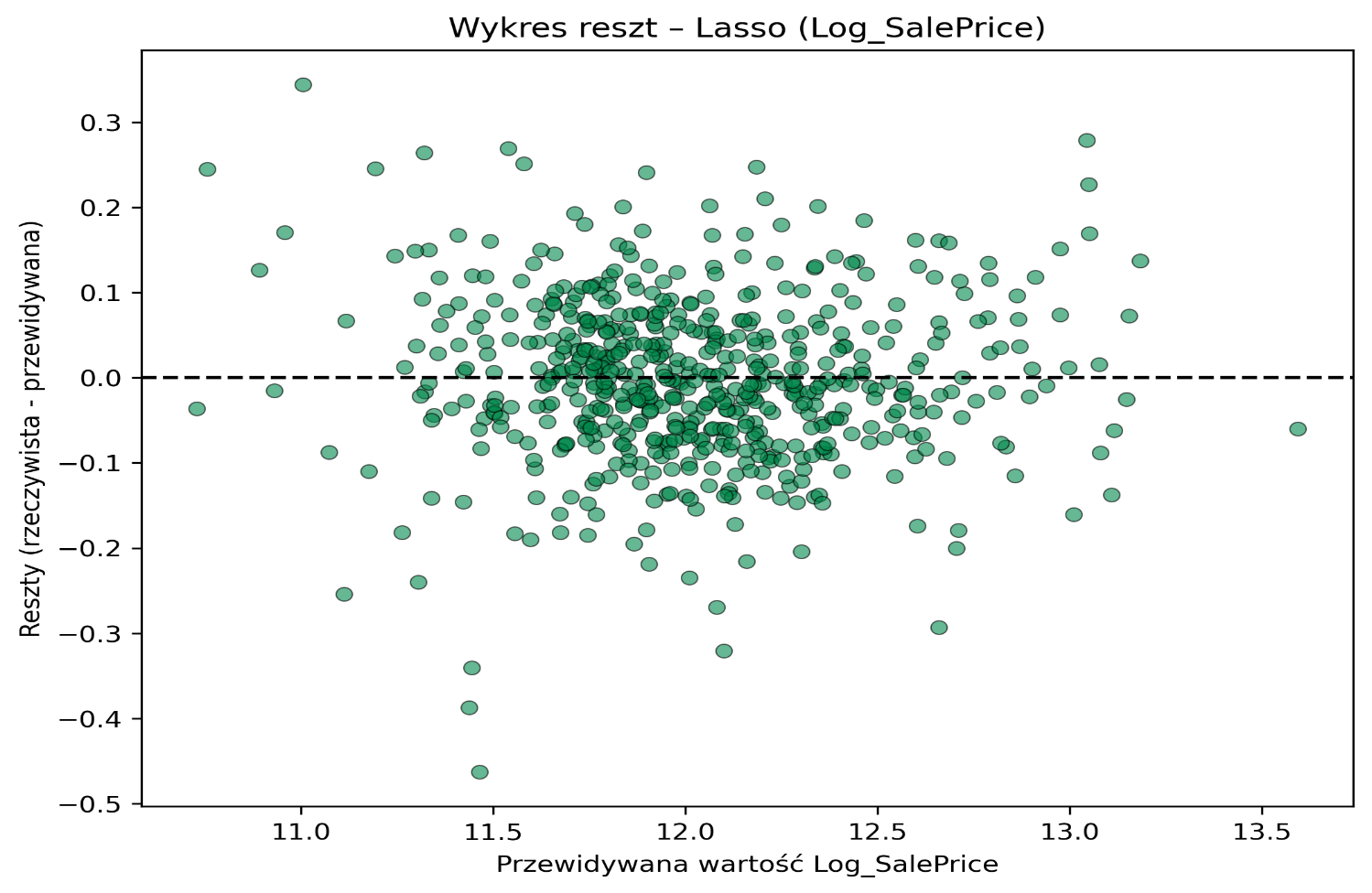
* + 1. **Regresja Lasso (Log\_SalePrice)**

Dla modelu Lasso z logarytmowaną zmienną celu najlepsza okazała się wartość alfa ~ 0,00020. Okazało się, że w przypadku logarytmicznego przekształcenia SalePrice już bardzo niewielka regularyzacja pozwala na uniknięcie przeuczenia. Model Lasso wyzerował znaczącą liczbę współczynników regresji. Ostateczny model wykorzystał jedynie 120 z 199 dostępnych cech, a pozostałe 79 wyzerował. Oznacza to, że logarytmiczna transformacja zmiennej celu jeszcze wyraźniej zaznaczyła istotność konkretnego podzbioru predyktorów niż w modelu Lasso dla SalePrice. Na zbiorze testowym osiągnięto R2 ~ 0,947 i RMSE ~ 0,0971 w skali logarytmicznej. Model Lasso dla Log\_SalePrice okazał się mieć zbliżone wyniki do modelu Ridge dla logartmizowanej zmiennej celu. Z całą pewnością warto zauważyć, że taki dobry wynik jest jeszcze wsparty redukcją liczby zmiennych, co w dużym stopniu wpłynie korzystnie na interpretację.



Wartości przewidywane modelu Lasso względem rzeczywistych wartości Log\_SalePrice.

Rysunek 5.12.



Wykres reszt dla modelu Lasso budowanego w oparciu o zmienną celu Log\_SalePrice.

Rysunek 5.13.

Modele OLS okazały się nierzetelne w przypadku kilku punktów danych, których diametralnie złe predykcje wpłynęły na ujemny R2 oraz zupełnie nierealistyczne wartości błędów RMSE w charakterze globalnym, ponieważ kilka obserwacji zniekształcało obraz całego modelu. Uznano, że usuwanie tych obserwacji nie jest dobrym rozwiązaniem, ponieważ wcześniejsze analizy nie wskazywały jednoznacznie, że nieruchomości te są obserwacjami odstającymi. Zamiast tego skorzystano z modeli wyposażonych w regularyzację. Poniżej znajduje się tabela podsumowująca podrozdział.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Skala zmiennej celu** | **Liczba cech** | **Cechy wykorzystane** | **Cechy wyzerowane** | **R² (test)** | **RMSE (test)** |
| OLS | SalePrice | 199 | 199 | 0 | silnie ujemny | silnie dodatni |
| Ridge | SalePrice | 199 | 199 | 0 | 0,9062 | 27 499 $ |
| Lasso | SalePrice | 199 | 140 | 59 | 0,9061 | 27 514 $ |
| OLS | Log\_SalePrice | 199 | 199 | 0 | silnie ujemny | silnie dodatni |
| Ridge | Log\_SalePrice | 199 | 199 | 0 | 0,9476 | 0,0971 |
| Lasso | Log\_SalePrice | 199 | 120 | 79 | 0,9475 | 0,0971 |

Podsumowanie zbudowanych modeli regresji liniowej OLS, Ridge i Lasso.

Rysunek 5.13.

6. Podsumowanie wyników i znaczenie praktyczne

Niniejszy rozdział zawiera podsumowanie wyników stworzonych w podrozdziale 5.5, modeli, porównanie modeli ze sobą oraz wytypowanie najlepszego modelu na zadanym zbiorze. Analizę porównawczą oparto o szereg wskaźników statystycznych, spośród których najważniejszymi są R2, RMSE, MAE i MAPE (%). Oprócz tych wskaźników wykorzystano jeszcze kryterium Akaike i kryterium Schwarza. Z oczywistych przyczyn, które opisane zostały w podrozdziale 5.5. porównaniu zostaną poddane tylko modele Ridge i Lasso w obu wersjach zmiennej celu. Porównanie wykonywane jest na bazie zbioru testowego.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Skala zmiennej celu** | **R**2 | **Adjusted R**2 | **RMSE** | **MAE** | **MAPE (%)** | **RSE** |
| Ridge | Log\_SalePrice | 0,9476 | 0,9205 | 0,0971 | 0,0731 | 0,6105 | 0,1196 |
| Lasso | Log\_SalePrice | 0,9475 | 0,9339 | 0,0971 | 0,074 | 0,6179 | 0,109 |
| Ridge | SalePrice | 0,9062 | 0,8578 | 27 499,84 | 17 745,67 | 10,0848 | 33 883,27 |
| Lasso | SalePrice | 0,9061 | 0,8765 | 27 514,86 | 17 742,23 | 10,0523 | 31 574,47 |

Porównanie współczynników oceny jakości model regresji ( z wykluczeniem OLS).

Rysunek 6.1.

* + 1. **Analiza porównawcza**

Najważniejszym kryterium porównawczym jest współczynnik determinacji. Wszystkie modele osiągają bardzo dobre rezultaty. Każdy model z regularyzacją L1 lub L2 osiąga wyniki ponad 90%. Dodatkowo, należy zaznaczyć, że chociaż wartości są wysokie, to modele wcale nie są przeuczone, gdyż podobne (lub nawet niższe wartości modele osiągały dla zbiorów treningowych). Najwyższe wartości R2 osiągnęły oba modele, gdzie zmienna celu jest wyrażona logarytmem – kolejno Ridge 0,9476 oraz Lasso 9475. Jak widać wartości te są niemal tożsame. Pokazuje to, że chociaż regularyzacja L1 i L2 działają inaczej, to najważniejsze cechy w obu modelach są tak samo istotne, a to przekłada się na zbliżone wartości R2, RMSE, MAE i pozostałych współczynników. Chcąc uzyskać tożsamą skalę porównania między modelami wykorzystano mechanizm backtransfer dla modeli z Log\_SalePrice. Tabela na rysunku 5.15 jednoznacznie potwierdza, że modele z przekształceniem są dokładniejsze. RMSE (USD) jest mniejsze o ok. 8000 $, natomiast MAE o ponad 4000 $. MAPE (%) też jednoznacznie potwierdza przewagę modeli budowanych w oparciu o logarytmiczną zmienną celu.  
Spośród tych dwóch modeli lepsze wyniki osiąga regularyzacja metodą Ridge. Po backtransformacji RMSE modelu Ridge nie przekłada się jednak na dużo lepszy wynik RMSE. Sprawdzono zatem, czy różnice w błędach na korzyść Ridge są istotne statystycznie przy pomocy testu Wilcoxona. Otrzymane wyniki (p-value = 0,0706) wskazuje, że przy standardowym poziomie istotności alfa = 0,05, mediana różnic bezwzględnych błędów między Ridge a Lasso nie jest istotna statystycznie.

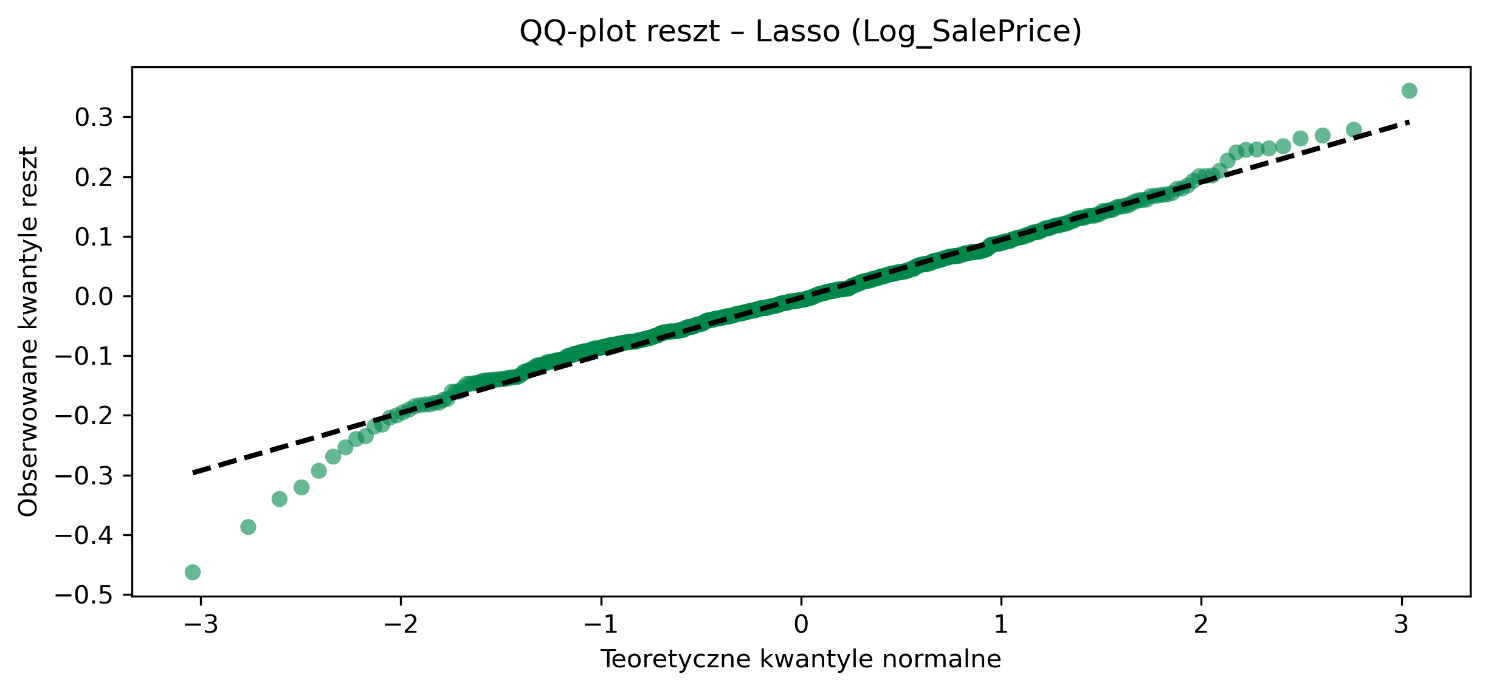
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Skala zmiennej celu** | **RMSE (USD)** | **MAE (USD)** | **MAPE (%)** | **RSE (USD)** |
| Ridge | SalePrice (backtransfer) | 19 655,91 | 13 281,01 | 7,375 | 24 218,56 |
| Lasso | SalePrice (backtransfer) | 19 860,71 | 13 461,64 | 7,4521 | 22 295,50 |
| Ridge | SalePrice | 27 499,84 | 17 745,67 | 10,0848 | 33 883,27 |
| Lasso | SalePrice | 27 514,86 | 17 742,23 | 10,0523 | 31 574,47 |

Porównanie współczynników oceny jakości model regresji (backtransfer).

Rysunek 6.2.

Biorąc pod uwagę to, że głównym celem było nie tylko osiągnięcie jak najdokładniejszego wyniku, ale również uzyskanie modelu łatwiejszego w interpretacji, wybrano model regresji Lasso jako ten najoptymalniejszy. Niewielkie różnice na jego niekorzyść przy R2, RMSE, MAE i MAPE (%) ustępują prostocie modelu. Model Lasso dla Log\_SalePrice wyzerował 79 cech, co znacznie uprościło interpretację modelu. Dla potwierdzenia powyższego dla obu modeli wyliczono wartości kryteriów AIC i BIC (w skali logarytmicznej, przed backtransferem). Oba silnie wskazują, że to model Lasso daje lepszy kompromis między jakością dopasowania a złożonością modelu:

* ΔAIC ≈ 157,4 na korzyść Lasso(Log)
* ΔBIC ≈ 502,9 na korzyść Lasso(Log)



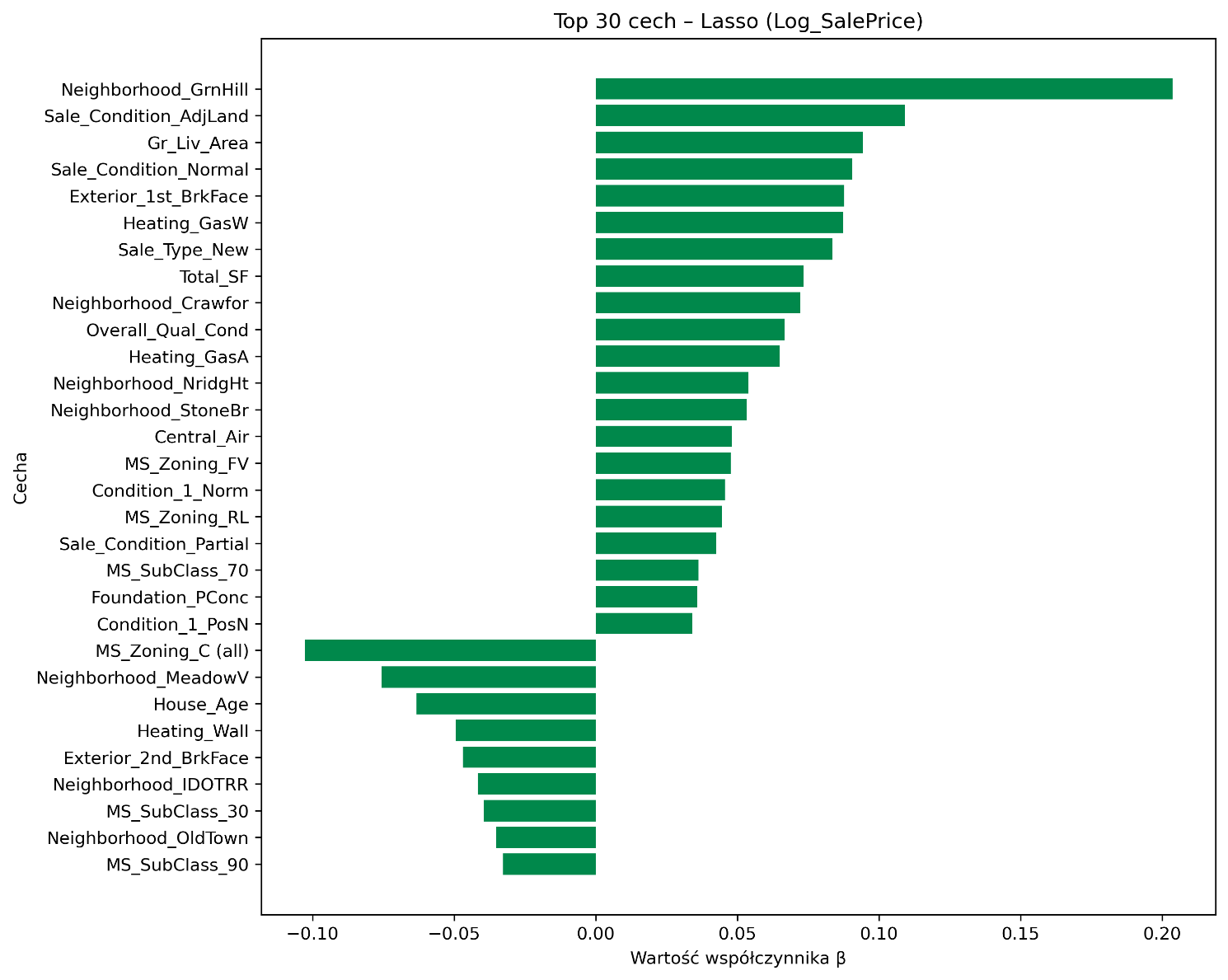
Rysunek 6.3.

Wykres kwantylowy Q-Q reszt dla modelu Lasso opartego o Log\_SalePrice.

Reszty modelu układają się wzdłuż prostej, co oznacza, że są zbliżone do wartości teoretycznych kwantyli normalnych. Wykres reszt z rysunku 5.13. pokazuje, że nie występuje heteroskedastyczność. Nie są złamane zasady konieczne do tworzenia modelu liniowej regresji.

* + 1. **Wnioski i rekomendacje**

Spośród 120 cech, które zostały wykorzystane do stworzenia modelu Lasso (Log\_SalePrice) wytypowano 30 takich, których zmiana ma największy wpływ na cenę nieruchomości. Poniżej zaprezentowano wykres najsilniejszych predyktorów.



Wykres 30 najsilniejszych predyktorów modelu Lasso dla Log\_SalePrice.

Rysunek 6.4.

Na powyższym wykresie najsilniejsze dodatnie predyktory to:

* Neighborhood\_GrnHill (0,2036) – lokalizacja domu w dzielnicy Green Hill wywiera największy wpływ na cenę. Jest to zielona, bezpieczna i zadbana dzielnica, w której często osiedlają się rodziny z dziećmi.
* SaleCondition\_AdjLand (0,1092) – poza nieruchomością sprzedawana była także przyległą działka.
* Gr\_Liv\_Area (0,0943) – naturalny predyktor, ponieważ powierzchnia mieszkalna (sf) przekłada się na cenę mieszkania.
* Sale\_Condition\_Normal (0,0905) – duże znaczenie ma również stan nieruchomości przy sprzedaży. Braki zniszczeń i niedoróbek wpływają na poziom ceny.
* Exterior1st\_BrkFace (0,0878) – wykończenie przednie domu przy pomocy cegły.

W następnej kolejności istotne są również np. Sale\_Type\_New, Total\_SF, czy Overall\_Qual\_Cond. Dobór tych cech nie jest przypadkowy. Większość z nich jest intuicyjna lub wynika z charakterystyki miasta Ames w Iowa.

Do najsilniejszych ujemnych predyktorów należą:

* **MS\_Zoning\_C(all)** (-0,1027)- Oznacza działki o przeznaczeniu mieszkaniowo‐handlowym Ich obecność w skutkach koreluje ze średnio niższą ceną domów, prawdopodobnie ze względu na ruchliwą infrastrukturę i mniejszą prywatność.
* **Neighborhood\_MeadowV** –(0,0755) - Dzielnica Meadow Village, częściej wybierana przez kupujących nastawionych na bardziej ekonomiczne zakupy, co skutkuje przeciętnie niższymi cenami.
* **House\_Age** (-0,0633) - Starsze domy mają niższą cenę – efekt
* **Heating\_Wall** (-0,0494) - Ogrzewanie piecami ściennymi jest mniej efektywne i mniej cenione przez nabywców, co obniża wartość nieruchomości.

Innymi istotnymi cechami wpływającymi negatywnie na cenę nieruchomości są m. in. Neighborhood\_IDOTRR (dzielnica bezpośrednio przy linii kolejowej) oraz MS\_SubClass\_30.

* + 1. **Interpretacja logarytmiczna**

Dla cechy OverallQual model oszacował współczynnik ~ 0,174, zatem:

Wzrost oceny jakości wykończenia (OverallQual) o jedno odchylenie standardowe wiąże się z przeciętnym wzrostem ceny nieruchomości o około 19 %.

* + 1. Potencjalne ulepszenia i perspektywy dalszego rozwoju

Przeprowadzone w niniejszej pracy analizy z wykorzystaniem regresji liniowej oraz metod Ridge i Lasso wskazały na ich użyteczność w predykcji i analizie cen nieruchomości. Pomimo tego, że praca szczegółowo zajmuje się omawianym problemem, istnieją obszary, które mogłyby stanowić jej ciekawe rozwinięcie. Poniżej umieszczono zestawienie zagadnień, które mogłyby wytyczać dalsze kierunki rozwoju pracy:

* + 1. **Zastosowanie modeli nieliniowych**

Chociaż regresja liniowa i jej warianty są skuteczne, warto byłoby spróbować bardziej zaawansowanych modeli, takich jak drzewa decyzyjne, lasy losowe czy sieci neuronowe. Modele te mogą lepiej radzić sobie z bardziej złożonymi, nieliniowymi zależnościami w danych, co mogłoby poprawić dokładność przewidywań.

* + 1. **Dokładniejsza optymalizacja hiperparametrów**

W pracy zastosowano standardowe techniki optymalizacji hiperparametrów. Kolejnym krokiem mogłoby być użycie bardziej zaawansowanych metod, takich jak optymalizacja bayesowska lub algorytmy ewolucyjne, które mogą zapewnić jeszcze lepszą precyzję wyników.

* + 1. **Analiza przestrzenna z wykorzystaniem GIS**

Przy prowadzeniu analiz związanych z rynkiem nieruchomości kluczowym aspektem jest lokalizacja. Dalsze badania mogły obejmować zaawansowaną analizę przestrzenną z wykorzystaniem narzędzi GIS oraz metod regresji przestrzennej.

* + 1. **Analiza trendów czasowych**

Przeprowadzenie analizy trendów cenowych w dłuższej perspektywie czasu umożliwiłoby badanie dynamiki zmian cen oraz przewidywanie ich przyszłego poziomu na podstawie danych historycznych.

* + 1. Bibliografia

[1] GUS – Informacja sygnalna: *„Wskaźniki cen lokali mieszkalnych w 4 kwartale 2023 r.”**Dostęp* kwiecień, 2025. <https://stat.gov.pl/obszary-tematyczne/ceny-handel/wskazniki-cen/wskazniki-cen-lokali-mieszkalnych-w-4-kwartale-2023-roku,12,22.html>

[2] NBP – *„Informacja o cenach mieszkań i sytuacji na rynku nieruchomości IV kw. 2022”* Dostęp kwiecień, 2025. <https://nbp.pl/rynek-nieruchomosci-w-iv-kwartale-2022-r/>

[3] Mooya, M. M. (2016). *„Real Estate Valuation Theory: A Critical Appraisal.”* Springer

[4] Fan, G.-Z., & Ong, S. E. (2019). *„Data-Driven Housing Price Prediction Models: A Global Overview”*. Journal of Real Estate Literature, 27(2), 203–228.

[5] Kauko, T., & Hooimeijer, P. (2018). *„Neural Network Modelling in Housing Market Research: An Overview”*. Real Estate Economics, 30(1), 117–132.

[6] Gujarati, D. N., & Porter, D. C. (2020). *„Econometrics by Example”*. Red Globe Press.

[7] Wooldridge, J. M. (2019). *„Introductory Econometrics: A Modern Approach”*. Cengage Learning.

[8] Rosen, S. (1974). *„Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition”*. Journal of Political Economy, 82(1), 34–55.

[9] Samuel, A.L. (1959). *„Some Studies in Machine Learning Using the Game of Checkers"*. IBM Journal of Research and Development, 3(3), 210–229. DOI: 10.1147/rd.33.0210.

[10] Draper, N. R., & Smith, H. (1998). *„Applied Regression Analysis (3rd ed.)”*. Wiley, (Rozdział 2: *„The Gauss–Markov Theorem”*).

[11] Hastie, T., Tibshirani, R. & Friedman, J. (2009). *„The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)”*. Springer. (Rozdział 3: *„Linear Methods for Regression”*)

[12] De Cock, D. (2011). *„Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project”*. Journal of Statistics Education Volume 19, Number 3(2011). Dostęp kwiecień, 2025.<https://jse.amstat.org/v19n3/decock.pdf> [Online]

[13] American Statistical Association (2011). *„Ames Housing Data Files”* (supplement to De Cock 2011). [Archiwalne] <https://ww2.amstat.org/publications/jse/v19n3/decock/data.html>

[14] American Statistical Association (2011). *„Ames Housing Data txt”*. [Archiwalne] <http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.txt>

[15] American Statistical Association (2011). *„Ames Housing Data xls”*. [Archiwalne] <http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.xls>

[16] OpenML (2025). Dane OpenML dataset #42165. Dostęp kwiecień, 2025. <https://www.openml.org/d/42165> [Online]

[17] Story County Assessor’s Office (2025). Dostęp kwiecień, 2025. [https://www.storycountyiowa.gov/89/Assessor-Property-Values](https://www.storycountyiowa.gov/89/Assessor-Property-Values%20) [Online]

[18] Beacon, Inc. (2025). Dostęp kwiecień, 2025. <https://beacon.schneidercorp.com> [Online]

[19] Maciej Kopczyński (2025). Kod źródłowy. GitHub repository: [https://github.com/Mkopczynski55/PracaInzynierska](https://github.com/Mkopczynski55/PracaInzynierska%20) [Online]

[20] Oracle Corporation (2025). *„MySQL 8.0 Reference Manual: Spatial Data Types.”* Dostęp: kwiecień 2025. <https://dev.mysql.com/doc/refman/8.0/en/spatial-types.html> [Online]