



Studium Licencjackie

Kierunek Metody Ilościowe w Ekonomii i Systemy Informacyjne

Specjalność Metody Analizy Decyzji

Imię i nazwisko autora

Marek Polit

Nr albumu 121564

Analiza wpływu planowanej linii metra M4 na rynek nieruchomości w Warszawie

**Praca licencjacka
pod kierunkiem naukowym
dr. Małgorzaty Wrzosek
Zakład Wspomagania i Analizy Decyzji**

Warszawa, lipiec 2025

Spis treści

Wprowadzenie	4
Rozdział I: Kontekst rynkowy i metody analizy rynku mieszkaniowego	5
1.1 Historia Warszawskiego Metra	5
1.2 Perspektywy rozwoju i znaczenie rynku mieszkaniowego w Warszawie	7
1.3 Charakterystyka kluczowych zmiennych różnicujących ceny mieszkań	9
1.4 Metody analizy danych na rynku nieruchomości	11
Rozdział II: Proces przygotowania danych do dalszej analizy	14
2.1 Pozyskanie i przetwarzanie danych	14
2.2 Eksploracja, filtrowanie oraz grupowanie danych	19
2.3 Wizualizacje rynku mieszkaniowego w Warszawie	25
Rozdział III: Metody oszacowania wpływu bliskości metra na Rynek Nieruchomości	30
3.1 Przegląd danych i budowanie modeli uczenia maszynowego	30
3.2 Metoda wyznaczania wpływu linii metra M4 na rynek nieruchomości	33
3.3 Wizualizacje przewidywanych zmian na rynku mieszkaniowym	37
Zakończenie	40
Bibliografia	42
Spis tabel	44
Spis rysunków	44
Streszczenie	45

Wprowadzenie

Celem niniejszej pracy jest predykcja zmian cenowych na rynku mieszkaniowym w Warszawie, które mogą nastąpić po wybudowaniu nowych stacji metra linii M4. Tematyka pracy zyskuje na istotności w obliczu narastających obaw młodego pokolenia o możliwość nabycia własnego mieszkania, szczególnie w kontekście trwającej w Polsce debaty nad wprowadzeniem podatku katastralnego. Z drugiej strony, przedmiot badań porusza równie ważny aspekt komunikacyjny, związany z możliwością korzystania z bardziej komfortowego środka transportu, co ma szczególne znaczenie dla rozwoju miasta. Przykładem może być Berlin, który na dzień 1 maja 2025 r. dysponuje dziewięcioma liniami metra (U-Bahn) oraz rozległą siecią szybkiej kolei miejskiej (S-Bahn). Taka infrastruktura pozwala mu objąć niemal dwukrotnie większy obszar niż Warszawa i w znacznym stopniu kształtuje tamtejszy rynek nieruchomości. W odniesieniu do stolicy Polski, obecnie trwają rozmowy na temat budowy trzech nowych linii metra, których zakończenie planuje się do roku 2050, przy czym na etapie opracowywania niniejszej pracy jedynie dla linii M4 ustalone precyzyjne lokalizacje wszystkich stacji. W związku z powyższym niniejsza praca przedstawia starannie opracowaną metodykę oceny wpływu odległości od stacji metra na ceny nieruchomości, opartą na analizie danych pochodzących z ogólnodostępnych źródeł internetowych. W rozdziale pierwszym omówiono teoretyczne podstawy modelowania cen – opisano zarówno koncepcje hedoniczne, jak i nowoczesne techniki uczenia maszynowego, podkreślając przy tym jednocześnie atuty Warszawy oraz przedstawiając krótki rys historyczny rozwoju Warszawskiego metra. Drugi rozdział poświęcono kompleksowemu przygotowaniu danych, obejmującego m.in. odpowiednie pozyskanie, przefiltrowanie oraz przetworzenie danych, tak aby stanowiły wiarygodne źródło informacji dla modeli. W rozdziale trzecim, skupiono się na praktycznej części badania – optymalizacji i porównaniu algorytmów regresyjnych, analizie ważności zmiennych oraz symulacji wpływu nowych stacji metra na prognozowane ceny, z uwzględnieniem odległości od metra, jako zmiennej różnicującej ceny mieszkań. Całe badanie służy weryfikacji hipotezy o istotnym wpływie odległości mieszkania od stacji metra, na cenę rynkową. Należy zaznaczyć, że ze względu na koncepcyjny charakter pracy konieczne jest cykliczne monitorowanie zmian rynkowych, a opracowana metodologia może posłużyć jedynie do estymacji potencjalnych fluktuacji cenowych. Choć na etapie tworzenia niniejszej analizy przyjęto oficjalne lokalizacje stacji M4, w przyszłości warto zweryfikować, czy plany nie zostały poddane modyfikacjom.

Rozdział I

Kontekst rynkowy i metody analizy rynku mieszkaniowego

Celem niniejszego rozdziału jest przedstawienie tła rynkowego niezbędnego do przeprowadzenia analizy wpływu planowanej linii metra M4 na rynek nieruchomości w Warszawie. Rozdział ten stanowi punkt wyjścia do dalszych analiz, dlatego skupiono się zarówno na charakterystyce rynkowej Warszawy, jak i na metodach ilościowych umożliwiających uchwycenie zależności między cechami mieszkań, a ich ceną.

W rozdziale zaprezentowano krótki zarys historii rozwoju sieci metra w Warszawie, ze szczególnym uwzględnieniem planowanej linii M4, która stanowi kluczowy element niniejszego badania. Następnie wskazano zalety i perspektywy ekspansji Warszawskiego rynku nieruchomości na tle innych większych miast w Polsce. Kolejne podrozdziały zawierają przegląd literatury pod kątem wyboru optymalnych modeli predykcyjnych, powszechnie stosowanych w analityce mieszkaniowej, a także charakterystyki zmiennych, które wykorzystywane były do tej pory w analizach przestrzennych.

1.1 Historia Warszawskiego Metra

Początki idei budowy metra w Warszawie sięgają okresu przedwojennego. Już w 1925 roku władze miasta stołecznego Warszawy, podjęły uchwałę o opracowaniu projektu kolei podziemnej, na wzór funkcjonującego w Paryżu metra (tzw. „Metropolitainu”). Dwa lata później władze zatwierdziły wstępny plan przebiegu metra i zleciły zapoczątkowanie prac nad eksploatacją geologiczną terenów Warszawy. Pierwowzór projektu zakładał stworzenie dwóch krzyżujących się linii biegących w kierunkach:

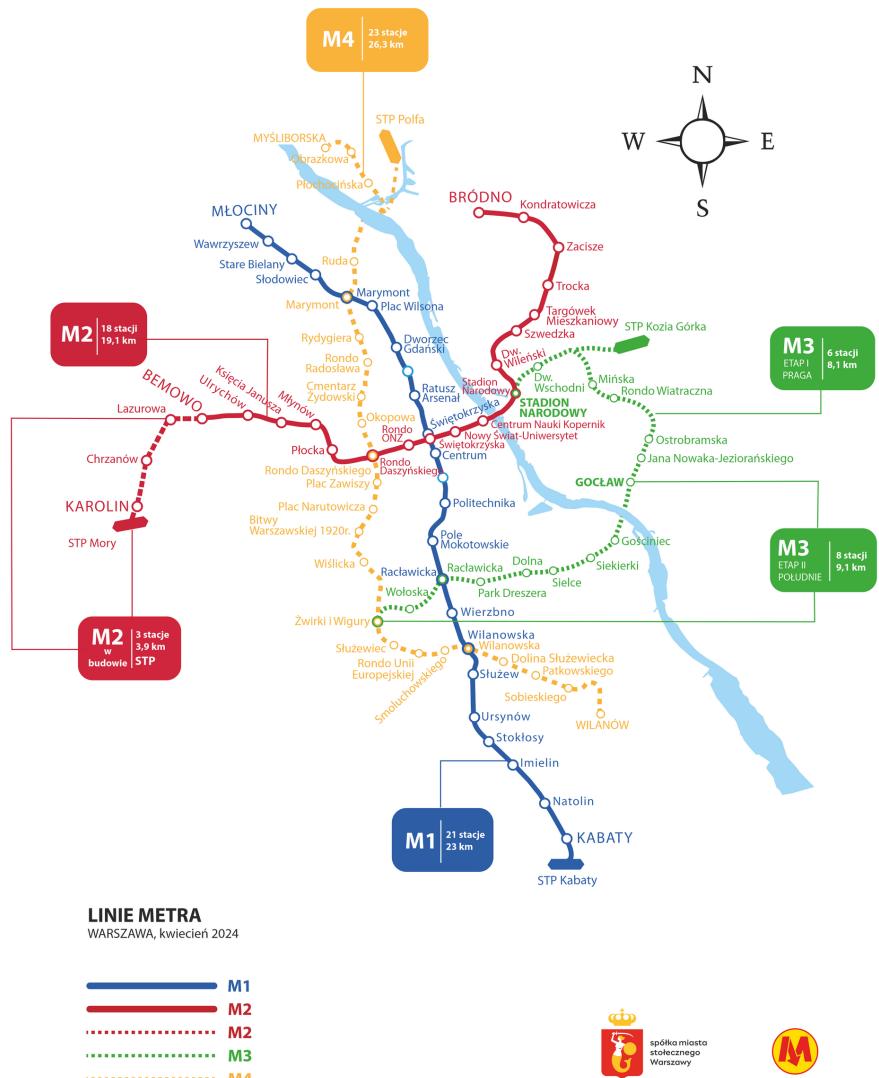
- Północ-Południe (przebiegająca między placem Unii Lubelskiej, a Muranowem)
- Wschód-Zachód (obejmująca trasę od Woli do Pragi)

Choć wykonano odwierty i opracowano szczegółowe koncepcje inżynieryjne, ówczesny kryzys gospodarczy uniemożliwił realizację tych założeń, a projekt na wiele lat porzucono.

Do prac nad koncepcją metra powrócono prawie dekadę później, tuż przed wybuchem II Wojny Światowej. W 1938 roku powstało Biuro Studiów Kolei Podziemnej, które przygotowało projekt sieci o długości 46 km. Niestety działania wojenne przerwały prace planistyczne, a znaczna część dokumentacji zaginęła podczas Powstania Warszawskiego. Od razu po zakończeniu wojny wznowiono pracę i opracowano nową trasę łączącą tym razem Służew z Młocinami oraz Wolę z Wawrem.

W latach 50. zapadła decyzja o realizacji tzw. „metra głębokiego”, w dużej mierze motywowana względami militarnymi i geopolitycznymi. Tak zbudowany tunel, umożliwiałby efektywny przejazd konwojom wojskowym, a ostatecznie mógłby być wykorzystany do budowy faktycznego środka komunikacyjnego. Budowę rozpoczęto, jednak po trzech latach – w 1953 roku – projekt wstrzymano z uwagi na jego wysokie koszty i zbyt duże trudności techniczne.

Koncepcja budowy metra powróciła w latach 70., jednak priorytety związane z inwestycjami drogowymi, sprawiły, że na metro nie wystarczyło już środków. Przełom nastąpił dopiero na początku lat 80., kiedy władze PRL uznały, że rozwój komunikacji zbiorowej jest konieczny dla funkcjonowania stolicy. W 1982 roku Rada Ministrów formalnie zatwierdziła budowę pierwszej linii metra. W roku następnym powołano Generalną Dyrekcję Budowy Metra i tak oto symbolicznie rozpoczęto inwestycję wbiciem pierwszego pala 15 kwietnia 1983 roku.



Rysunek 1: Schemat warszawskiego metra – stan na kwiecień 2024 r.

Źródło: Metro Warszawskie (2024).

Mimo oficjalnego poparcia i entuzjazmu społecznego, realizacja inwestycji napotykała liczne trudności – ograniczone finansowanie, nieregularne dostawy materiałów, a także brak doświadczenia w zakresie technologii tunelowych. Konieczne było dostosowanie oraz rozwój metod inżynierijnych oraz współpraca z polskimi uczelniami i technikami w celu szkolenia przyszłych operatorów metra.

Choć projekt pierwszej linii metra wielokrotnie opóźniano i ograniczano, ostatecznie w 1995 roku uruchomiono pierwszy odcinek z Kabat do Politechniki. Kolejne stacje oddawano sukcesywnie: Centrum (1998), Ratusz (2001), Dworzec Gdańskiego (2003), Plac Wilsona (2005) i Ślądowiec (2008). Finalne otwarcie całej linii M1 miało miejsce w październiku 2008 roku.

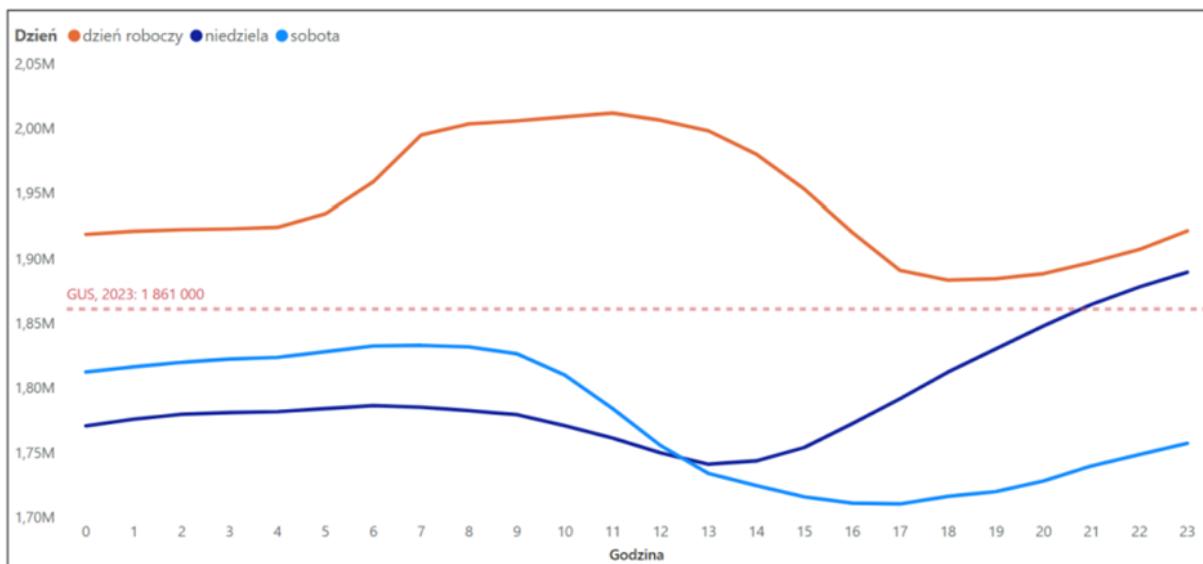
Rozbudowę systemu kontynuowano w XXI wieku. W 2015 roku otwarto centralny odcinek drugiej linii metra (M2), przecinającej Wisłę, a następnie rozszerzono ją w kierunkach wschodnim i zachodnim. Obecnie trwają prace nad trzecią (M3) i czwartą (M4) linią metra, które mają uzupełnić siatkę połączeń, ze szczególnym uwzględnieniem obszarów Gocławia i Wilanowa. Szczególne znaczenie w tej pracy przypisuje się planowanej linii M4, której trasa będzie przebiegać przez rozwijające się rejony mieszkaniowe i może istotnie wpłynąć na ich atrakcyjność inwestycyjną (Historia budowy metra, 2024). Dokładny przebieg obecnych połączeń podziemnych przedstawiony jest ciągłymi liniami (Rysunek 1). Liniami przerywanymi zaznaczono natomiast planowane linie metra M3 i M4 oraz przedłużenie linii M2, której nowy odcinek zakończy się pętlą na Karolinie.

1.2 Perspektywy rozwoju i znaczenie rynku mieszkaniowego w Warszawie

Rynek nieruchomości w Warszawie stanowi istotny element rozwoju stolicy z perspektywy wzrostu kapitału ludzkiego. Z podsumowania badania „Property Index 2023” wynika, że pomimo trudnej sytuacji makroekonomicznej w Polsce, liczba nowych mieszkań oddanych do użytku była jedną z najwyższych w Europie (Deloitte, 2023). Najwięcej mieszkań przypadających na 1000 mieszkańców stolic oddano do użytku kolejno we: Francji, w Polsce oraz w Danii.

Warszawski rynek odnotowywał również na przestrzeni ostatniej dekady znaczący wzrost demograficzny. Na podstawie badania telemetrycznego przeprowadzonego na zlecenie Biura Strategii i Analiz Urzędu m.st. Warszawy w 2023 roku, oszacowano rzeczywistą liczbę mieszkańców stolicy przy użyciu danych dotyczących logowań kart SIM do sieci komórkowej. Zgromadzony materiał wskazuje na wyraźny wzrost liczby osób faktycznie przebywających w granicach administracyjnych miasta, co przekłada się na rosnące potrzeby mieszkaniowe w stolicy. Wykres zawarty w badaniu (Rysunek 2), przedstawia liczbę urządzeń zalogowanych do sieci telekomunikacyjnej na terenie Warszawy w październiku 2023 roku. Wybrany miesiąc miał na celu ograniczenie wpływu czynników sezonowych, takich jak okres wakacyjny czy święta.

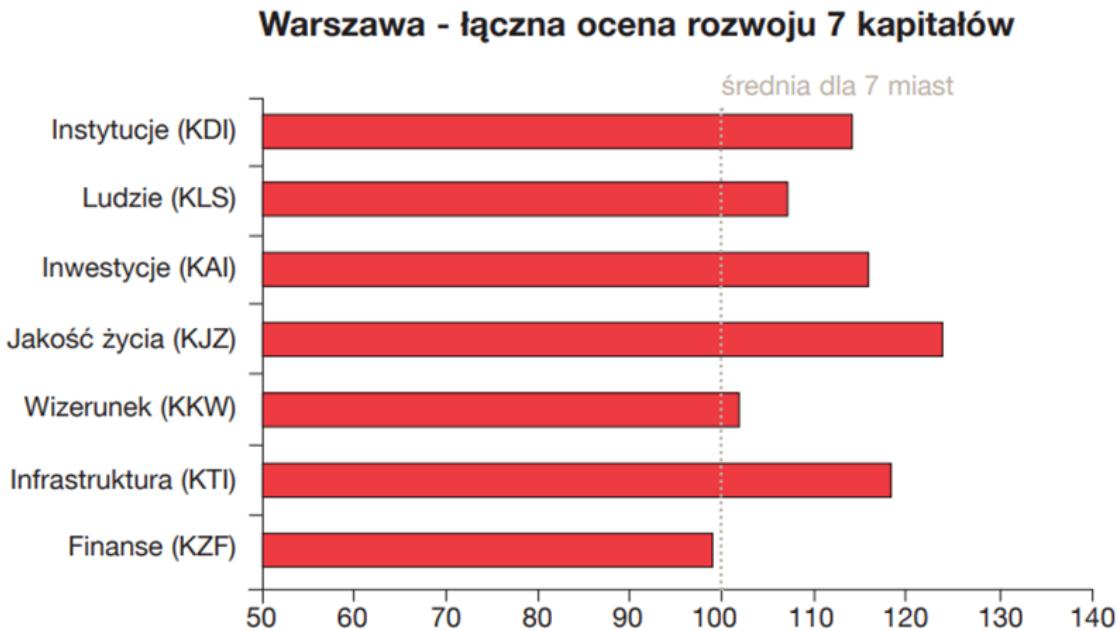
Linia trendu z podziałem na dni robocze oraz weekendowe, implikuje istotny wzrost liczby aktywnych urządzeń w godzinach 6:00-16:00 oraz porównywalnie niższe wartości zalogowanych telefonów w okresie końca tygodnia. Według tych oszacowań, Warszawę zamieszkuje obecnie ponad 150 tys. rezydentów więcej, niż dekadę wcześniej (1 715 tys. w 2013 roku). (Biuro Strategii i Analiz Urzędu m.st. Warszawy 2024, s. 5).



Rysunek 2: Wykres aktywności telekomunikacyjnej w obrębie Warszawy

Źródło: Biuro Strategii i Analiz Urzędu m.st. Warszawy (2024, s. 5)

Warszawa wyróżnia się także w pozostałych kluczowych obszarach rozwoju, nie tylko pod względem demograficznym. Według raportu, porównującego czynniki kapitałowe, w których punktem odniesienia był standaryzowany wskaźnik dla największych miast w Polsce, Warszawa cechuje się wysokim potencjałem ekspansyjnym. Zjawisko to może wynikać z migracji wewnętrznych, w ramach których największe miasta działają jak ośrodki grawitacyjne, przyciągające ludność z mniej zaludnionych regionów kraju (w tym z mniejszych miast). Wykres kolumnowy (Rysunek 3) przedstawia porównanie siedmiu kluczowych obszarów rozwoju miejskiego, w tym kapitału instytucjonalnego, ludzkiego, inwestycyjnego, infrastrukturalnego, jakości życia, wizerunkowego oraz finansowego. Analiza tych danych wskazuje, że Warszawa osiąga wysokie wyniki w niemal wszystkich wymienionych aspektach, z nieznacznym odchyleniem od średniej jedynie w zakresie kapitału finansowego. Niższy wynik w tym obszarze wynika najprawdopodobniej z obowiązku odprowadzania subwencji wyrównawczej, mającej na celu wsparcie mniej zamożnych samorządów i ograniczenie nadmiernej koncentracji środków w jednym ośrodku miejskim.



Rysunek 3: Porównanie poziomu rozwoju kapitałów Warszawy na tle największych miast w Polsce

Źródło: PwC (2019, s. 7)

Dalsze fragmenty raportu zwracają uwagę na szczególne atuty stolicy w kontekście rynku mieszkaniowego. Wśród nich wymienia się przede wszystkim:

- Dobry wizerunek i dużą atrakcyjność inwestycyjną
- Liczny kapitał ludzki
- Wysoki potencjał instytucjonalno-kulturowy
- Zapewnienie solidnego stanu bezpieczeństwa

Wszystkie powyższe informacje dowodzą, że Warszawski rynek nieruchomości jest dosyć dynamicznym sektorem z dużymi perspektywami rozwojowymi oraz rosnącym zarówno popytem, jak i podażą mieszkań.

1.3 Charakterystyka kluczowych zmiennych różnicujących ceny mieszkań

Analiza cen na rynku nieruchomości wymaga interdyscyplinarnego podejścia do modelowania, łączącego ekonometrię i geoinformatykę. Badania koncentrujące się na cenie mieszkania powinny uwzględnić zarówno czynniki endogeniczne (metraż, wyposażenie, stan techniczny), jak i egzogeniczne (odległość od metra, szkoły, czy najbliższego parku). Z perspektywy wyróżnienia zmiennych dla budowania modelu predykcyjnego, trudniejsze jest wyłonienie charakterystyk egzogenicznych, gdyż wymagają one zaawansowanej inżynierii danych w celu ich

pozyskania. Przykładowymi takimi czynnikami, są zmienne wykorzystane w analizie na temat teoretycznych i metodycznych aspektów wyznaczania indeksów cen na rynku nieruchomości (Trojanek 2018, s. 192). Wśród istotnych zmiennych, które wymagały oszacowania odległości mieszkania od określonych punktów przestrzennych, znalazły się między innymi:

- Logarytm odległości do najbliższego Jeziora
- Logarytm odległości do najbliższego przystanku autobusowego
- Logarytm odległości do najbliższego przystanku tramwajowego
- Logarytm odległości do najbliższych terenów zielonych
- Logarytm odległości do najbliższej szkoły podstawowej
- Logarytm odległości do najbliższego centrum handlowego

Zastosowanie transformacji logarytmicznej przez autora mogło wynikać z potrzeby dokładniejszego odwzorowania relacji między odlegością a ceną nieruchomości. Wskazane przekształcenie pozwala również na lepsze uchwycenie nieliniowych zależności, jednocześnie redukując wpływ obserwacji odstających. Ponadto, zastosowanie takiej transformacji mogło pozytywnie wpływać na jakość budowanych modeli, zarówno poprzez lepsze spełnienie ich klasycznych założeń, jak i ułatwienie procesu uczenia się. W przypadku czynników endogenicznych mieszkania, które w bezpośredni sposób mogą wpływać na jego wartość, warto zwrócić uwagę na elementy wyposażenia oraz inne udogodnienia, takie jak n.p. zapewnienie ochrony. Zmienne, które mogą mieć istotne znaczenie w kontekście różnicowania cen, to między innymi:

- powierzchnia,
- obecność windy,
- liczba sypialni, łazienek i toalet,
- obecność garażu i liczba miejsc parkingowych,
- data publikacji ogłoszenia,
- obecność ochrony,
- obecność pralni i jadalni,
- numer piętra,
- wiek budynku,
- typ kuchenki (gazowa, elektryczna).

Źródło: Vargas-Calderón, Camargoc (2020, s. 6).

W sytuacji, gdy dane wykorzystywane w analizie pochodzą z różnych okresów, zaasadne jest uwzględnienie wpływu zmienności rynkowej na ceny nieruchomości. Pominięcie tego aspektu mogłoby prowadzić do zafałszowania wyników, zwłaszcza w przypadku analiz obejmujących dłuższy horyzont czasowy. Jednym z rozwiązań jest wprowadzenie zmiennych czasowych, takich jak rok lub miesiąc publikacji ogłoszenia, co umożliwia modelowi uchwycenie zarówno długoterminowych trendów, jak i efektów sezonowości. Z praktycznego punktu widzenia, istotne jest także odpowiednie przekształcenie zmiennej wyjaśnianej. Ze względu na silną korelację pomiędzy ceną a powierzchnią mieszkania, powszechnie stosowaną praktyką jest wykorzystanie ceny za metr kwadratowy zamiast całkowitej wartości transakcyjnej. Pozwala to ograniczyć ryzyko nadmiernego wpływu zmiennej powierzchni na wyniki modelu oraz zmniejszyć wariancję oszacowań, co sprzyja bardziej miarodajnej ocenie pozostałych cech istotnych dla kształtowania cen.

1.4 Metody analizy danych na rynku nieruchomości

Jednym z klasycznych podejść stosowanych w analizie cen nieruchomości jest wykorzystanie metod hedonicznych. W praktyce opierają się one na konstrukcji modelu regresji, w którym cena mieszkania wyjaśniana jest za pomocą zbioru jego cech, takich jak lokalizacja, metraż czy dostępne udogodnienia. Estymacja (np. liniowa) parametrów odbywa się z wykorzystaniem metody najmniejszych kwadratów, co umożliwia określenie wpływu poszczególnych zmiennych – zarówno pozytywnego, jak i negatywnego – na wartość nieruchomości (Belniak, Wieczorek 2017, s. 60–62).

Choć metoda ta znajduje szerokie zastosowanie ze względu na swoją intuicyjność i prostotę, nie jest pozbawiona ograniczeń. Modele liniowe są szczególnie wrażliwe na nieliniowości w danych oraz obecność obserwacji odstających, które mogą znacząco zaburzyć trafność estymacji (Das et al. 2020, s. 4). Jednym z możliwych sposobów radzenia sobie z tym problemem jest zastosowanie drzew regresyjnych, które nie tylko potrafią lepiej odwzorować złożone zależności między zmiennymi, ale również oferują większą przejrzystość interpretacyjną. Niemniej jednak, tego rodzaju modele mają również swoje ograniczenia – w szczególności cechują się dużą wrażliwością na dane uczące, co może skutkować wysoką zmiennością wyników oraz problemami z uogólnianiem predykcji (Kok et al. 2017, s. 205).

W odpowiedzi na ograniczenia klasycznych modeli liniowych, coraz częściej w literaturze przedmiotu wykorzystywane są algorytmy uczenia maszynowego, które umożliwiają modelowanie bardziej złożonych i nieliniowych zależności pomiędzy zmiennymi. Przykładem może być badanie przeprowadzone przez zespół badawczy z Departamentu Informatyki w Pakistanie, w którym porównano skuteczność jedenastu algorytmów predykcyjnych w kontekście prognozowania cen nieruchomości w Islamabadzie. Wyniki analizy wykazały, że jednym z najskutecznie-

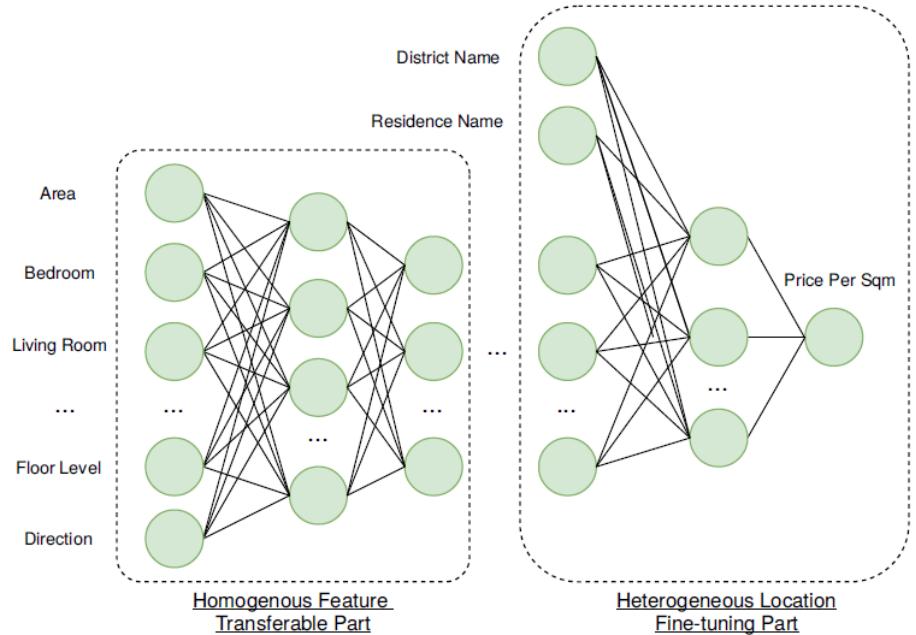
teczniejszych podejść była regresja wektorów nośnych (ang. Support Vector Regression, SVR), która osiągnęła najwyższą dokładność szacowania teoretycznych cen mieszkań (Imran et al. 2021, s. 11).

Kolejnym algorytmem o potwierdzonej skuteczności jest regresja oparta na lasach losowych (Random Forest Regression), charakteryzująca się wysoką odpornością na przeuczenie modelu oraz stabilnością predykcji. W badaniu przeprowadzonym w Mumbaju, technika ta została zastosowana w zestawieniu z innymi metodami uczenia maszynowego i wykazała relatywnie niskie błędy oszacowań w prognozowaniu cen mieszkań (Borde et al. 2017, s. 1825).

Wśród metod zespołowych (ang. ensemble methods), szczególne miejsce zajmuje algorytm wzmacniania gradientowego (Gradient Boosting), który buduje kolejne drzewa decyzyjne w sposób sekwencyjny, korygując błędy popełnione przez wcześniejsze modele poprzez optymalizację funkcji straty. Nowoczesna implementacja tej techniki – XGBoost – uwzględnia dodatkowo regularyzację L1 i L2, co zwiększa odporność modelu na nadmierne dopasowanie. Zastosowanie tego rozwiązania w badaniu przeprowadzonym w Bogocie (Kolumbia) pozwoliło na uzyskanie niskiego średniego błędu prognoz – na poziomie 8,9% w odniesieniu do średniej ceny za metr kwadratowy mieszkania (Vargas-Calderón, Camargoc 2020, s. 8).

Z uwagi na rosnącą dostępność danych oraz postęp w obszarze sztucznej inteligencji, coraz większym zainteresowaniem cieszą się również modele głębokiego uczenia (ang. deep learning). Ich główną zaletą jest zdolność do modelowania bardzo złożonych, nieliniowych zależności pomiędzy zmiennymi, co czyni je szczególnie przydatnymi w kontekście zróżnicowanego i niejednorodnego rynku nieruchomości. W przeciwieństwie do klasycznych modeli regresyjnych, sieci neuronowe nie wymagają ręcznego definiowania formy funkcji zależności, co umożliwia im automatyczne wykrywanie wzorców i relacji obecnych w danych. Choć interpretacja działania poszczególnych neuronów i warstw sieci może być trudna, modele tego typu wykazują wysoką skuteczność także w przypadku danych zawierających szумy czy wartości odstające (Peterson, Flanagan 2009, s. 159).

Jednym z kluczowych wyzwań w modelowaniu cen nieruchomości, szczególnie w ujęciu międzyregionalnym, jest ograniczona zdolność algorytmów predykcyjnych do uogólniania wyników na obszary o odmiennych charakterystykach lokalizacyjnych. Problem ten związany jest z dużym zróżnicowaniem przestrzennym, które wpływa na interpretację i znaczenie poszczególnych zmiennych w różnych częściach miasta lub kraju. W odpowiedzi na tę trudność, Guo et al.(2018, s. 4–6) zaproponowali nowatorskie podejście oparte na rozdzieleniu zmiennych wejściowych na dwie grupy: cechy homogeniczne i heterogeniczne.



Rysunek 4: Wizualizacja proponowanego modelu strumienia przetwarzania danych
 Źródło: Guo, Lin, Ma, Bal, Li (2018, s. 6)

W tym ujęciu, cechy homogeniczne to zmienne bezpośrednio związane z charakterystyką samej nieruchomości, takie jak powierzchnia, liczba łazienek czy piętro, na którym znajduje się lokal. Z kolei cechy heterogeniczne odnoszą się do otoczenia nieruchomości, czyli elementów zależnych od lokalizacji – przykładowo, dzielnica, osiedle lub inne uwarunkowania przestrzenne. Autorzy opracowali architekturę modelu, w której dane przetwarzane są równolegle w dwóch osobnych ścieżkach (ang. pipelines), z których każda obsługuje jedną grupę cech, przy wykorzystaniu technik głębokiego uczenia. W opisywanym podejściu dane wejściowe są najpierw wykorzystywane do wytrenowania modelu w oparciu o cechy homogeniczne, a następnie dostrajane z wykorzystaniem cech heterogenicznych, specyficznych dla danej lokalizacji. Taki schemat umożliwia na przykład trenowanie modelu na danych z Warszawy, przy czym część wspólna – obejmująca cechy charakterystyczne dla każdej nieruchomości (takie jak metraż, liczba pokoi, piętro) – może zostać przeniesiona na inne miasta. Następnie model jest dostosowywany do uwarunkowań lokalnych, np. specyfiki rynku we Wrocławiu, poprzez integrację zmiennych przestrzennych. Co istotne, taka konstrukcja modelu pozwala na osiąganie wysokiej precyzji predykcji również w miastach, w których dostępność danych jest ograniczona. Według autorów, zastosowanie podejścia z rozdzieleniem cech umożliwia osiągnięcie porównywalnych rezultatów predykcyjnych przy wykorzystaniu jedynie 20% danych, które byłyby wymagane w przypadku tradycyjnego modelu bez zastosowania struktury typu *pipeline* (Guo et al. 2018, s. 11).

W kolejnym rozdziale skupiono się już na opracowaniu konkretnego podejścia modelowego oraz konstrukcji zbioru danych, na podstawie którego możliwe będzie przeprowadzenie analizy empirycznej wpływu planowanej linii metra M4 na ceny mieszkań w stolicy.

Rozdział II

Proces przygotowania danych do dalszej analizy

Rozdział II przedstawia metodykę zbierania i przetwarzania danych, w celu przygotowania ich do budowy modeli uczenia maszynowego. W podrozdziale 2.1 omówiono proces pozyskiwania i przetwarzania danych, obejmujący web scraping ofert z portalu Otodom, ekstrakcję kluczowych atrybutów, geokodowanie adresów oraz inne aspekty inżynierii cech. Podrozdział 2.2 poświęcono eksploracji oraz oczyszczaniu zbioru, w celu usunięcia potencjalnie niemiarodajnych obserwacji. W podrozdziale 2.3 zaprezentowano wizualizacje wyników grupowania, tak by zobrazować lepiej kontekst teoretyczny rynku mieszkaniowego w Warszawie.

2.1 Pozyskanie i przetwarzanie danych

W ramach badania jako podstawowe źródło danych wykorzystano ogłoszenia zamieszczone na portalu Otodom. W pierwszym etapie, w celu zgromadzenia danych, zdefiniowano funkcję, która przyjmuje w argumencie adres URL pojedyńczej oferty i zwraca poszczególne zmienne w formie listy. Przykładowa struktura zastosowania tej funkcji w środowisku *Jupyter Notebook* w języku *Python* może wyglądać następująco:

Komórka wejściowa:

```
web_scrape_otodom_offer_as_list("https://www.otodom.pl/pl/oferta/3-pokojowe-mieszkanie-58m2-balkon-XXXXXXX")
```

Komórka wyjściowa:

```
[ 'Cena: 800 000 zł',
  'Tytuł: 3-pokojowe mieszkanie 58m2 + balkon Bez Prowizji',
  'Opis: Mieszkanie 3 pokojowe z widokiem na Park ...',
  'Powierzchnia: 58 m2',
  "Liczba pokoi: '8'",
  'Piętro: 1/4',
  'Czynsz: brak danych',
  'Rynek: pierwotny',
  'Rok budowy: 2000',
  'Informacje dodatkowe: internet, telewizja kablowa, telefon',
  'Stan wykończenia: do wykończenia',
  'Winda: nie',
  'Adres: ul. Jakościowa 4, Mokotów, Mazowieckie' ]
```

Kolejnym etapem zbierania danych było opracowanie funkcji, która, wykorzystując skonfigurowany sterownik, uruchamia przeglądarkę internetową, przechodzi kolejno przez strony z ofertami sprzedaży mieszkań i z każdej z nich wyodrębnia adresy URL poszczególnych ogłoszeń. Po zakończeniu tego procesu, funkcja zamyka przeglądarkę, a zgromadzone linki stanowią bazę danych, z której można pobierać szczegółowe informacje o ofertach. Przykładowa struktura kilku pierwszych ofert, może wyglądać następująco:

Komórka wejściowa:

```
oferty
```

Komórka wyjściowa:

```
[ 'https://www.otodom.pl/pl/oferta/trzy-pokoje-wesola-FISCX',
  'https://www.otodom.pl/pl/oferta/80m2-ogrod-wawa-DSAKXS',
  'https://www.otodom.pl/pl/oferta/ursus-vita-GFUSXA',
  'https://www.otodom.pl/pl/oferta/2-pokojowe-garaz-HFCSSAWE',
  'https://www.otodom.pl/pl/oferta/wyjatkowy-dom-HGGFSER' ]
```

Ostatnim procesem było zaimplementowanie mechanizmu umożliwiającego pobranie oraz integrację szczegółowych informacji o ofertach do jednej, spójnej bazy danych. W tym celu zbudowano funkcję integrującą listę *oferty* oraz funkcję *web_scrape_otodom_offer_as_list(URL)*, która w sposób iteracyjny wyciąga z każdej z ofert zmienne, a następnie dodaje je jako nowe wiersze w zbiorze rekordów *mieszkania*.

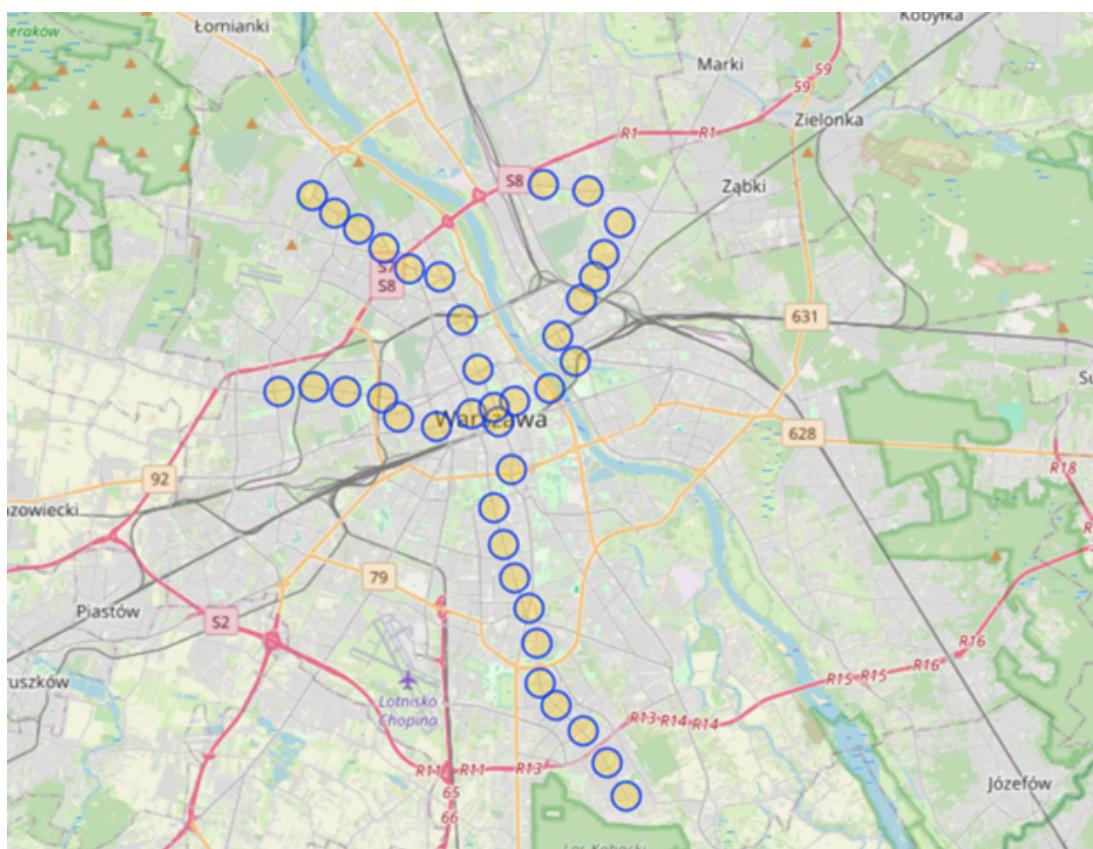
Cena	Tytuł	Opis	Pow.	Pokoi	Piętro	Czynsz	Rynek	Rok	Info.	Stan	Winda	Adres
725 000 zł	Stylowe, 3-pokojowe mieszkanie	Mieszkanie stylowe,...	49,89 m ²	[‘3’]	3/5	600 zł	wtórny	2022	telewizja, internet	do zamieszkania	tak	ul. Chełmżyńska 10
680 000 zł	M2 10 min od Złotej	Dla rodziny / studentów	37,7 m ²	[‘2’]	8/12	brak	wtórny	1968	winda, oddz. kuchnia	do remontu	tak	ul. Fasolowa 12/5
830 000 zł	3 Pokoje i loggia	3-pokojowe mieszkanie	50 m ²	[‘3’]	4/4	1 050 zł	wtórny	1973	piwnica, oddz. kuchnia	do remontu	nie	ul. Czerwiakowska 15
1 551 200 zł	2-pokojowe mieszkanie	Mieszkanie z dobrym układem	56 m ²	[‘2’]	–	500 zł	wtórny	2019	monitoring, ochrona	do wykończenia	tak	ul. Grzybowska 24
825 000 zł	Służew nad Dolinką	Piękny widok na park	46 m ²	[‘2’]	7/10	600 zł	pierwotny	1979	winda, balkon, piwnica	do zamieszkania	tak	ul. Mokotowska 78

Tabela 1: Przykładowe oferty mieszkań z portalu Otodom

Źródło: Opracowanie własne

Podczas ściągania poszczególnych ofert, weryfikowana jest również kompletność danych tzn. jeżeli liczba elementów w liście nie odpowiada oczekiwanej liczbie atrybutów, funkcja

ta pobiera jedynie informacje dostępne, a nieopublikowane atrybuty oznacza jako brak danych. Ponadto, jeżeli cena oferty jest określona jako "Zapytaj o cenę", następuje pominięcie przetwarzania danej oferty, gdyż zmienna oznaczająca cenę jest kluczowa dla rozwiązywanego problemu. Aby zapewnić bezpieczeństwo danych w trakcie długotrwałego procesu przetwarzania, co tysiąc przetworzonych ofert tworzona jest kopia zapasowa bazy danych w postaci pliku CSV. Po zakończeniu iteracji cała zaktualizowana baza jest zapisywana do ostatecznego pliku, co stanowi fundament dla dalszej analizy rynku nieruchomości zgodnie z założonymi celami badawczymi.



Rysunek 5: Stacje metra w Warszawie
Źródło: Overpass Turbo, dostęp: 12.04.2025.

W kolejnym etapie, wykorzystując metodę *geocode* udostępnioną przez obiekt klienta Google Maps, wysyłane jest żądanie do usługi geokodowania. Jeżeli wyniki zapytania są prawidłowe, funkcja wyodrębnia wartości szerokości i długości geograficznej. W przypadku braku wyników lub wystąpienia błędu, funkcja zwraca komunikat *brak danych*. Uzyskane zmienne zostały zintegrowane z głównym zbiorem danych, co umożliwiło ich kompleksowe wykorzystanie w dalszych etapach analizy. Po wyznaczeniu współrzędnych geograficznych ofert możliwe stało się określenie odległości od najbliższych obiektów przestrzennych, które zgodnie z założeniami badania - mogą mieć wpływ na cenę za metr kwadratowy. Do tego celu wyko-

rzystano narzędzie Overpass Turbo - przeglądarkową aplikację służącą do pobierania danych z projektu OpenStreetMap za pomocą ustrukturyzowanych kwerend. Na Rysunku 5 przedstawiono wizualizację wyniku wyszukiwania po wpisaniu zapytania dotyczącego lokalizacji stacji metra.

Dane OSM, tworzone i utrzymywane przez społeczność użytkowników na całym świecie, zawierają szczegółowe informacje o elementach takich jak parki, granice administracyjne oraz inne obiekty i elementy przestrzenne. Ponadto w Overpass Turbo wykorzystano funkcję eksportu danych do formatu GeoJSON, co znaczco ułatwiło wizualizację np. granic administracyjnych Warszawy.

Po zimportowaniu tabel z narzędzia Overpass Turbo, zbudowano funkcję, która znajduje odległości od następujących obiektów:

- najbliższej stacji metra linii M1
- najbliższej stacji metra linii M2
- najbliższego supermarketu
- najbliższego obszaru zielonego

Wykonanie zapytań o odległość w sposób iteracyjny, czyli osobno dla każdej nieruchomości i każdego punktu referencyjnego — skutkowałoby nieproporcjonalnie wysoką złożonością obliczeniową całego procesu. Z racji tego, wykorzystano algorytm przeszukiwania najmniejszych odległości o nazwie BallTree. Metodyka działania tego algorytmu polega na dzieleniu danych na hierarchiczną strukturę „kul” (w sensie geometrycznym) otaczających grup punktów, co umożliwia szybkie wyszukiwanie najbliższych sąsiadów poprzez eliminację dużych obszarów przestrzennych. Po zimplementowaniu danych z zapytań Google Maps API oraz algorytmu BallTree, główny zbiór danych został powiększony o 6 nowych zmiennych:

...	Szerokość geograficzna	Długość geograficzna	Odległość do metra M1 (m)	Odległość do metra M2 (m)	Odległość do większego sklepu (m)	Odległość do terenu zielonego (m)
...	52,2516	21,0389	2671	360	184	74
...	52,2228	21,0112	559	1385	64	159
...	52,1861	21,0648	2877	6394	305	304
...	52,2331	20,9413	4283	704	291	91

Tabela 2: Przykładowe wiersze nowo wprowadzonych atrybutów geolokalizacyjnych

Ostatnim etapem z zakresu ekstrakcji zmiennych, było wyodrębnienie poszczególnych zmiennych, tak aby przyjmowały wartości liczbowe. W przypadku zmiennej *cena*, niezbędna była również konwersja niektórych jednostek, gdyż te były podane w dolarach lub euro. W takich przypadkach, obliczono nowe wartości cen mieszkań po kursie z dnia, w którym dane były

zbierane (23/12/2024). Z kolumny *informacje dodatkowe* wyciągnięto wszystkie udogodnienia, a następnie kolejno zamieniono na wartości binarne w nowych kolumnach.

Po dogłębnej analizie ekstrakcyjnej wyróżniono następujące kolumny:

- *Cena za m²* (zmienna ciągła)
- *Cena* (zmienna ciągła)
- *Opis* (format tekstowy)
- *Powierzchnia* (zmienna ciągła)
- *Liczba pokoi* (zmienna dyskretna)
- *Piętro* (zmienna dyskretna)
- *Czynsz* (zmienna ciągła)
- *Rynek* (zmienna kategoryczna)
- *Rok budowy* (zmienna ciągła)
- *Stan wykończenia* (zmienna kategoryczna)
- *Dzielnica* (zmienna kategoryczna)
- Zmienne binarne (1 = obecność udogodnienia): *Winda, Balkon, Domofon, Drzwi antywłamaniowe, Mieszkanie dwupoziomowe, Garaż, Internet, Klimatyzacja, Kuchenka, Łódówka, Meble, Miejsce parkingowe, Monitoring, Ochrona, Oddzielna kuchnia, Ogródek, Okna antywłamaniowe, Piekarnik, Piwnica, Pomieszczenie użytkowe, Pralka, Rolety antywłamaniowe, System alarmowy, Taras, Telefon, Telewizja kablowa, Telewizor, Teren zamknięty, Videofon, Zmywarka*
- *Długość geograficzna* (zmienna ciągła)
- *Szerokość geograficzna* (zmienna ciągła)
- *Odległość do najbliższej stacji metra M1* (zmienna ciągła)
- *Odległość do najbliższej stacji metra M2* (zmienna ciągła)
- *Odległość do najbliższego większego sklepu* (zmienna ciągła)
- *Odległość do najbliższego obszaru zielonego* (zmienna ciągła)

Po fazie inżynierii cech ostateczna liczba wszystkich mieszkań wynosiła około 14 tys.

2.2 Eksploracja, filtrowanie oraz grupowanie danych

Analizę eksploracyjną danych rozpoczęto od identyfikacji braków w pierwotnym zbiorze, w którym największą liczbę niekompletnych obserwacji (oprócz kluczowych zmiennych) odnotowano dla zmiennych takich jak *czyszcz*, *stan wykończenia*, *dzielnica*, *rok budowy* oraz *piętro*. Po usunięciu wierszy, w których nie było wartości *ceny za metr kwadratowy* lub dokładnego adresu nieruchomości, liczba braków uległa istotnemu zmniejszeniu, co przedstawiono poniżej:

- *Czyszcz*: redukcja liczby braków z 5995 do 4678
- *Stan wykończenia*: redukcja z 2231 do 1619
- *Rok budowy*: redukcja z 1256 do 846
- *Piętro*: redukcja z 987 do 646
- *Dzielnica*: redukcja z 2364 do 52

Ze względu na dużą skalę deficytów w zmiennej dotyczącej czyszcza, kolumna ta została całkowicie usunięta ze zbioru. W dalszych etapach zastosowano imputację medianową dla zmiennych ilościowych – rok budowy i piętro, natomiast brakujące dane w zmiennej jakościowej stan wykończenia uzupełniono przez wprowadzenie dodatkowej kategorii *brak informacji*. Rekordy bez wskazania dzielnicy zostały wyeliminowane, ponieważ brak tej informacji wskazywał na umiejscowienie nieruchomości poza obszarem Warszawy. Dodatkowo oceniono współzależności między zmiennymi poprzez wyodrębnienie par o największych wartościach bezwzględnych współczynnika korelacji.

Wśród par zmiennych, wykazujących bardzo silne zależności, znalazły się między innymi:

- *Wideofon i domofon*: korelacja 1,0
- *Drzwi antywłamaniowe i okna antywłamaniowe*: korelacja 1,0
- *Garaż i miejsce parkingowe*: korelacja 1,0
- *Ochrona i monitoring*: korelacja 1,0
- *Telewizja kablowa i internet*: korelacja 0,85
- *Piekarnik i lodówka*: korelacja 0,86

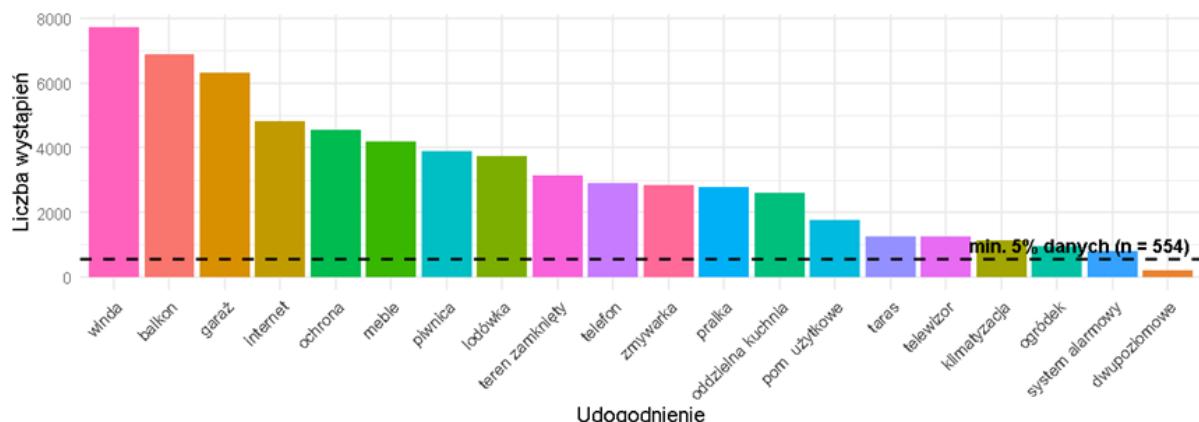
Ze względu na przekroczenie przyjętego progu korelacji rzędu 0,8 - podjęto decyzję o eliminacji zmiennych: *wideofon*, *drzwi antywłamaniowe*, *miejsce parkingowe*, *monitoring*, *telewizja kablowa* oraz *piekarnik*. Jednocześnie postanowiono zachować w analizie te zmienne, których współczynniki korelacji nie przekraczały przyjętego progu, lecz znajdowały się w jego bezpośrednim sąsiedztwie — tak, aby nie pozbywać się modelu potencjalnie istotnych informacji.

Po przeprowadzeniu wstępnej analizy braków danych oraz eliminacji zmiennych o wysokich współczynnikach korelacji, przystąpiono do etapów filtrowania oraz wstępnego grupowania danych, co miało na celu dalsze poprawienie jakości zbioru oraz ograniczenie wpływu wartości odstających na ostateczne wyniki analizy rynku mieszkaniowego.

W ramach tego etapu zastosowano następujące kryteria:

- Skrajne wartości *ceny za metr kwadratowy*, odpowiadające górnemu i dolnemu 1% obserwacji, wyeliminowano, usuwając rekordy, dla których wartość tej zmiennej wynosiła poniżej 10 416 PLN lub powyżej 42 076 PLN.
- Dla zmiennej odpowiadającej *cenie* zastosowano mniej restrykcyjną filtrację na poziomie 0,5%, co skutkowało eliminacją obserwacji dotyczących mieszkań o cenach mniejszych niż 382 000 PLN oraz przekraczających 4 877 600 PLN.
- Analogiczne kryteria przyjęto w przypadku zmiennej *powierzchni* – wyeliminowano rekordy odpowiadające nieruchomościom o powierzchni mniejszej niż 20 m² oraz większej niż 170 m².
- W zmiennej *rok budowy* usunięto wiersze, których wartość była poza przedziałem od 1860 do 2027 roku.
- Zmienna *liczba pokoi* została przekształcona poprzez pogrupowanie wszystkich obserwacji wskazujących wartość większą lub równą 5 - do wspólnej kategorii oznaczonej jako „5+”, co miało na celu uwzględnienie niskiej częstotliwości występowania mieszkań charakteryzujących się większą liczbą pokoi.

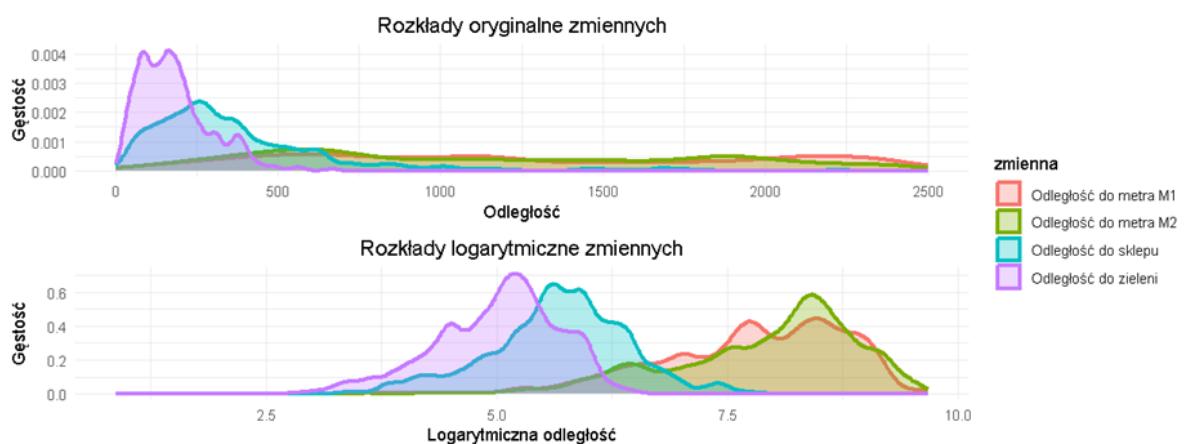
Takie podejście umożliwiło stworzenie bardziej stabilnego i reprezentatywnego zbioru danych, który stanowi solidną podstawę do dalszych etapów grupowania i filtracji zmiennych.



Rysunek 6: Liczba mieszkań z poszczególnymi udogodnieniami

Źródło: Opracowanie własne

Na zaprezentowanym wykresie (Rysunek 6) przedstawiono częstotliwość występowania poszczególnych udogodnień, uszeregowanych od najpopularniejszych do najrzadszych. Można zaobserwować, że do najczęściej oferowanych udogodnień należą między innymi *winda*, *balkon* czy *garaż*, natomiast stosunkowo rzadziej spotykane są na przykład *klimatyzacja* czy *system alarmowy*. Analiza wykazała również, że niektóre zmienne, takie jak mieszkania *dwupoziomowe*, występują znacznie poniżej przyjętego progu minimalnego, wynoszącego 5% wszystkich obserwacji (w niniejszym przypadku odpowiadającego około 554 rekordom). Z tego powodu kategoria mieszkania *dwupoziomowe* została wyeliminowane z dalszych etapów analizy, aby uniknąć błędów związanych z nadmierną dysproporcją w klasach zmiennej (tzw. problem niezrównoważenia danych). Dzięki takiemu podejściu możliwe było zachowanie spójności i wiarygodności uzyskanych wyników, a także skoncentrowanie się na charakterystykach częściej występujących na rynku mieszkaniowym.



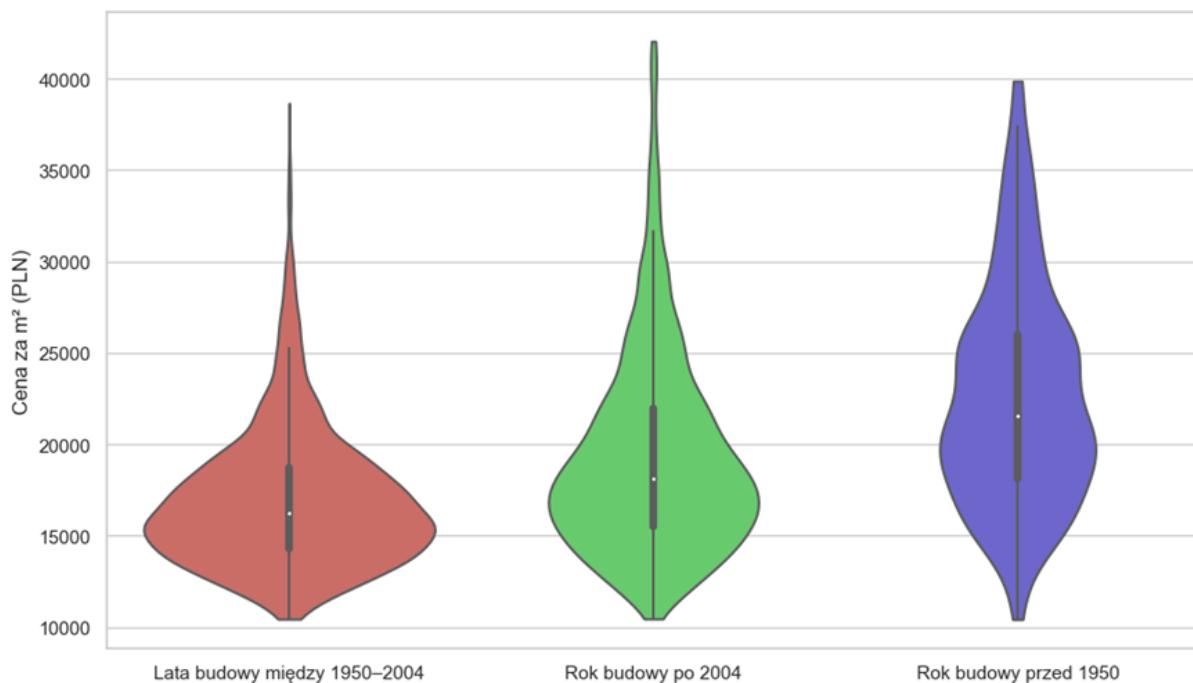
Rysunek 7: Porównanie rozkładów zmiennych odległości

Źródło: Opracowanie własne

Na kolejnych wykresach (Rysunek 7) przedstawiono rozkłady gęstości czterech zmiennych związanych z odlegością od wybranych punktów infrastruktury (stacje metra linii M1 i M2, najbliższy sklep oraz tereny zielone). Górnny wykres odzwierciedla pierwotne wartości, ujawniając silną prawostronną skośność w przypadku odległości od sklepu oraz od zieleni. Po pozostałe dwie zmienne (odległość do metra linii M1 i M2) przyjmują rozkład bardziej zbliżony do jednostajnego, gdyż – po przycięciu skali do 2500 metrów – obserwacje te rozkładają się dość jednorodnie. Taka obserwacja sugeruje dużą liczbę supermarketów oraz terenów zielonych na obszarze Warszawy, co znajduje potwierdzenie w fakcie, iż infrastruktura metra, licząca obecnie 38 stacji, jest kilkukrotnie mniej liczna niż wspomniane punkty użytkowe. Na dolnym wykresie zastosowano transformację $\log(1+x)$ parametrów, której celem jest zmniejszenie wpływu wartości odstających i zbliżenie rozkładów do bardziej symetrycznej formy. Przekształcone wartości posiadają mniej skośne rozkłady, co może ułatwiać dalsze etapy modelowania statystycznego

i uczenia maszynowego, zapobiegając problemom związanym z dominacją skrajnie dużych odległości w analizie. Ujednolicone pod względem kształtu rozkłady zwiększą czytelność danych i pozwalają na lepsze odwzorowanie różnic w odległościach pomiędzy poszczególnymi punktami użyteczności miejskiej.

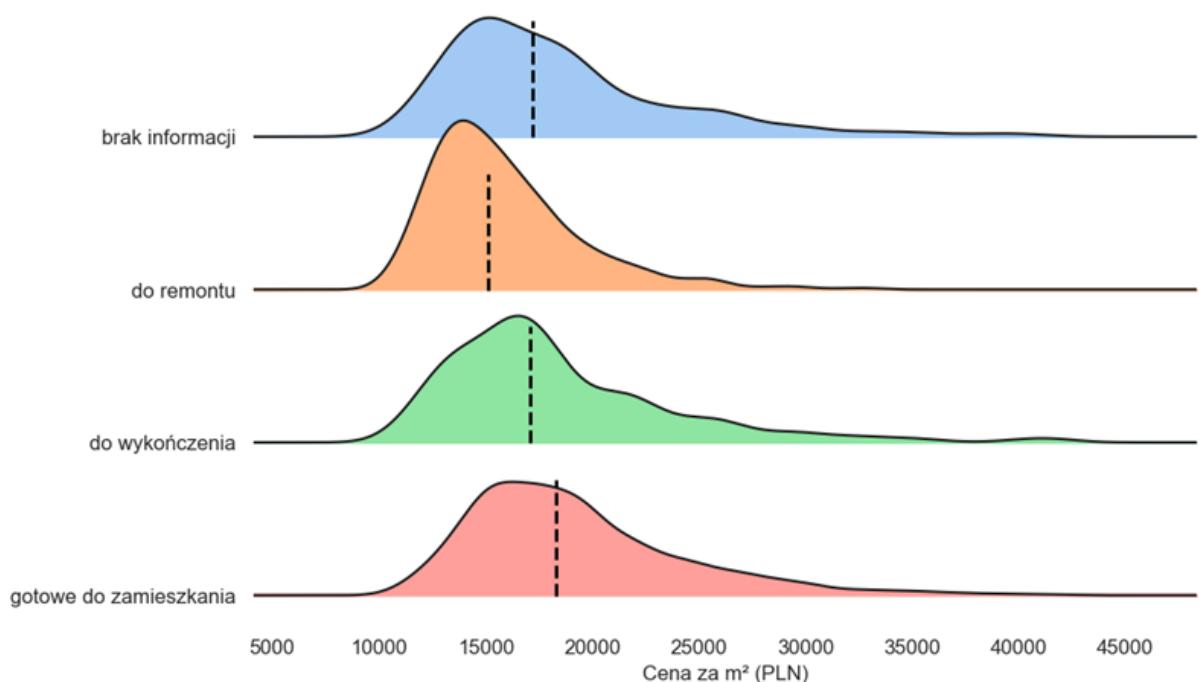
Zmienna *rok budowy* – po wcześniejszym usunięciu skrajnych wartości – obejmowała zakres lat od 1860 do 2027. Aby możliwie najtrątniej uchwycić różnice w cenach za metr kwadratowy w zależności od wieku nieruchomości, wykorzystano drzewo regresyjne, dzielące zmienną ciągłą na mniejsze przedziały tak, aby zmaksymalizować różnice między średnimi cenami w poszczególnych segmentach. W rezultacie uzyskano trzy grupy: mieszkania zbudowane przed rokiem 1950, między 1950 a 2004 rokiem oraz po roku 2004.



Rysunek 8: Rozkład ceny za metr kwadratowy względem pogrupowanej zmiennej *rok budowy*
 Źródło: Opracowanie własne

Zaprezentowany powyżej wykres wiolinowy (Rysunek 8) przedstawia rozkłady cen za metr kwadratowy w trzech kategoriach. Wyniki wskazują na wyraźne zróżnicowanie pomiędzy najstarszym segmentem a lokalami młodszymi, co można wyjaśnić zarówno odmiennymi standardami technicznymi, jak i częstszym umiejscowieniem starszych budynków w centralnych częściach miasta, co przekłada się na wyższe ceny. Z kolei porównanie dwóch pozostałych przedziałów (1950–2004, powyżej 2004) ujawnia mniejsze, aczkolwiek w dalszym ciągu istotne różnice w poziomie cen, które mogą wynikać z postępu technologicznego czy nowocześniejszych projektów architektonicznych w nieruchomościach wybudowanych po 2004 roku. Ponadto, rozkłady w tych kategoriach cechują się wyższą koncentracją wartości w porówna-

niu z nieruchomościami sprzed 1950 roku, co wskazuje, że starsze mieszkania odznaczają się zdecydowanie większą zmiennością cen za metr kwadratowy.



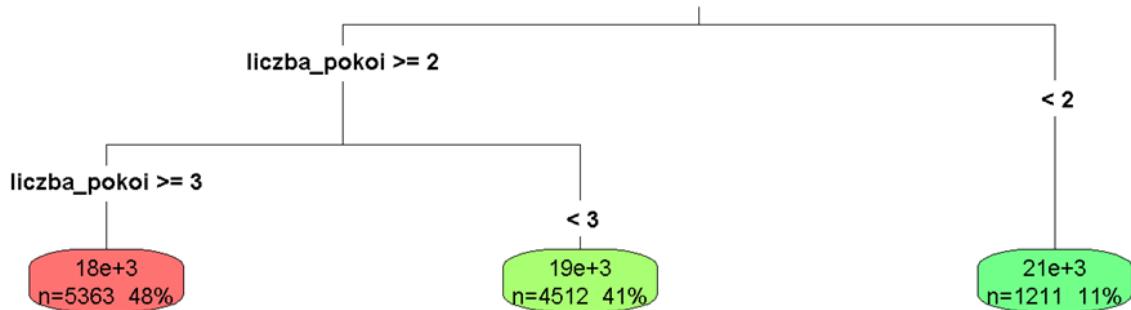
Rysunek 9: Rozkład ceny za metr kwadratowy względem zmiennej *stan wykończenia* z zaznaczoną medianą

Źródło: Opracowanie własne

Niezbędna, w kontekście analizy danych, stała się również transformacja braków danych w zmiennej *stan wykończenia*. W takich przypadkach zmienna przyjmowała nową kategorię o nazwie "brak informacji". Wykres (Rysunek 9) ilustruje rozkłady cen za metr kwadratowy w czterech kategoriach, z zaznaczeniem mediany cen przy użyciu przerywanej linii. Analiza ujawnia wyraźne różnice pomiędzy poszczególnymi kategoriami. Rozkład cen w grupie „do remontu” przesunięty jest w kierunku niższych wartości, co przekłada się na zauważalnie niższą medianę. Z kolei w przypadku mieszkań „gotowych do zamieszkania” obserwuje się najwyższą medianę oraz przesunięcie rozkładu w stronę wyższych cen, co odzwierciedla ich wyższy standard. Kategoria „do wykończenia” plasuje się pomiędzy wspomnianymi skrajnościami, wskazując na istnienie gradacji cen w zależności od stanu technicznego lokalu. Warto również zwrócić uwagę na „brak informacji” – rozkład tej grupy może być mniej miarodajny ze względu na możliwie niejednorodny stan techniczny lokali, a tym samym szeroką rozpiętość cenową. Wizualizacja potwierdza istotny wpływ stanu wykończenia na cenę nieruchomości, ukazując, że bardziej zaawansowane etapy prac remontowo-budowlanych (bądź ich brak) mogą znacząco kształtować ostateczną wartość mieszkania.

Po wstępny przekształceniu zmiennej *liczba pokoi*, w ramach którego wszystkie war-

tości równe lub większe niż 5 zostały połączone w jedną kategorię ‘5+’, zastosowano drzewo regresyjne, tak aby wyodrębnić główne podgrupy różnicujące cenę za metr kwadratowy mieszkania.



Rysunek 10: Drzewo regresyjne: Podział liczby pokoi na 3 podgrupy

Źródło: Opracowanie własne

Zaprezentowany na wykresie podział pokazuje, że model w pierwszej kolejności rozdzielił mieszkania jednopokojowe (tzw. kawalerki) od lokali wielopokojowych, uwypuklając fakt, iż niewielkie powierzchnie często wiążą się z wyższymi cenami jednostkowymi – co może wynikać z ograniczonej podaży niedużych lokali na rynku i dużego popytu wśród na bywów. Następnie, w ramach segmentu obejmującego lokale wielopokojowe, drzewo dokonało dalszego podziału na mieszkania dwupokojowe oraz grupę mieszkań posiadających trzy lub więcej pokoi.

W efekcie powstały trzy kategorie:

- Mieszkania jednopokojowe (11% obserwacji, średnia około 21 tys. PLN za m²),
- Mieszkania dwupokojowe (41% obserwacji, średnia około 19 tys. PLN za m²),
- Mieszkania z trzema lub większą liczbą pokoi (48% obserwacji, średnia około 18 tys. PLN za m²).

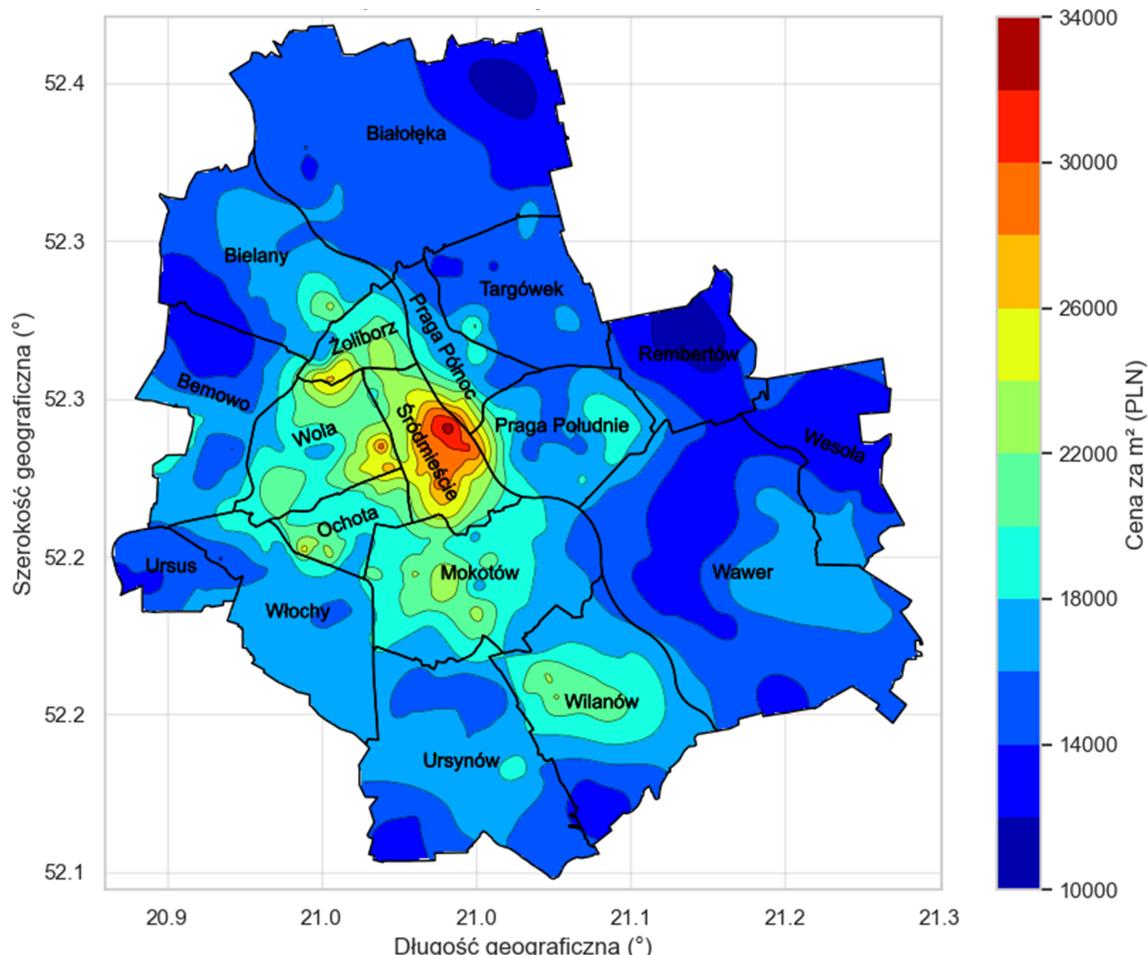
Na podstawie przeprowadzonych testów t-studenta oceniono również, czy różnice średnich cen za metr kwadratowy pomiędzy poszczególnymi grupami zmiennych binarnych są istotne statystycznie. Rezultaty testów pokazują, że dla znacznej części analizowanych cech zaobserwowano statystycznie istotne różnice ($p < 0,05$), choć w niektórych przypadkach różnice wskazywały na konieczność porzucenia zmiennych. Przykładowo, w analizie zmiennej *rynek* średnie wartości dla grup wtórnej i pierwotnej wynosiły odpowiednio 18 639,52 PLN oraz 18 637,70 PLN, co przekłada się na bardzo niewielką różnicę średnich (1,82 PLN) oraz statystykę testową równą 0,015 ($p\text{-value} = 0,988$). Podobne wyniki uzyskano dla zmiennych *kuchenka*, *domofon*, *okna antywłamaniowe* oraz *rolety antywłamaniowe*, gdzie wartości p prze-

kraczały przyjęty poziom istotności, co wskazuje na brak statystycznie istotnych różnic pomiędzy analizowanymi grupami.

W związku z powyższym, zdecydowano o eliminacji tych zmiennych z dalszej analizy. Dodatkowo, zmienna kategoryczna *piętro* została wykluczona z modelu, ponieważ nie wykazywała istotnego wpływu na różnicowanie cen za metr kwadratowy. Takie podejście umożliwiło uproszczenie modelu oraz skoncentrowanie się wyłącznie na zmiennych wykazujących istotny wpływ na wartość nieruchomości, co przyczynia się do trafniejszego odwzorowania mechanizmów rynkowych.

2.3 Wizualizacje rynku mieszkaniowego w Warszawie

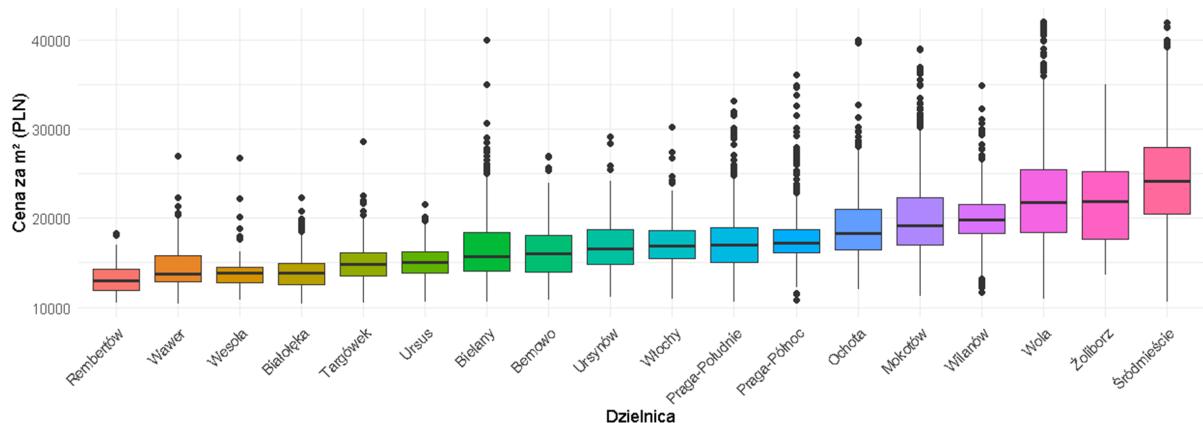
Po wstępnym grupowaniu i filtracji zbadano, jak kształtują się rozkłady cen za metr kwadratowy na terenie Warszawy, aby lepiej zrozumieć specyfikę obszarów silnie wpływających na wartość nieruchomości. Wizualizacja umożliwia identyfikację regionów charakteryzujących się szczególnie wysokimi lub niskimi cenami, co z kolei może wpływać na strategie wyceny oraz dalsze prognozy rynkowe.



Rysunek 11: Mapa izolinii cenowych w Warszawie

Źródło: Opracowanie własne

Na mapie (Rysunek 11) zastosowano interpolację cen za metr kwadratowy, dla losowo wybranych 50 obserwacji z każdej dzielnicy, czyli łącznie 900 obserwacji.¹ Takie podejście umożliwiło wyrównanie wpływu nierównomiernej liczby ofert w poszczególnych obszarach – podczas gdy niektóre dzielnice mogły dysponować nawet kilkuset ofertami, inne posiadały ich znacznie mniej, co mogłoby zniekształcać ogólną analizę. Interpolacja pozwoliła na uzyskanie bardziej spójnego obrazu, który dobrze oddaje lokalne różnice w wartościach ceny za metr kwadratowy. Wizualizacja wyraźnie wskazuje istnienie kilku kluczowych wzorców, które cechują Warszawski rynek mieszkaniowy. Po pierwsze, obszary położone w centrum miasta charakteryzują się najwyższymi cenami – jest to zgodne z intuicyjnym przypuszczeniem, że bliskość centralnych punktów Warszawy, rozwinięta infrastruktura oraz prestiż lokalizacji przekładają się na wyższe stawki za nieruchomości. Po drugie, zauważalny jest także wyraźny gradient cenowy w obrębie miasta – dzielnice położone na zachodzie wykazują tendencję do utrzymywania wyższych cen w porównaniu z obszarami wschodnimi. Taki podział może odzwierciedlać różnice w poziomie usług, dostępności udogodnień oraz ogólnym standardzie życia między tymi częściami miasta. Ostatecznie, na obrzeżach Warszawy ceny za metr kwadratowy są znacznie niższe niż w centrum, co dodatkowo potwierdza wpływ lokalizacji oraz odległości od centralnych obszarów na wycenę nieruchomości.



Rysunek 12: Cena za metr kwadratowy w poszczególnych dzielnicach

Źródło: Opracowanie własne

Na przedstawionym wykresie typu boxplot (Rysunek 12) zaprezentowano rozkład cen za metr kwadratowy w różnych dzielnicach Warszawy. Taka forma wizualizacji pozwoliła zobrazować znaczące dzielnice na rynku mieszkaniowym, co w dużej mierze ułatwia ocenę podziału dokonanego przez algorytm.

¹Z każdej z 18 dzielnic pobrano po 50 obserwacji, co łącznie daje $18 \cdot 50 = 900$ przypadków wykorzystanych do interpolacji.

Po zbudowaniu drzew regresyjnych, algorytm w celu maksymalizacji różnic średnich między grupami, podzielił zmienną *dzielnicą* na 3 zasadnicze podgrupy:

- **Podgrupa 1 – Dzielnice peryferyjne:**

W tej grupie znalazły się dzielnice o najniższych wartościach porządkowych (*Rembertów, Białołęka, Wesoła, Wawer, Targówek, Ursus*). Obszary te charakteryzują się stosunkowo najniższymi cenami za metr kwadratowy, co jest zgodne z ich położeniem oddalonym od centralnych punktów miasta oraz ograniczonym dostępem do rozbudowanej infrastruktury.

- **Podgrupa 2 – Dzielnice o średnim zasięgu:**

Do tej kategorii zakwalifikowano dzielnice o wartościach porządkowych średniego poziomu, takie jak *Bemowo, Bielany, Ursynów, Włochy, Praga-Południe, Praga-Północ oraz Ochota*. Obserwowany na wykresie rozkład cen w tej grupie wykazuje umiarkowane wartości, z większą rozpiętością, co odzwierciedla zróżnicowane cechy tych obszarów pod względem dostępu do udogodnień oraz jakości życia.

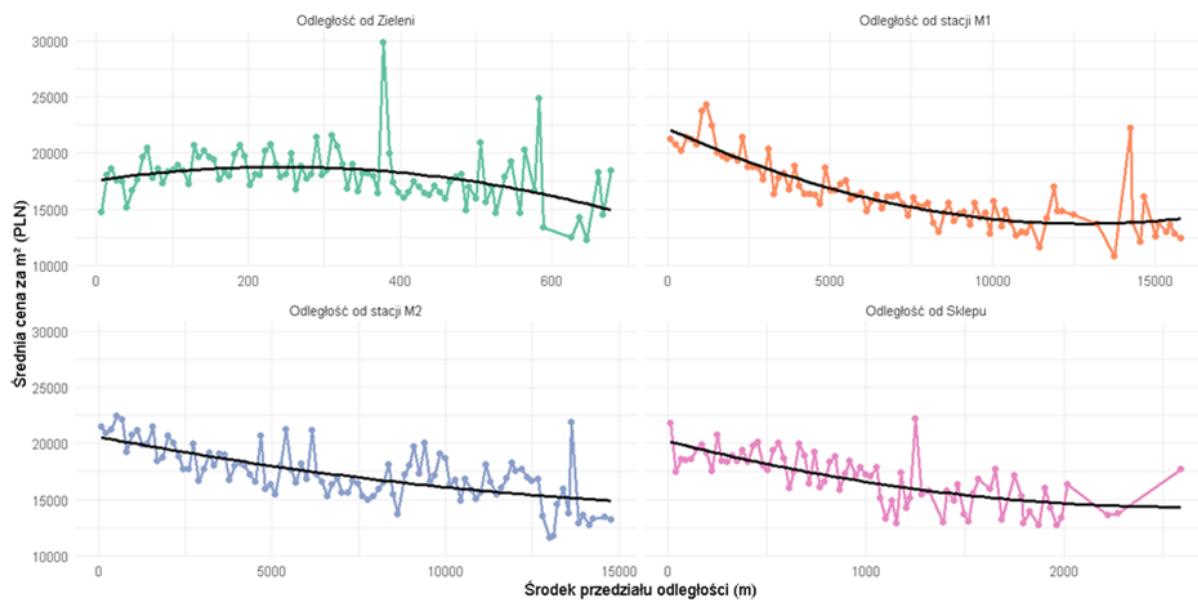
- **Podgrupa 3 – Dzielnice centralne i prestiżowe:**

Do tej grupy zaliczono dzielnice o najwyższych wartościach porządkowych, reprezentowane między innymi przez *Wilanów, Mokotów, Żoliborz, Wolę oraz Śródmieście*. W tych obszarach ceny za metr kwadratowy są najwyższe, co wynika z doskonałej lokalizacji, dobrze rozwiniętej infrastruktury oraz wysokiego standardu oferowanych nieruchomości.



Rysunek 13: Chmura najczęściej występujących w opisach ogłoszeń słów
 Źródło: Opracowanie własne

Na kolejnym rysunku zaprezentowano chmurę słów (Rysunek 13), w której wielkość czcionki poszczególnych wyrażeń odzwierciedla częstotliwość ich występowania w opisach ofert sprzedaży mieszkań. Zauważalna dominacja takich słów jak „centrum”, „minut” czy „spacerem” wskazuje na wagę, jaką sprzedający przywiązuje do podkreślania lokalizacji oraz dogodności komunikacyjnych – często podkreślana jest możliwość szybkiego dotarcia do kluczowych punktów miasta bądź fakt, że wiele miejsc jest dostępnych pieszo. Duża częstotliwość wyrazów związanych z samą nieruchomościami, takich jak „mieszkania”, „budynek” czy „pokój”, jest naturalna ze względu na specyfikę ofert, jednak pozwala również wychwycić najczęściej eksponowane zalety wnętrza (np. obecność „salonu” czy „balkonu”). W połączeniu z wcześniejszymi analizami dotyczącymi rozkładu cen oraz lokalizacji, chmura słów dostarcza kolejnego kontekstu, ukazując, w jaki sposób sprzedający starają się wyróżnić swoje oferty. Wyrazy odwołujące się do „parków”, „okolicy” czy „zieleni” sugerują, że bliskość terenów rekreacyjnych jest istotna dla potencjalnych nabywców i stanowi argument przekładający się na atrakcyjność ofert. Takie podejście dowodzi, że poza ceną i parametrami ściśle technicznymi, kluczowe są także czynniki związane z komfortem życia i udogodnieniami dostępnymi w sąsiedztwie – co wyraźnie koresponduje z dotychczasowymi spostrzeżeniami dotyczącymi istotności lokalizacji oraz odległości od kluczowych punktów infrastruktury.



Rysunek 14: Średnia cena w zależności od odległości z trendem wielomianowym

Źródło: Opracowanie własne

Na ostatnim wykresie (Rysunek 14) ukazano, w jaki sposób średnia cena za metr kwadratowy zmienia się wraz ze wzrostem odległości od wybranych udogodnień, takich jak stacja metra (linie M1 i M2), tereny zielone czy najbliższy sklep. Dane zostały pogrupowane w 100 równych przedziałów odległości dla każdej zmiennej, co pozwoliło na zagregowanie obserwa-

cji i zredukowanie wpływu pojedynczych punktów odstających. Nałożona na każdy wykres krzywa trendu – będąca wielomianem drugiego stopnia – umożliwia uchwycenie nieliniowych zależności pomiędzy ceną a odległością. W kilku przypadkach widoczna jest tendencja ścisłe malejąca, szczególnie wyraźna w niewielkiej odległości od kluczowych punktów, co potwierdza hipotezę o istotnym wpływie bliskości infrastruktury na wycenę nieruchomości. Jednocześnie w dalszych odległościach obserwowany trend może ulegać stabilizacji lub nawet nieznacznym wahaniom, wskazując na mniejsze znaczenie dodatkowych kilometrów oddalenia. Tak opracowana wizualizacja stanowi cenne uzupełnienie wcześniejszych analiz rozkładu cen i potwierdza, że obecność atrakcyjnych udogodnień w niewielkiej odległości od lokali często przekłada się na wyraźnie wyższe stawki za metr kwadratowy.

W ten sposób przetworzone dane posłużą w kolejnym rozdziale do zbudowania modelu uczenia maszynowego przewidującego cenę mieszkań, a także testu tego zbioru na fikcyjnym zbiorze danych, na podstawie którego możliwe będzie oszacowanie wpływu odległości od metra na cenę za metr kwadratowy mieszkania.

Rozdział III

Metody oszacowania wpływu bliskości metra na Rynek Nieruchomości

Celem ostatniego rozdziału jest przedstawienie procesu budowy modeli uczenia maszynowego do prognozowania ceny za metr kwadratowy mieszkania oraz ocena wpływu planowanej linii metra M4 na rynek nieruchomości w Warszawie. W podrozdziale 3.1 opisano przygotowanie i transformację danych, wybór cech oraz optymalizację hiperparametrów dla wybranych algorytmów regresji. W podrozdziale 3.2 przedstawiono metodę wyznaczania wpływu odległości od stacji metra na cenę, z wykorzystaniem ważności cech w modelach drzewiastych oraz teoretycznych scenariuszy zmieniających odległość. Natomiast w podrozdziale 3.3 zaprezentowano wizualizacje w postaci map gradientowych ilustrujące prognozowane zmiany cen mieszkań w przestrzeni miejskiej.

3.1 Przegląd danych i budowanie modeli uczenia maszynowego

Po zakończeniu etapu kompleksowego przygotowania danych, przystąpiono do konstrukcji modeli uczenia maszynowego. Celem było prognozowanie ceny za metr kwadratowy na podstawie szeregu dostępnych atrybutów. Aby umożliwić algorytmom przetwarzanie cech nominalnych, każdą zmienną kategoryczną przekształcono do postaci binarnej z pominięciem jednej kategorii dla każdej zmiennej, tak aby wyeliminować zjawisko współliniowości danych. Finalny zbiór objaśniający składa się z 32 zmiennych, które podzielono na następujące grupy:

- **Udogodnienia mieszkania (zmienne binarne)** — winda, garaż, internet, klimatyzacja, lodówka, meble, ochrona, balkon, oddzielna kuchnia, ogródek, piwnica, pomieszczenie użytkowe, pralka, system alarmowy, taras, telefon, telewizor, teren zamknięty, zmywarka;
- **Stan wykończenia (zmienna kategoryczna zakodowana binarnie)** — w modelu wykorzystano trzy zmienne binarne: „do remontu”, „do wykończenia” oraz „do zamieszkania”;
- **Liczba pokoi (zmienna kategoryczna zakodowana binarnie)** — w modelu wykorzystano dwie zmienne binarne: mieszkania 1-pokojowe oraz 2-pokojowe;
- **Rok budowy (zmienna kategoryczna zakodowana binarnie)** — w modelu wykorzystano dwie zmienne binarne: budynki wybudowane przed 1950 rokiem oraz po 2005 roku;
- **Dzielnica (zmienna kategoryczna zakodowana binarnie)** — w modelu wykorzystano dwie zmienne binarne: dzielnice peryferyjne oraz dzielnice o średnim zasięgu;

- **Odległości (zmienne ciągłe)** — logarytmy odległości od stacji metra M1, metra M2, terenów zielonych oraz najbliższego sklepu.

W kolejnym kroku dla każdego z zestawów cech przeprowadzono trening modeli, aby ocenić ich zdolność do trafnego przewidywania ceny za metr kwadratowy. Dzięki zastosowaniu tak skonstruowanego zbioru zmiennych możliwe stało się uchwycenie zarówno jakościowych, jak i przestrzennych aspektów nieruchomości. Spośród wyróżnionych w pierwszym rozdziale modeli wybrano do analizy: *Support Vector Regression*, *XGBoost*, *Random Forest* oraz *Deep Neural Network*. Wcześniej jednak dokonano optymalizacji hiperparametrów metodą szukania po siatce (ang. Grid Search), co umożliwiło znalezienie odpowiednich parametrów dla każdego z modeli:

- **Support Vector Regression:**

- Współczynnik kary za błędy ($C = 10\ 000$);
- Epsilon ($\varepsilon = 1000$);
- Jądro ($kernel = rbf$);
- Współczynnik gamma ($gamma = scale$);

- **Random Forest:**

- Liczba drzew ($n_estimators = 150$);
- Maksymalna głębokość drzew ($max_depth = 20$);
- Liczba cech rozważanych przy podziale ($max_features = \sqrt{n}$);

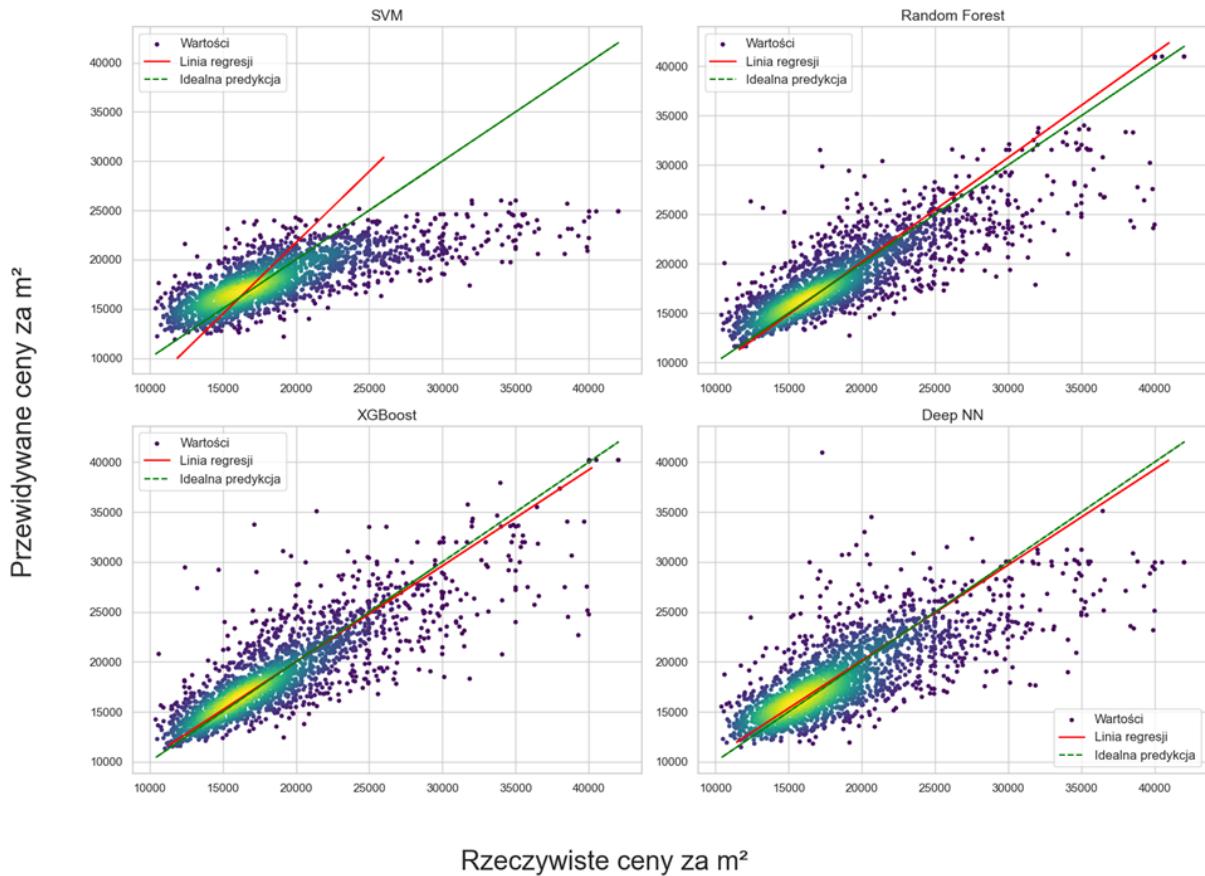
- **XGBoost:**

- Szybkość uczenia ($learning_rate = 0,1$);
- Maksymalna głębokość drzew ($max_depth = 7$);
- Liczba estymatorów ($n_estimators = 200$);

- **Deep Neural Network:**

- Pierwsza warstwa ukryta – 128 neuronów, aktywacja ReLU;
- Współczynnik dropout ($dropout = 0,3$);
- Druga warstwa ukryta – 64 neurony, aktywacja ReLU;
- Współczynnik dropout ($dropout = 0,2$);
- Trzecia warstwa ukryta – 32 neurony, aktywacja ReLU;
- Warstwa wyjściowa – 1 neuron, aktywacja liniowa.

Po zakończeniu optymalizacji hiperparametrów przeprowadzono szczegółową ocenę jakości wytrenowanych modeli. Na Rysunku 15 przedstawiono porównanie wartości rzeczywistych oraz przewidywanych cen za metr kwadratowy dla wcześniejszych czterech algorytmów.

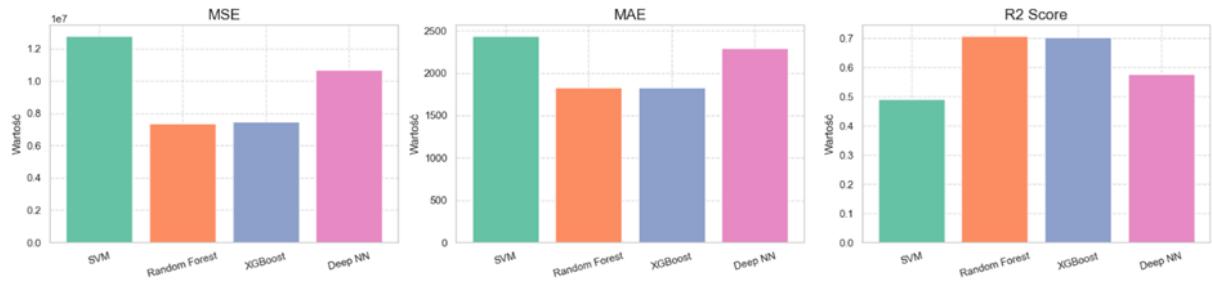


Rysunek 15: Średnia cena w zależności od odległości z trendem wielomianowym

Źródło: Opracowanie własne

Analiza rozkładu punktów wokół prostej $y = x$ wskazuje na ogólną zgodność przewidywań z rzeczywistością w przypadku prawie wszystkich modeli. Największe odchylenia obserwuje się w przypadku *SVR*, gdyż linia trendu nie ma zbliżonego nachylenia do idealnej linii predykcyjnej, co może sugerować niedoszacowanie skrajnych obserwacji. Modele drzewiaste (*Random Forest* i *XGBoost*) oraz *Deep Neural Network* charakteryzują się bardziej zwartą chmurą punktów i nachyleniem regresji zbliżonym do jednostkowego, co świadczy o wyższej trafności ich prognoz.

W kolejnej sekcji przedstawiono kwantyfikację powyższych obserwacji przy użyciu metryk MSE, MAE oraz R^2 , umożliwiając precyzyjne porównanie skuteczności poszczególnych algorytmów. W tym przypadku (Rysunek 16), predykcje *Deep Neural Network* okazały się być znacznie mniej trafne, niż sugerowałby to poprzedni wykres.



Rysunek 16: Porównanie metryk regresji dla modeli

Źródło: Opracowanie własne

Z analizy metryk wynika, że metody drzewiaste wykazują najlepszy kompromis między dokładnością a stabilnością. *Random Forest* i *XGBoost* osiągają najniższe wartości MSE oraz MAE i jednocześnie najwyższe R^2 (0,70). Pośrednie rezultaty oszacowania błędów oraz współczynnika determinacji w przypadku *Deep Neural Network*, wynikają najprawdopodobniej z relatywnie niskiej ilości danych jak na model, który służy głównie znajdowaniu bardzo skomplikowanych nieliniowych zależności. Najsłabsze wyniki modelu *Support Vector Regression* potwierdziły się najwyższymi metrykami błędów oszacowania.

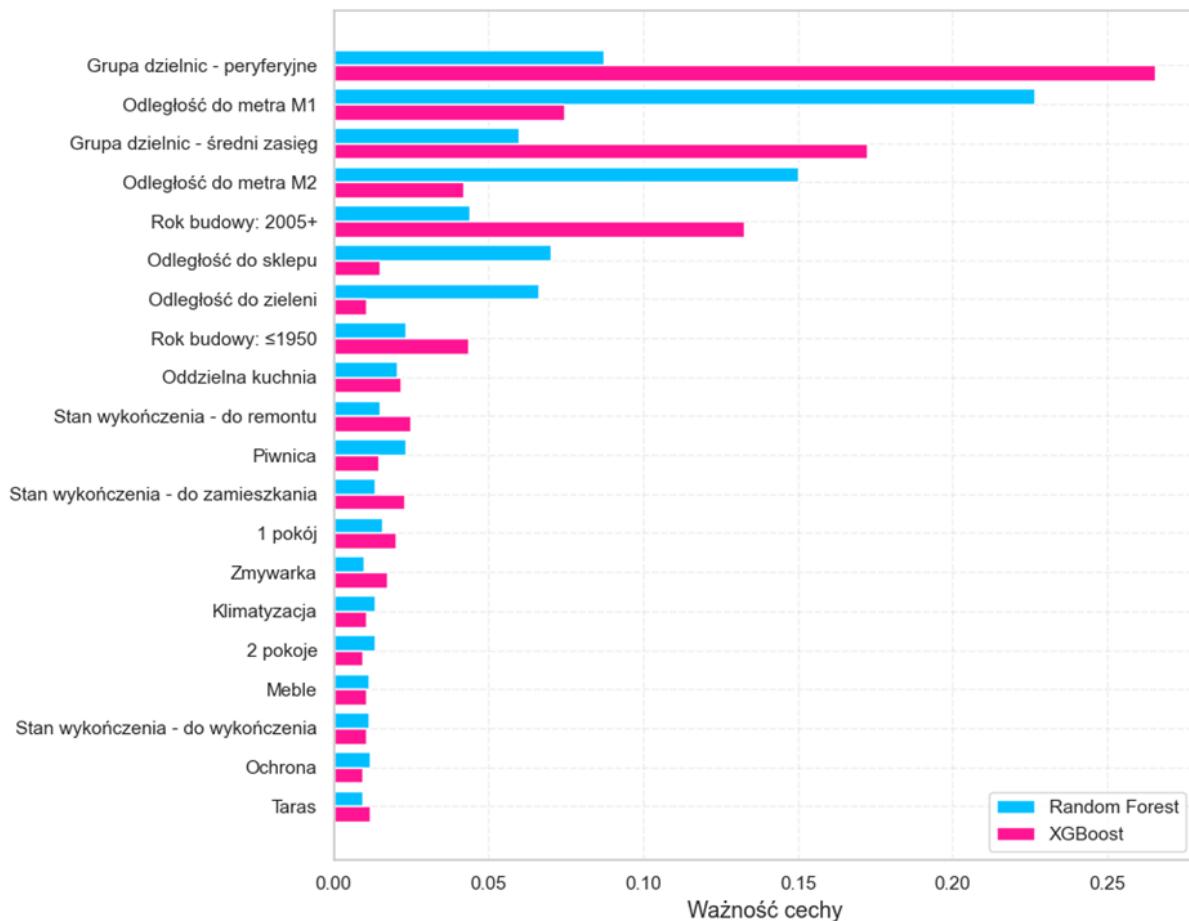
Z tego względu w kolejnych etapach analizy rozważane są tylko modele drzewiaste z grupy esemble, czyli *Random Forest* i *XGBoost*, aby skupić się na kluczowych wnioskach dotyczcych analizy.

3.2 Metoda wyznaczania wpływu linii metra M4 na rynek nieruchomości

Znaczącą zaletą modeli drzewiastych jest ich zdolność do interpretowania ważności poszczególnych zmiennych dla modelu. W kontekście budowy modeli predykcyjnych, umiejętność wyodrębnienia najważniejszych cech może prowadzić do uproszczenia modelu - poprzez skupienie się na kluczowych czynnikach determinujących zmienną objaśnianą. W zakresie ekonomicznym, pozwala to na klarowną komunikację wyników z inwestorami lub zarządem, poprzez zrozumiałe zobrazowanie zmiennych mających realny wpływ na np. wynik finansowy. W tym przypadku, porównanie ważności cech pozwala również na określenie, czy faktycznie: "*Odległość mieszkania od stacji metra jest determinantą jego ceny za metr kwadratowy?*" oraz tego jak silna jest zależność pomiędzy tymi cechami na tle innych parametrów.

Następujący wykres (Rysunek 17) przedstawia porównanie ważności cech dla dwóch najlepszych modeli wyjaśniających *cenę za metr kwadratowy*. Analiza porównawcza doprowadziła do sformułowania istotnych wniosków:

- Niewielka liczba zmiennych odpowiada za wysoką skuteczność predykcyjną modeli. (co odzwierciedla rozkład zgodny z zasadą Pareto)



Rysunek 17: Porównanie ważności cech: Random Forest vs XGBoost

Źródło: Opracowanie własne

- Model wskazuje na szczególne znaczenie zmiennych takich jak: *dzielnicą, odległość od metra, sklepu, zieleni, rok budowy*
- zmienne binarne udogodnień, stanu wykończenia, liczby pokoi mają znacznie mniejszą zdolność predykcyjną (prawdopodobnie służą jako korekta przy końcowych liściach drzew)
- Zmienna *odległość od metra* ma dużo większe znaczenie predykcyjne w modelu Random Forest w porównaniu do XGboost

W dalszej części analizy, po zbudowaniu i ocenie modeli predykcyjnych, przygotowano teoretyczny zbiór zmiennych. W pierwszym kroku określona została charakterystyka „najpopularniejszego mieszkania” na podstawie oryginalnego zbioru danych. W przypadku zmiennych binarnych, dla każdej z nich wybrano kategorię dominującą — wskazującą wartość (0 lub 1), która najczęściej występowała w danych. Natomiast dla zmiennych ciągłych przyjęto średnią wartość danej cechy w zbiorze. Jedyną zmienną, która została celowo zmodyfikowana, była

odległość od stacji metra. Dla potrzeb analizy przypisano jej wartości w postaci ciągu arytmetycznego, obejmującego liczby od 1 do 15 878². Zadbano również o to, aby przygotowany zbiór nie zawierał sprzecznych wartości zmiennych. Przykładowo, jeżeli lokalizacja „najpopularniejszego typu” mieszkania wskazywała na dzielnicę peryferyjną, to zgodnie z zasadami logiki zmienna określająca przynależność do dzielnicy o średnim zasięgu przyjmowała wartość 0. W celu lepszego zobrazowania struktury danych, poniżej przedstawiono teoretyczny zbiór fikcyjnie wygenerowanych obserwacji:

...	dziel_peryferyjna	dziel_sredni_zasieg	odl_metro_M1	odl_metro_M2	odl_zielen	odl_sup_market
...	0	1	1	2683	149	291
...	0	1	2	2683	149	291
...	0	1	3	2683	149	291
...	0	1	4	2683	149	291
...	0	1	5	2683	149	291
⋮	⋮	⋮	⋮	⋮	⋮	⋮
...	0	1	15874	2683	149	291
...	0	1	15875	2683	149	291
...	0	1	15876	2683	149	291
...	0	1	15877	2683	149	291
...	0	1	15878	2683	149	291

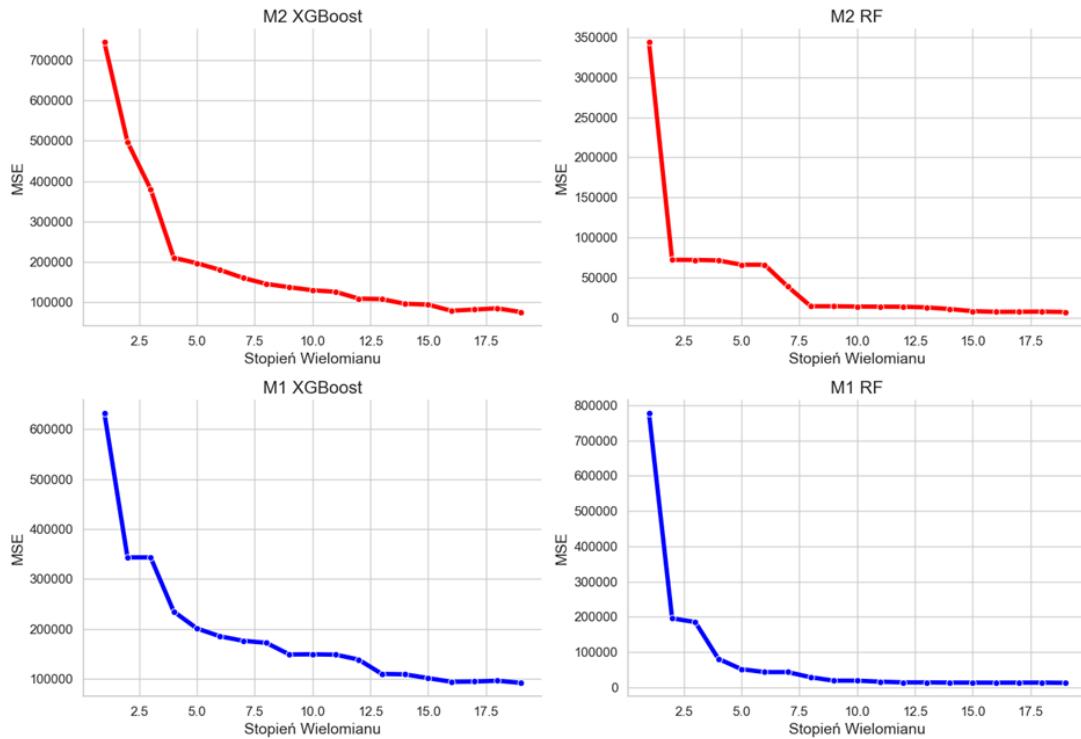
Tabela 3: Przykładowy fragment teoretycznego zbioru danych

Źródło: Opracowanie własne

Analogicznie stworzono zbiór dla zmiennej *odleglosc do metra M2*. Przyjęta konstrukcja zbioru teoretycznego pozwala na pełną kontrolę nad analizowanymi cechami nieruchomości oraz ograniczenie wpływu potencjalnych czynników zakłócających. W rezultacie możliwe stało się przeprowadzenie symulacji, której wyniki w sposób klarowny odzwierciedlają zależności między odległością od metra, a prognozowaną ceną za metr kwadratowy mieszkania (*ceteris paribus*).

W przygotowanych w powyższy sposób zbiorach zastosowano wcześniej skonstruowane modele predykcyjne, uzyskując tym samym teoretyczne oszacowania ceny za metr kwadratowy mieszkania w zależności od zmiennej określającej odległość od stacji metra. Następnie, w celu przybliżenia otrzymanych zależności odpowiednią funkcją interpolacyjną, przeprowadzono analizę wpływu stopnia wielomianu interpolacyjnego na wartość średniego błędu kwadratowego estymacji. Panel wykresów przedstawiony poniżej (Rysunek 18) ilustruje zależność pomiędzy stopniem wielomianu interpolacyjnego a wartością błędu średniokwadratowego przybliżenia.

²Wartość ta została oszacowana na podstawie maksymalnej odległości między planowaną lokalizacją stacji metra M4 a granicą administracyjną Warszawy.



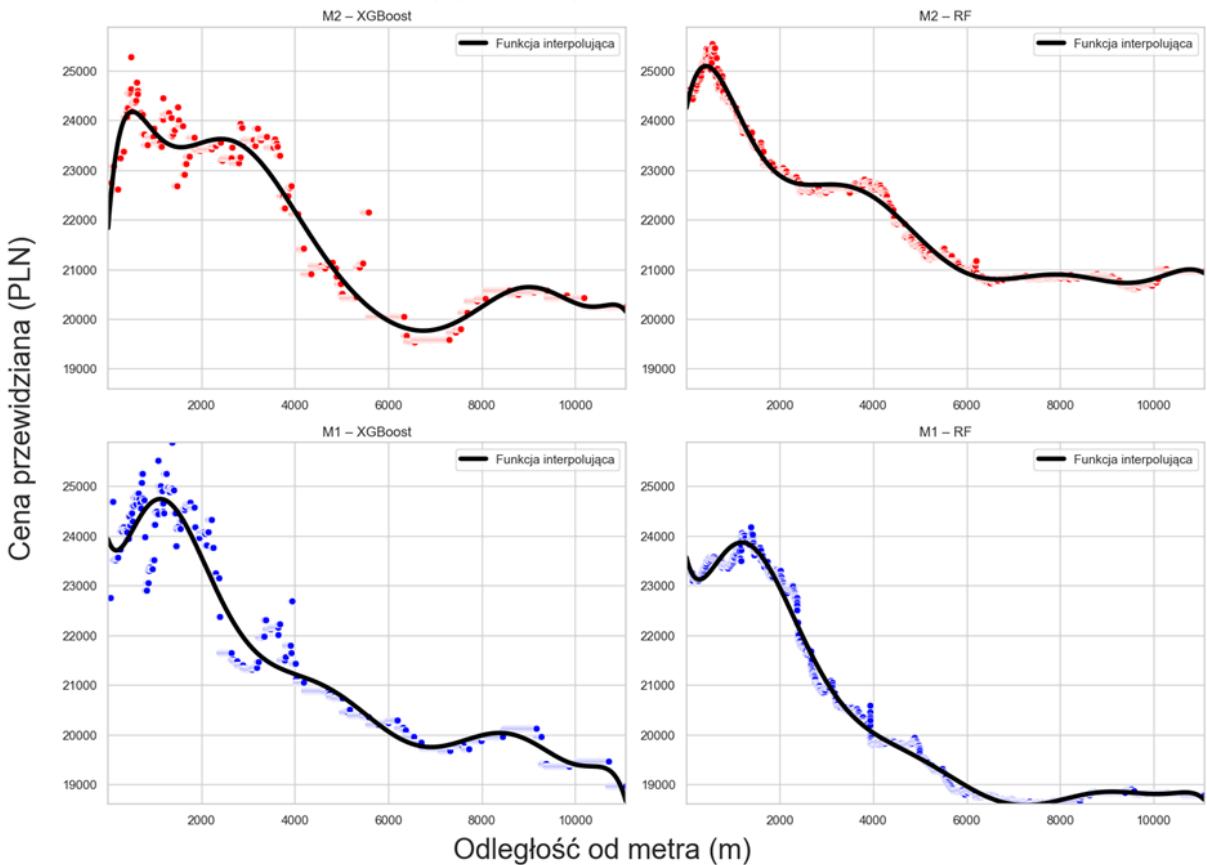
Rysunek 18: Porównanie wartości błędu średniokwadratowego dla różnych modeli

Źródło: Opracowanie własne

Analiza wyników wskazuje, że już przy zastosowaniu wielomianu stopnia piątego funkcja interpolacyjna w sposób satysfakcyjny odwzorowuje wpływ odległości od stacji metra na cenę za metr kwadratowy. Jednocześnie zaobserwowano, że od stopnia trzynastego wzwyż dalsze zwiększanie stopnia wielomianu prowadzi jedynie do nieznacznej redukcji wartości błędu średniokwadratowego. Z uwagi na rosnącą złożoność obliczeniową dla wyższych stopni wielomianu, zdecydowano się na przybliżenie wpływu odległości od metra za pomocą funkcji interpolacyjnej stopnia dziesiątego:

$$f(x) = a_0x^{10} + a_1x^9 + a_2x^8 + a_3x^7 + a_4x^6 + a_5x^5 + a_6x^4 + a_7x^3 + a_8x^2 + a_9x + a_{10}$$

Po estymacji wartości parametrów funkcji interpolacyjnych dla każdego modelu oraz każdej odległości od metra, możliwe stało się graficzne przedstawienie kształtu przebiegu otrzymanych funkcji. W ten sposób uzyskano intuicyjny obraz wpływu bliskości infrastruktury transportowej na wartość nieruchomości. Na kolejnym panelu (Rysunek 19) przedstawiono wyniki oszacowania dla każdego ze zbiorów (M1 i M2) oraz dla każdego modelu (XGBoost i Random Forest) z uwzględnieniem oszacowanej funkcji interpolacyjnej. Ilustracja uwypukla główną różnicę w oszacowaniach obu modeli - Random Forest cechuje się dużo stabilniejszymi prognozami, gdzie każda kolejna jest relatywnie małym odchyleniem od poprzedniej.



Rysunek 19: Porównanie wpływu odległości od metra na cenę za metr kwadratowy

Źródło: Opracowanie własne

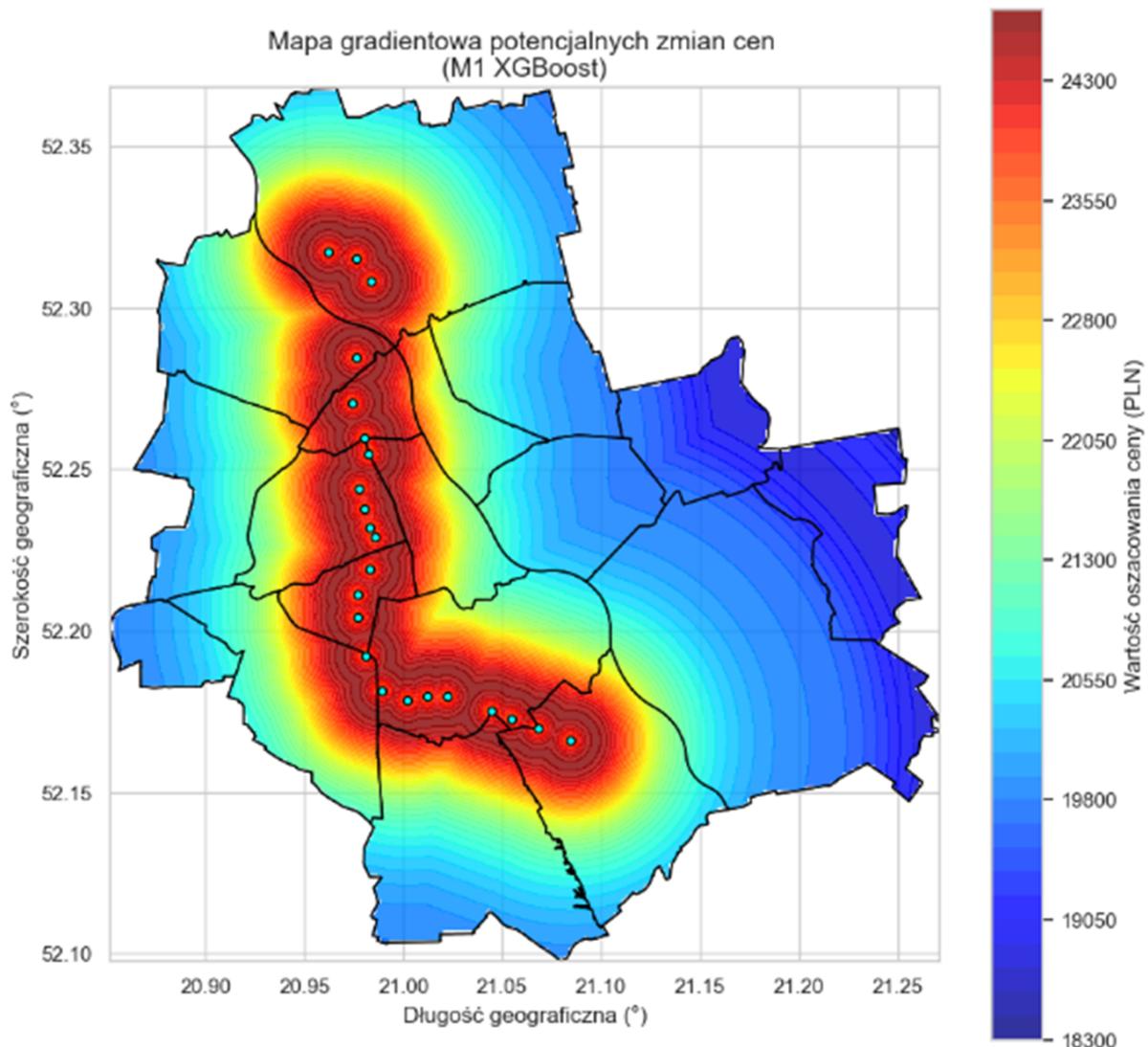
W przypadku modelu XGBoost oszacowania cechują się wyższą heteroskedastycznością, szczególnie wśród odległości do 6 tys. metrów. Przybliżenie przebiegu funkcji dla metra M1 wskazuje na dosyć szybki spadek wartości nieruchomości po przekroczeniu progu 2 tys. metrów. Z drugiej strony, funkcja interpolująca dla linii M2 charakteryzuje się relatywnie łagodniejszym spadkiem, który rozpoczyna się głównie w okolicach 4000 metrów od stacji metra.

Przebieg tych funkcji posłuży w ostatniej części pracy jako punkt wyjściowy do obrzutowania potencjalnej zmiany cen w przyszłości po wybudowaniu linii M4.

3.3 Wizualizacje przewidywanych zmian na rynku mieszkaniowym

Po przeprowadzeniu procesu modelowania, istotnym elementem analizy stało się zwrócenie uwagi na uzyskane wyniki prognoz dotyczących zmian cen mieszkań w przestrzeni miejskiej. W celu przygotowania wizualizacji przedstawiających te zmiany, wykorzystano dane dotyczące dokładnych lokalizacji planowanych stacji metra M4 w Warszawie, udostępnione na oficjalnej stronie Metra Warszawskiego.

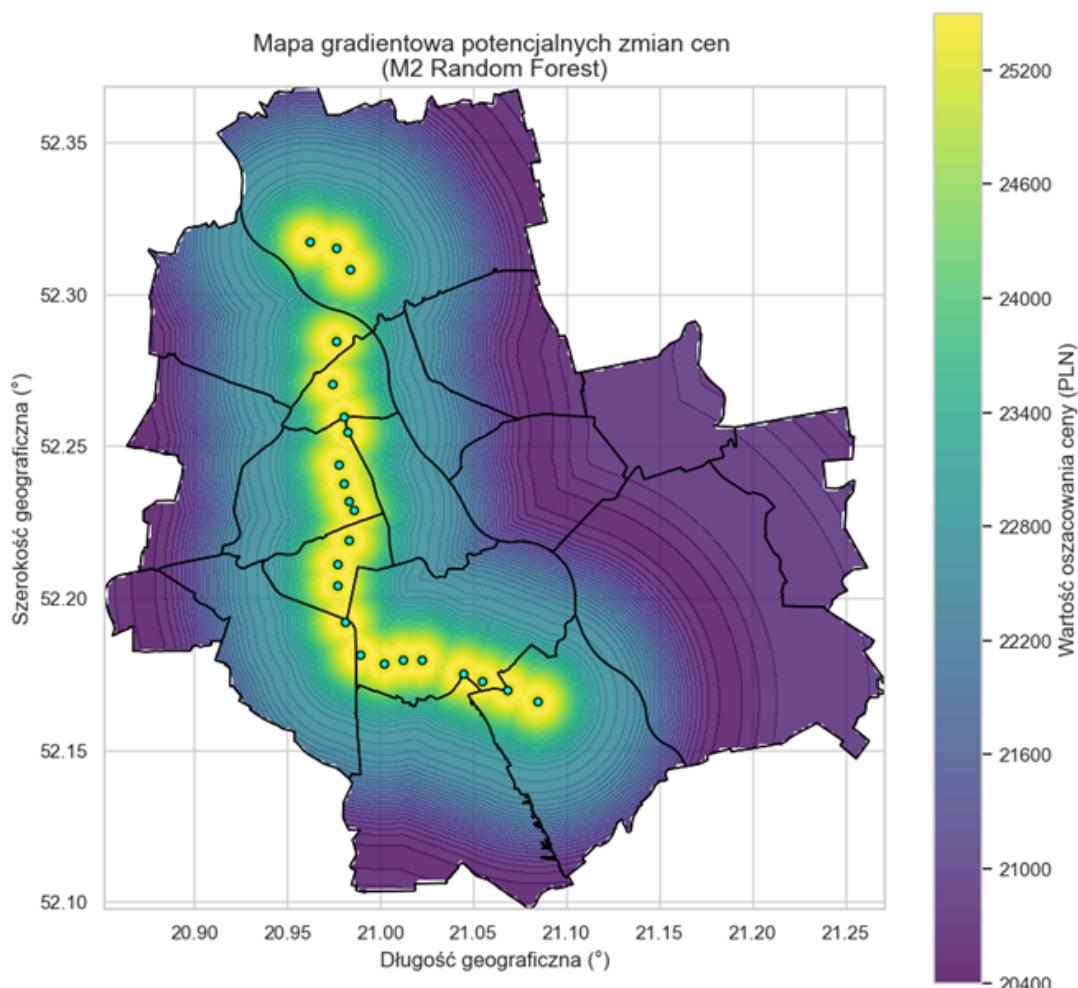
Po wczytaniu współrzędnych geograficznych stacji, przyjęto podejście analogiczne do modelowania propagacji fali – każda stacja została potraktowana jako niezależne źródło wpływu na ceny nieruchomości, emitujące "falę" o kształcie opisany wcześniejszymi oszacowanymi funkcjami interpolacyjnymi. W miejscach, w których oddziaływanie kilku stacji nakładało się na siebie, przyjęto zasadę wyboru wartości maksymalnej – oznacza to, że dla każdego punktu na mapie przypisywana była najwyższa przewidywana wartość wzrostu ceny wynikająca z oddziaływania najbliższej stacji. W efekcie takiego podejścia uzyskano płynne, gradientowe mapy potencjalnych zmian cen mieszkań za metr kwadratowy, umożliwiające intuicyjną ocenę przestrzennych wzorców przewidywanych wartości. Mapy przedstawione w tym rozdziale stanowią bezpośrednią ilustrację wyników przeprowadzonych symulacji i umożliwiają wskazanie obszarów, które w największym stopniu mogą odczuć wzrosty cen wynikające z realizacji nowych inwestycji w infrastrukturę transportową.



Rysunek 20: Mapa gradientowa linii M4 z użyciem modelu (M1 XGBoost)

Źródło: Opracowanie własne

Pierwsza z zaprezentowanych map gradientowych (Rysunek 20) przedstawia prognozowane zmiany cen mieszkań w Warszawie po wybudowaniu nowych stacji linii M4, gdyby cena zachowywała się tak jak oszacowanie modelu *XGBoost* przy użyciu zbioru teoretycznego ze zmieniającą się odlegością od stacji linii M1. Widoczny wpływ na ceny obejmuje stosunkowo szerokie obszary, ze szczególnym nasileniem po zachodniej stronie miasta. Obszary o najwyższych prognozowanych wzrostach cen oznaczone są intensywnymi barwami czerwieni i pomarańcza, wskazującymi lokalizacje o największym potencjale wzrostu wartości nieruchomości. Rozkład przestrzenny zmian wskazuje, że największe wzrosty cen spodziewane są w dzielnicach takich jak Wola, Bemowo oraz Bielany. Najsilniejszy efekt widoczny jest w bezpośrednim sąsiedztwie planowanych stacji metra oraz wzduż ich głównych ciągów komunikacyjnych. Szczególną uwagę należy zwrócić na nowo planowane miejsca stacji, które znajdują się w dalszej odległości od obecnie istniejących sieci metra M1 i M2. Prognozy mogą cechować się wyższą trafnością w obszarach położonych w południowo-zachodniej części Warszawy, w szczególności na pograniczu Mokotowa i Ursynowa, w rejonach granicznych Włoch i Mokotowa, a także na terenie Ochoty.



Rysunek 21: Mapa gradientowa linii M4 z użyciem modelu (M2 Random Forest)

Źródło: Opracowanie własne

Druga z opracowanych grafik (Rysunek 21), tym razem przy użyciu modelu *Random Forest* oraz teoretycznego zbioru danych ze zmieniającą się jedynie odlegością od stacji linii M2, cechuje się dużo szybszym spadkiem cenowym (oznaczonym przejściem z żółci do zielonej) wraz ze zwiększeniem odległości od stacji metra. Jednocześnie, obszary o umiarkowanym wpływie lokalizacji na wartość nieruchomości (zielone strefy) pozostają stosunkowo rozległe – obejmują niemal całą zachodnią część Warszawy oraz fragment jej północnych rejonów. Warto przy tym zwrócić uwagę, że rozległy obszar o średnich i wysokich estymowanych wartościach cenowych w pełni pokrywa centralne dzielnice miasta – w szczególności Śródmieście i dzielnice do niej sąsiadujące po stronie zachodniej. Spośród wszystkich dzielnic wcześniej zaklasyfikowanych jako centralne i prestiżowe, jedynie Wilanów nie został w pełni objęty obszarem średnich lub wysokich estymowanych cen. Jednocześnie do tej grupy dołączyła Ochota, która pierwotnie została zakwalifikowana jako dzielnica o średnim zasięgu – co dodatkowo potwierdza istotny wpływ lokalizacji względem infrastruktury metra na wycenę mieszkań.

Tym samym zakończono analizę przestrzennych wzorców zmian cen, uzyskanych przy wykorzystaniu symulacyjnych zbiorów danych oraz wybranych modeli uczenia maszynowego. Wnioski płynące z przeprowadzonych analiz stanowią podstawę do sformułowania końcowych refleksji i podsumowania najistotniejszych obserwacji.

Zakończenie

Przeprowadzona analiza jednoznacznie potwierdza, że odległość od stacji metra stanowi jeden z najistotniejszych czynników wpływających na kształtowanie się cen nieruchomości mieszkaniowych w Warszawie. Wyniki pracy wskazują, iż planowana rozbudowa linii metra — w szczególności budowa linii M4 — może przyczynić się do istotnych zmian w strukturze cenowej rynku mieszkaniowego, zwłaszcza w tych obszarach miasta, które obecnie charakteryzują się relatywnie niskim poziomem dostępności transportu podziemnego. Efekt ten jest szczególnie zauważalny w zachodnio-południowej części Warszawy, gdzie prognozowany wzrost cen mieszkań w promieniu do 2000 metrów od nowo planowanych stacji metra może być znaczący. Co istotne, choć ceny mieszkań w obszarach już obsługiwanych przez istniejące linie metra również mogą rosnąć, tempo tych wzrostów będzie najprawdopodobniej stabilniejsze.

Z kolei, nieobjęte planami rozbudowy metra rejony — w szczególności wschodnio-południowa część Warszawy — mogą doświadczać względnej stagnacji lub nawet pogłębiającej się luki w wartości nieruchomości w porównaniu do obszarów objętych inwestycjami infrastrukturalnymi. Sytuacja ta może doprowadzić do dalszego zróżnicowania przestrzennego cen mieszkań w stolicy, co w dłuższej perspektywie może oddziaływać na zjawiska społeczno-gospodarcze takie jak migracje wewnętrzmięskie oraz lokalne konflikty spowodowane nierów-

nomiernym rozłożeniem dostępności komunikacyjnej w obrębie Warszawy.

Warto podkreślić, że zastosowana w pracy metodologia — oparta na ustandaryzowanym procesie gromadzenia, oczyszczania i wzbogacania danych, a także zaawansowanym modelowaniu predykcyjnym, pozwoliła nie tylko na uchwycenie istniejących zależności przestrzennych, ale również na wygenerowanie wiarygodnych prognoz zmian cenowych w różnych częściach miasta. Kluczową zaletą opracowanego rozwiązania jest jego pełna replikowalność: każda część strumienia danych, od ekstrakcji i przetwarzania informacji po finalne modelowanie i wizualizację, została zaprojektowana w sposób umożliwiający łatwe odświeżanie analiz w przyszłości. Oznacza to, że w miarę postępu prac budowlanych nad linią M4 oraz ujawniania kolejnych planów dotyczących linii M3, cały model może zostać ponownie uruchomiony na aktualizowanych danych, dostarczając tym samym aktualnych i trafnych prognoz.

Wyniki tej pracy mogą okazać się szczególnie użyteczne dla deweloperów, inwestorów oraz władz samorządowych. Możliwość identyfikacji przyszłych „hotspotów” inwestycyjnych w oparciu o analizę infrastruktury transportowej stwarza szansę na efektywniejsze planowanie nowych inwestycji mieszkaniowych, a także na prowadzenie bardziej świadomej polityki mieszkaniowej przez administrację publiczną. Dla nabywców mieszkań informacje te mogą stanowić realne wsparcie w podejmowaniu decyzji zakupowych o charakterze zarówno konsumpcyjnym, jak i inwestycyjnym.

Podsumowując, rozbudowa warszawskiego metra, a w szczególności linia M4, będzie miała wyraźny wpływ na ceny nieruchomości mieszkaniowych w stolicy. Stworzone w ramach pracy narzędzie predykcyjne pozwala nie tylko na zrozumienie tych zjawisk, ale również na ich bieżące monitorowanie i aktualizację. W świetle rosnących wyzwań związanych z urbanizacją, dostępnością mieszkań oraz presją demograficzną, takie analityczne podejście do rozwiązania tego problemu dostarczyło cennych informacji o kształtującym się rynku mieszkaniowym w Warszawie

Bibliografia

- [1] Trojanek R., 2018, Teoretyczne i metodyczne aspekty wyznaczania indeksów cen na rynku mieszkaniowym, <https://wydawnictwo.ue.poznan.pl/books/978-83-7417-984-3/978-83-7417-984-3.pdf> [dostęp 01.05.2025]
- [2] Vargas-Calderón V., Camargoc J.E., 2020, Towards robust and speculation-reduction real estate pricing models based on a data-driven strategy, https://www.researchgate.net/publication/357918691_Towards_robust_and_speculation-reduction_real_estate_pricing_models_based_on_a_data-driven_strategy [dostęp 01.05.2025]
- [3] Belniak S., Wieczorek D., 2017, Wycena nieruchomości metodą cen hedonicznych – procedura i zastosowanie, <https://sciendo.com/article/10.4467/2353737XCT.17.087.6563> [dostęp 01.05.2025]
- [4] Das S.S.S., Ali M.E., Li Y.-F., Kang Y.-B., Sellis T., 2021, Boosting house price predictions using geo-spatial network embedding, https://www.researchgate.net/publication/357918691_Towards_robust_and_speculation-reduction_real_estate_pricing_models_based_on_a_data-driven_strategy [dostęp 01.05.2025]
- [5] Kok N., Koponen E.-L., Martínez-Barbosa C.A., 2017, Big Data in Real Estate? From Manual Appraisal to Automated Valuation, https://sustainable-finance.nl/upload/researches/Kok-et-al_Big-Data-in-Real-Estate.pdf [dostęp 01.05.2025]
- [6] Imran I., Zaman U., Waqar M., Zaman A., 2021, Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data, https://www.researchgate.net/publication/353371025_Using_Machine_Learning_Algorithms_for_Housing_Price_Prediction_The_Case_of_Islamabad_Housing_Data [dostęp 01.05.2025]
- [7] Borde S., Rane A., Shende G., Shetty S., 2017, Real Estate Investment Advising Using Machine Learning, <https://www.irjet.net/archives/V4/i3/IRJET-V4I3499.pdf> [dostęp 01.05.2025]
- [8] Peterson S., Flanagan A.B., 2009, Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal, https://www.researchgate.net/publication/46526860_Neural_Network_Hedonic_Pricing_Models_in_Mass_Real_Estate_Appraisal [dostęp 01.05.2025]
- [9] Guo Y., Lin S., Ma X., Bal J., Li C.-T., 2018, Homogeneous Feature Transfer and Heterogeneous Location Fine-tuning for Cross-City Property Appraisal Frame-

- work, https://www.researchgate.net/publication/331126953_Homogeneous_Feature_Transfer_and_Heterogeneous_Location_Fine-Tuning_for_Cross-City_Property_Appraisal_Framework_16th_Australasian_Conference_AusDM_2018_Bahrurst_NSW_Australia_November_28-30_2018_Revis [dostęp 01.05.2025]
- [10] Metro Warszawskie sp. z o.o., 2024, Historia budowy metra, <https://metro.waw.pl/metro-warszawskie/linia-m3/aktualnosci-m3/> [dostęp 01.05.2025]
- [11] Deloitte Polska, 2023, Property Index 2023, <https://www.deloitte.com/pl/pl/Industries/real-estate/research/raport-Property-Index-2023.html> [dostęp 01.05.2025]
- [12] Biuro Strategii i Analiz Urzędu m.st. Warszawy, 2024, Raport z badania telemetrycznego liczby mieszkańców Warszawy, <https://um.warszawa.pl/documents/55043703/0/Raport+ludno%C5%9B%C4%87+Warszawy.pdf/813f08fa-7b63-f672-60a3-8cee7d096a9a?t=1721891916986> [dostęp 01.05.2025]
- [13] PricewaterhouseCoopers, 2019, Raport na temat wielkich miast Polski – Warszawa, <https://www.pwc.pl/pl/sektor-publiczny/raporty-warszawa-pol.pdf> [dostęp 01.05.2025]

Spis tabel

1	Przykładowe oferty mieszkań z portalu Otodom	15
2	Przykładowe wiersze nowo wprowadzonych atrybutów geolokalizacyjnych	17
3	Przykładowy fragment teoretycznego zbioru danych	35

Spis rysunków

1	Schemat warszawskiego metra – stan na kwiecień 2024 r.	6
2	Wykres aktywności telekomunikacyjnej w obrębie Warszawy	8
3	Porównanie poziomu rozwoju kapitałów Warszawy na tle największych miast w Polsce	9
4	Wizualizacja proponowanego modelu strumienia przetwarzania danych	13
5	Stacje metra w Warszawie	16
6	Liczba mieszkań z poszczególnymi udogodnieniami	20
7	Porównanie rozkładów zmiennych odległości	21
8	Rozkład ceny za metr kwadratowy względem pogrupowanej zmiennej <i>rok budowy</i>	22
9	Rozkład ceny za metr kwadratowy względem zmiennej <i>stan wykończenia z za-</i> <i>znaczoną medianą</i>	23
10	Drzewo regresyjne: Podział liczby pokoi na 3 podgrupy	24
11	Mapa izolinii cenowych w Warszawie	25
12	Cena za metr kwadratowy w poszczególnych dzielnicach	26
13	Chmura najczęściej występujących w opisach ogłoszeń słów	27
14	Średnia cena w zależności od odległości z trendem wielomianowym	28
15	Średnia cena w zależności od odległości z trendem wielomianowym	32
16	Porównanie metryk regresji dla modeli	33
17	Porównanie ważności cech: Random Forest vs XGBoost	34
18	Porównanie wartości błędu średniokwadratowego dla różnych modeli	36
19	Porównanie wpływu odległości od metra na cenę za metr kwadratowy	37
20	Mapa gradientowa linii M4 z użyciem modelu (M1 XGBoost)	38
21	Mapa gradientowa linii M4 z użyciem modelu (M2 Random Forest)	39

Streszczenie

Niniejsza praca koncentruje się na predykcji zmian cen mieszkań w Warszawie, które mogą nastąpić w wyniku budowy nowych stacji metra linii M4. Temat ma istotne znaczenie społeczno-gospodarcze, zważywszy na rosnące ceny nieruchomości oraz problemy z dostępnością mieszkań. Główna hipoteza badawcza zakłada, że im mniejsza odległość od nowej stacji metra M4, tym wyższa będzie cena za metr kwadratowy mieszkania. Badanie obejmuje zarówno zmienne endogeniczne (jak liczba pokoi czy rok budowy), jak i egzogeniczne (odległość od infrastruktury miejskiej).

W pierwszym rozdziale przedstawiono kontekst historyczny warszawskiego metra oraz charakterystykę lokalnego rynku nieruchomości, a także dokonano przeglądu literatury w zakresie metod predykcji cen – od regresji hedonicznej po modele uczenia maszynowego.

Drugi rozdział poświęcono przygotowaniu danych: pobrano oferty mieszkań z portalu Otodom, przetworzono je (w tym geokodowanie adresów i obliczenie odległości od stacji metra, sklepów i terenów zielonych), a następnie oczyszczono oraz przekształcono zmienne do postaci umożliwiającej modelowanie.

W ostatnim rozdziale zaimplementowano cztery modele regresji: SVR, Random Forest, XGBoost i sieć neuronową. Po ich przetestowaniu wybrano dwa najlepiej działające (Random Forest i XGBoost), a następnie stworzono profile symulacyjne przeciętnego mieszkania, aby przeanalizować wpływ odległości od stacji metra na cenę. Wyniki interpolowano i zwizualizowano w formie map gradientowych, które wskazują na potencjalny wzrost cen, szczególnie w dzielnicach zachodnio-południowych Warszawy. Analiza potwierdza badaną hipotezę, że budowa linii M4 może znacząco wpływać na strukturę cen mieszkań w stolicy.